Extreme Event Modelling

Liwei Wu, SID: 52208712

Department of Mathematics

City University of Hong Kong

Supervisor: Dr. Xiang Zhou

March 31, 2014

Contents

Abstract

In the **Theory and Methods** Part of this paper, we discussed the basic theory and methods of Extreme Value Modelling. Both the Generalized Extreme Value (GEV) distribution and the Generalized Pareto Distribution (GPD) are introduced. Correspondingly, there are two methods to model the Extremes, i.e. the Block Maxima method and the Threshold method. Then in the **Application** Part, we applied these methods to fit the data to obtain the corresponding return level. Firstly, we applied the Block Maxima Method to the simulated normal data. Then we applied the Threshold Method to the Dow Jones Index data and the Hong Kong climate data and reached informative conclusion based on the return levels we obtained.

KEY WORDS: Extreme value theory; Generalized Pareto distribution; Dow Jones Index; Hong Kong climate data; Threshold method.

1 Introduction

Extreme Events are of great significance in daily life and extreme event modelling techniques are widely used in many disciplines. In real life, we are sometimes interested in getting to know what the extreme levels can be within a certain period of time. For example, we want to predict how large the 100-year flood can be if we only have the past 10 years of data available, such that we can make necessary preparations. In this case, we have to apply the Extreme Event Modelling techniques to make a reasonable predication. There are essentially two methods to model the extremes. The first one is called the Block Maximal method and the second one is called the Threshold method.

The models we are going to introduce in the paper focus on the statistical behavior of

$$M_n = \max\{X_1, X_2, ..., X_n\},$$

where $X_1, X_2, ..., X_n$, is a sequence of independent random variables having a common distribution function F. In Applications, the X_i usually represents values of a process measured on a regular time-scale – perhaps hourly measurements of sea-level, or daily mean temperatures – so that 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.

If we do not have access to each X_i and only know the Block Maxima, then we have no choice but to use the Block Maxima method. However, if each X_i is known, using the Threshold method will allow us to make better use of the available data. In this paper, we want to introduce the theory and methods of Extreme Value Modelling and show examples of how to apply these two methods to real data. Therefore, the paper is structured as below. In Section (2.1), we discussed the theoretical foundation of the Block Maxima method. In Section (2.2), the threshold method is discussed to model the threshold excesses. Then, in Section (3.1), we applied the Block Maxima method to the simulated normal data and made an inference of the return levels using the 95% confidence intervals. In Section (3.2) and (3.3), we applied the Threshold method to the Dow Jones Index data and Hong Kong climate data respectively and obtained the corresponding return levels with the 95% confidence intervals. In Section (4), the R codes written to produce the results in Section (3) are given. Finally, in Section (5), we have the summary and outlook: a new method using Bayesian hierarchical model is introduced for the spatial data.

2 Theory and Methods

2.1 Asymptotic Models

2.1.1 Extremal Types Theorem

In theory the distribution of M_n can be derived exactly for all values of n:

$$\begin{aligned} Pr\left(M_n \leq z\right) &= Pr\left(X_1 \leq z, ..., X_n \leq z\right) \\ &= Pr\left(X_1 \leq z\right) \times ... \times Pr\left(X_n \leq z\right) \\ &= \left(F\left(z\right)\right)^n \end{aligned}$$

However, this is not immediately helpful in practice, since the distribution function F is unknown. One possibility is to use standard statistical techniques to estimate F from observed data, and then to substitute this estimate into equations above. Unfortunately, very small discrepancies in the estimate of F can lead to substantial discrepancies for F^n.

An alternative approach is to accept that F is unknown, and to look for approximate families of models for F^n, which can be estimated on the basis of the extreme data only. This is similar to the usual practice of approximating the distribution of sample means by the normal distribution, as justified by the central limit theorem.

We proceed by looking at the behavior of F^n as n $\rightarrow \infty$. But this alone is not enough: for any $z < z+$, where $z+$ is the upper end-point of F, $F^n\left(z\right) \rightarrow 0$ as $n \rightarrow \infty$, so that the distribution of M_n degenerates to a point mass on $z+$. This difficulty is avoided by allowing a linear renormalization of the variable M_n:

$$M_n^* = \frac{M_n - b_n}{a_n}$$

for sequences of constants $\{a_n > 0\}$ and $\{b_n\}$. Appropriate choices of the $\{a_n\}$ and $\{b_n\}$ stabilize the location and scale of M_n^* as n increases, avoiding the difficulties that arise with the variable M_n. We therefore seek limit distributions for M_n^*, with appropriate choices of $\{a_n\}$ and $\{b_n\}$, rather than M_n.

Theorem 2.1 *If there exist sequences of constants* $\{a_n\}$ *and* $\{b_n\}$ *such that*

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z) \ as\, n \to \infty$$

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

I

$$G(z) = \exp\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\}$$

II

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

III

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

for parameters $a > 0$, b *and, in the case of families II and III,* $\alpha > 0$.

The rescaled sample maxima $(M_n - b_n)/a_n$ converge in distribution to a variable having a distribution within one of the families labeled I and II and III. Collectively, these three classes of distribution are termed the extreme value distributions, with types I, II and III widely known as the Gumbel, Fréchet and Weibull families respectively. Each family has a location and scale parameter, b and a respectively; additionally, the Fréchet and Weibull families have a shape parameter α. Theorem 2.1 implies that, when M_n can be stabilized with suitable sequences a_n and b_n, the corresponding normalized variable M_n^* has a limiting distribution that must be one of the three types of extreme value distribution.The remarkable feature of this result is that the three types of extreme value distributions are the only possible limits for the distributions of the M_n^*, regardless of the distribution F for the population. It is in this sense that the theorem provides an extreme value analog of the central limit theorem.

2.1.2 The Generalized Extreme Value Distribution

In early applications of extreme value theory, it was usual to adopt one of the three families, and then to estimate the relevant parameters of that distribution. But there are two weaknesses: first, a technique is required to choose which of the three families is most appropriate for the data at hand; second, once

such a decision is made, subsequent inferences presume this choice to be correct, and do not allow for the uncertainty such a selection involves, even though this uncertainty may be substantial.

A better analysis is offered by a reformulation of the models in Theorem 2.1. It is straightforward to check that the Gumbel, Fréchet and Weibull families can be combined into a single family of models having distribution functions of the form

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

defined on the set $\{z : 1+\xi(z-\mu)/\sigma > 0\}$, where the parameters satisfy $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$. This is the **generalized extreme value** (GEV) family of distributions. The model has three parameters: a location parameter μ; a scale parameter, σ; and a shape parameter, ξ. The type II and type III classes of extreme value distribution correspond respectively to the cases $\xi > 0$ and $\xi < 0$ in this parameterization. The subset of the GEV family with $\xi = 0$ is interpreted as the limit of (1) as $\xi \to 0$, leading to the Gumbel family with distribution function

$$G(z) = \exp\left[-\exp\left\{-\frac{z-\mu}{\sigma}\right\}\right], \; -\infty < z < \infty \tag{2}$$

The unification of the original three families of extreme value distribution into a single family greatly simplifies statistical implementation. Through inference one, the data themselves determine the most appropriate type of tail behavior, and there is no necessity to make subjective a priori judgements about which individual extreme value family to adopt.

Theorem 2.2 *If there exist sequences of constants* $\{a_n > 0\}$ *and* $\{b_n\}$ *such that*

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z) \;\; as \; n \to \infty \tag{3}$$

for a non-degenerate distribution function G, then G is a member of the GEV family

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

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

2.1.3 Examples

A few examples will be given to illustrate how careful choice of normalizing sequences does lead to a limit distribution within the GEV family, as implied by **Theorem 2.1**.

First Example

If $X_1, X_2, ...$ is a sequence of independent standard exponential exp(1) variables, $F(x) = 1 - e^{-x}$ for $x > 0$. In this case, letting $a_n = 1$ and $b_n = n$,

$$\begin{aligned} Pr\{(M_n - b_n)/a_n \leq z\} &= F^n(z + \log n) \\ &= \left[1 - e^{-(z+\log n)}\right]^n \\ &= \left[1 - n^{-1}e^{-z}\right]^n \\ &\to \exp\left(-e^{-z}\right) \end{aligned}$$

as $n \to \infty$, for each fixed $z \in \mathbb{R}$ Hence, with the chosen a_n and b_n, the limit distribution of M_n as $n \to \infty$ is the Gumbel distribution, corresponding to $\xi = 0$ in the GEV family.

Second Example

If $X_1, X_2, ...$ is a sequence of independent standard Fréchet variables, $F(x) = \exp(-1/x)$ for $x > 0$. Letting $a_n = n$ and $b_n = 0$,

$$\begin{aligned} Pr\{(M_n - b_n)/a_n \leq z\} &= F^n(nz) \\ &= [\exp\{-1/(nz)\}]^n \\ &= \exp(-1/z) \end{aligned}$$

as $n \to \infty$, for each fixed $z > 0$. Hence, the limit in this case - which is an exact result for all n, because of the max-stability of F - is also the standard Fréchet distribution: $\xi = 1$ in the GEV family.

Third Example

If $X_1, X_2, ...$ are a sequence of independent uniform $U(0, 1)$ variables, $F(x) = x$ for $0 \leq x \leq 1$. For fixed

$z < 0$, suppose $n > -z$ and let $a_n = 1/n$ and $b_n = 1$. Then,

$$\begin{aligned} Pr\left\{\left(M_n - b_n\right)/a_n \leq z\right\} &= F^n\left(n^{-1}z + 1\right) \\ &= \left(1 + \frac{z}{n}\right)^n \\ &= e^z \end{aligned}$$

as $n \to \infty$. Hence, the limit distribution is of Weibull type, with $\xi = -1$ in the GEV family.

Fourth Example

If $X_1, X_2, ...$ are a sequence of independent normal random variables, let $a_n = \frac{1}{\sqrt{2\ln n}}$ and $b_n = \sqrt{2\ln n} - \frac{\ln\ln n + \ln(4\pi)}{2\sqrt{2\ln n}}$. Then,

$$Pr\left\{\left(M_n - b_n\right)/a_n \leq z\right\} = e^{-e^{-x}}$$

as $n \to \infty$. Hence, the limit distribution is of Gumbel type, corresponding to $\xi = 0$ in the GEV family.

