Nonparametric Smoothing in Extreme Value Theory

Author
John-Craig Clur

Supervisor
Professor Linda Haines

Submitted to the Department of Statistical Science
in partial fulfilment of the requirements for the degree of

Master of Science in Financial Mathematics
at the
University of Cape Town

November 22, 2010

Abstract

This mini-dissertation investigates the modelling of non-stationary sample extremes using a roughness penalty approach, in which smoothed natural cubic splines are fitted to the location and scale parameters of the generalized extreme value distribution and the distribution of the r largest order statistics. Estimation is performed by implementing a Fisher scoring algorithm to maximize the penalized log-likelihood function. The approach provides a flexible framework for exploring smooth trends in sample extremes, with the benefit of balancing the trade-off between 'smoothness' and adherence to the underlying data by simply changing the smoothing parameter. To evaluate the overall performance of the extreme value theory methodology in smoothing extremes a simulation study was performed. Two real data sets, namely extreme motor and property reinsurance claims in selected African countries, were also introduced in order to illustrate the approach of modelling the trend in extremes over time.

Plagiarism Declaration

1. This dissertation is my own work. It has not been submitted before for any degree or examination to any other University.

2. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work

3. Each significant contribution to, and quotation in, this dissertation from the work of other people has been cited and referenced.

Signature: ________________________

Date: ________________________

Acknowledgements

My sincere thanks to my supervisor, Professor Linda Haines, for her invaluable advice and direction, for her commitment and for her continuous support and help.

I am extremely grateful to the Swiss Reinsurance Company Ltd for providing the data used in the application section of this mini-dissertation.

I would also like to thank my parents, Brenda and Norman Clur, my aunt and uncle, Sheila and Eric Clur, and my sister, Alison de Lorm, for their unconditional love and moral support.

A special thanks to my best friend Tracy Jensen for always being there.

Lastly, I would like to thank Shivani Ramjee, Dave Strugnell, Professor Iain MacDonald and Professor Rob Dorrington for their continuous support.

Contents

List of Figures

List of Tables

Glossary of Terms

Abbreviation or Symbol	Explanation
$\rightarrow$	Tends to
$\xrightarrow{P}$	Convergence in probability
$\Lambda(x)$	Standard Gumbel distribution function
$\Theta_\alpha(x)$	Standard Fréchet distribution function
$\Phi_\alpha(x)$	Standard Weibull distribution function
$\Gamma(x)$	Gamma function
$\Gamma^{(m)}(x)$	mth derivative of the gamma function
$\Psi(x)$	Digamma function
$\Psi'(x)$	1st derivative of the digamma function
μ	Location parameter for GEV distribution
σ	Scale parameter for GEV distribution
ξ	Shape parameter for GEV distribution
z_p	The extreme quantile also referred to as the return level
$z^*_{\alpha/2}$	The $(1-\alpha/2)$ quantile of the standard normal distribution
$l(\mu,\sigma,\xi;1)$	Log-likelihood function for the GEV distribution
$l(\mu,\sigma,\xi;r)$	Log-likelihood function for the r largest order statistics with $\xi \neq 0$
$l(\mu,\sigma,0;1)$	Log-likelihood function for the Gumbel distribution
$l(\mu,\sigma,0;r)$	Log-likelihood function for the r largest order statistics with $\xi = 0$
$l^{PLL}(\mu,\sigma,\xi;r)$	Penalized log-likelihood function
GEV	Generalized extreme value
MPLE	Maximum penalized likelihood estimate
NCS	Natural cubic spline

Chapter 1

Introduction

1.1 Background

During one's lifetime, extraordinary situations arise that have never been experienced before, and are unlikely to occur again. For example, the Shaanxi earthquake in 1556, the China floods in 1931, the Bhola cyclone in 1970, the Europe heat wave in 2003 and the 2004 Indian Ocean tsunami. These extreme weather events not only cause widespread damage and loss of life, but also have a devastating financial impact. In addition to extreme weather events, extreme events are also experienced in the financial markets, for example the Wall Street Crash in 1929 and the market crash on 19 October 1987 known to the financial world as Black Monday.

In 1974, Munich Reinsurance Company (Munich Re) set up a department called Geo Risks Research. One of the main responsibilities of this department is to analyse natural catastrophes and loss potentials around the world. The data collected now forms the backbone of the NatCatservice® database. Certain statistics published by Munich Re, which have been taken from the NatCatservice® database, are shown in Figures 1.1 and 1.2, whilst the size of the losses from the 10 costliest natural catastrophes are shown in Table 1.1. It is interesting to observe that there is an increasing trend in both the number of natural catastrophes and overall losses arising from natural catastrophes over time. Furthermore, Table 1.1 identifies the amount of loss which was insured, highlighting the fact that individuals and companies have grossly underestimated the risks and potential financial effects of these extreme events.

NatCatSERVICE

Significant natural catastrophes 1980 – 2009

10 costliest natural catastrophes ordered by overall losses

Munich RE

Period	Event	Affected Area	Overall losses	Insured losses	Fatalities
			US$ m, original values		
25-30.8.2005	Hurricane Katrina	USA: LA, New Orleans, Slidell; MS, Biloxi, Pascagoula, Waveland, Gulfport	125,000	62,000	1,300
17.1.1995	Earthquake	Japan: Prefecture Hyogo, Kobe, Osaka, Kyoto	100,000	3000	6,400
12.5.2008	Earthquake	China: Sichuan, Mianyang, Beichuan, Wenchuan, Shifang, Chengdu, Guangyuan, Ngawa, Ya'an	85,000	300	84,000
17.1.1994	Earthquake	USA: Northridge, Los Angeles, San Fernando Valley, Ventura, Orange	44,000	15,300	60
6-14.9.2008	Hurricane Ike	USA. Cuba. Haiti. Dominican Republic. Turks and Caicos Islands. Bahamas	38,000	18,500	170
May-September 1998	Floods	China: Jangtsekiang, Songhua Jiang	30,700	1,000	4,200
23.10.2004	Earthquakes	Japan: Honshu, Niigata, Ojiya, Tokyo, Nagaoka, Yamakoshi	28,000	760	50
23-27.8.1992	Hurricane Andrew	USA: FL, Homestead; LA. Bahamas	26,500	17,000	60
27.6-13.8.1996	Floods	China: Guizhou, esp. Guiyang; Zhejiang; Sichuan; Hunan; Anhui; Jiangxi; Hubei; Guangxi; Jiangsu	24,000	445	3,050
7-21.9.2004	Hurricane Ivan	USA. Trinidad and Tobago. Venezuela. Colombia. Mexico	23,000	13,800	130

Table 1.1: The 10 costliest natural catastrophes from 1980-2009 ordered by overall losses as provided by Munich Reinsurance Company

It is imperative that insurance companies can correctly assess the potential losses that can arise due to extreme and thus rare events, in order to put aside necessary reserves and purchase appropriate levels of reinsurance. Failure to do so could result in companies running into major financial difficulties. The analysis and modelling surrounding extreme events, especially when assessing the nature of possible future extreme events that could be more extreme than those already observed, is therefore an important task, not only for insurance companies, but also for risk managers, engineers and scientists.

One method of modelling these extreme events, which provides the necessary mathematical and statistical foundations, is extreme value theory. Extreme value theory aims to model the tail behaviour of a distribution explicitly, which is where extremes occur. A mathematical and statistical introduction to extreme value theory can be found in Leadbetter *et al.* (1983), Coles (2001) and de Haan and Ferreira (2006). Embrechts *et al.* (1997) and Beirlant *et al.* (2005) provide further information on the statistical aspects and applications of extreme value

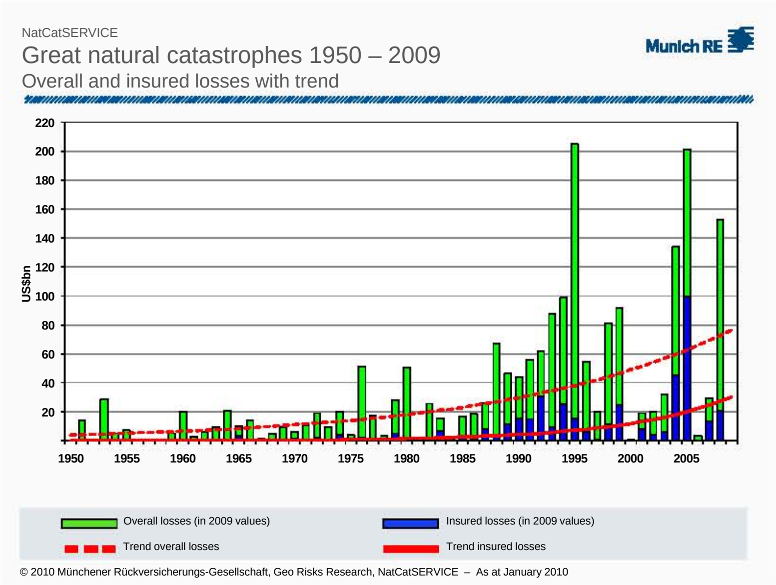

Figure 1.1: The overall and insured losses of great natural catastrophes from 1950 - 2009 as provided by Munich Reinsurance Company

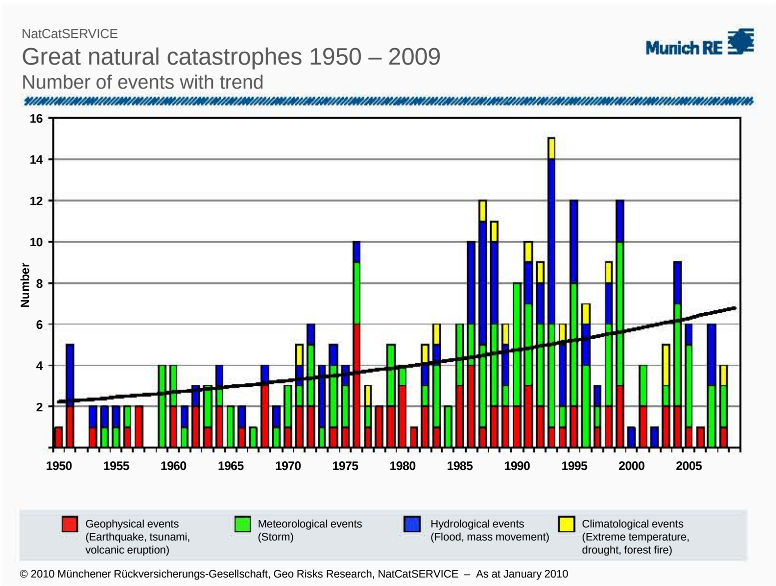

Figure 1.2: The number of great natural catastrophes from 1950 - 2009 as provided by Munich Reinsurance Company

theory. In particular, Embrechts *et al.* (1997) emphasize the application of extreme value theory to insurance and finance. On the other hand, Beirlant *et al.* (2005) demonstrate the wide range of possible applications of extreme value theory including hydrology, meteorology, geology, metallurgy, insurance and finance.

1.2 Rationale for Smoothing Extremes

Time series data of real-world phenomena such as share prices, exchange rates or weather patterns are inherently non-stationary, that is, a random process whose statistical properties vary with time. Within a financial context, non-stationarity is often apparent due to changes in business cycles, interest rates, inflation and technological advancements.

When considering insurance losses, it is evident from Figures 1.1 and 1.2, that large losses arising from natural catastrophes are becoming more frequent and more severe over time. In general, when considering the insurance cycle, large insurance losses may become more or less severe over time, or for that matter, more or less frequent (Chavez-Demoulin and Embrechts, 2004). It is therefore important to be able to account for non-stationarity when attempting to model these extremes.

Coles (2001) provides a good introduction to modelling non-stationary extremes and highlights the fact that the use of parametric techniques to model the scale and shape parameters are becoming standard practice. However, the application of nonparametric techniques allows for smoothing methods to be incorporated and can therefore be seen as preferable to parametric techniques (Laurini and Pauli, 2009). Another advantage of using nonparametric techniques is that it provides a flexible framework for exploring the trends in extremes, with no restrictions placed on potential trends.

1.3 Aims of the mini-dissertation

The aims of this mini-dissertation are as follows:

1. To provide an introduction to classical extreme value theory, which summarizes the main results and, more importantly, which draws together a

number of key results that are scattered in the literature.

2. To introduce both the basic concepts of a natural cubic spline and a roughness penalty and also to discuss how they are used in nonparametric regression.

3. To investigate the performance of modelling non-stationary extremes, using the roughness penalty approach, on simulated data and to illustrate the effects of changing the sample size, the number of extremes and the smoothing parameters on the maximum penalized likelihood estimates.

4. To investigate the application of smoothing extremes on a set of real data.

1.4 Outline of the mini-dissertation

An introduction to the main results from classical extreme value theory, including the Fisher-Tippett Theorem, are presented in Chapter 2. Furthermore, the main properties of the generalized extreme value distribution and the asymptotic distribution for the r largest order statistics are considered. Chapter 3 moves away from extreme value theory to discuss the idea of spline smoothing, with the focus being placed on the natural cubic spline and the use of the integrated squared second derivative as the roughness penalty. The concepts from the Chapters 2 and 3 are then combined together in Chapter 4, in order to consider the methodology for smoothing extremes by penalizing the log-likelihood with a roughness penalty. Chapter 5 begins with an outline for simulating extreme random variates from both the generalized extreme value distribution and the asymptotic distribution for the r largest order statistics. After this outline, a simulation study to investigate the performance of using the roughness penalty approach in modelling non-stationary extremes is reported. The roughness penalty approach in modelling non-stationary extremes is then illustrated on two real datasets in Chapter 6, more specifically reinsurance claims incurred by the Swiss Reinsurance Company Ltd in selected African countries. In the closing chapter, Chapter 7, the mini-dissertation concludes with a brief discussion highlighting the aims which have been achieved and recommendations for future research.

Chapter 2

Classical Extreme Value Theory

The aim of this chapter is to introduce the basic concepts of classical extreme value theory and to present a summary of results that are scattered in the literature. Focus is placed on the asymptotic models for block maxima, for the kth largest order statistic in a block and the r largest order statistics in a block, considering the properties and the statistical inference of the various models. For the purpose of this mini-dissertation, the modelling of extremes using threshold excess models or point processes has not been considered. For information regarding the threshold excess models and the point process characterization of extremes refer to Coles (2001).