2.1.4 Modelling extremes using Block Maxima

The apparent difficulty that the normalizing constants $\{a_n > 0\}$ and $\{b_n\}$ will be unknown in practice is easily resolved. Assuming (3), for large enough n,

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \approx G\left(z\right)$$

Equivalently,

$$\begin{aligned} P(M_n \leq z) &\approx G((z - b_n)/a_n) \\ &= G^*(z) \end{aligned}$$

where G^* is another member of the GEV family. In other words, if Theorem 2.2 enables approximation of the distribution of M_n^* by a member of the GEV family for large n, the distribution of M_n itself can also be approximated by a different member of the same family.

A series of independent observations X_1, X_2...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, in which case n is the number of observations in a year and the block maxima are annual maxima. Estimates of extreme quantiles of the annual maximum distribution are then obtained by inverting (2):

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-log\,(1-p)\}^{-\xi}\right], & \text{for } \xi \neq 0; \\ \mu - \sigma log\,\{-log\,(1-p)\}, & \text{for } \xi = 0. \end{cases} \tag{4}$$

where $G(z_p) = 1 - p$.

Definition A **return period** also known as a recurrence interval is an estimate of the likelihood of an event, such as an earthquake, flood or a river discharge flow to occur. It is a statistical measurement typically based on historic data denoting the average recurrence interval over an extended period of time, and is usually used for risk analysis (e.g. to decide whether a project should be allowed to go forward in a zone of a certain risk, or to design structures to withstand an event with a certain return period).

Definition The **return level** is associated with the corresponding return period and indicates the level the maxima can reach within such a return period (e.g. the 100-year flood return level).

In common terminology, Z_p is the return level associated with the return period $1/p$, since to a reasonable degree of accuracy, the level z_p is expected to be exceeded on average once every $1/p$ years. Define $y_p = -log(1-p)$, then (4) will become,

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - y_p^{-\xi}\right], & \text{for } \xi \neq 0; \\ \mu - \sigma log\, y_p, & \text{for } \xi = 0. \end{cases}$$

The **return level plot** is the graph in which if Z_p is plotted against y_p on a logarithmic scale - or equivalently, if Z_p is plotted against $log\, y_p$. If $\xi = 0$, the plot is linear. If $\xi < 0$, the plot is convex with asymptotic limit as $p \to 0$ at $\mu - \sigma/\xi$. If $\xi > 0$, the plot is concave and has no finite bound.

Motivated by Theorem 2.2, the GEV provides a model for the distribution of block maxima. Its application consists of blocking the data into blocks of equal length, and fitting the GEV to the set of block maxima. But in implementing this model for any particular dataset, the choice of block size can

be critical. The choice amounts to a trade-off between bias and variance: blocks that are too small mean that approximation by the limit model in Theorem 2.2 is likely to be poor, leading to bias in estimation and extrapolation; large blocks generate few block maxima, leading to large estimation variance. We now simplify notation by denoting the block maxima $Z_1, ..., Z_m$. These are assumed to be independent variables from a GEV distribution whose parameters are to be estimated. If the X_i are independent then the Z_i are also independent. Under the assumption that $Z_1, ..., Z_m$ are independent variables having the GEV distribution, the log-likelihood for the GEV parameters when $\xi \neq 0$ is

$$l(\mu, \sigma, \xi) = -m \log(\sigma) - (1 + 1/\xi) \sum_{i=1}^{m} \log\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{m}\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]^{(-1/\xi)}, \quad (5)$$

provided that

$$1 + \xi\left(\frac{z_i - \mu}{\sigma}\right) > 0, \forall i = 1, .., m \quad (6)$$

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

$$l(\mu, \sigma) = -m\ \log\sigma - \sum_{i=1}^{m}\left(\frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^{m}\exp\left\{-\left(\frac{z_i - \mu}{\sigma}\right)\right\} \quad (7)$$

Maximization of the pair of Equations.(5) and (7) with respect to the parameter vector (μ, σ, ξ) leads to the maximum likelihood estimate with respect to the entire GEV family.

After obtaining the maximum likelihood estimates of the GEV parameters, we can substitute them into (4) and obtain the maximum likelihood estimate of z_p for $0 < p < 1$, the 1/p return level, as

$$\hat{z_p} = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left[1 - y_p^{-\hat{\xi}}\right] & \text{for } \hat{\xi} \neq 0 \\ \hat{\mu} - \hat{\sigma}\log y_p & \text{for } \hat{\xi} = 0 \end{cases}$$

where $y_p = -\log(1 - p)$. Furthermore, by the delta method, we can obtain the variance of the maximum likelihood estimate

$$Var(\hat{z_p}) \approx \nabla z_p^T V \nabla z_p$$

where V is the variance-covariance matrix of $\left(\hat{\mu}, \hat{\sigma}, \hat{\xi}\right)$ and

$$\begin{aligned}\nabla z_p^T &= \left[\begin{array}{ccc} \frac{\partial z_p}{\partial \mu} & \frac{\partial z_p}{\partial \sigma} & \frac{\partial z_p}{\partial \xi} \end{array}\right] \\ &= \left[\begin{array}{ccc} 1 & -\xi^{-1}\left(1-y_p^{-\xi}\right) & \sigma\xi^{-2}\left(1-y_p^{-\xi}\right) - \sigma\xi^{-1}y_p^{-\xi}\log y_p \end{array}\right]\end{aligned}$$

evaluated at $\left(\hat{\mu}, \hat{\sigma}, \hat{\xi}\right)$.

2.1.5 Modelling Checking

It is important to check the validity of an extrapolation based on a GEV model. To achieve the goal, we can use the probability plot, quantile plot, return level plot, and density plot.

- A probability plot is a comparison of the empirical and fitted distribution functions. With ordered block maximum data $z_{(1)} \leq z_{(2)} \leq ... \leq z_{(m)}$, the empirical distribution function evaluated at $z_(i)$ is given by

  $$\tilde{G}\left(z_{(i)}\right) = i/(m+1)$$

  By substitution of parameter estimates into (1), the corresponding model based estimates are

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

  If the GEV model is working well,

  $$\tilde{G}\left(z_{(i)}\right) \approx \hat{G}\left(z_{(i)}\right)$$

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

  $$\left\{\left(\tilde{G}\left(z_{(i)}\right), \hat{G}\left(z_{(i)}\right)\right), i = 1, ..., m\right\}$$

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

- Similarly, a quantile plot consists

  $$\left\{\left(\hat{G}^{-1}\left(i/(m+1)\right), z_i\right), i = 1, ..., m\right\}$$

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

- The return level plot summarises the fitted model and consists of the locus of points

$$\{(\log\ y_p, \hat{z}_p)\,, 0 < p < 1\}$$

Confidence intervals can be added to the plot to increase its informativeness. Empirical estimates of the return level function can also be added, enabling the return level plot to be used as a model diagnostic.

- For completeness, an equivalent diagnostic based on the density function is a comparison of the probability density function of a fitted model with a histogram of the data. This is generally less informative than the previous plots, since the form of a histogram can vary substantially with the choice of grouping intervals.

It is worth noting that the first two methods work for any continuous distribution. The third one is a special modelling checking method for extreme value analysis.

2.1.6 GEV distribution for Minima

Definition

$$\tilde{M}_n = \min\{X_1, ..., X_n\}$$

and

$$\tilde{\mu} = -\mu$$

Theorem 2.3 *If there exist sequences of constants* $\{a_n > 0\}$ *and* $\{b_n\}$ *such that*

$$P\left(\frac{\tilde{M}_n - b_n}{a_n} \leq z\right) \to \tilde{G}\,(z)\ \ as\, n \to \infty \tag{8}$$

for a non-degenerate distribution function $\tilde{G}$*, then* $\tilde{G}$ *is a member of the GEV family of distributions for minima:*

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

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

In situations where it is appropriate to model block minima, the GEV distribution for minima can be applied directly. An alternative is to exploit the duality between the distributions for maxima and minima. Given data $z_1, ..., z_m$ that are realizations from the GEV distribution for minima, with parameters $(\tilde{\mu}, \sigma, \xi)$, this implies fitting the GEV distribution for maxima to the data $-z_1, ..., -z_m$ The maximum likelihood estimate of the parameters of this distribution corresponds exactly to that of the required GEV distribution for minima, apart from the sign correction for μ only: $\hat{\tilde{\mu}} = -\hat{\mu}$.

2.2 Threshold Models

2.2.1 Introduction and Motivation

Modelling only block maxima is a wasteful approach to extreme value analysis if other data on extremes are available. Therefore, if an entire time series of, say, hourly or daily observations is available, then we can make better use of the data by avoiding altogether the procedure of blocking. Let X_1, X_2, ... be a sequence of independent random variables with common distribution function F, and let

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

It is natural to regard those of the X_i exceeding some high threshold u as extreme events. Denoting an arbitrary term in the X_i sequence by X, it follows that a description of the stochastic behaviour of extreme events is given by the conditional probability

$$Pr\{X > u + y | X > u\} = \frac{1 - F(u + y)}{1 - F(u)}, y > 0$$

If the parent distribution F were known, then the distribution of threshold exceedances would also be known. In practical applications, this is not the case. Therefore, we seek approximations that broadly applicable for high values of the threshold. The results are given in the following theorem.

2.2.2 The Generalized Pareto Distribution

Theorem 2.4 *Denote an arbitrary term in the X_i sequence by X, and suppose that F satisfies Theorem 2.2, so that for large n,*

$$P\left(M_n \leq z\right) \approx G(z),$$

where

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

for some $\mu, \sigma > 0, \xi$. Then, for large enough u, the distribution function of $(X-u)$, conditional on $X > u$, is approximately

$$H(y) = 1 - \left(1+\frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} \tag{9}$$

defined on $\{y : y > 0$ and $(1+\xi y/\tilde{\sigma}) > 0\}$, where

$$\tilde{\sigma} = \sigma + \xi(u-\mu)$$

The family of distributions defined by Equation (9) is called the generalized Pareto family. Theorem 2.4 implies that, if block maxima have approximating distribution G, then threshold excesses have a corresponding approximate distribution within the generalized Pareto family. Moreover, the parameters of the generalized Pareto distribution of threshold excesses are uniquely determined by those of the associated GEV distribution of block maxima.

2.2.3 Threshold Selection

The raw data consist of a sequence of independent and identically distributed measurement $x_1, ..., x_n$. Extreme events are identified by defining a high threshold u, for which the exceedances are $\{x_i : x_i > u\}$. Label these exceedances by $x_{(1)}, ..., x_{(k)}$, and define threshold excesses by $y_j = x_{(j)} - u$, for $j = 1, ..., k$. By Theorem 2.4, the Y_i may be regarded as independent realizations of a random variable whose distribution can be approximated by a member of the generalized Pareto family. Inference consists of fitting the generalized Pareto family to the observed threshold exceedances, followed by model verification and extrapolation.

This approach contrasts with the block maxima approach through the characterization of an observation as extreme if it exceeds a high threshold. But the issue of threshold choice is analogous to the

choice of block size in the block maxima approach, implying a balance between bias and variance. In this case, too low a threshold is likely to violate the asymptotic basis of the model, leading to bias; too high a threshold will generate few excesses with which the model can be estimated, leading to high variance. The standard practice is to adopt as low a threshold as possible, subject to the limit model providing a reasonable approximation. Two methods are available for this purpose: one is an exploratory technique carried out prior to model estimation; the other is an assessment of the stability of parameter estimates, based on the fitting of models across a range of different thresholds.

The first method is based on the mean of the generalized Pareto distribution. If Y has a generalized Pareto distribution with parameters σ and ξ, then

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

provided $\xi < 1$. When $\xi \geq 1$ the mean is infinite. Now, suppose the generalized Pareto distribution is valid as a model for the excesses of a threshold u_0 generated by a series $X_1, ..., X_n$, of which an arbitrary term is denoted X. Then, we have,

$$E(X - u_0 | X > u_0) = \frac{\sigma_{u_0}}{1-\xi}$$

provided $\xi < 1$, where we adopt the convention of using σ_{u_0} to denote the scale parameter corresponding to excesses of the threshold u_0. σ_{u_0} can be defined mathematically too as below:

$$\sigma_{u_0} = \sigma + \xi(u_0 - \mu)$$

But if the generalized Pareto distribution is valid for excesses of the threshold u_0, it should equally be valid for all thresholds $u > u_0$, subject to the appropriate change of scale parameter to σ_{u_0}. Hence, for $u > u_0$,

$$\begin{aligned} E(X - u | X > u) &= \frac{\sigma_u}{1-\xi} \\ &= \frac{\sigma_{u_0} + \xi(u - u_0)}{1-\xi} \end{aligned}$$

So, for $u > u_0$, $E(X - u | X > u)$ is a linear function of u. Furthermore, $E(X - u | X > u)$ is simply the mean of the excesses of the threshold u, for which the sample mean of the threshold excesses of u provides an empirical estimate. These estimates are expected to change linearly with u, at levels of u for which the generalized Pareto model is appropriate. This leads to the following procedure. The locus

of points

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} \left(x_{(i)} - u \right) \right) : u < x_{max} \right\}$$

where $x_{(1)}, ..., x_{(n_u)}$ consist of the n_u observations that exceed u, and x_{max} is the largest of the x_i, is termed the **mean residual life plot**. Above a threshold u_0 at which the generalized Pareto distribution provides a valid approximation to the excess distribution, the mean residual life plot should be approximately linear in u. Confidence intervals can be added to the plot based on the approximate normality of sample means.

The second method for threshold selection is to estimate the model at a range of thresholds. Above a level u_0 at which the asymptotic motivation for the generalized Pareto distribution is valid, estimates of the shape parameter ξ should be approximately constant, while estimates of σ_u should be linear in u. Since sometimes the first method is difficult to interpret, it could be a better way to just look for the stability while varying the threshold of the fitted GPD. The theoretical basis is explained as follows.

By the Theorem 2.4, if a generalized Pareto distribution is a reasonable model for excesses of a threshold u_0, then excesses of a higher threshold u should also follow a generalized Pareto distribution. The shape parameters of the two distributions are identical. However, denoting by σ_u the value of the generalized Pareto scale parameter for a threshold of $u > u_0$, it follows

$$\sigma_u = \sigma_{u_0} + \xi(u - u_0).$$

We can see the scale parameter changes with respect to u. In order to overcome the difficulty, we can reparameterize the generalized Pareto scale parameter as

$$\sigma^* = \sigma_u - \xi u,$$

which is now constant with respect to u. Consequently, estimates of both σ^* and ξ should be constant above u_0, if u_0 is a valid threshold for excesses to follow the generalized Pareto distribution. Therefore, it makes sense to plot both $\hat{\sigma}^*$ and $\hat{\xi}$ against u with their corresponding confidence intervals, and select u_0 as the lowest value of u for which the estimates remain near-constant.

2.2.4 Model Checking

After obtaining the proper threshold of the fitted GPD, we need to assess the quality of the fitted generalized Pareto model. It can be done using probability plots, quantile plots, return level plots, and density plots.

With the threshold u, threshold excesses $y_{(1)} \leq .... \leq y_{(k)}$ and an estimated model $\hat{H}$, the probability plot is given by

$$\left\{ \left(i/\left(k+1\right), \hat{H}\left(y_{(i)}\right) \right); i = 1, ..., k \right\}$$

where

$$\hat{H}\left(y\right) = 1 - \left(1 + \frac{\hat{\xi} y}{\hat{\sigma}} \right)^{-1/\hat{\xi}},$$

when $\hat{\xi} \neq 0$.

The quantile plot is given in a similar way: when $\hat{\xi} \neq 0$,

$$\left\{ \left(\hat{H}^{-1}\left(i/\left(k+1\right)\right), y_{(i)} \right), i = 1, ..., k \right\}$$

where

$$\hat{H}^{-1}\left(y\right) = u + \frac{\hat{\sigma}}{\hat{\xi}} \left[y^{-\hat{\xi}} - 1 \right]$$

According to the theory, if excesses of u fits into the GPD model, both the probability and quantile plots should consist of points that are approximately linear.

The return level plot is given by $\{(m, \hat{x}_m)\}$ for large values of m, where

$$\hat{x}_m = u + \frac{\hat{\sigma}}{\hat{\xi}} \left[\left(m\hat{\zeta}_u \right)^{\hat{\xi}} - 1 \right]$$

and

$$\hat{\zeta}_u = Pr\{X > u\}$$

Lastly, the density function of the fitted GPD model can be compared to a histogram of the threshold exceedances.

3 Application

3.1 Simulation of data from normal distribution

3.1.1 Introduction

We now consider the data $x_{1,1}, x_{1,2}, ..., x_{1,n}, x_{2,1}, x_{2,2}, ..., x_{2,n}, x_{3,1}, ..., x_{m,1}, ..., x_{m,n}$ from standard normal distribution: $\mathcal{N}(0,1)$. The data set is generated from R. Dividing the data into m blocks, we have n data points in each block. We now simplify notation by denoting the block maxima $Z_1, Z_2, ..., Z_m$. These are assumed to be independent variables from a GEV distribution who'se parameters are to be estimated given that n is large enough. Since X_i are independent, then the Z_i are also independent. We know that the log-liklihood for the GEV parameters when $\xi \neq 0$ is as follows:

$$l(\mu, \sigma, \xi) = -m \log(\sigma) - (1 + 1/\xi) \sum_{i=1}^{m} \log\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{m}\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]^{(-1/\xi)},$$

provided that

$$1 + \xi\left(\frac{z_i - \mu}{\sigma}\right) > 0, \forall i = 1, .., m$$

3.1.2 maximal likelihood estimates and corresponding confidence intervals

in this specific case, we have $m = 200$ and $n = 500$. Using numerical optimisation algorithms, we can obtain the maximal likelihood estimate with respect to the entire GEV family. Using packages in R, we have the estimates as follows: $\mu = 2.886553$, with a 95% confidence interval $[2.837136, 2.93597]$, $\sigma = 0.3216$, with a 95% confidence interval $[0.286984, 0.3562161]$, $\xi = -0.08799757$, with a 95% confidence interval $[-0.1778897, 0.001894544]$. Notice that the confidence interval of ξ contains 0, which means the Gumbel Distribution could be a more accurate model in the entire GEV family. Also, it makes sense that confidence intervals for return levels obtained by fitting against Gumbel Distribution are narrower than those obtained by fitting against a member of general GEV distribution. All of the four diagnostic plots above: probability plot, quantile plot, return level plot and density plot, provides evidence that GEV is a good fit to the block maxima.