The structure of this chapter is as follows. Section 1 introduces the main ideas behind classical extreme value theory, including the Fisher-Tippet Theorem and a brief explanation as to how the extreme value distributions can be combined into one distribution known as the Generalized Extreme Value distribution. Section 2 looks at how the ideas of modelling maximum values in a block can be extended to modelling the minimum values. The properties of the Generalized Extreme Value distribution are then summarized in Section 3, while statistical inference for the Generalized Extreme Value distribution is discussed in Section 4, focusing on maximum likelihood estimation and inference for return levels. Finally, asymptotic models for the kth largest order statistic and r largest order statistics in a block, including their properties and statistical inference, are presented in Section 5 and Section 6.

2.1 Asymptotic Distributions for Maxima

Suppose that $X_1, X_2, \ldots, X_n$ is a sequence of independent and identically distributed random variables having a common distribution function F. Classical extreme value theory is mainly concerned with the distribution function for the maximum $M_n = \max(X_1, X_2, \ldots, X_n)$ and its properties as $n \to \infty$. The distribution function for M_n is easily found to be

$$Pr\,(M_n \le x) = Pr\,(X_1 \le x, \ldots, X_n \le x) = F^n(x).$$

Since extremes occur in the upper or lower tails of a distribution, it is important to try to characterize the tail behaviour of the distribution F. To do this, one needs to consider the asymptotic behaviour of M_n and how this is related to the distribution function F near the end points of the tails as $n \to \infty$.

As pointed out by Coles (2001) and Embrechts *et al.* (1997), the disadvantage of M_n in its current form is that its distribution function will degenerate to a point mass on the right end point $x_+ = sup\{x \in \mathbb{R} : F(x) < 1\}$, in other words $M_n \xrightarrow{P} x_+$ as $n \to \infty$. This provides no further information as to the asymptotic distribution of M_n. The solution to this problem is to normalize the maximum. By choosing an appropriate sequence of normalizing constants $\{a_n > 0\}$ and $\{b_n \in \mathbb{R}\}$, a linear transformation of M_n can be found which stabilizes the location and scale of M_n as n increases, such that

$$Pr\left(\frac{M_n - b_n}{a_n} \le x\right) = F^n(a_n x + b_n) \to G(x). \tag{2.1}$$

where G is a non-degenerate distribution function.

One of the key results in classical extreme value theory is that there are only three possible limiting distributions G for the normalized maximum. This result was derived by Fisher and Tippett (1928) and is presented in the Fisher-Tippett Theorem, also referred to as the Extremal Types Theorem.

Theorem 1. (Fisher-Tippett Theorem)
If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n \in \mathbb{R}\}$ such that

$$Pr\left(\frac{M_n - b_n}{a_n} \le x\right) \xrightarrow{P} G(x) \quad as \quad n \to \infty,$$

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

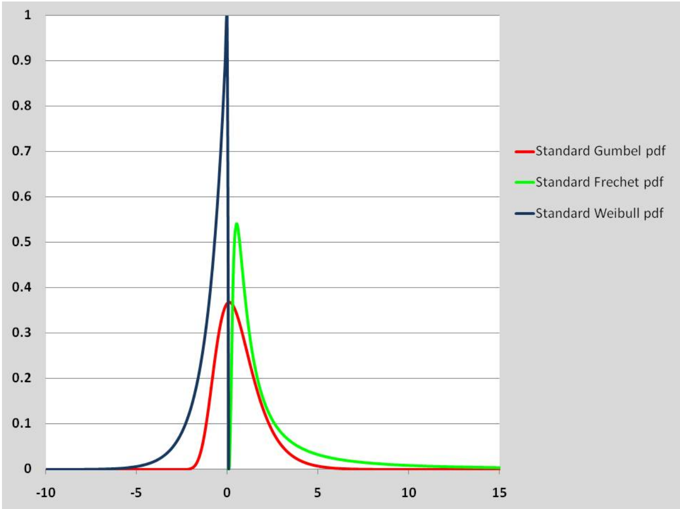

Figure 2.1: Distribution functions of the standard extreme value distributions

following distributions:

$$\text{Gumbel:} \quad \Lambda(x) = \exp\left\{-\exp\left[-\left(\frac{x-b}{a}\right)\right]\right\}, \qquad -\infty < x < \infty \tag{2.2}$$

$$\text{Fréchet:} \quad \Theta_\alpha(x) = \begin{cases} 0, & \text{if } x \leq b \\ \exp\left\{-\left(\frac{x-b}{a}\right)^{-\alpha}\right\}, & \text{if } x > b \end{cases} \tag{2.3}$$

$$\text{Weibull:} \quad \Phi_\alpha(x) = \begin{cases} \exp\left\{-\left[-\left(\frac{x-b}{a}\right)\right]^{\alpha}\right\}, & \text{if } x < b \\ 1 & \text{if } x \geq b \end{cases} \tag{2.4}$$

for parameters $a > 0$ and $b \in \mathbb{R}$ and for $\alpha > 0$ in the case of the Fréchet and Weibull distributions.

For a sketch of the proof of the Fisher-Tippett Theorem see Embrechts *et al.* (1997) and for a more complete proof see Leadbetter *et al.* (1983). The distribution functions $\Lambda(x), \Theta_\alpha(x)$ and $\Phi_\alpha(x)$ as presented in the Fisher-Tippett Theorem are called the extreme value distributions and are also commonly referred to as Extreme Value Type I, Type II and Type III distributions respectively. The standard extreme value distributions are special cases with $a = 1$ and $b = 0$ and their densities are shown in Figure 2.1.

The possible non-degenerate distribution functions G that occur in the limit of (2.1) form a class of max-stable distributions, that is the distribution function G is such that, for each $n = 2, 3, \ldots$, there are constants $\{a_n > 0\}$ and $\{b_n \in \mathbb{R}\}$

such that $G^n(a_n x + b_n) = G(x)$. One can also show that a distribution is max-stable if, and only if, it is one of the three extreme value distributions.

The Fisher-Tippett Theorem has one major implication, that regardless of the distribution function F, the extreme value distributions are the only possible limit distributions for a maximum for which the scale and the location have been stabilized.

If (2.1) holds for some sequences of constants $\{a_n > 0\}$ and $\{b_n \in \mathbb{R}\}$, then F is said to belong to the domain of attraction of G, written $F \in D(G)$. For example, some of the distributions $F \in D(\Phi_\alpha)$ include the beta and uniform distributions, $F \in D(\Theta_\alpha)$ include the Cauchy and log-gamma distributions, and $F \in D(\Lambda)$ include the gamma, normal, log-normal and exponential distributions. However, it should be pointed out that there are certain cases where the non-degenerate limiting distribution G for the maximum does not exist under any linear normalization. Two common examples are when the sequence of independent and identically distributed random variables X_i for $i = 1, 2, \ldots$ are from a Poisson or a geometric distribution. For more information regarding conditions under which $F \in D(G)$, refer to Leadbetter *et al.* (1983) or Embrechts *et al.* (1997).

The extreme value distributions are closely linked from a mathematical point of view. If X is a random variable such that $X > 0$, one can show that the extreme value distributions are related as follows:

$$\text{“}\, X \text{ has df } \Theta_\alpha \quad \Longleftrightarrow \quad \ln X^\alpha \text{ has df } \Lambda \quad \Longleftrightarrow \quad -X^{-1} \text{ has df } \Phi_\alpha \,\text{”}$$

where df refers to the distribution function (Embrechts *et al.*, 1997). Furthermore, the three extreme value distributions can be combined into one distribution known as the Generalized Extreme Value (GEV) distribution. The GEV distribution is given by

$$G(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \tag{2.5}$$

where the support is defined by $1 + \xi\,(x-\mu)/\sigma > 0$ and the parameters satisfy $-\infty < \mu < \infty, \sigma > 0$ and $-\infty < \xi < \infty$. The parameters μ, σ and ξ represent the location, scale and shape parameters for the GEV distribution.

The Fréchet distribution specified in (2.3) corresponds to the GEV distribution when $\xi > 0$ and setting $\xi = 1/\alpha, \sigma = b/\alpha$ and $\mu = a + b$. The Weibull

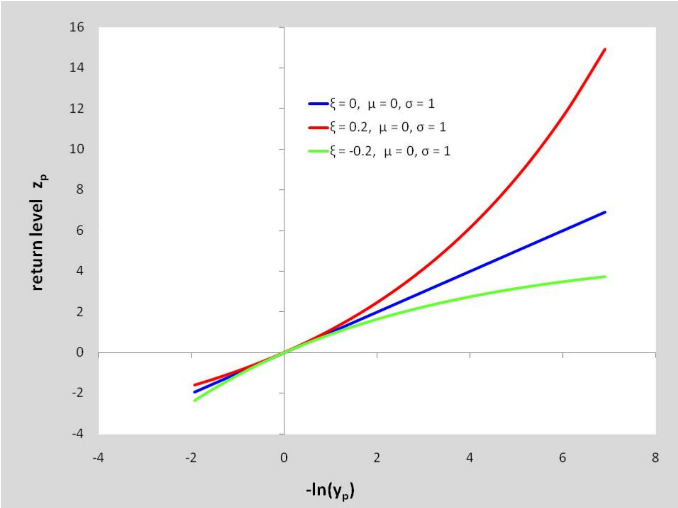

Figure 2.2: Example of Return Level Plots

distribution specified in (2.4) corresponds to the GEV distribution when $\xi < 0$ and setting $\xi = -1/\alpha, \sigma = b/\alpha$ and $\mu = a - b$. The Gumbel distribution corresponds to the case when $\xi = 0$ and is interpreted as the limit of (2.5) as $\xi \to 0$ to give (2.2) with $\sigma = b$ and $\mu = a$.

The extreme quantile z_p of the GEV distribution, associated with the upper tail probability p, that is $G(z_p) = 1 - p$, is obtained by inverting equation (2.5), so that

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-\ln(1-p)\}^{-\xi}\right], & \text{for} \quad \xi \neq 0 \\ \mu - \sigma \ln\{-\ln(1-p)\}, & \text{for} \quad \xi = 0 \end{cases} . \tag{2.6}$$

The extreme quantile z_p is more commonly referred to as the return level z_p associated with the return period $1/p$. When considering annual data, one would expect the the return level z_p to be exceeded every $1/p$ years by an annual maximum (Coles, 2001). By defining $y_p = -\ln(1-p)$, and plotting z_p against $-\ln(y_p)$ one obtains a return level plot. Return level plots are often used for model presentation and validation. As can be seen from Figure 2.2, the return level plot is linear when $\xi = 0$, concave when $\xi < 0$ and convex when $\xi > 0$.

2.2 Asymptotic Distributions for Minima

Extremes happen at both ends of the spectrum, in both large and small quantities. One must therefore be able to model minima as well as maxima. As in the previous section, let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables having a common continuous distribution function F. The minimum m_n is given by $m_n = \min(X_1, X_2, \ldots, X_n)$.

One method of modelling the minimum m_n is to exploit the relationship between maximum and minimum, where

$$m_n = \min(X_1, X_2, \ldots, X_n) = -\max(-X_1, -X_2, \ldots, -X_n) = -\tilde{M}_n.$$

The results for block maxima can therefore easily be extended to block minima using this relationship.

Another approach is to model the minimum directly by fitting the appropriate limiting distribution for minimum. To determine the distribution function for the minimum m_n note that $Pr(m_n \leq x) = Pr(-\tilde{M}_n \leq x) = 1 - Pr(\tilde{M}_n \leq -x)$. The GEV distribution for a minimum is therefore given by

$$G(x) = 1 - \exp\left\{ -\left[1 - \xi\left(\frac{x+\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \right\}, \tag{2.7}$$

where the support is defined by $1 - \xi\,(x+\mu)\,/\sigma > 0$ and the parameters satisfy $-\infty < \mu < \infty, \sigma > 0$ and $-\infty < \xi < \infty$.

Due to the fact that minima and maxima are closely related, minima will not be considered any further. For further details on minima see Leadbetter *et al.* (1983) and Coles (2001).

2.3 Properties of the GEV Distribution

When modelling the GEV distribution and considering its properties, one needs to separate the shape parameter ξ into two scenarios, one when $\xi \neq 0$, that is for the Fréchet and Weibull distributions, and the other when $\xi = 0$, that is for the Gumbel distribution.

2.3.1 Properties when $\xi = 0$

Within the GEV distribution, the Gumbel distribution corresponds to the case when $\xi = 0$. The distribution function for the Gumbel is obtained by considering

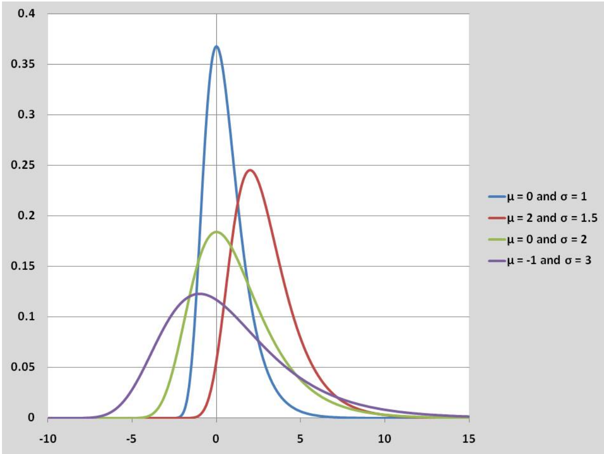

Figure 2.3: Densities of different Gumbel distributions

the limit of (2.5) as $\xi \to 0$, leading to

$$G(x) = \exp\left[-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right] \tag{2.8}$$

where $-\infty < x < \infty$. Differentiating with respect to x gives the probability density function

$$g(x) = \frac{1}{\sigma}\exp\left[-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right]\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right].$$

Several Gumbel densities are shown in Figure 2.3 to illustrate the effects of changing the location and scale parameters.

In order to find the characteristic function, the Gumbel random variable X is first transformed into a standard Gumbel random variable using $Z = \frac{X-\mu}{\sigma}$. The density of Z is then given by

$$p(z) = e^{-e^{-z}}e^{-z}$$

for $-\infty < z < \infty$. The characteristic function of Z is easily determined to be

$$\varphi_Z(t) = E\left[e^{itZ}\right] = \int_{-\infty}^{\infty} e^{itz}e^{-e^{-z}}e^{-z}dz = \int_0^{\infty} u^{-it}e^{-u}du = \Gamma(1-it)$$

where $i = \sqrt{-1}$ and the characteristic function of $X = \mu + \sigma Z$, which follows the Gumbel distribution, is therefore given by

$$\varphi_X(t) = E\left[e^{itX}\right] = e^{it\mu}\Gamma(1-it\sigma).$$

This can be used to determine the moment generating function by considering $\varphi_X(-it)$ to give

$$M_X(t) = E\big[e^{tx}\big] = e^{t\mu}\Gamma(1 - t\sigma).$$

The moment generating function can then be used to obtain the mean and variance of X. By considering the first and second derivatives of $M_X(t)$

$$M'_X(t) = \mu e^{t\mu}\Gamma(1 - t\sigma) - \sigma e^{t\mu}\Gamma^{(1)}(1 - t\sigma)$$
$$M''_X(t) = \mu^2 e^{t\mu}\Gamma(1 - t\sigma) - 2\mu\sigma e^{t\mu}\Gamma^{(1)}(1 - t\sigma) + \sigma^2 e^{t\mu}\Gamma^{(2)}(1 - t\sigma)$$

and evaluating these derivatives at $t = 0$, the first two moments of X are obtained as

$$E[X] = \mu - \sigma\Gamma^{(1)}(1) = \mu + \sigma\gamma$$
$$E[X^2] = \mu^2 - 2\sigma\mu\Gamma^{(1)}(1) + \sigma^2\Gamma^{(2)}(1) = (\mu + \sigma\gamma)^2 + \sigma^2\frac{\pi^2}{6}$$

where $\gamma = 0.5772156649...$ is Euler's constant. It then follows that the variance is equal to

$$Var(X) = \sigma^2\Psi'(1) = \sigma^2\frac{\pi^2}{6}.$$

A more general expectation for the standard Gumbel random variable $Z = \frac{X-\mu}{\sigma}$ given by Smith (1986) is

$$E[Z^m \exp(-cZ)] = (-1)^m\Gamma^{(m)}(1 + c) \tag{2.9}$$

for $c > -1$. This result is used later when determining certain expectations in the Fisher scoring algorithm. Furthermore, by setting $c = 0$ in equation (2.9), it is possible to obtain moments for the standard Gumbel random variable Z, that is

$$E[Z^m] = (-1)^m\Gamma^{(m)}(1),$$

which can then be used to determine the moments of the Gumbel random variable X. For completeness, the median of the Gumbel distribution is given by

$$x_{(m)} = \mu - \sigma\Big(\ln(\ln 2)\Big)$$

and the mode by μ. A summary of the properties for the Gumbel distribution is given in Table 2.1.

Distribution Function	$\exp\left[-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right]$
Density Function	$\frac{1}{\sigma}\exp\left[-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right]\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]$
Characteristic Function	$e^{it\mu}\Gamma(1-it\sigma)$
Expected Value	$\mu+\sigma\gamma$
Variance	$\sigma^2\frac{\pi^2}{6}$
Median	$\mu-\sigma\Big(\ln(\ln 2)\Big)$
Mode	μ

Table 2.1: Summary of the properties for the Gumbel distribution

2.3.2 Properties when $\xi \neq 0$

The shape parameter ξ has an impact on the support and tail behaviour of the GEV distribution. The support of the GEV distribution is given by $1 + \xi\left(\frac{x-\mu}{\sigma}\right) > 0$. When $\xi > 0$, the distribution is skewed to the right and the support is bounded to the left with $\mu - \sigma/\xi < x$. When $\xi < 0$, the distribution is skewed to the left and the support is bounded to the right with $x < \mu - \sigma/\xi$. To see the effect that the shape parameter has on the GEV distribution, consider Figures 2.4 and 2.5. The effects of changing the scale and shape parameter are shown in Figure 2.6.

The probability density function for the GEV is obtained by differentiating $G(x)$ in expression (2.5) to give

$$g(x) = \frac{1}{\sigma}\exp\left\{-\left[1+\xi\Big(\frac{x-\mu}{\sigma}\Big)\right]^{-\frac{1}{\xi}}\right\}\left[1+\xi\Big(\frac{x-\mu}{\sigma}\Big)\right]^{-\left(\frac{1}{\xi}+1\right)}$$

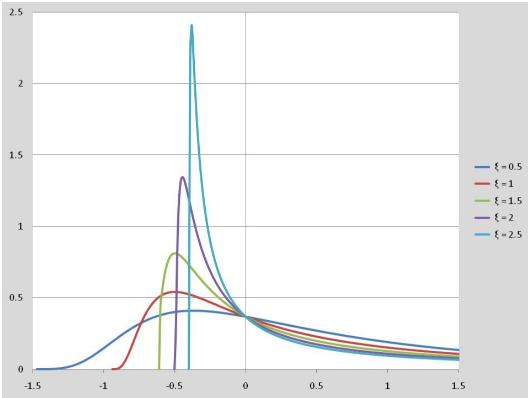

Figure 2.4: GEV density functions with $\mu = 0$, $\sigma = 1$ and $\xi > 0$

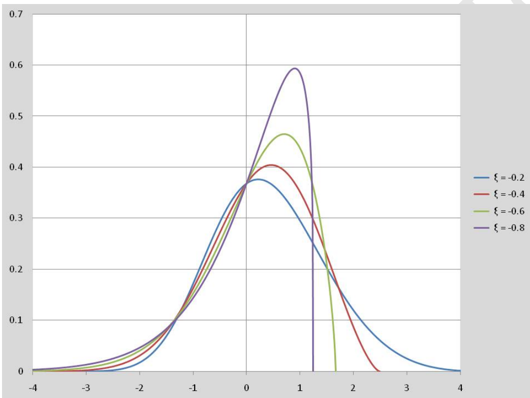

Figure 2.5: GEV density functions with $\mu = 0$, $\sigma = 1$ and $\xi < 0$

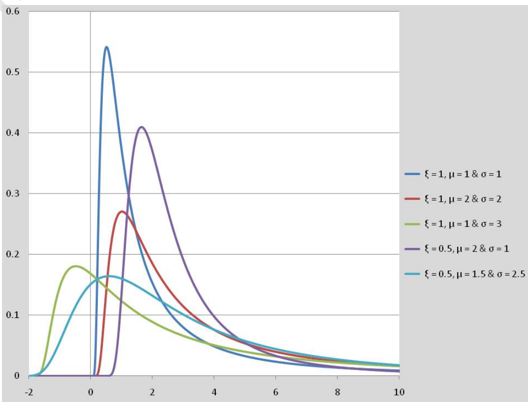

Figure 2.6: GEV density functions with $\xi > 0$

with support defined by $1+\xi\left(\frac{x-\mu}{\sigma}\right)>0$ and the parameters satisfy $-\infty<\mu<\infty, \sigma>0$ and $-\infty<\xi<\infty$. One method of calculating the moments for the GEV distribution is to transform the GEV random variable X into a standard Gumbel random variable. This is achieved by setting

$$e^{-W}=\left[1+\xi\Big(\frac{X-\mu}{\sigma}\Big)\right]^{-\frac{1}{\xi}},$$

which implies that $W=\frac{1}{\xi}\ln\left[1+\xi\Big(\frac{X-\mu}{\sigma}\Big)\right]$ and $X=\mu-\frac{\sigma}{\xi}+\frac{\sigma}{\xi}\,e^{W\xi}$. After the transformation, it is easily seen that W is a standard Gumbel random variable. One can then use the moment generating function of the standard Gumbel, that is $M_W(t)=\Gamma(1-t)$, to determine the moments of the GEV distribution. The first two moments are calculated as follows

$$\begin{aligned} E[X] &= E\left[\mu-\frac{\sigma}{\xi}+\frac{\sigma}{\xi}\,e^{W\xi}\right] \\ &= \mu-\frac{\sigma}{\xi}+\frac{\sigma}{\xi}\,M_W(\xi)=\mu-\frac{\sigma}{\xi}+\frac{\sigma}{\xi}\,\Gamma(1-\xi) \qquad (2.10) \end{aligned}$$

and

$$\begin{aligned} E[X^2] &= E\left[\left(\mu-\frac{\sigma}{\xi}\right)^2+2\left(\mu-\frac{\sigma}{\xi}\right)\frac{\sigma}{\xi}\,e^{W\xi}+\frac{\sigma^2}{\xi^2}\,e^{2W\xi}\right] \\ &= \left(\mu-\frac{\sigma}{\xi}\right)^2+2\left(\mu-\frac{\sigma}{\xi}\right)\frac{\sigma}{\xi}\,M_W(\xi)+\frac{\sigma^2}{\xi^2}\,M_W(2\xi) \\ &= \left(\mu-\frac{\sigma}{\xi}\right)^2+2\left(\mu-\frac{\sigma}{\xi}\right)\frac{\sigma}{\xi}\,\Gamma(1-\xi)+\frac{\sigma^2}{\xi^2}\,\Gamma(1-2\xi). \end{aligned}$$

Higher moments can be calculated in a similar way. It is worthwhile noting that the first moment only exists when $\xi<1$ and the second moment only exists when $\xi<\frac{1}{2}$. The existence of the n^{th} moment $E[X^n]$ will only exists if $\xi<\frac{1}{n}$. Therefore, if $\xi<0$ the GEV distribution has finite moments and if $\xi>0$ the GEV distribution has finite moments of order less than $\frac{1}{\xi}$. Using the first two moments, one can then calculate the variance of the GEV variable X as follows

$$Var(X) = E[X^2]-(E[X])^2=\frac{\sigma^2}{\xi^2}\left[\Gamma(1-2\xi)-\Big(\Gamma(1-\xi)\Big)^2\right].$$

For $b\in\mathbb{R}$ and positive integers c and m, Tawn (1988) gives a more general expectation for a standard GEV random variable $Z=\frac{X-\mu}{\sigma}$ as

$$E[Y]=(-\xi)^{c-m}\sum_{p=0}^{m}(-1)^p\binom{m}{p}\Gamma^{(c)}(2+b\,\xi-p\,\xi) \qquad (2.11)$$

where $Y=Z^m(1+\xi Z)^{-\left(\frac{1}{\xi}+b\right)}\left(\ln(1+\xi Z)\right)^c$. The general expectation given by equation (2.11) is used to determine the expectations of the second derivative of

Distribution Function	$\exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$
Density Function	$\frac{1}{\sigma}\exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\left(\frac{1}{\xi}+1\right)}$
Expected Value	$\mu-\frac{\sigma}{\xi}+\frac{\sigma}{\xi}\Gamma(1-\xi)$
Variance	$\frac{\sigma^2}{\xi^2}\left[\Gamma(1-2\xi)-\left(\Gamma(1-\xi)\right)^2\right]$
Median	$\mu+\frac{\sigma}{\xi}[(\ln 2)^{-\xi}-1]$
Mode	$\mu+\frac{\sigma}{\xi}[(1+\xi)^{-\xi}-1]$

Table 2.2: Summary of the properties for the GEV distribution

the likelihood function in the Fisher scoring algorithm. In addition, the general expectation given by equation (2.11) can also be used to determine the moments of the GEV distribution. By setting $c = 0$ and $b = -\frac{1}{\xi}$ this implies that $Y = Z^m$ and

$$E[Y] = E[Z^m] = (-\xi)^{-m}\sum_{p=0}^{m}(-1)^p\binom{m}{p}\Gamma(1-p\,\xi).$$

For completeness, the median of the GEV distribution is given by

$$x_{(m)} = \mu + \frac{\sigma}{\xi}[(\ln 2)^{-\xi} - 1]$$

and the mode by

$$\mu + \frac{\sigma}{\xi}[(1+\xi)^{-\xi} - 1].$$

A summary of the properties for the GEV distribution is given in Table 2.2.

2.4 Modelling Extremes

In practice, a random sample $y_1, y_2, \ldots, y_{nN}$ of observed data is available. These observations are subdivided into consecutive blocks of length n. This then generates a series of sample block maxima $x_1, x_2, \ldots, x_N$, that is $x_i = \max(y_{n(i-1)+1}, y_{n(i-1)+2}, \ldots, y_{ni})$ for $i = 1, 2, \ldots, N$, which can be used to fit the GEV distribution.

The choice of block size n is critical as it amounts to a trade-off between bias and variance (Coles, 2001). The block size n needs to be large enough so that the asymptotic result of the Fisher-Tippet Theorem will hold, at least approximately. If the block size n is too small, the approximation will be poor, leading to parameter estimates that are biased. If the block size n is too large, few sample block maxima will be generated. This will increase the uncertainty in the parameter estimates, resulting in a larger variance.

Once the data has been blocked and the sample block maxima found, various methods are available to estimate the parameters μ, σ and ξ in the GEV model. These include the method of moments, the method of probability-weighted moments, the elemental percentile method, the quantile least squares method, and the maximum likelihood method.

In this mini-dissertation, our attention will be restricted to the maximum likelihood method. Details regarding the other methods can be found in Castillo *et al.* (2005). In addition, Embrechts *et al.* (1997) provides a detailed explanation relating to the estimation of the shape parameter ξ using Pickard's Estimator, Hill's Estimator and the Deckers-Einmahl-de Haan Estimator.