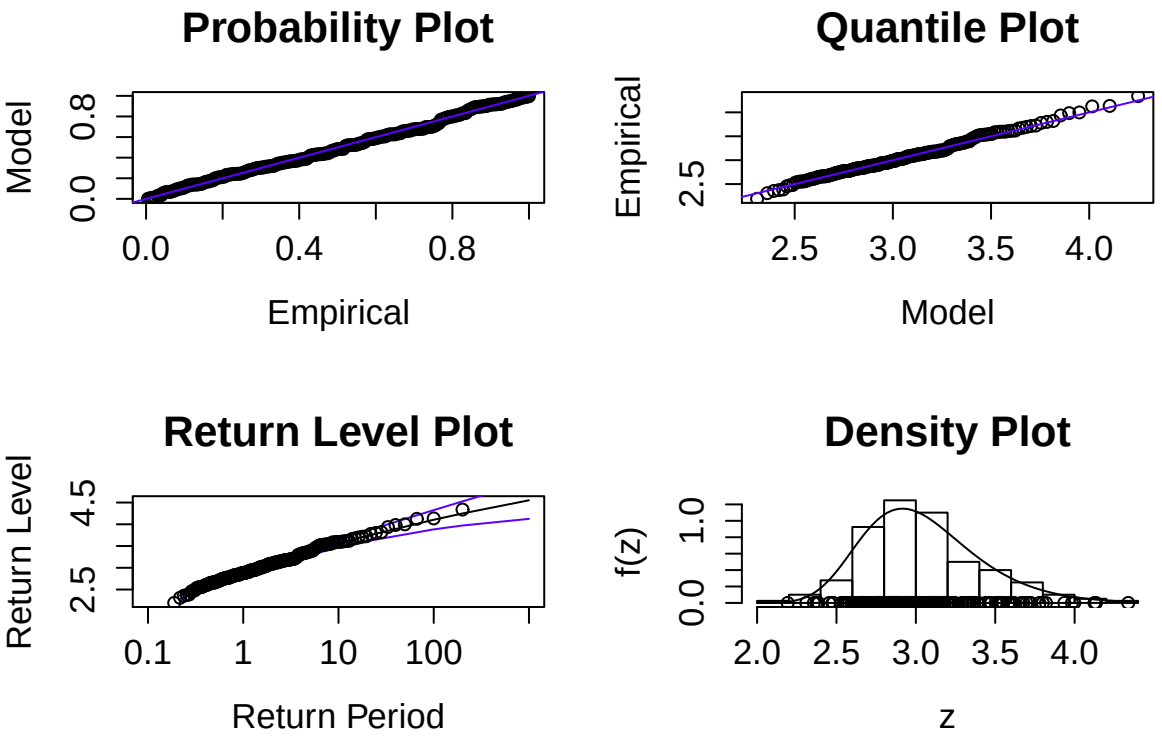

Figure 1: GEV fit

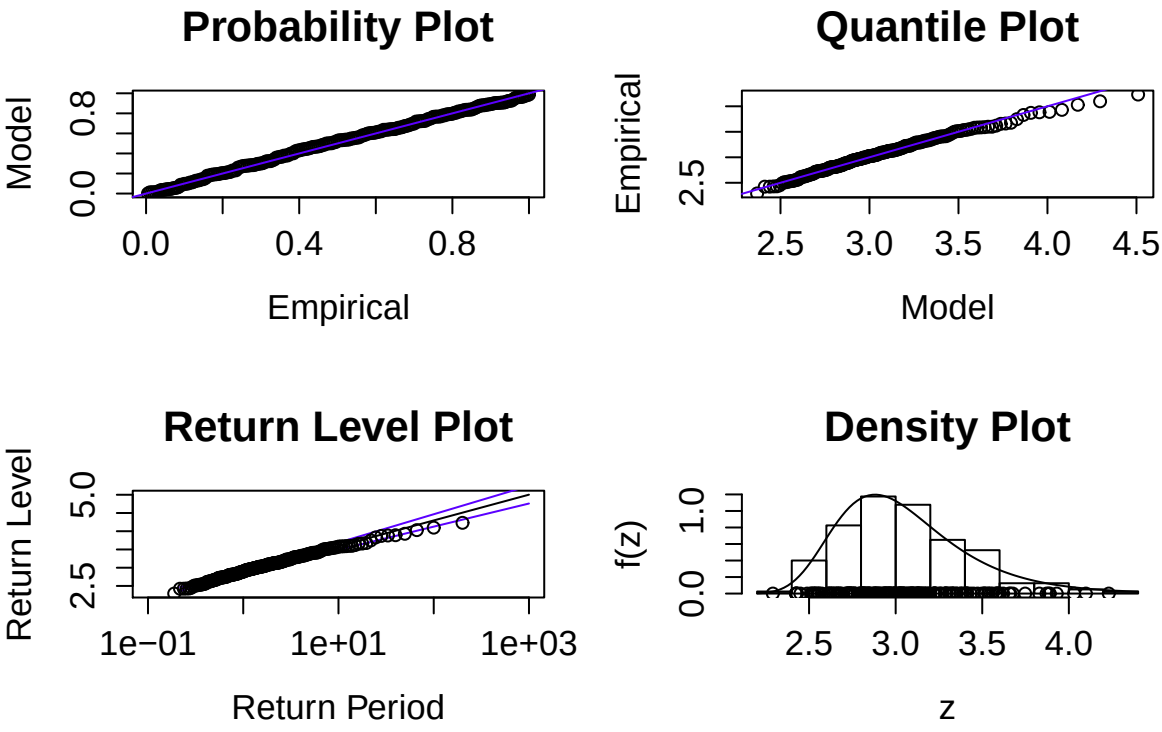

Figure 2: Gumbel fit

3.1.3 Inference for return levels

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

$$\hat{z_p} = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left[1 - y_p^{-\hat{\xi}}\right] & \text{for } \hat{\xi} \neq 0 \\ \hat{\mu} - \hat{\sigma}\log y_p & \text{for } \hat{\xi} = 0 \end{cases}$$

where $y_p = -\log(1-p)$. Furthermore, by the delta method, we have

$$Var(\hat{z_p}) \approx \nabla z_p^T V \nabla z_p$$

where V is the variance-covariance matrix of $\left(\hat{\mu}, \hat{\sigma}, \hat{\xi}\right)$ and

$$\begin{aligned} \nabla z_p^T &= \begin{bmatrix} \frac{\partial z_p}{\partial \mu} & \frac{\partial z_p}{\partial \sigma} & \frac{\partial z_p}{\partial \xi} \end{bmatrix} \\ &= \begin{bmatrix} 1 & -\xi^{-1}\left(1 - y_p^{-\xi}\right) & \sigma\xi^{-2}\left(1 - y_p^{-\xi}\right) - \sigma\xi^{-1} y_p^{-\xi}\log y_p \end{bmatrix} \end{aligned}$$

evaluated at $\left(\hat{\mu}, \hat{\sigma}, \hat{\xi}\right)$. Apart from the delta method, the profile likelihood function can also be used to obtain a more accurate confidence interval of z_p by expressing μ with z_p, σ, ξ and substituting into the GEV model. It is usually long return periods, corresponding to small values of p, that are of greatest interest.

For the specific example above, estimates and confidence intervals for returns levels can be obtained. For example, when $p = 1/10$, the estimate for the 10-year return level $\hat{z}_{0.1}$is 3.536803, with a 95% confidence interval $[3.453434, 3.641219]$, and similarly when $p = 1/100$, the estimate for the 100-year return level $\hat{z}_{0.01}$is 4.136239, with a 95% confidence interval $[3.955557, 4.447778]$.

3.2 Application in stock market

3.2.1 Introduction

Sometimes people are interested in the extreme event in stock markets, such as how large the daily loss percentage would be within 100 years. Such problems can be modelled and solved using Threshold Method from the Extreme Event Modelling. An example with detailed procedures is given in this section to illustrate how to obtain 100-year return level of daily loss percentage in stock market.

Since the Dow Jones Index data is readily available from the Internet, we can just use it to measure the daily loss percentage in the stock market in US. We obtained the daily Dow Jones Index data from 1896-05-26 to 2013-08-02, totalling 31960 data points. After inputting the csv file containing the data into R, we can use the existing R package "ismev" and "extRemes" to choose the proper threshold, fit the data into the Generalized Pareto Distribution (GPD), and obtain the 100-year return level of daily loss percentage in stock market with the corresponding 95% confidence interval.

3.2.2 Selecting the proper threshold

Let us first have a brief look at the data points in the plot below.

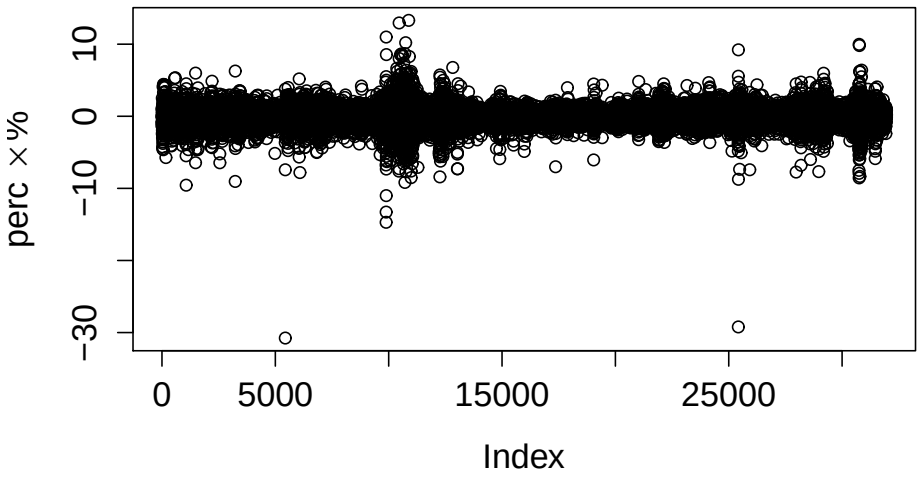

Figure 3: data points of daily percentage change of Dow Jones Index

In the plot, the index axis shows the index of 31960 data points, which are numbered in the time order. The Perc index shows the daily percentage change for each data point. A positive number indicates the daily gain while a negative number indicates the daily loss. Since there are very few data points (only about 3% out of all data points) with the daily loss greater than 2%, it is safe to assume that the event of reasonably large daily loss percentage occurs can be classified into extreme events and that the threshold models method from the extreme event modelling can be used. We use the threshold method instead of the Block Maxima method, because we want to make the most use of the available data points.

The first procedure when using the threshold method is to select the proper threshold. As stated before in the subsection (2.3.2), two methods will be used before making a final decision.

The first method is to select a proper threshold such that the mean residue life plot should be approximately linear above the selected proper threshold u_0.

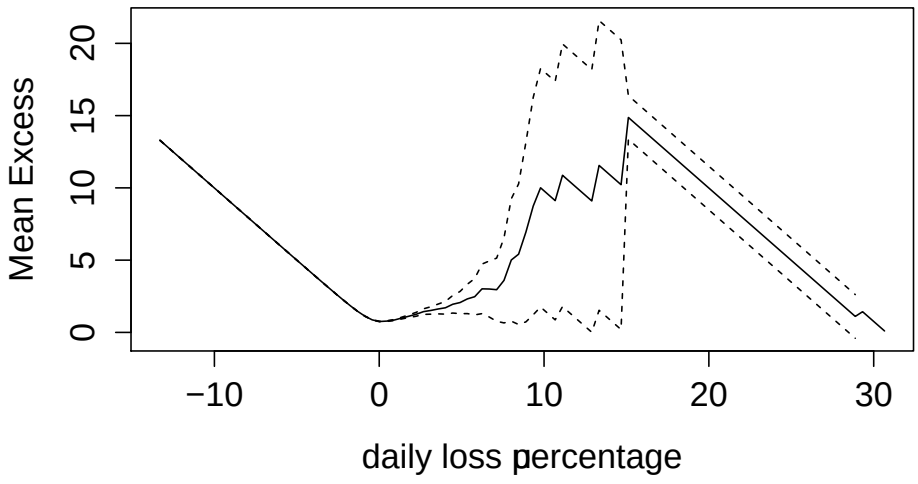

Figure 4: mean residue life plot

In the mean residue life plot above, the information above the daily loss percentage $u = 10$ is not very accurate due to very few points (actually only 5) with daily loss greater than 10%. Therefore, we should ignore that part of the plot and conclude that the proper threshold u_0 should satisfy $u_0 \geq 3$, since it is easy to see that the plot is approximately linear above $u = 3$.

To further explore what the proper threshold should be, the second method is used: look for the

stability of parameters σ^* and ξ while varying the threshold of the fitted GPD.

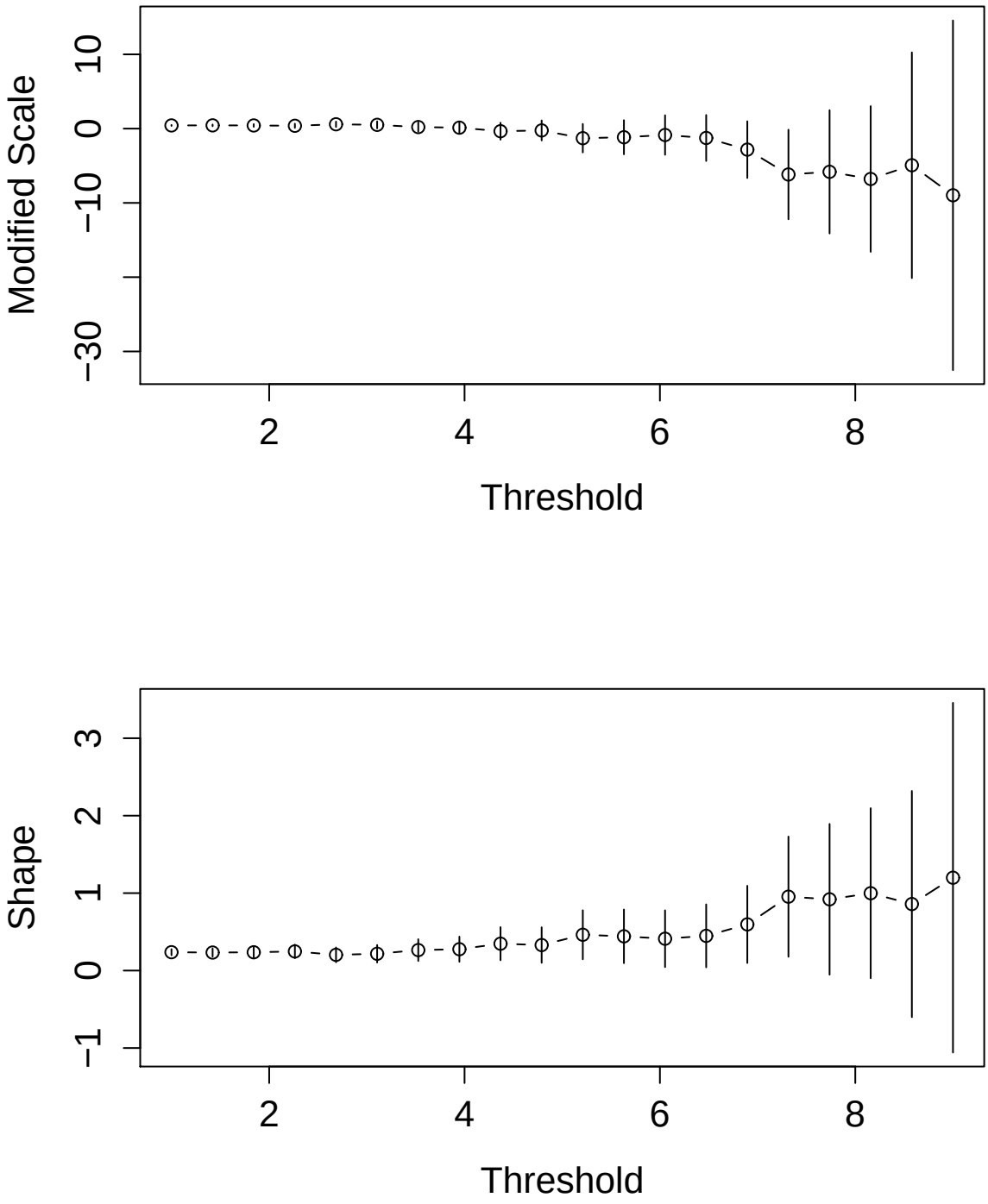

Figure 5: parameter estimates against threshold

From the plot above, we can see that the estimated parameters are more or less stable when $u \geq 5$.

Therefore, the selected threshold of $u = 5$ appears reasonable.

3.2.3 Obtaining the 100-year return level of daily loss percentage

Firstly, we need to fit the data points into the GPD with the selected threshold $u = 5$. The maximum likelihood estimates in this case are

$$\left(\hat{\sigma}, \hat{\xi}\right) = (1.2694261, 0.3737744)$$

with a corresponding maximised log-likelihood of -143.4933. The variance-covariance matrix is calculated as

$$\begin{bmatrix} 0.04395658 & -0.01503806 \\ -0.01503806 & 0.01762163 \end{bmatrix}$$

leading to standard errors of 0.2096583 and 0.1327465 for $\hat{\sigma}$ and $\hat{\xi}$ respectively. In particular, it follows that a 95% confidence interval for ξ is obtained as $0.3737744 \pm 1.96 \times 0.1327465 = [0.1135913, 0.6339575]$. Since the maximum likelihood estimate $\xi > 0$, it leads to an unbounded distribution in the return level plot and the evidence for this is pretty strong, since the 95% interval for ξ is exclusively in the positive domain.

Since there are 89 exceedances of the threshold $u = 5$ in the complete set of 31960 observations, the maximum likelihood estimate of the exceedance probability is $\hat{\zeta_u} = 89/31960 = 0.002784731$, with approximate variance $Var\left(\hat{\zeta_u}\right) = \hat{\zeta_u}\left(1 - \hat{\zeta_u}\right)/31960 = 8.688912 \times 10^{-8}$. Hence, the complete variance-covariance matrix for $\left(\hat{\zeta}, \hat{\sigma}, \hat{\xi}\right)$ is

$$V = \begin{bmatrix} 8.688912 \times 10^{-8} & 0 & 0 \\ 0 & 0.04395658 & -0.01503806 \\ 0 & -0.01503806 & 0.01762163 \end{bmatrix}$$

We focus on the 100-year return level $\hat{x}_m$, thus here number of observations $m = 365 \times 100$. From the theory part, we know that

$$x_m = u + \frac{\sigma}{\xi}\left[(m\zeta_u)^{\xi} - 1\right]$$

and

$$Var\left(\hat{x}_m\right) \approx \nabla x_m^T V \nabla x_m,$$

where

$$\nabla x_m^T = \left[\begin{array}{ccc} \frac{\partial x_m}{\partial \zeta_u}, & \frac{\partial x_m}{\partial \sigma}, & \frac{\partial x_m}{\partial \xi} \end{array} \right]$$
$$= \left[\begin{array}{ccc} \sigma m^{\xi} \zeta_u^{\xi-1}, & \xi^{-1} \left\{ (m\zeta_u)^{\xi} - 1 \right\}, & -\sigma\xi^{-2} \left\{ (m\zeta_u)^{\xi} - 1 \right\} + \sigma\xi^{-1} (m\zeta_u)^{\xi} \log(m\zeta_u) \end{array} \right]$$

evaluated at $\left(\hat{\zeta}, \hat{\sigma}, \hat{\xi}\right)$. Thus, we can substitute into the formulas above and obtain $\hat{x}_m = 20.710722$ and $Var(\hat{x}_m) = 27.80601$, leading to a 95% confidence interval for x_m of $[10.375366, 31.04608]$.

On the next page, we draw the diagnostic plots for the fitted GPD with the threshold $u = 5$ and the return level plot for the model. Since out of the diagnostic plots the probability plot and quantile plot are approximately linear and the straight line fits almost all the data points, it is safe to conclude that the chosen GPD with the threshold $u = 5$ fits the data points pretty well and that the model we chose is valid. We also draw the return level plot separately so that the plot is larger and clearer to see. You can easily find the 100-year return level $\hat{x}_m$ in the plot.