2.4.1 Maximum Likelihood Estimation

Given a series of observed sample block maxima $x_1, x_2, \ldots, x_N$ which are assumed to to be independent realizations from a random variable each having a GEV distribution with probability density function $f(x; \mu, \sigma, \xi)$, the likelihood function is given by

$$L(\mu, \sigma, \xi \,|\, x_1, x_2, \ldots, x_N) = f(x_1, x_2, \ldots, x_N; \mu, \sigma, \xi) = \prod_{i=1}^{N} f(x_i; \mu, \sigma, \xi)$$

and the log-likelihood function by

$$l(\mu,\sigma,\xi\,|\,x_1,x_2,\ldots,x_N) = \ln L(\mu,\sigma,\xi|\,x_1,x_2,\ldots,x_N) = \sum_{i=1}^{N} \ln f(x_i;\mu,\sigma,\xi).$$

For convenience, the log-likelihood function is denoted by $l(\mu,\sigma,\xi)$. The maximum likelihood estimators $\hat{\mu},\hat{\sigma}$ and $\hat{\xi}$ are the parameter values that maximize the log-likelihood function. In the case of the GEV when $\xi \neq 0$, the log-likelihood function is given explicitly as

$$l(\mu,\sigma,\xi) \;=\; \sum_{i=1}^{N}\Big(-\ln(\sigma) - \Big[1+\xi\Big(\frac{x_i-\mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}} - \Big(\frac{1}{\xi}+1\Big)\ln\Big[1+\xi\Big(\frac{x_i-\mu}{\sigma}\Big)\Big]\Big) \tag{2.12}$$

with the restriction that $1+\xi\left(\frac{x_i-\mu}{\sigma}\right) > 0$ for $i = 1,2,\ldots,N$. Differentiating equation (2.12) with respect to μ,σ and ξ yields the likelihood equations for the GEV case:

$$\frac{\partial l(\mu,\sigma,\xi)}{\partial\mu} \;=\; \sum_{i=1}^{N}\left(-\frac{1}{\sigma}\Big[1+\xi\Big(\frac{x_i-\mu}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} + \frac{1}{\sigma}(1+\xi)\Big[1+\xi\Big(\frac{x_i-\mu}{\sigma}\Big)\Big]^{-1}\right) = 0$$

$$\begin{aligned}\frac{\partial l(\mu,\sigma,\xi)}{\partial\sigma} \;=\; \sum_{i=1}^{N}\Bigg(&-\frac{1}{\sigma}\left(\frac{x_i-\mu}{\sigma}\right)\Big[1+\xi\Big(\frac{x_i-\mu}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \\ &+\left(\frac{1+\xi}{\sigma}\right)\left(\frac{x_i-\mu}{\sigma}\right)\Big[1+\xi\Big(\frac{x_i-\mu}{\sigma}\Big)\Big]^{-1} - \frac{1}{\sigma}\Bigg) = 0\end{aligned}$$

$$\begin{aligned}\frac{\partial l(\mu,\sigma,\xi)}{\partial\xi} \;=\; \sum_{i=1}^{N}\Bigg(&-\frac{1}{\xi^2}\Big[1+\xi\Big(\frac{x_i-\mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}}\ln\Big(1+\xi\Big(\frac{x_i-\mu}{\sigma}\Big)\Big) \\ &+\frac{1}{\xi}\left(\frac{x_i-\mu}{\sigma}\right)\Big[1+\xi\Big(\frac{x_i-\mu}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} + \frac{1}{\xi^2}\ln\Big[1+\xi\Big(\frac{x_i-\mu}{\sigma}\Big)\Big] \\ &-\Big(\frac{1}{\xi}+1\Big)\Big(\frac{x_i-\mu}{\sigma}\Big)\Big[1+\xi\Big(\frac{x_i-\mu}{\sigma}\Big)\Big]^{-1}\Bigg) = 0\end{aligned}$$

In the case of the GEV when $\xi = 0$, the Gumbel log-likelihood function is given by

$$l(\mu,\sigma,0) = -N\ln(\sigma) - \sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right) - \sum_{i=1}^{N}\exp\Big[-\Big(\frac{x_i-\mu}{\sigma}\Big)\Big] \tag{2.13}$$

and the likelihood equations are

$$\frac{\partial l(\mu,\sigma,0)}{\partial \mu} = \sum_{i=1}^{N} \frac{1}{\sigma}\Big\{1-\exp\Big[-\Big(\frac{x_i-\mu}{\sigma}\Big)\Big]\Big\} = 0$$

$$\frac{\partial l(\mu,\sigma,0)}{\partial \sigma} = \frac{1}{\sigma}\sum_{i=1}^{N}\Big\{-1+\frac{x_i-\mu}{\sigma}\Big(1-\exp\Big[-\Big(\frac{x_i-\mu}{\sigma}\Big)\Big]\Big)\Big\} = 0.$$

The likelihood equations for both the GEV and Gumbel case have no analytical solution. Therefore, the maximum likelihood estimates $\hat{\mu}, \hat{\sigma}$ and $\hat{\xi}$ must be obtained using standard numerical optimization algorithms such as Quasi-Newton numerical maximization. For the case when $\xi \neq 0$, constrained optimization must be implemented with the constraint that $1+\xi\left(\frac{x_i-\mu}{\sigma}\right) > 0$ for $i = 1, 2, \ldots, N$. With modern technology and programs such as Matlab and R, this is generally not a serious problem.

When using the maximum likelihood approach, one must be aware of the regularity conditions that are required to ensure that the maximum likelihood estimators for a parameter θ have the usual asymptotic properties, that is $\hat{\theta} \sim N\left(\theta, I(\theta)^{-1}\right)$ where $I(\theta)^{-1}$ is the inverse of the Fisher information matrix. Due to the fact that the support of the GEV distribution depends on the model parameters μ, σ and ξ, the regularity conditions are not necessarily satisfied and therefore the maximum likelihood estimators may not have the usual asymptotic properties.

Smith (1985) found that the asymptotic properties of the maximum likelihood estimates for the GEV distribution are dependant on the shape parameter ξ. When $\xi > -0.5$, the maximum likelihood estimators have the usual asymptotic properties but when $\xi < -0.5$ they do not have the standard asymptotic properties. However, the maximum likelihood estimators can generally be obtained when $-1 < \xi < -0.5$ but often do not exist when $\xi < -1$.

2.4.2 Inference for Return Levels

One desirable property of maximum likelihood estimates is the invariance property. The invariance property states that if $\hat{\theta}$ is the maximum likelihood estimate of θ, then $h(\hat{\theta})$ is the maximum likelihood estimate of $h(\theta)$. Therefore, the maximum likelihood estimate of the extreme quantile z_p is obtained by substituting the maximum likelihood estimates $\hat{\mu}, \hat{\sigma}$ and $\hat{\xi}$ into equation (2.6), so that

$$\hat{z}_p = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left[1 - \{-\ln(1-p)\}^{-\hat{\xi}}\right], & \text{for} \quad \hat{\xi} \neq 0 \\ \hat{\mu} - \hat{\sigma}\ln\{-\ln(1-p)\}, & \text{for} \quad \hat{\xi} = 0 \end{cases} .$$

In addition, the variance of $\hat{z}_p$ can be obtained by the delta method and therefore, using a normal approximation,

$$\hat{z}_p \sim N\left(z_p\,,\, \nabla z_p^T V \nabla z_p\right)$$

where V is the variance-covariance matrix of $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ and $\nabla z_p^T = \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \xi}\right]$ are evaluated at the maximum likelihood estimates $\hat{\mu}, \hat{\sigma}$ and $\hat{\xi}$ (Coles, 2001). Using these results, an approximate $100(1-\alpha)\%$ confidence interval for the extreme quantile z_p, is given by

$$z_p \in \left(\hat{z}_p \pm z^*_{\alpha/2}\sqrt{\nabla z_p^T V \nabla z_p}\right)$$

where $z^*_{\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution.

2.5 The kth Largest Extreme

Extremes are normally considered to be rare events. However, there are situations where additional extreme events may have occurred within a particular period that are possibly more extreme than the extremes in other periods. One therefore needs be be able to extend the model for block maxima to include other extreme order statistics.

2.5.1 Asymptotic Distributions for the kth Largest Extreme

As before, let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables having a common continuous distribution function F. Define the kth largest order statistic by

$$M_n^{(k)} = k\text{th largest of}\,\{X_1, X_2, \ldots, X_n\}$$

so that $M_n^{(n)} \leq M_n^{(n-1)} \leq \ldots \leq M_n^{(1)}$. The distribution function for the kth largest order statistic is given in the following theorem.

Theorem 2. (Coles, 2001)

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n \in \mathbb{R}\}$ such that

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

for some non-degenerate distribution function G, so that G is the GEV distribution function given by (2.5), then, for fixed k,

$$Pr\left(\frac{M_n^{(k)} - b_n}{a_n} \leq x\right) \xrightarrow{P} G_k(x) \quad as \quad n \to \infty,$$

on $\{z : 1 + \xi(x - \mu)/\sigma > 0\}$, where

$$G_k(x) = \exp\{-\tau(x)\} \sum_{s=0}^{k-1} \frac{\tau(x)^s}{s!} \tag{2.14}$$

with

$$\tau(x) = \left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}$$

The above result can be restated to give the asymptotic distribution of $M_n^{(k)}$ in terms of that of the maximum M_n (Leadbetter *et al.*, 1983). Suppose there exist a sequence of constants $\{a_n > 0\}$ and $\{b_n \in \mathbb{R}\}$ such that,

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

for some non-degenerate distribution function G. Then, for each $k = 1, 2, \ldots,$

$$Pr\left(\frac{M_n^{(k)} - b_n}{a_n} \leq x\right) \xrightarrow{P} G(x) \sum_{s=0}^{k-1} \frac{(-\ln G(x))^s}{s!} \tag{2.15}$$

when $G(x) > 0$ and zero when $G(x) = 0$. The proof can be found in Leadbetter *et al.* (1983). Two interesting observations can be drawn from the result of (2.15). Firstly, the limiting distribution for the kth largest order statistic is based on the same distribution function G as that of the maximum. Secondly, the same normalizing constants $\{a_n > 0\}$ and $\{b_n \in \mathbb{R}\}$ are used for all k, including the case for the maximum when $k = 1$ (Leadbetter *et al.*, 1983).

2.5.2 Properties of the kth Largest Extreme

As with the properties for the GEV distribution, it is important to consider the cases when $\xi \neq 0$ and when $\xi = 0$ separately. In addition, by setting $k = 1$, the same properties as for the GEV distribution will be obtained.

For the case when $\xi \neq 0$, Tawn (1988) gives the expectation

$$E\big[Y\big] = \frac{(-\xi)^{c-m}}{\Gamma(j)} \sum_{p=0}^{m} (-1)^p \binom{m}{p} \Gamma^{(c)}(j + b\,\xi - p\,\xi + 1) \tag{2.16}$$

where $Y = \left(Z^{(k)}\right)^m \left(1 + \xi Z^{(k)}\right)^{-\left(\frac{1}{\xi}+b\right)} \left(\ln\left(1 + \xi Z^{(k)}\right)\right)^c$ and $Z^{(k)}$ is the standard kth largest order statistic given by $Z^{(k)} = \frac{X^{(k)}-\mu}{\sigma}$. The first two moments of the standard kth largest order statistic when $\xi \neq 0$ are obtained by setting $m = 1$, $b = -\frac{1}{\xi}$, $c = 0$ in equation (2.16) so that

$$\begin{aligned} E\left[Z^{(k)}\right] &= -\frac{1}{\xi}\left(1 - \frac{\Gamma(k-\xi)}{\Gamma(k)}\right) \\ E\left[\left(Z^{(k)}\right)^2\right] &= \frac{1}{\xi^2}\left(1 - 2\frac{\Gamma(k-\xi)}{\Gamma(k)} + \frac{\Gamma(k-2\xi)}{\Gamma(k)}\right). \end{aligned}$$

These can then be used to determine the first two moments of the kth largest order statistic $X^{(k)}$ as

$$\begin{aligned} E\left[X^{(k)}\right] &= \mu + \sigma E\left[Z^{(k)}\right] = \mu - \frac{\sigma}{\xi}\left(1 - \frac{\Gamma(k-\xi)}{\Gamma(k)}\right) \\ E\left[\left(X^{(k)}\right)^2\right] &= E\left[\left(\mu + \sigma Z^{(k)}\right)^2\right] = \mu^2 + 2\mu\sigma E\left[Z^{(k)}\right] + \sigma^2 E\left[\left(Z^{(k)}\right)^2\right] \\ &= \left(\mu - \frac{\sigma}{\xi}\right)^2 + 2\left(\mu - \frac{\sigma}{\xi}\right)\frac{\sigma}{\xi}\frac{\Gamma(k-\xi)}{\Gamma(k)} + \frac{\sigma^2}{\xi^2}\frac{\Gamma(k-2\xi)}{\Gamma(k)}. \end{aligned}$$

The variance of the kth largest order statistic $X^{(k)}$ is then easily calculated to be

$$\begin{aligned} Var\left[X^{(k)}\right] &= E\left[\left(X^{(k)}\right)^2\right] - \left(E\left[X^{(k)}\right]\right)^2 \\ &= \frac{\sigma^2}{\xi}\left(\frac{\Gamma(k-2\xi)}{\Gamma(k)} - \left(\frac{\Gamma(k-\xi)}{\Gamma(k)}\right)^2\right). \end{aligned}$$

When $\xi = 0$, it is possible to derive the density function of the kth largest order statistic which is given by

$$f\left(x^{(k)}\right) = \frac{1}{\sigma\Gamma(k)} \exp\left[-\exp\left[-\left(\frac{x^{(k)}-\mu}{\sigma}\right)\right]\right] \exp\left[-k\left(\frac{x^{(k)}-\mu}{\sigma}\right)\right]$$

where the parameters satisfy $-\infty < \mu < \infty$ and $\sigma > 0$. Smith (1986) gives the expectation

$$E\left[\left(Z^{(k)}\right)^m \exp\left(-\rho Z^{(k)}\right)\right] = \frac{(-1)^m \Gamma^{(m)}(k+\rho)}{\Gamma(k)} \tag{2.17}$$

where $Z^{(k)} = \frac{X^{(k)}-\mu}{\sigma}$ is the standard kth largest order statistic with $\xi = 0$. The first two moments of $Z^{(k)}$ are determined by setting $\rho = 0$ and $m = 1, 2$ in equation (4.10) to give

$$E\left[Z^{(k)}\right] = \frac{(-1)\Gamma^{(1)}(k)}{\Gamma(k)} = -\Psi(k)$$
$$E\left[\left(Z^{(k)}\right)^2\right] = \frac{\Gamma^{(2)}(k)}{\Gamma(k)} = \Psi'(k) + (\Psi(k))^2$$

and these can then be used to determine the first two moments of the kth largest order statistic $X^{(k)}$ as

$$\begin{aligned} E\left[X^{(k)}\right] &= E\left[\mu + \sigma Z^{(k)}\right] = \mu - \sigma\Psi(k) \\ E\left[\left(X^{(k)}\right)^2\right] &= E\left[\left(\mu + \sigma Z^{(k)}\right)^2\right] = \mu^2 + 2\mu\sigma E\left[Z^{(k)}\right] + \sigma^2 E\left[\left(Z^{(k)}\right)^2\right] \\ &= (\mu - \sigma\Psi(k))^2 + \sigma^2\Psi'(k). \end{aligned}$$

The variance of the kth largest order statistic with $\xi = 0$ is then given by

$$Var\left(X^{(k)}\right) = E\left[\left(X^{(k)}\right)^2\right] - \left(E\left[X^{(k)}\right]\right)^2 = \sigma^2\Psi'(k).$$

The moment generating function for $Z^{(k)}$ is obtained by setting $m = 0$ in equation (4.10), so that

$$E\left[e^{\rho Z^{(k)}}\right] = \frac{\Gamma(k-\rho)}{\Gamma(k)}$$

and therefore the moment generating function for the kth largest order statistic $X^{(k)}$ is given by

$$E\left[e^{\rho X^{(k)}}\right] = e^{\rho\mu} E\left[e^{\sigma\rho Z^{(k)}}\right] = e^{\rho\mu}\frac{\Gamma(k-\sigma\rho)}{\Gamma(k)}. \tag{2.18}$$

The moment generating function in (2.18) provides a second method of obtaining the moments of the kth largest order statistic $X^{(k)}$.