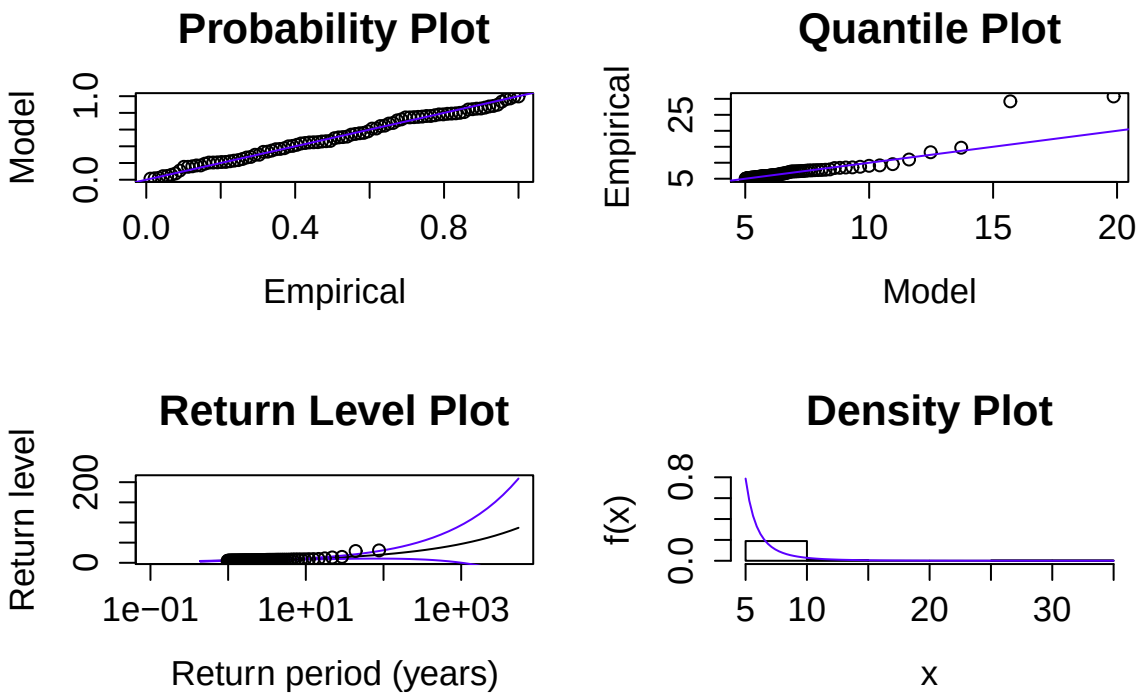

Figure 6: Diagnostic plots

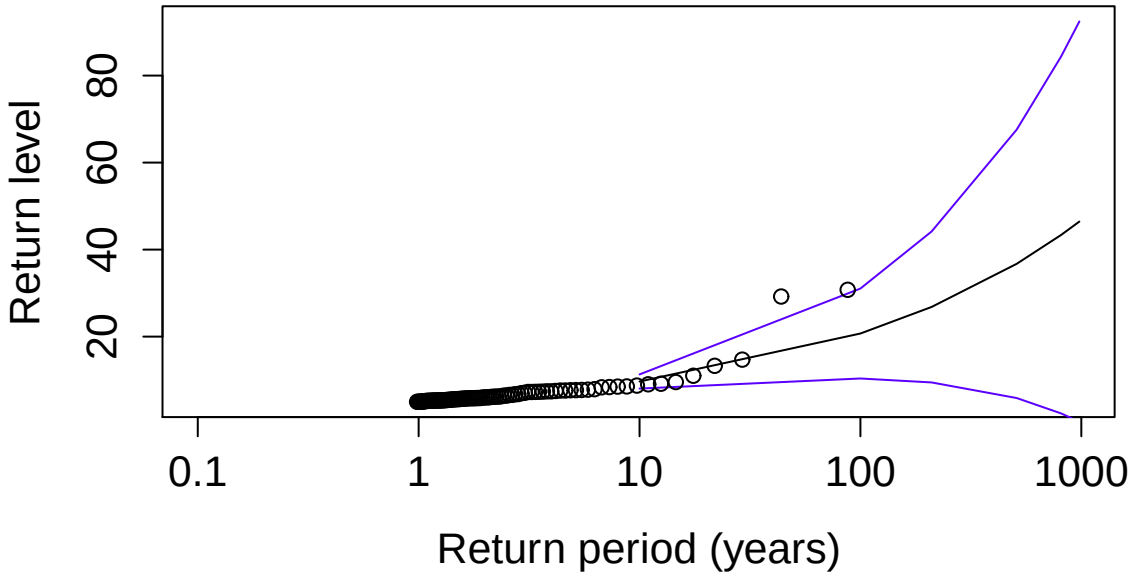

Figure 7: return level plot

3.2.4 Conclusion

We know from the part above that the 95% confidence interval for the 100-year return level of daily loss percentage x_m is $[10.375366, 31.04608]$. It just means: we are 95% confident to say that the maximum daily loss percentage within a hundred years will fall between 10.375366 and 31.04608. Although the range is pretty wide, we can still conclude that the probability of the event that daily loss exceeds 10% occurring within a hundred years is about 0.95, which just means that every 100 years stock market crash with daily loss more than 10% will occur with a very high probability. Therefore, perhaps some precautions can be done to get prepared for the possible stock market crash given that we already know how large the maximum daily loss percentage can be within a hundred years.

In particular, the same statistical method used here for Dow Jones Index can be applied in other stock market around the world, such as Hong Kong, China, Japan and so on. The comparisons of the 100-year return level among these different stock markets would be interesting. In addition, the analysis of data at financial turmoil period could reveal extreme patterns of financial market under huge uncertainty.

3.3 Application in Hong Kong climate data

3.3.1 Introduction

It is a long tradition to use the extreme value analysis to study climate and meteorology problem. Here we study the daily maximum temperature in Hong Kong. Since the Hong Kong climate data is publicly available from the Hong Kong Observatory website: `http://www.weather.gov.hk/cis/data_e.htm`, we can easily have access to the daily data from 1997 till now with detailed information, including mean pressure, max/mean/min air temperature, mean relative humility and so on. However, we will only focus on the maximum air temperature, because our goal is to explore how high the maximum air temperature can be within a hundred years and within a hundred and fifty years respectively. The results can be significant in real life applications, since we then know what the worst situation we are facing is and can take measures accordingly.

We will basically use the similar method we used in the subsection (3.2) **Application in stock market**: Threshold Method from the Extreme Value Modelling. However, in this case it is a little bit more complicated since we have to take into account the non-stationarity of the data points, since obviously the daily maximum air temperature will depend on the month of the year and tend to cluster together. To overcome the difficulty, we will only select the data points from June, July and August, since by observation only the data points in these three months will likely become the maximum air temperature throughout the year. We draw the plot of the data points. In the Figure 8, the x-axis shows the index of 1564 data points, which are numbered in the time order. The y-index shows the daily maximum air temperature for each data point. From the Figure 8, we can easily see that non-stationarity of the data points we chose can be ignored. Therefore, we can simply use the traditional method used before. We obtained the daily maximum air temperature data in June, July and August from 1997 till 2013, totalling 1564 data points. We can use the existing R package "ismev" and "extRemes" to choose the proper threshold, fit the data into the Generalized Pareto Distribution (GPD), and obtain the 100-year and 150-year return level of daily maximum air temperature in Hong Kong with the corresponding 95% confidence interval.

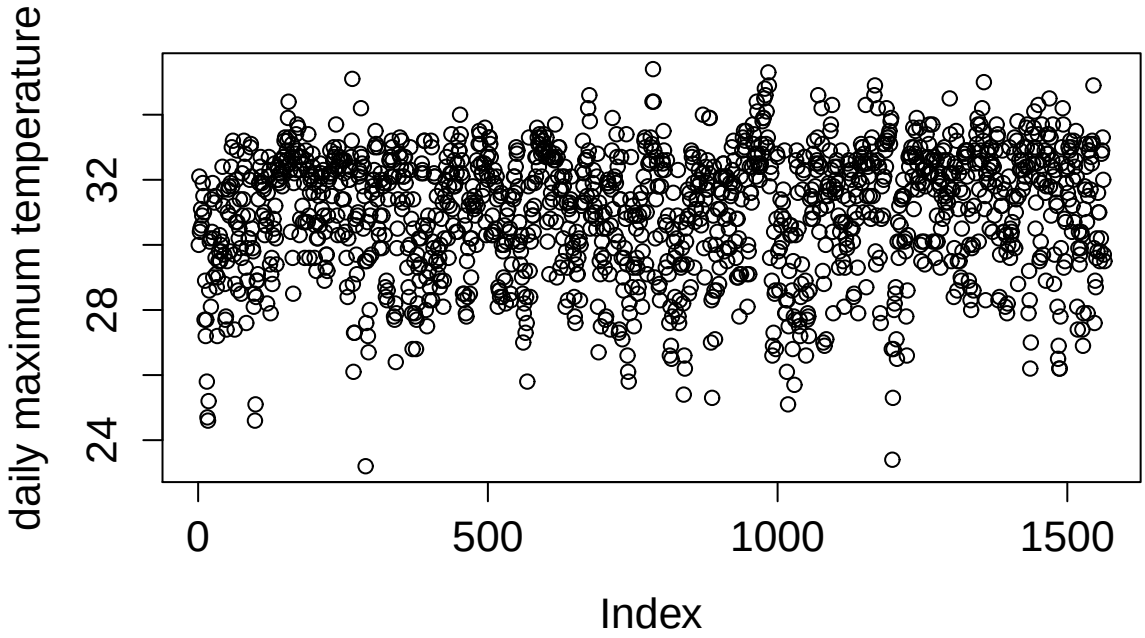

Figure 8: data points of daily maximum temperature of Hong Kong climate data

3.3.2 Selecting the proper threshold

Based on the Figure 8, it is safe to assume that the event of high daily maximum air temperature occurs can be classified into extreme events and that the threshold models method from the extreme event modelling can be used. We use the threshold models method instead of the Block Maxima method, because we want to make the most use of the available data points.

The first procedure when using the threshold models method is to select the proper threshold. As stated before in the theory part, two methods will be used before making a final decision.

The first method is the mean residue life plot should be approximately linear above the proper threshold u_0.

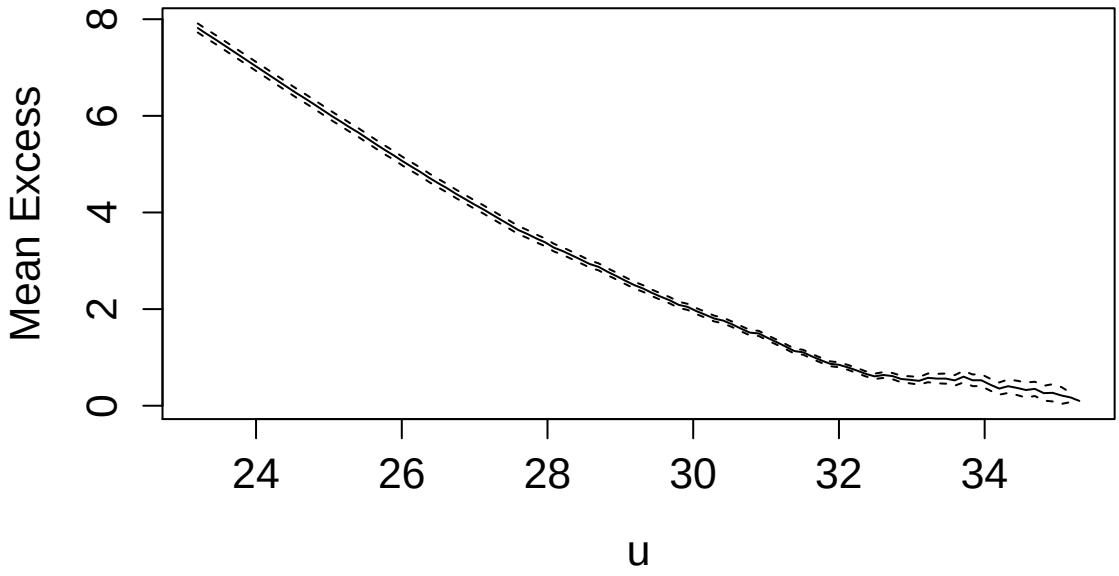

Figure 9: mean residue life plot

In the mean residue life plot above, it is approximately linear above $u = 32$. Therefore, we can conclude that the proper threshold u_0 should satisfy $u_0 \geq 32$.

To further explore what the proper threshold should be, the second method is used: look for the stability of parameters σ^* and ξ while varying the threshold of the fitted GPD.

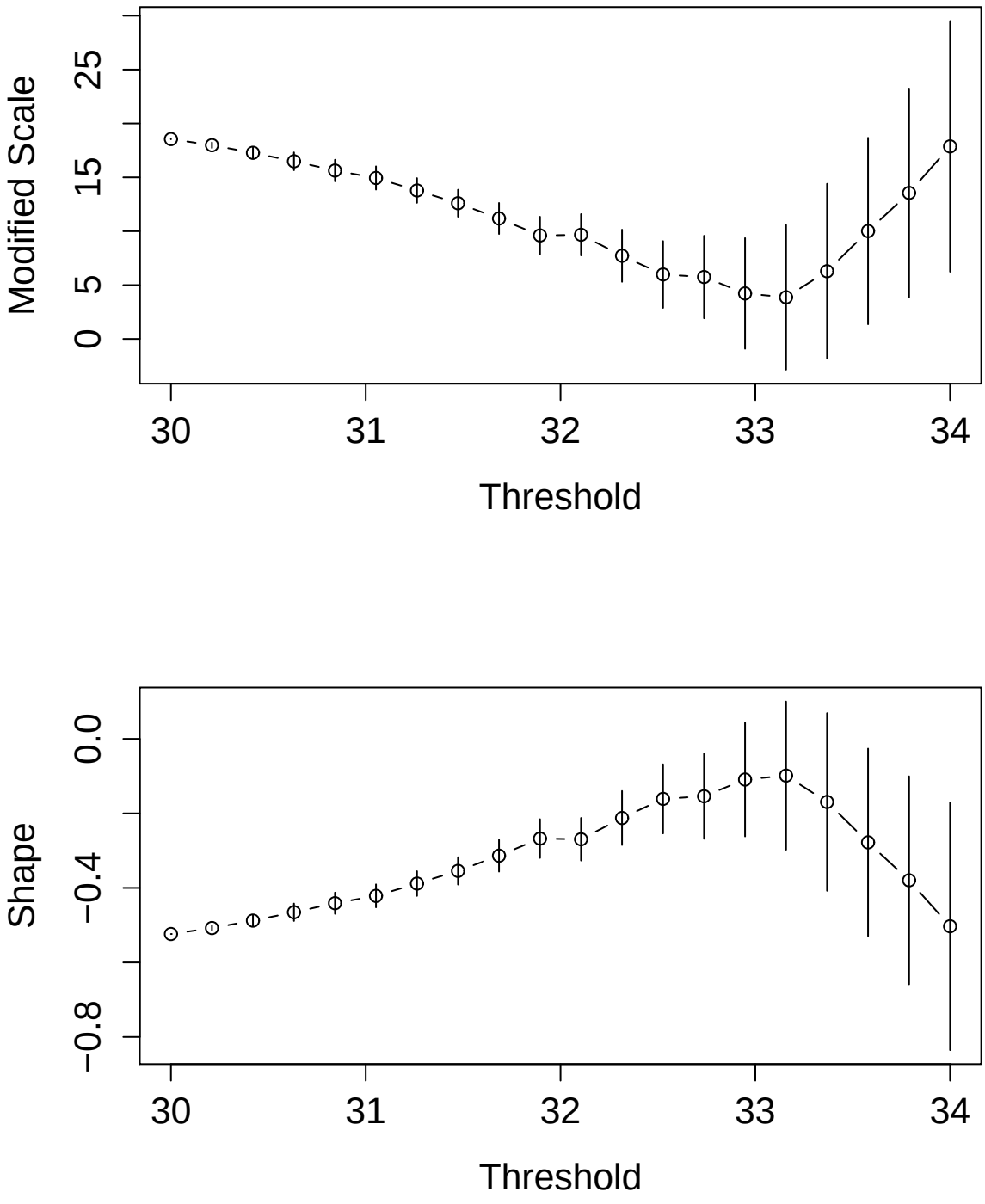

Figure 10: parameter estimates against threshold

From the plot above, we can see that the estimated parameters are more or less stable when $u \geq 32$. Therefore, the selected threshold of $u = 32$ appears reasonable.

3.3.3 Obtaining the 100-year and 150-year return level of daily maximum air temperature

Firstly, we need to fit the data points into the GPD with the selected threshold $u = 32$. The maximum likelihood estimates in this case are

$$\left(\hat{\sigma}, \hat{\xi}\right) = (1.0722162, -0.2809016)$$

with a corresponding maximised log-likelihood of -449.7361. The variance-covariance matrix is calculated as

$$\begin{bmatrix} 0.002685927 & -0.0011527629 \\ -0.001152763 & 0.0007207715 \end{bmatrix}$$

, leading to standard errors of 0.05182593 and 0.02684719 for $\hat{\sigma}$ and $\hat{\xi}$ respectively. In particular, it follows that a 95% confidence interval for ξ is obtained as $-0.2809016 \pm 1.96 \times 0.02684719 = [-0.3335221, -0.2282811]$. Since the maximum likelihood estimate $\xi < 0$, it leads to an bounded distribution in the return level plot and the evidence for this is pretty strong, since the 95% interval for ξ is exclusively in the negative domain.

Since there are 570 exceedances of the threshold $u = 32$ in the complete set of 1564 observations, the maximum likelihood estimate of the exceedance probability is $\hat{\zeta_u} = 570/1564 = 0.3644501$, with approximate variance $Var\left(\hat{\zeta_u}\right) = \hat{\zeta_u}\left(1 - \hat{\zeta_u}\right)/1564 = 1.480986 \times 10^{-4}$. Hence, the complete variance-covariance matrix for $\left(\hat{\zeta}, \hat{\sigma}, \hat{\xi}\right)$ is

$$V = \begin{bmatrix} 1.480986 \times 10^{-4} & 0 & 0 \\ 0 & 0.002685927 & -0.0011527629 \\ 0 & -0.001152763 & 0.0007207715 \end{bmatrix}$$

We firstly calculate the 100-year return level $\hat{x}_m$, thus here number of observations $m = 365 \times 100$. From the theory part, we know that

$$x_m = u + \frac{\sigma}{\xi}\left[(m\zeta_u)^{\xi} - 1\right]$$

and

$$Var\left(\hat{x}_m\right) \approx \nabla x_m^T V \nabla x_m,$$

where

$$\begin{aligned}\nabla x_m^T &= \left[\begin{array}{ccc} \frac{\partial x_m}{\partial \zeta_u}, & \frac{\partial x_m}{\partial \sigma}, & \frac{\partial x_m}{\partial \xi} \end{array} \right] \\ &= \left[\begin{array}{ccc} \sigma m^{\xi} \zeta_u^{\xi-1}, & \xi^{-1}\left\{(m\zeta_u)^{\xi}-1\right\}, & -\sigma\xi^{-2}\left\{(m\zeta_u)^{\xi}-1\right\}+\sigma\xi^{-1}(m\zeta_u)^{\xi}\log(m\zeta_u) \end{array} \right]\end{aligned}$$

evaluated at $\left(\hat{\zeta}, \hat{\sigma}, \hat{\xi}\right)$. Thus, we can substitute into the formulas above and obtain $\hat{x}_m = 35.55201$ and $Var(\hat{x}_m) = 0.02606413$, leading to a 95% confidence interval for x_m of $[35.23559, 35.86844]$.

Similarly we can calculate the 150-year return level $\hat{x}_m$, thus here number of observations $m = 365 \times 150$. We obtain $\hat{x}_m = 35.58055$, leading to a 95% confidence interval for x_m of $[35.25327, 35.90782]$.

On the next page, we draw the diagnostic plots for the fitted GPD with the threshold $u = 32$ and the return level plot for the model. Since out of the diagnostic plots the probability plot and quantile plot are approximately linear and the straight line fits almost all the data points, it is safe to conclude that the chosen GPD with the threshold $u = 32$ fits the data points pretty well and that the model we chose is valid. We also draw the return level plot separately so that the plot is larger and clearer to see. You can easily find the 100-year and 150-year return level $\hat{x}_m$ in the plot.