2.6 The r Largest Extremes

The advantage of considering the distribution of the kth largest order statistic in isolation is that one can readily obtain its properties. However, given that $M_n^{(k)} \leq M_n^{(k-1)}$ for $k = 2, 3, \ldots, n$, $M_n^{(k-1)}$ will not be independent of $M_n^{(k)}$ and the distribution for $M_n^{(k)}$ will influence the distribution of $M_n^{(k-1)}$. As such, when modelling the r largest extremes, one needs to be able to specify the joint distribution of the r largest order statistics $\left(M_n^{(1)}, M_n^{(2)}, \ldots, M_n^{(r)}\right)$.

2.6.1 Asymptotic Joint Distributions for the r Largest Extremes

Suppose there exist a sequence of constants $\{a_n > 0\}$ and $\{b_n \in \mathbb{R}\}$. Then Tawn (1988) gives the joint distribution function for the r largest order statistics as

$$P\left(\frac{M_n^{(1)} - b_n}{a_n} < x^{(1)}, \frac{M_n^{(2)} - b_n}{a_n} < x^{(2)}, \ldots, \frac{M_n^{(r)} - b_n}{a_n} < x^{(r)}\right)$$
$$\xrightarrow{p} \sum_{s_1=0}^{1} \sum_{s_2=0}^{2-s_1} \cdots \sum_{s_{r-1}=0}^{r-1-s_1-\ldots s_{r-2}} \frac{(\lambda_2 - \lambda_1)^{s_1}}{s_1!} \cdots \frac{(\lambda_r - \lambda_{r-1})^{s_{r-1}}}{s_{r-1}!} e^{-\lambda_r}$$

where $\lambda_i = \lambda(x^{(i)}) = \left(1 + \xi x^{(i)}\right)^{-\frac{1}{\xi}}$ and $x_i^{(r)} < x_i^{(r-1)} < \ldots < x_i^{(1)}$ with support $1 + \xi\left(x^{(k)} - \mu\right)/\sigma > 0$ for $k = 1, 2, \ldots, r$. Leadbetter *et al.* (1983) emphasize that the limiting joint distribution function for the r largest order statistics has a very complicated form. However, for the case when r = 2, a relatively simple function arises and is given by

$$P\left(X^{(1)} \leq x^{(1)}, X^{(2)} \leq x^{(2)}\right) = G\left(x^{(2)}\right)\left[\ln G\left(x^{(1)}\right) - \ln G\left(x^{(2)}\right) + 1\right]$$

where G is the GEV distribution function given by (2.5).

Compared to the complicated form of the limiting joint distribution function for the r largest order statistics, the limiting joint density function has a fairly simple form. Coles (2001) gives the limiting asymptotic joint density function of the r largest order statistics with $\xi \neq 0$ as

$$\begin{aligned} f(x^{(1)}, \ldots, x^{(r)}) &= \exp\left[-\left[1 + \xi\left(\frac{x^{(r)} - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right] \prod_{k=1}^{r} \left\{\frac{1}{\sigma}\left[1 + \xi\left(\frac{x^{(k)} - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1}\right\} \\ &= G\left(x^{(r)}\right) \prod_{k=1}^{r} \frac{g\left(x^{(k)}\right)}{G\left(x^{(k)}\right)} \end{aligned} \tag{2.19}$$

where $x^{(r)} \leq x^{(r-1)} \leq \ldots \leq x^{(1)}$ with support $1 + \xi\left(x^{(k)} - \mu\right)/\sigma > 0$ for $k = 1, 2, \ldots, r$ and the parameters satisfy $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$.

When $\xi = 0$, the joint density for the r largest order statistics is interpreted as the limit of (2.19) as $\xi \to 0$ to give

$$f(x^{(1)}, \ldots, x^{(r)}) = \exp\left[-\exp\left[-\left(\frac{x^{(r)} - \mu}{\sigma}\right)\right]\right] \prod_{k=1}^{r} \left\{\frac{1}{\sigma} \exp\left[-\left(\frac{x^{(k)} - \mu}{\sigma}\right)\right]\right\}. \tag{2.20}$$

In the case of $r = 1$, (2.19) and (2.20) reduce to the densities of the GEV and Gumbel distributions respectively.

2.6.2 The r Largest Standard Extremes

When considering the the standard extreme value distributions, that is the distributions specified by (2.2), (2.3) and (2.4) with $a = 1$ and $b = 0$, David and Nagaraja (2003) showed that the limiting joint distribution for the r largest standard extremes corresponds to the joint distribution of

$$\lambda^{-1}(Y_1), \lambda^{-1}(Y_1 + Y_2), \ldots, \lambda^{-1}(Y_1 + Y_2 + \ldots + Y_r) \tag{2.21}$$

where $\lambda^{-1}(y)$ is an inverse function and Y_i are independent exponential variates with mean 1. For $y > 0$, $\lambda^{-1}(y) = y^{-\frac{1}{\alpha}}$ for the standard Fréchet distribution, $\lambda^{-1}(y) = -y^{\frac{1}{\alpha}}$ for the standard Weibull distribution and $\lambda^{-1}(y) = -\ln y$ for the standard Gumbel distribution. This result is utilized later in the simulation study to simulate the r largest order statistics.

2.6.3 Modelling the r Largest Extremes

In practice, a random sample $y_1, y_2, \ldots, y_{nN}$ of observed data is available. These observations are subdivided into consecutive blocks of length n. From each block the r largest observations are extracted, leading to a series containing the r largest order statistics $M_i^{(r)} = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(r)})$ for $i = 1, 2, \ldots, N$, where $x_i^{(k)} = k$th largest of $\big(y_{n(i-1)+1}, y_{n(i-1)+2}, \ldots, y_{ni}\big)$ and $x_i^{(r)} < x_i^{(r-1)} < \ldots < x_i^{(1)}$. The series $(M_1^{(r)}, M_2^{(r)}, \ldots, M_N^{(r)})$ can then be used to fit the distribution for the r largest order statistics.

The choice of block size n and number of order statistics r used for each block is critical, as it amounts to a trade-off between bias and variance (Coles, 2001). As with the GEV model, the block size n needs to be large enough so that the asymptotic results hold at least approximately. Furthermore, the number of order statistics r must be small in comparison to the number of observations within each block for the asymptotic results to hold at least approximately (Tawn, 1988).

Once the data has been blocked and the r largest order statistics extracted from each block, the parameters μ, σ and ξ can be estimated by maximizing the likelihood function. When $\xi \neq 0$, the likelihood function is given by

$$L(\mu, \sigma, \xi, r) = \prod_{i=1}^{N} \left(\exp\left[-\Big[1+\xi\Big(\frac{x_i^{(r)} - \mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}} \right] \prod_{k=1}^{r} \frac{1}{\sigma} \Big[1+\xi\Big(\frac{x_i^{(k)} - \mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}-1} \right)$$

with the restriction that $1 + \xi(x_i^{(k)} - \mu)/\sigma > 0$ for $k = 1, 2, \ldots, r$ and $i = 1, 2, \ldots, N$. The log-likelihood function is then given by

$$l(\mu,\sigma,\xi,r) = \sum_{i=1}^{N} \left[-\left[1+\xi\left(\frac{x_i^{(r)}-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} - r\ln(\sigma) - \sum_{k=1}^{r}\left(\frac{1}{\xi}+1\right)\ln\left[1+\xi\left(\frac{x_i^{(k)}-\mu}{\sigma}\right)\right]\right].$$

Furthermore, when $\xi = 0$ the likelihood function is given by

$$L(\mu,\sigma,0,r) = \prod_{i=1}^{N}\left(\exp\left[-\exp\left[-\left(\frac{x_i^{(r)}-\mu}{\sigma}\right)\right]\right] \prod_{k=1}^{r} \frac{1}{\sigma}\exp\left[-\left(\frac{x_i^{(k)}-\mu}{\sigma}\right)\right]\right)$$

and the log-likelihood function by

$$l(\mu,\sigma,0,r) = \sum_{i=1}^{N}\left[-\exp\left[-\left(\frac{x_i^{(r)}-\mu}{\sigma}\right)\right] - r\ln(\sigma) - \sum_{k=1}^{r}\left(\frac{x_i^{(k)}-\mu}{\sigma}\right)\right].$$

In addition, when $\xi \neq 0$ the maximum likelihood estimates $\hat{\mu}, \hat{\sigma}$ and $\hat{\xi}$ are found by constrained optimization and when $\xi = 0$ by direct optimization.

Chapter 3

Nonparametric Smoothing

In this chapter, the ideas and concepts behind nonparametric regression using the roughness penalty approach are introduced. Most of the definitions and theorems have been drawn from Green and Silverman (1994) and have been presented in an attempt to introduce the idea of 'smoothing' data.

The layout of this chapter is as follows. Firstly, cubic splines and natural cubic splines are defined and certain optimal features of natural cubic splines are briefly presented in Section 1. Then Section 2 examines how natural cubic splines can be represented mathematically, focusing mainly on the value-second derivative representation. The idea of how cubic splines are 'smoothed' by penalizing the residual sum of squares with a roughness penalty is discussed in Section 3, while different methods of choosing the smoothing parameter are discussed in Section 4. Finally, in Section 5, the idea of penalizing the log-likelihood function with a roughness penalty is introduced.

3.1 Introduction to Cubic Splines

Suppose $t_1, t_2, \ldots, t_n$ are real numbers on some interval $[a, b]$ satisfying $a < t_1 < t_2 < \ldots < t_n < b$. A function $g(t)$ defined on $[a, b]$ is said to be a cubic spline if the following two conditions are satisfied:

1. $g(t)$ is a cubic polynomial on each of the intervals $(a, t_1), (t_1, t_2), \ldots, (t_n, b)$.

2. The polynomial pieces fit together at the points t_i (called knots) in such a way that $g(t), g'(t)$ and $g''(t)$ are continuous at each knot t_i, and hence on the whole of $[a, b]$.

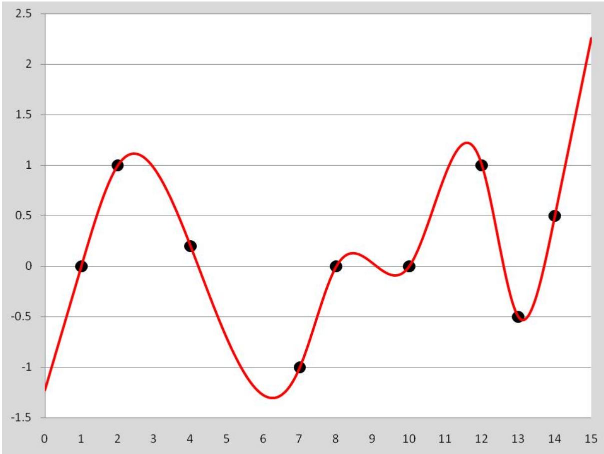

Figure 3.1: Example of a natural cubic spline fitted to a random sample of points on the interval [0,15]

A natural cubic spline (NCS) is a cubic spline $g(t)$ that satisfies a third condition, known as the natural boundary condition. The natural boundary condition is that the second and third derivatives of $g(t)$ are zero at the end points of the interval $[a, b]$, that is $g''(a) = g''(b) = g'''(a) = g'''(b) = 0$. Since $g(t)$ is a cubic polynomial, the natural boundary conditions imply that $g(t)$ is linear on the two extreme intervals (a, t_1) and (t_n, b). An example of a NCS is shown in Figure 3.1.

Natural cubic splines have two very important features. Firstly, given points (t_i, y_i), for $i = 1, \ldots, n$ with $t_1 < t_2 < \ldots < t_n$, it is possible to obtain a unique NCS, $g(t)$, with knots at the points t_i, that perfectly interpolates the points (t_i, y_i) such that $g(t_i) = y_i$ for $i = 1, \ldots, n$. Secondly, one can prove mathematically that on some interval $[a, b]$, the NCS $g(t)$ will minimize the value of $\int_a^b \{f''(x)\}^2 dx$ among all functions $f(t)$ that are differentiable on $[a, b]$ and have an absolutely continuous first derivative, that is $\int_a^b \{g''(x)\}^2 dx \leq \int_a^b \{f''(x)\}^2 dx$. This is sometimes referred to as the minimum curvature property.

3.2 Mathematical Representation of a Natural Cubic Spline

There are various different mathematical representations for a NCS, each with their own advantages and disadvantages. For example, one can use the NCS conditions directly to obtain the following mathematical representation. Firstly, $g(t)$ is a continuous cubic polynomial on each interval (t_i, t_{i+1}) for $i = 0, 1, \ldots, n$ with $t_0 = a$ and $t_{n+1} = b$. This implies that for $t_i \leq t \leq t_{i+1}$

$$g(t) = d_i(t - t_i)^3 + c_i(t - t_i)^2 + b_i(t - t_i) + a_i$$

for given constants a_i, b_i, c_i and d_i. In addition, the continuity condition implies that $g(t_{i+1}) = a_{i+1} = d_i(t_{i+1} - t_i)^3 + c_i(t_{i+1} - t_i)^2 + b_i(t_{i+1} - t_i) + a_i$ and the natural boundary conditions imply that $d_0 = c_0 = d_n = c_n = 0$. Although, this representation has the advantage of explaining the definition of a NCS from a mathematical perspective, it is not computationally efficient.

Natural cubic splines also have a natural representation in terms of the radial basis function (Ruppert *et al.*, 2003), where the cubic spline is defined as

$$g(t) = \beta_0 + \beta_1 t + \sum_{i=1}^{n} \beta_{1i}|t - t_i|^3.$$

A more suitable representation of a NCS, which is both mathematically efficient and tractable, is known as the value-second derivative representation (Green and Silverman, 1994).

3.2.1 Value-Second Derivative Representation

The value-second derivative representation of a NCS $g(t)$, with knots $t_1 < t_2 < \ldots < t_n$, depends on two column vectors, $\mathbf{g}$ and $\boldsymbol{\gamma}$, and two tridiagonal matrices, $\mathbf{Q}$ and $\mathbf{R}$, which are defined below. For simplicity, define $g_i = g(t_i)$ and $\gamma_i = g''(t_i)$ for $i = 1, 2, \ldots, n$. Due to the natural boundary conditions, $\gamma_1 = 0$ and $\gamma_n = 0$. The two column vectors are then defined by $\mathbf{g} = (g_1, g_2, \ldots, g_n)^T$ and $\boldsymbol{\gamma} = (\gamma_2, \gamma_3, \ldots, \gamma_{n-1})^T$.

For the two tri-diagonal matrices, let the spacing between successive knots be given by $h_i = t_{i+1} - t_i$ for $i = 1, 2, \ldots, n-1$. The matrix $\mathbf{Q}$ is then defined as an $n \times (n-2)$ matrix with entries q_{ij}, for $i = 1, 2, \ldots, n$ and $j = 2, 3, \ldots, n-1$, given by

$$q_{j-1,j} = h_{j-1}^{-1}, \quad q_{j,j} = -h_{j-1}^{-1} - h_j^{-1}, \quad q_{j+1,j} = h_j^{-1}$$

for $j = 2, \ldots, n-1$, and $q_{i,j} = 0$ for $|i - j| \geq 2$. In addition, the matrix $\mathbf{R}$ is defined to be an $(n-2) \times (n-2)$ matrix with entries r_{ij} for $i = 2, \ldots, n-1$

and $j = 2, \ldots, n-1$ with

$$r_{i,i} = \frac{1}{3}(h_{i-1} + h_i) \quad \text{for } i = 2, \ldots, n-1,$$

$$r_{i,i+1} = r_{i+1,i} = \frac{1}{6}h_i \quad \text{for } i = 2, \ldots, n-2$$

and

$$r_{i,j} = 0 \quad \text{for } |i-j| \geq 2.$$

Both $\mathbf{R}$ and $\mathbf{Q}$ are tridiagonal matrices and hence banded matrices with bandwidth 3. It is also worth pointing out that $\mathbf{R}$ is symmetric, strictly diagonal dominant and also strictly positive-definite. This implies that $\mathbf{R}^T = \mathbf{R}$, $|r_{i,i}| > \sum_{j \neq i} |r_{j,i}|$, and for some vector $\mathbf{v} \neq 0, \mathbf{v}^T\mathbf{R}\mathbf{v} > 0$.

Due to the non-standard numbering of the matrices, the matrices $\mathbf{Q}$ and $\mathbf{R}$ are specified below in a more convenient format.

$$\mathbf{Q} = \begin{pmatrix} h_1^{-1} & 0 & \cdots & 0 & 0 \\ -h_1^{-1} - h_2^{-1} & h_2^{-1} & \cdots & 0 & 0 \\ h_2^{-1} & -h_2^{-1} - h_3^{-1} & \cdots & 0 & 0 \\ 0 & h_3^{-1} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & h_{n-3}^{-1} & 0 \\ 0 & 0 & \cdots & -h_{n-3}^{-1} - h_{n-2}^{-1} & h_{n-2}^{-1} \\ 0 & 0 & \cdots & h_{n-2}^{-1} & -h_{n-2}^{-1} - h_{n-1}^{-1} \\ 0 & 0 & \cdots & 0 & h_{n-1}^{-1} \end{pmatrix}$$

$$\mathbf{R} = \begin{pmatrix} \frac{1}{3}(h_1 + h_2) & \frac{1}{6}h_2 & \cdots & 0 & 0 \\ \frac{1}{6}h_2 & \frac{1}{3}(h_2 + h_3) & \cdots & 0 & 0 \\ 0 & \frac{1}{6}h_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{6}h_{n-3} & 0 \\ 0 & 0 & \cdots & \frac{1}{3}(h_{n-3} + h_{n-2}) & \frac{1}{6}h_{n-2} \\ 0 & 0 & \cdots & \frac{1}{6}h_{n-2} & \frac{1}{3}(h_{n-2} + h_{n-1}) \end{pmatrix}$$

The value-second derivative representation can now be stated in the following theorem.

Theorem 3. (Green and Silverman, 1994)
The vectors $\mathbf{g}$ and $\boldsymbol{\gamma}$ specify a natural cubic spline $g(t)$ if and only if the condition

$$\mathbf{Q}^T\mathbf{g} = \mathbf{R}\boldsymbol{\gamma} \tag{3.1}$$

is satisfied.

If the condition (3.1) is satisfied, the vectors $\mathbf{g}$ and $\boldsymbol{\gamma}$ can then be used to specify the value of the cubic spline $g(t)$ at any point in the interval $[a, b]$. Using the fact that $g(t)$ is a cubic polynomial and $g''(t)$ a linear function on each interval $[t_i, t_{i+1}]$ for $i = 1, 2, \ldots, n-1$, Green and Silverman (1994) show that it is possible to derive explicit formulae for $g(t)$, so that for $t_i \leq t \leq t_{i+1}$

$$\begin{aligned} g(t) &= \frac{(t-t_i)g_{i+1} + (t_{i+1}-t)g_i}{h_i} \\ &\quad -\frac{1}{6}(t-t_i)(t_{i+1}-t)\left[\left(1+\frac{t-t_i}{h_i}\right)\gamma_{i+1} + \left(1+\frac{t_{i+1}-t}{h_i}\right)\gamma_i\right]. \end{aligned} \tag{3.2}$$

In addition, due to the natural boundary conditions, $g(t)$ will be linear on the two extreme intervals. As such, for $a \leq t \leq t_1$

$$g(t) = g_1 + (t-t_1)\left(\frac{g_2-g_1}{t_2-t_1} - \frac{1}{6}(t_2-t_1)\gamma_2\right) \tag{3.3}$$

and for $t_n \leq t \leq b$

$$g(t) = g_n + (t-t_n)\left(\frac{g_n-g_{n-1}}{t_n-t_{n-1}} + \frac{1}{6}(t_n-t_{n-1})\gamma_{n-1}\right). \tag{3.4}$$

3.3 Smoothing Cubic Splines

Regression analysis is a statistical tool primarily used in modelling the relationship between response and explanatory variables, the aim being to determine a mathematical model that will best fit the data. One popular method used to determine the best model is least squares regression, which aims to minimize the residual sum of squares $\sum_{i=1}^{n}\{y_i - f(t_i)\}^2$. However, when $f(t)$ is allowed to be any curve, any attempt to minimize the residual sum of squares would serve no purpose. This is because it is always possible to choose $f(t)$ in such a way that it will interpolate the data, resulting in the residual sum of squares being zero. In addition, $f(t)$ will not be unique and may display too much rapid variation.

One method of overcoming this problem is to penalize the residual sum of squares with a roughness penalty. The penalized sum of squares, for a twice-differentiable function $f(t)$ defined on $[a, b]$ with a smoothing parameter $\alpha \geq 0$, is defined by

$$S(f) = \sum_{i=1}^{n}\{y_i - f(t_i)\}^2 + \alpha\int_a^b \{f''(x)\}^2\,dx \tag{3.5}$$

The addition of the roughness penalty to the residual sum of squares ensures a trade-off between the goodness-of-fit to the data, quantified by the residual sum of squares $\sum_{i=1}^{n}\{y_i - f(t_i)\}^2$, and the 'smoothness' of the curve, quantified by the roughness penalty $\int_a^b \{f''(x)\}^2 dx$. The smoothing parameter α represents the 'rate of exchange' between the residual sums of squares and the 'smoothness' of the curve.

The problem of minimizing equation (3.5) over all sufficiently smooth curves on $[a, b]$ has a unique solution $\hat{f}(t)$ which is a NCS $g(t)$ with knots at the points t_i. This arises due to the minimum curvature property of a NCS. Equation (3.5) can therefore be re-written as

$$S(g) = \sum_{i=1}^{n}\{y_i - g(t_i)\}^2 + \alpha \int_a^b \{g''(x)\}^2 dx. \tag{3.6}$$

Since $g(t)$ is a NCS, the roughness penalty will satisfy

$$\int_a^b g''(t)^2 dt = \boldsymbol{\gamma}^T \mathbf{R} \boldsymbol{\gamma} = \mathbf{g}^T \mathbf{K} \mathbf{g}$$

where $\mathbf{K} = \mathbf{Q}\mathbf{R}^{-1}\mathbf{Q}^T$ (Green and Silverman, 1994). Note that $\mathbf{K}$ is symmetric as $\mathbf{K}^T = \left(\mathbf{Q}\mathbf{R}^{-1}\mathbf{Q}^T\right)^T = (\mathbf{Q}^T)^T\mathbf{R}^{-1}\mathbf{Q}^T = \mathbf{Q}\mathbf{R}^{-1}\mathbf{Q}^T = \mathbf{K}$. The penalized sum of squares can therefore be re-written in matrix notation as follows:

$$\begin{aligned}
S(\mathbf{g}) &= (\mathbf{y}-\mathbf{g})^T(\mathbf{y}-\mathbf{g}) + \alpha\mathbf{g}^T\mathbf{K}\mathbf{g} \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T(\mathbf{I}+\alpha\mathbf{K})^{-1}\mathbf{y} + \mathbf{y}^T(\mathbf{I}+\alpha\mathbf{K})^{-1}\mathbf{y} \\
&\quad -\mathbf{y}^T\mathbf{g} - \mathbf{g}^T\mathbf{y} + \mathbf{g}^T(\mathbf{I}+\alpha\mathbf{K})\mathbf{g} \\
&= \mathbf{g}^T(\mathbf{I}+\alpha\mathbf{K})\left\{\mathbf{g} - (\mathbf{I}+\alpha\mathbf{K})^{-1}\mathbf{y}\right\} \\
&\quad -\mathbf{y}^T(\mathbf{I}+\alpha\mathbf{K})^{-1}(\mathbf{I}+\alpha\mathbf{K})\{\mathbf{g} - (\mathbf{I}+\alpha\mathbf{K})^{-1}\mathbf{y}\} \\
&\quad +\mathbf{y}^T\mathbf{y} - \mathbf{y}^T(\mathbf{I}+\alpha\mathbf{K})^{-1}\mathbf{y} \\
&= \{\mathbf{g} - (\mathbf{I}+\alpha\mathbf{K})^{-1}\mathbf{y}\}^T(\mathbf{I}+\alpha\mathbf{K})\{\mathbf{g} - (\mathbf{I}+\alpha\mathbf{K})^{-1}\mathbf{y}\} \qquad (3.7) \\
&\quad +\mathbf{y}^T\mathbf{y} - \mathbf{y}^T(\mathbf{I}+\alpha\mathbf{K})^{-1}\mathbf{y}
\end{aligned}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ and $\mathbf{g} = (g_1, g_2, \ldots, g_n)^T$. From equation (3.7), it follows immediately that $S(\mathbf{g})$ has a unique minimum obtained by setting

$$\mathbf{g} = \mathbf{A}(\alpha)\mathbf{y} = (\mathbf{I}+\alpha\mathbf{K})^{-1}\mathbf{y}. \tag{3.8}$$

The matrix $\mathbf{A}(\alpha) = (\mathbf{I}+\alpha\mathbf{K})^{-1}$, which maps the observed values y_i to their 'predicted values' $\hat{y}_i = g(t_i)$, is often referred to as the hat matrix or the smoother matrix. The trace of the hat matrix, $\text{tr}\{\mathbf{A}(\alpha)\}$, represents the degrees of freedom for the model. There also exists an inverse relationship between the smoothing parameter α and the degrees of freedom, so that, as the smoothing parameter increases, the degrees of freedom decrease.

In order to specify the entire smoothed cubic spline, $\mathbf{g}$ is found from (3.8) and then $\boldsymbol{\gamma} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{g}$ can be solved. Then, using equations (3.2) - (3.4), the value of the smoothed cubic spline $g(t)$ can be specified at any point in the interval $[a, b]$. A more efficient method of solving for $\mathbf{g}$ and $\boldsymbol{\gamma}$ is to use the Reinsch algorithm. For details on the Reinsch algorithm refer to Green and Silverman (1994).

To illustrate the ideas of smoothed cubic splines, 100 points (t_i, y_i) on the interval $[0, 20]$ were simulated from the model $y(t) = \sin(2\pi t) + \varepsilon$ where $\varepsilon \sim IID(0, \sigma^2)$, various smoothed cubic splines were fitted to the simulated data and these are shown in Figure 3.2. In addition, Figure 3.2 captures the effect of changing the smoothing parameter α. When $\alpha = 0$, a NCS is obtained that perfectly interpolates the data. When α is small, the main component of $S(\mathbf{g})$ in equation (3.6) will be the residual sum of squares, resulting in a very close fit to the data. However, as α increases in size, more emphasis is placed on the roughness penalty and less on the residual sum of squares, resulting in a smoother curve. In the extreme, as $\alpha \to \infty$, $S(\mathbf{g})$ will be dominated by the roughness penalty term, resulting in a curve showing little or no curvature, in other words the curve will approach a straight line.

It possible to extend the penalized sum of squares, defined by equation (3.5), into a more general form in which the residuals are weighted. Suppose that $w_1, w_2, \ldots, w_n$ are strictly positive weights. The penalized weighted sum of squares is then defined by

$$S_W(f) = \sum_{i=1}^{n} w_i \{y_i - f(t_i)\}^2 + \alpha \int_a^b \{f''(x)\}^2 \, dx. \tag{3.9}$$

Once again, there exists a unique solution to the problem of minimizing equation (3.9) over all sufficiently smooth curves on $[a, b]$, given by a NCS $g(t)$ with knots at the points t_i. Equation (3.9) can therefore be re-written in matrix notation as

$$S_W(\mathbf{g}) = (\mathbf{y} - \mathbf{g})^T \mathbf{W} (\mathbf{y} - \mathbf{g}) + \alpha \mathbf{g}^T \mathbf{K} \mathbf{g},$$