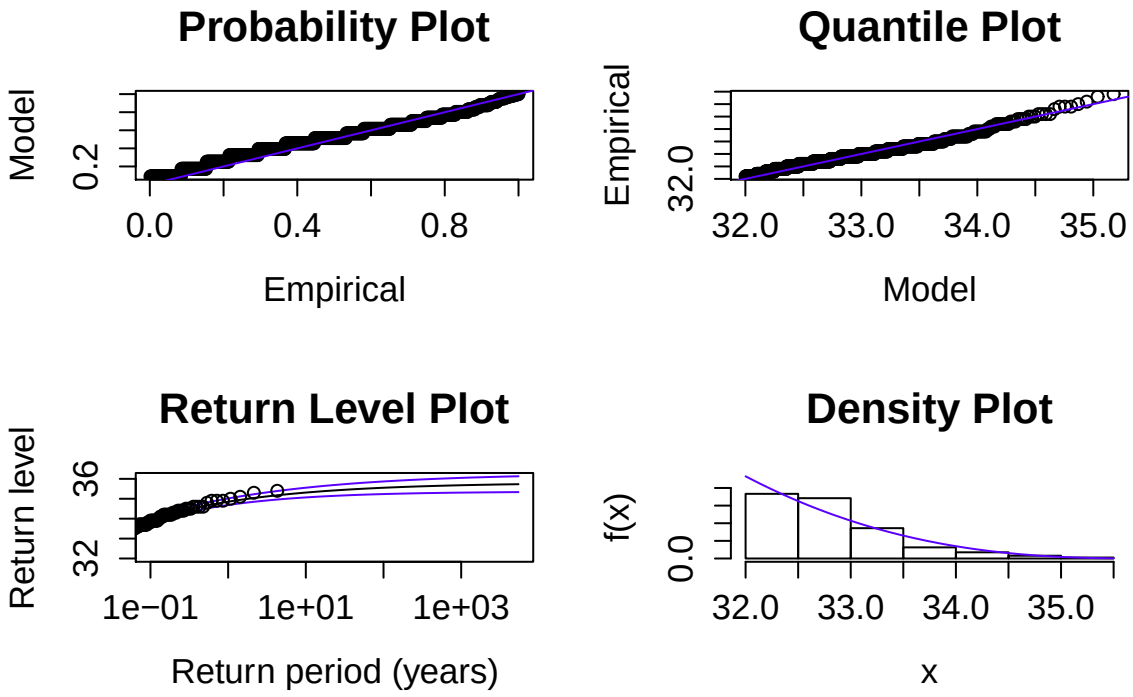

Figure 11: Diagnostic plots

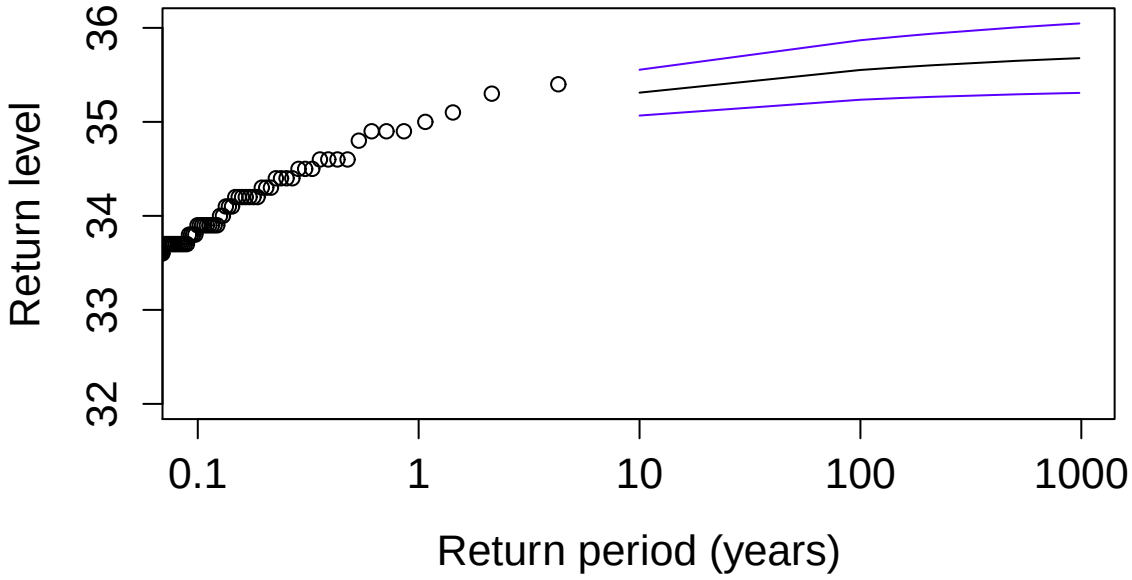

Figure 12: return level plot

3.3.4 Conclusion

We know from the part above that the 95% confidence interval for the 100-year return level of daily maximum air temperature x_m is $[35.23559, 35.86844]$. It just means: we are 95% confident to say that the maximum daily loss percentage within a hundred years will fall between 35.23559 and 35.86844. Similarly, the 150-year return level of daily maximum air temperature x_m is $[35.25327, 35.90782]$, which means: we are 95% confident to say that the maximum daily loss percentage within a hundred and fifty years will fall between 35.25327 and 35.90782. Although the range is pretty wide, we can still conclude that the maximum air temperature will most likely never reach 36 degrees in Hong Kong even within 150 years. Therefore, we know what we are up to and take some precautionary measures in the event of extremely high air temperature.

4 R codes

4.1 Simulation of data from normal distribution

```
n=500

m=100000/n

Z=rep(0,m)

x=rnorm(n=n*m,mean=0,sd=1)

for (i in 1:m){

  Z[i]=max(x[((i-1)*n+1):(i*n)])

}

Z=sort(x=Z,decreasing=F)


ml=gev.fit(Z)

mu=ml$mle[1]

sigma=ml$mle[2]

xi=ml$mle[3]

covariance_matrix=ml$cov

mu_lb=mu-qnorm(p=0.975)*sqrt(covariance_matrix[1,1])

mu_ub=mu+qnorm(p=0.975)*sqrt(covariance_matrix[1,1])

sigma_lb=sigma-qnorm(p=0.975)*sqrt(covariance_matrix[2,2])

sigma_ub=sigma+qnorm(p=0.975)*sqrt(covariance_matrix[2,2])

xi_lb=xi-qnorm(p=0.975)*sqrt(covariance_matrix[3,3])

xi_ub=xi+qnorm(p=0.975)*sqrt(covariance_matrix[3,3])


plot(ml)

gum.diag(gum.fit(Z))

para=c(mu,mu_lb,mu_ub,sigma,sigma_lb,sigma_ub,xi,xi_lb,xi_ub)
```

4.2 Application in stock market

```
table=read.csv("DJIA.csv",stringsAsFactors=FALSE,header=T,sep=";")

which(table$VALUE=="#N/A")

table=table[-which(table$VALUE=="#N/A"),]

write.csv(table,file="modified.csv")

table$VALUE=as.numeric(table$VALUE)

perc=rep(x=0,times=length(table$VALUE))

for (i in 2:length(perc)){

  perc[i]=(table$VALUE[i]-table$VALUE[i-1])/table$VALUE[i]

}

perc=100*perc

plot(perc,ylab= expression("perc "%*% "%"))

table$PERC=perc

table$DATE[which(table$PERC==min(perc))]

table$DATE[which(table$PERC==max(perc))]

write.csv(table,file="modified.csv")


loss=-perc

mrl.plot(loss)

title(xlab="daily loss percentage")

gpd.fitrange(loss,1,9,nint=20)

loss.gpd=gpd.fit(loss,5)

z=return.level(z=loss.gpd, conf = 0.05, rperiods= c(10,100,210,510,810,980), make.plot = TRUE)

z$return.level

z$confidence.delta

gpd.diag(loss.gpd)
```

4.3 Application in Hong Kong climate data

```
y=read.csv("dailyhk.csv",sep=";")
plot(y$Max.Temp..deg..C.,ylab="daily maximum temperature")
mrl.plot(y$Max.Temp..deg..C.)
gpd.fitrange(y$Max.Temp..deg..C., 30, 34, nint = 20)
b.gpd=gpd.fit(y$Max.Temp..deg..C.,32)
gpd.diag(b.gpd)
s=return.level(z=b.gpd, conf = 0.05, rperiods= c(10,100,150,210,510,810,980), make.plot = TRUE)
s$return.level
s$confidence.delta
```

5 Summary and Outlook

In the **Theory and Methods** Part of this paper, we discussed the basic theory and methods of Extreme Value Modelling. Both the Generalized Extreme Value (GEV) distribution and the Generalized Pareto Distribution (GPD) are introduced. Correspondingly, there are two methods to model the Extremes, i.e. the Block Maxima method and the Threshold method. Then in the **Application** Part, we applied these methods to fit the data to obtain the corresponding return level. Firstly, we applied the Block Maxima Method to the simulated normal data. Then we applied the Threshold Method to the Dow Jones Index data and the Hong Kong climate data and reached informative conclusion based on the return levels we obtained.

In the Hong Kong climate data case, we assume that the data are obtained from one station of the same location. However, in real world application, it is possible that the data are obtained from a number of stations of different locations. If so, we can use the Bayesian hierarchical model for spatial extremes to produce a map characterising extreme behaviour across a geographic region. Such methodology is discussed at length in the paper (2). The basic idea of the Bayesian hierarchical model is discussed in the rest of this section.

To produce the return level map, both the exceedances and their rate of occurrence must be modelled,

and we construct separate hierarchical models for each. Hierarchical models allow one to statistically model a complex process and its relationship to observations in several simple components. There are three layers in both of our hierarchal models. The base layer (Data Layer) models the data (either exceedance amounts or number of exceedances) at each station. The second layer (Process Layer) models the latent process that drives the climatological extreme precipitation for the region. The third layer (Priors) consists of the prior distributions of the parameters that control the latent process.

The inference for the parameters in our models θ given the stations data $Z(x)$ comes from the Bayes rule:

$$p(\theta|Z(x)) \propto p(Z(x)|\theta)p(\theta)$$

where p denotes a probability density. Based on the conditional distribution of our hierarchical model, we can get:

$$p(\theta|Z(x)) \propto p_1(Z(x)|\theta_1)p_2(\theta_1|\theta_2)p_3(\theta_2)$$

where p_i is the density associated with level i of the hierarchical model and depends on parameters θ_i.

Based on the equation above, we can obtain the posterior distributions of $\sigma(x)$, $\xi(x)$, and $\zeta(x)$ by using MCMC algorithms. Then, the return level posterior distribution as well as the return level maps can be produced accordingly.

6 References

1. Stuart Coles, *An Introduction to Statistical Modeling of Extreme Values* (New York: Springer, 2001).

2. Daniel Cooley, Douglas Nychka & Philippe Naveau (2007) Bayesian Spatial Modeling of Extreme Precipitation Return Levels, Journal of the American Statistical Association, 102:479, 824-840, DOI: 10.1198/016214506000000780