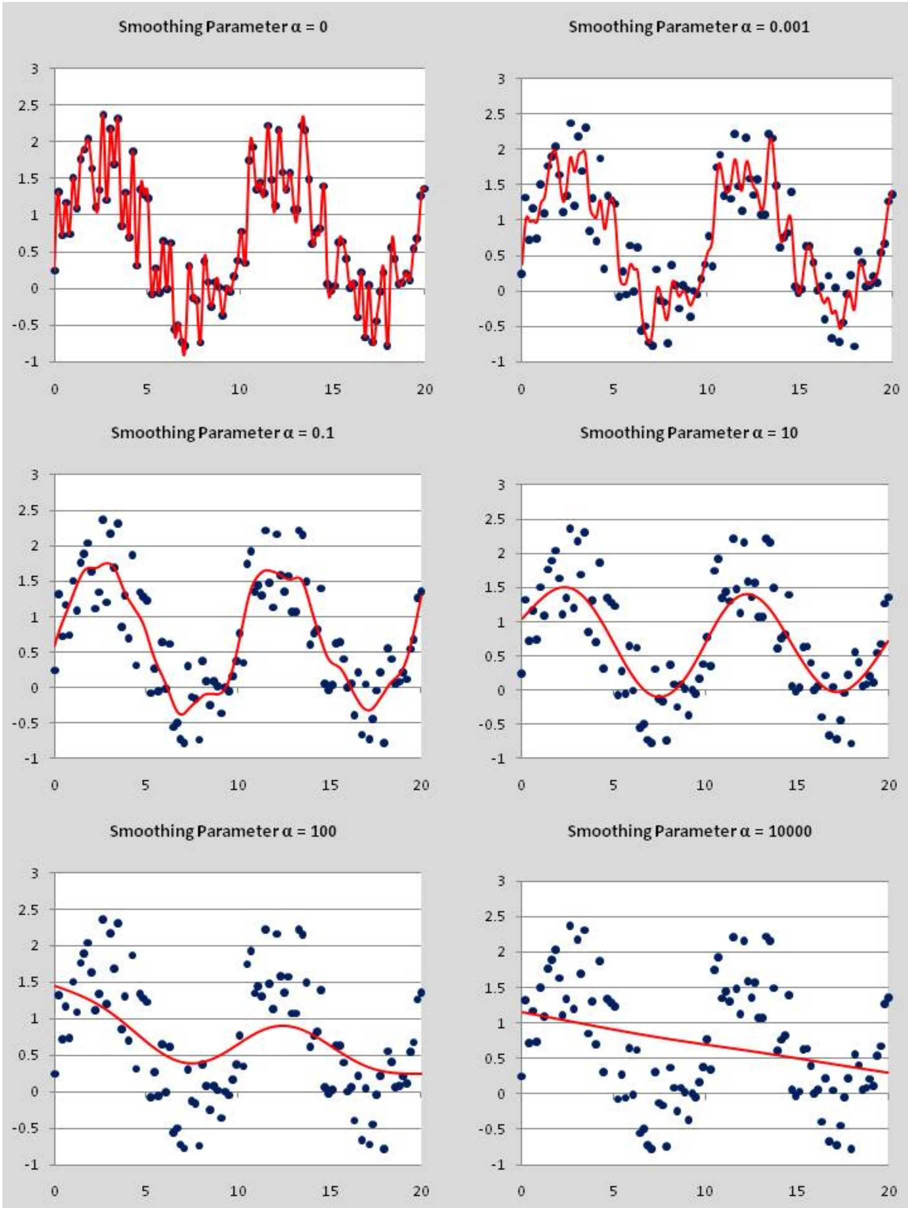

Figure 3.2: Various smoothed cubic splines with different smoothing parameters α fitted to simulated data

where $\mathbf{W}$ is a diagonal matrix with diagonal elements w_i and $S(\mathbf{g})$ attains a minimum on setting $\mathbf{g} = \mathbf{A}(\alpha)\mathbf{y} = (\mathbf{W} + \alpha\mathbf{K})^{-1}\mathbf{W}\mathbf{y}$.

3.4 Selecting the Smoothing Parameter

The smoothing parameter α represents a trade-off between the goodness-of-fit and smoothness of the curve. The aim in this section is to describe various methods for choosing an optimal smoothing parameter which achieves a desired balance between the goodness of fit and smoothness of a curve.

There are two basic approaches to selecting the smoothing parameter. The first is to apply a subjective approach and select the smoothing parameter manually. The second approach is to apply some form of automatic selection criteria such as Cross-Validation (CV), Generalized Cross-Validation (GCV) or Akaike's information criterion (AIC).

3.4.1 Cross-Validation

The motivation behind cross-validation techniques is to evaluate the predictive power of the model. This is done by partitioning the data set into two subsets. The first set, referred to as the calibration sample, is used to determine the model. The second set, referred to as the validation sample, is then used to determine how well the model predicts the data.

Within the context of smoothing splines, the approach of 'leave-one-out' cross-validation is usually employed. Given a dataset with observations y_i at distinct points t_i, for $i = 1, \ldots, n$, the calibration sample, constructed by leaving out the i^{th} observation, is used to determine the smoothed NCS, $\hat{g}^{(-i)}(t;\alpha)$, that minimizes

$$\sum_{j \neq i}^{n} \{y_j - g(t_j)\}^2 + \alpha \int_a^b \{g''(x)\}^2 \, dx.$$

The NCS $\hat{g}^{(-i)}(t;\alpha)$ can then be used on the validation sample, (t_i, y_i), to determine how well the model predicts the observation by considering the squared residual $\left\{y_i - \hat{g}_i^{(-i)}(t_i;\alpha)\right\}^2$. This process is repeated for all the observations. The cross-validation score function can then be calculated as the average of the squared residuals as

$$CV(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i - \hat{g}_i^{(-i)}(t_i; \alpha) \right\}^2 . \tag{3.10}$$

The smoothing parameter α that minimizes the cross-validation score function, $CV(\alpha)$, is regarded as the optimal smoothing parameter.

Equation (3.10) is not the most computationally efficient means of calculating the cross-validation score function, $CV(\alpha)$. With a little algebra and basic mathematical manipulation, it can be shown that the cross-validation score satisfies

$$CV(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{g}(t_i; \alpha)}{1 - A_{ii}(\alpha)} \right)^2 \tag{3.11}$$

where $\hat{g}(t; \alpha)$ is the smoothed NCS calculated from the full data set and $A(\alpha)$ is the hat matrix (Green and Silverman, 1994). The advantage of using equation (3.11) is that only one NCS needs to be determined and the cross-validation score can be calculated using only the ordinary residuals and the diagonal elements of the hat matrix. This reduces the computation burden that arises when using the cross-validation function represented by (3.10).

3.4.2 Other Automatic Model Selection Criteria

One disadvantage of using cross-validation is the enormous amount of computing required, which becomes evident as the size of the dataset increases. One method of reducing the amount of computation required is to use the generalized cross-validation criterion, which is given by

$$GCV(\alpha) = \frac{1}{n\left(1 - n^{-1}\text{tr}A(\alpha)\right)^2} \sum_{i=1}^{n} \left\{ y_i - \hat{g}(t_i; \alpha) \right\}^2 .$$

where $\hat{g}(t; \alpha)$ is the smoothed NCS calculated from the full data set.

An alternative method to cross-validation is to choose the degrees of freedom, df, that minimize an information criterion, such as the the well-known Akaike's information criterion AIC $= -2\ln(L) + 2df$ (Akaike, 1973). The degrees of freedom for a smoothed NCS, that is $\text{tr}\{\mathbf{A}(\alpha)\}$, decrease from n when $\alpha = 0$ to 2 as $\alpha \to \infty$ (Green and Silverman, 1994). AIC is asymptotically equivalent to cross-validation and has the advantage of being less computationally demanding (Chavez-Demoulin and Davison, 2005).

3.5 Penalizing the Log-Likelihood

Given data pairs (t_i, y_i) for $i = 1, 2, \ldots, n$, a natural way to view the approach of minimizing the penalized sum of squares, given by equation (3.6), is as a method for fitting a model of the form

$$y = g(t) + \text{error} \tag{3.12}$$

to the observed data, where $g(t)$ is a NCS.

One possible extension to the model (3.12) is to assume that the observations are independent random variates from a known parametric family of distributions F, that is

$$y_i \sim F(g(t_i), \phi)$$

for $i = 1, 2, \ldots, n$ where $g(t_i)$ is a 'smooth' time-varying parameter and ϕ is a vector of constant parameters. One method of determining estimates for $g(t)$ and ϕ is to maximize the log-likelihood function $l(g(t), \phi | y_1, y_2, \ldots, y_n)$. However, when $g(t)$ is allowed to be any curve, any attempt to maximize the log-likelihood function would serve no purpose. This is because it is always possible to choose $g(t)$ and ϕ in such a way that the fitted values will interpolate the observed data. One method of overcoming this problem is to penalize the log-likelihood with the roughness penalty $\int g''(t)^2 dt$ to obtain the penalized log-likelihood, defined by

$$l^{PLL}(g(t), \phi | y_1, y_2, \ldots, y_n) = l(g(t), \phi | y_1, y_2, \ldots, y_n) - \frac{\lambda}{2} \int g''(t)^2 dt,$$

in which case it will be optimal to choose $g(t)$ to be a NCS. The ideas behind maximizing the penalized log-likelihood function, within the context of classical extreme value theory, are discussed in more detail in the next chapter.

Chapter 4

Smoothing Extremes

The aim of this chapter is to combine concepts from the previous two chapters and to consider smoothing extremes by penalizing the log-likelihood function with a roughness penalty. The structure of the chapter is as follows. Firstly, the topic of non-stationary extremes is introduced in Section 1. This is followed by a discussion on smoothing non-stationary extremes using the penalized log-likelihood function in Section 2. Then the ideas underpinning the use of the Fisher scoring algorithm to maximize the penalized log-likelihood function are presented in Section 3. Finally, in Section 4, the choice of selecting the optimal smoothing parameters using cross-validation is briefly discussed.

4.1 Non-Stationary Extremes

In Chapter 2, the modelling of extremes using the GEV distribution and the distribution for the r largest order statistics assumed that the underlying data were stationary. These models can be extended to account for data that are non-stationary by allowing the location and scale parameters, μ and σ respectively, to vary over time. It is often difficult to estimate the shape parameters ξ with accuracy. Consequently, any attempt to try to model the shape parameter ξ as a function of time is usually regarded as unrealistic (Coles, 2001).

Leadbetter *et al.* (1983) explain that when a trend or seasonal component is present in the underlying data, it is still possible to determine the actual asymptotic distribution of the maximum as the limit laws still hold, at least approximately, provided that the normalizing constants $\{a_n > 0\}$ and $\{b_n \in \mathbb{R}\}$ are adjusted.

When modelling the maxima of a process whose statistical properties are allowed to change over time, a non-stationary GEV distribution can be used. The distribution function is given by

$$G(x,t) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu(t)}{\sigma(t)}\right)\right]^{-\frac{1}{\xi}}\right\} \tag{4.1}$$

where the support is defined by $1 + \xi\,(x - \mu(t))\,/\sigma(t) > 0$ and the parameters satisfy $-\infty < \mu(t) < \infty, \sigma(t) > 0$ and $-\infty < \xi < \infty$ (Coles, 2001). The probability density function for the non-stationary GEV is then obtained by differentiating $G(x,t)$ in expression (4.1) with respect to x to give

$$g(x,t) = \frac{1}{\sigma(t)}\exp\left\{-\left[1+\xi\left(\frac{x-\mu(t)}{\sigma(t)}\right)\right]^{-\frac{1}{\xi}}\right\}\left[1+\xi\left(\frac{x-\mu(t)}{\sigma(t)}\right)\right]^{-\left(\frac{1}{\xi}+1\right)}.$$

In the case when $\xi = 0$, corresponding to the non-stationary Gumbel distribution, the distribution function is obtained by considering the limit of (4.1) as $\xi \to 0$, leading to

$$G(x,t) = \exp\left[-\exp\left[-\left(\frac{x-\mu(t)}{\sigma(t)}\right)\right]\right]$$

with the probability density function given by

$$g(x,t) = \frac{1}{\sigma(t)}\exp\left\{-\exp\left[-\left(\frac{x-\mu(t)}{\sigma(t)}\right)\right]\right\}\exp\left[-\left(\frac{x-\mu(t)}{\sigma(t)}\right)\right]$$

for $-\infty < \mu(t) < \infty, \sigma(t) > 0$ and $-\infty < \xi < \infty$. These ideas are also easily extended to the distribution for the r largest order statistics, where the probability density function for the non-stationary r largest order statistics is specified by replacing μ and σ in equations (2.19) and (2.20) by $\mu(t)$ and $\sigma(t)$ respectively.

Early applications of this approach include adopting parametric functions for $\mu(t)$ whilst keeping the scale parameter constant. For example, Smith (1986) looked at a linear trend, a quadratic trend and a linear trend plus a sinusoidal component for $\mu(t)$ and applied the distribution function for the non-stationary r largest order statistics with $\xi = 0$ to the Venice sea-level data. Tawn (1988) on the other hand looked at a linear and a quadratic trend for $\mu(t)$ and applied the distribution function for the non-stationary r largest order statistics with $\xi \neq 0$ to the sea levels at Lowestoft and Great Yarmouth.

One of the disadvantages of using parametric techniques is the lack of flexibility offered by the models. This lead to the use of nonparametric techniques to model $\mu(t)$ and $\sigma(t)$, which are more flexible and have the benefit of incorporating smoothing methods into the model. For example, Rosen and Cohen (1996) applied the non-stationary Gumbel distribution and distribution for the non-stationary r largest order statistics with $\xi = 0$ to Venice sea level data, using smoothed NCS's to model both $\mu(t)$ and $\sigma(t)$. A similar approach was followed by Pauli and Coles (2001), who considered the application of the non-stationary GEV distribution and the distribution for the r largest order statistics with $\xi \neq 0$ to athletics data for the women's 1500m and 3000m events and temperature data at Oxford and Worthing, using smoothed NCS to model $\mu(t)$ whilst keeping the scale parameter constant. More recent applications of smoothing sample extremes include the use of generalized additive models with spline smoothers, see Chavez-Demoulin and Embrechts (2004), Chavez-Demoulin and Davison (2005), Wand and Padoan (2008) and Laurini and Pauli (2009).

The approach followed in the present study is similar to that of Rosen and Cohen (1996) and Pauli and Coles (2001), where a smoothed NCS is fitted to the location and scale parameters $\mu(t)$ and $\sigma(t)$ of the non-stationary GEV distribution and of the non-stationary distribution for the r largest order statistics.

4.2 Modelling Non-Stationary Extremes

As with the modelling of stationary extremes, in practice, a random sample $y_1, y_2, \ldots, y_{nN}$ of observed data is available. These observations are subdivided into consecutive blocks of length n. From each block the largest r observations are extracted, leading to a series containing the r largest order statistics $M_i^{(r)} = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(r)})$ for $i = 1, 2, \ldots, N$, where $x_i^{(k)} =$ kth largest of $\left(y_{n(i-1)+1}, y_{n(i-1)+2}, \ldots, y_{ni}\right)$ and $x_i^{(r)} < x_i^{(r-1)} < \ldots < x_i^{(1)}$. In addition, each block is assumed to correspond to a specified time period so that the r largest order statistics $M_i^{(r)}$ are observed sequentially in time such that $M_i^{(r)}$ is recorded at time t_i, where $t_1 < t_2 < \ldots < t_N$. Furthermore, it is assumed that these extremes can be modelled using the distribution for the non-stationary r largest order statistics and that the asymptotic results will hold at least approximately.

The Log-Likelihood Function

When $\mu(t)$ and $\sigma(t)$ are specified parametric functions, for example $\mu(t) = \beta_0 + \beta_1 t$ and $\sigma(t) = \exp(\beta_2 + \beta_3 t)$ where $\beta_0, \beta_1, \beta_2$ and β_3 are unknown parameters, one method of estimating the parameters is to maximize the log-likelihood function. Now, assuming that $\mu(t)$ and $\sigma(t)$ are parametric functions describing the location and the scale parameter over time, let the vectors $\boldsymbol{\mu} = (\mu(t_1), \mu(t_2), \ldots, \mu(t_N))^T$ and $\boldsymbol{\sigma} = (\sigma(t_1), \sigma(t_2), \ldots, \sigma(t_N))^T$ denote the values of the corresponding parametric functions at the observed time points t_i for $i = 1, 2, \ldots, N$. Then, for $r \geq 1$, the log-likelihood function for the non-stationary distribution of the r largest order statistics with $\xi \neq 0$ is given by

$$\begin{aligned} l(\boldsymbol{\mu}, \boldsymbol{\sigma}, \xi; r) &= -\sum_{i=1}^{N} \Big[1 + \xi\Big(\frac{x_i^{(r)} - \mu(t_i)}{\sigma(t_i)}\Big)\Big]^{-\frac{1}{\xi}} - \sum_{i=1}^{N} r \ln(\sigma(t_i)) \\ &\quad - \sum_{i=1}^{N}\sum_{k=1}^{r} \Big(\frac{1}{\xi} + 1\Big) \ln \Big[1 + \xi\Big(\frac{x_i^{(k)} - \mu(t_i)}{\sigma}\Big)\Big] \end{aligned} \tag{4.2}$$

provided that $1 + \xi(x_i^{(k)} - \mu(t_i))/\sigma(t_i) > 0$ for $i = 1, 2, \ldots, N$ and $k = 1, 2, \ldots, r$. Furthermore, when $\xi = 0$ the log-likelihood function is given by

$$\begin{aligned} l(\boldsymbol{\mu}, \boldsymbol{\sigma}, 0; r) &= -\sum_{i=1}^{N} \exp\Big[-\Big(\frac{x_i^{(r)} - \mu(t_i)}{\sigma(t_i)}\Big)\Big] - \sum_{i=1}^{N} r \ln(\sigma(t_i)) \\ &\quad - \sum_{i=1}^{N}\sum_{k=1}^{r} \Big(\frac{x_i^{(k)} - \mu(t_i)}{\sigma(t_i)}\Big). \end{aligned} \tag{4.3}$$

When $r = 1$, the log-likelihood functions (4.2) and (4.3) are equivalent to the log-likelihood function for the non-stationary GEV distribution and the log-likelihood function for the non-stationary Gumbel distribution respectively.

The Penalized Log-Likelihood Function

When attempting to maximize the log-likelihoods (4.2) and (4.3) over all smooth functions $\mu(t)$ and $\sigma(t)$, no unique solution will be available. This is because it is always possible to choose $\mu(t)$ and $\sigma(t)$ such that the model fits the data exactly. One method of obtaining a unique solution for $\mu(t)$ and $\sigma(t)$ is to penalize the log-likelihood with a roughness penalty for each of the smooth functions $\mu(t)$ and $\sigma(t)$. The resulting penalized log-likelihood is then given by

$$l^{\text{PLL}}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \xi; r) = l(\boldsymbol{\mu}, \boldsymbol{\sigma}, \xi; r) - \frac{\lambda}{2} \int \mu''(t)^2 dt - \frac{\alpha}{2} \int \sigma''(t)^2 dt \tag{4.4}$$

where λ and α are smoothing parameters.

The penalized log-likelihood approach allows a trade-off between high values of the log-likelihood, representing a very close fit to the data, with the smoothness of the fitted curves $\mu(t)$ and $\sigma(t)$. These are often seen as conflicting objectives. The size of λ and α therefore govern the relative importance attached to these conflicting objectives. Small values of α and λ imply that $l^{\mathrm{PLL}}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \xi; r)$ is dominated by the log-likelihood, resulting in a closer fit to the data. On the other hand, large values of α and λ imply that the $l^{\mathrm{PLL}}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \xi; r)$ is dominated by the roughness penalties, resulting in smoother curve estimates. In the extreme, as $\alpha \to \infty$ and $\lambda \to \infty$, the $l^{\mathrm{PLL}}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \xi; r)$ is maximized when the value of the roughness penalties tend to zero. This is achieved by allowing $\mu(t)$ and $\sigma(t)$ to become linear functions, for which the integrals of $\mu''(t)$ and $\sigma''(t)$ are zero.

Green and Silverman (1994) showed that, when using the integrated squared second derivative as the choice of roughness penalty, it is optimal to choose $\mu(t)$ and $\sigma(t)$ to be NCS's with knots $t_1, t_2, \ldots, t_N$ to maximize the penalized log-likelihood (4.4). Consequently, allowing $\mu(t) = g_\mu(t)$ and $\sigma(t) = g_\sigma(t)$ to be NCS's, the penalized log-likelihood (4.4) can therefore be re-written as

$$\begin{aligned} l^{\mathrm{PLL}}(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r) &= l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r) - \frac{\lambda}{2}\int g_\mu''(t)^2 dt - \frac{\alpha}{2}\int g_\sigma''(t)^2 dt \\ &= l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r) - \frac{\lambda}{2}\mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu - \frac{\alpha}{2}\mathbf{g}_\sigma^T \mathbf{K} \mathbf{g}_\sigma. \end{aligned} \tag{4.5}$$

where $\mathbf{g}_{\boldsymbol{\mu}}^T = (g_\mu(t_1), g_\mu(t_2), \ldots, g_\mu(t_N))$, $\mathbf{g}_{\boldsymbol{\sigma}}^T = (g_\sigma(t_1), g_\sigma(t_2), \ldots, g_\sigma(t_N))$ and $\mathbf{K} = \mathbf{Q}\mathbf{R}^{-1}\mathbf{Q}^T$. The NCS's $\hat{g}_\mu(t)$ and $\hat{g}_\sigma(t)$ that maximize the penalized log-likelihood in (4.5) are called maximum penalized likelihood estimates (MPLE's).

The penalized log-likelihood in (4.5) applies to the general case for modelling extremes when both the location and scale parameters depend on time. When the scale parameter is assumed to be constant, that is $\sigma(t) = \sigma$, the penalized log-likelihood is modified so that

$$l^{\mathrm{PLL}}(\mathbf{g}_{\boldsymbol{\mu}}, \sigma, \xi; r) = l(\mathbf{g}_{\boldsymbol{\mu}}, \sigma, \xi; r) - \frac{\lambda}{2}\mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu.$$

4.3 The Fisher Scoring Algorithm

One method of maximizing the penalized log-likelihood in (4.5) is to use the Fisher scoring algorithm. The Fisher scoring algorithm involves an iterative process similar to the Newton-Raphson algorithm, using the expected information matrix rather than the observed information matrix, in other words using the expected value of the negative Hessian. The general Fisher scoring algorithm, for a likelihood function $l(\boldsymbol{\theta})$, is given by

$$\boldsymbol{\theta}^{new} = \boldsymbol{\theta}^{old} + \left\{E\Big(-\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big)\right\}^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}^{old}} \left(\frac{\partial l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^{old}}$$

where $\boldsymbol{\theta}$ is the vector of parameters, $\boldsymbol{\theta}^{old}$ is the vector of initial trial estimates, $\boldsymbol{\theta}^{new}$ is the vector of updated trial estimates and the first derivative of the likelihood function and the expected information matrix are evaluated at $\boldsymbol{\theta}^{old}$. The iterative process starts with an initial estimate θ^{old} to obtain an updated estimate $\boldsymbol{\theta}^{new}$. The process is then repeated by replacing $\boldsymbol{\theta}^{old}$ by $\boldsymbol{\theta}^{new}$ until convergence is obtained, normally chosen so that $|\boldsymbol{\theta}^{new} - \boldsymbol{\theta}^{old}| < 0.0001$.

4.3.1 Fisher Scoring when $\xi \neq 0$

To apply the Fisher scoring algorithm to the penalized log-likelihood $l^{\text{PLL}}(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi, r)$ in (4.5), let $\boldsymbol{\theta} = (\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi)^T$. The Fisher scoring algorithm can then be written as

$$\begin{pmatrix} \mathbf{g}_{\boldsymbol{\mu}}^{new} - \mathbf{g}_{\boldsymbol{\mu}}^{old} \\ \mathbf{g}_{\boldsymbol{\sigma}}^{new} - \mathbf{g}_{\boldsymbol{\sigma}}^{old} \\ \xi^{new} - \xi^{old} \end{pmatrix} = \left\{ E \begin{pmatrix} -\frac{\partial^2 l^{\text{PLL}}}{\partial \mathbf{g}_{\boldsymbol{\mu}} \partial \mathbf{g}_{\boldsymbol{\mu}}^T} & -\frac{\partial^2 l^{\text{PLL}}}{\partial \mathbf{g}_{\boldsymbol{\mu}} \partial \mathbf{g}_{\boldsymbol{\sigma}}^T} & -\frac{\partial^2 l^{\text{PLL}}}{\partial \mathbf{g}_{\boldsymbol{\mu}} \partial \xi} \\ -\frac{\partial^2 l^{\text{PLL}}}{\partial \mathbf{g}_{\boldsymbol{\sigma}} \partial \mathbf{g}_{\boldsymbol{\mu}}^T} & -\frac{\partial^2 l^{\text{PLL}}}{\partial \mathbf{g}_{\boldsymbol{\sigma}} \partial \mathbf{g}_{\boldsymbol{\sigma}}^T} & -\frac{\partial^2 l^{\text{PLL}}}{\partial \mathbf{g}_{\boldsymbol{\sigma}} \partial \xi} \\ -\frac{\partial^2 l^{\text{PLL}}}{\partial \xi \partial \mathbf{g}_{\boldsymbol{\mu}}^T} & -\frac{\partial^2 l^{\text{PLL}}}{\partial \xi \partial \mathbf{g}_{\boldsymbol{\sigma}}^T} & -\frac{\partial^2 l^{\text{PLL}}}{\partial \xi \partial \xi} \end{pmatrix} \right\}^{-1} \begin{pmatrix} \frac{\partial l^{\text{PLL}}}{\partial \mathbf{g}_{\boldsymbol{\mu}}} \\ \frac{\partial l^{\text{PLL}}}{\partial \mathbf{g}_{\boldsymbol{\sigma}}} \\ \frac{\partial l^{\text{PLL}}}{\partial \xi} \end{pmatrix}$$

where the derivatives and elements of the expected information matrix are evaluated at $\mathbf{g}_{\boldsymbol{\mu}}^{old}$, $\mathbf{g}_{\boldsymbol{\sigma}}^{old}$ and ξ^{old}. By writing out the penalized log-likelihood in terms of the log-likelihood $l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r)$ and the roughness penalties $\frac{\lambda}{2}\mathbf{g}_{\boldsymbol{\mu}}^T\mathbf{K}\mathbf{g}_{\boldsymbol{\mu}}$ and $\frac{\alpha}{2}\mathbf{g}_{\boldsymbol{\sigma}}^T\mathbf{K}\mathbf{g}_{\boldsymbol{\sigma}}$, the Fisher scoring algorithm can then be re-arranged and written as

$$\begin{pmatrix} W_\mu + \lambda\mathbf{K} & W_{\mu\sigma} & W_{\mu\xi} \\ W_{\sigma\mu} & W_\sigma + \alpha\mathbf{K} & W_{\sigma\xi} \\ W_{\xi\mu} & W_{\xi\sigma} & W_\xi \end{pmatrix} \begin{pmatrix} \mathbf{g}_{\boldsymbol{\mu}}^{new} - \mathbf{g}_{\boldsymbol{\mu}}^{old} \\ \mathbf{g}_{\boldsymbol{\sigma}}^{new} - \mathbf{g}_{\boldsymbol{\sigma}}^{old} \\ \xi^{new} - \xi^{old} \end{pmatrix} = \begin{pmatrix} u_\mu - \lambda\mathbf{K}\mathbf{g}_{\boldsymbol{\mu}}^{old} \\ u_\sigma - \alpha\mathbf{K}\mathbf{g}_{\boldsymbol{\sigma}}^{old} \\ u_\xi \end{pmatrix}$$

where

$$u_\mu = \frac{\partial l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r)}{\partial \mathbf{g}_{\boldsymbol{\mu}}} \qquad u_\sigma = \frac{\partial l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r)}{\partial \mathbf{g}_{\boldsymbol{\sigma}}}$$

$$u_\xi = \frac{\partial l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r)}{\partial \xi} \qquad W_\xi = E\bigg(-\frac{\partial^2 l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r)}{\partial \xi^2}\bigg)$$

$$W_\mu = E\bigg(-\frac{\partial^2 l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r)}{\partial \mathbf{g}_{\boldsymbol{\mu}} \partial \mathbf{g}_{\boldsymbol{\mu}}^T}\bigg) \qquad W_\sigma = E\bigg(-\frac{\partial^2 l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r)}{\partial \mathbf{g}_{\boldsymbol{\sigma}} \mathbf{g}_{\boldsymbol{\sigma}}^T}\bigg)$$

$$W_{\mu\sigma} = E\bigg(-\frac{\partial^2 l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r)}{\partial \mathbf{g}_{\boldsymbol{\mu}} \partial \mathbf{g}_{\boldsymbol{\sigma}}^T}\bigg) \qquad W_{\sigma\mu} = E\bigg(-\frac{\partial^2 l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r)}{\partial \mathbf{g}_{\boldsymbol{\sigma}} \partial \mathbf{g}_{\boldsymbol{\mu}}^T}\bigg)$$

$$W_{\mu\xi} = E\bigg(-\frac{\partial^2 l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r)}{\partial \mathbf{g}_{\boldsymbol{\mu}} \partial \xi}\bigg) \qquad W_{\xi\mu} = E\bigg(-\frac{\partial^2 l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r)}{\partial \xi \partial \mathbf{g}_{\boldsymbol{\mu}}^T}\bigg)$$

$$W_{\xi\sigma} = E\bigg(-\frac{\partial^2 l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r)}{\partial \xi \partial \mathbf{g}_{\boldsymbol{\sigma}}^T}\bigg) \qquad W_{\sigma\xi} = E\bigg(-\frac{\partial^2 l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r)}{\partial \mathbf{g}_{\boldsymbol{\sigma}} \partial \xi}\bigg)$$

are all evaluated at $\mathbf{g}_{\boldsymbol{\mu}}^{old}$, $\mathbf{g}_{\boldsymbol{\sigma}}^{old}$ and ξ^{old}. The Fisher scoring algorithm therefore involves $2N+1$ equations where the updating trial estimates $\mathbf{g}_{\boldsymbol{\mu}}^{new}$, $\mathbf{g}_{\boldsymbol{\sigma}}^{new}$ and ξ^{new} can be expressed as follows

$$\begin{aligned}
\mathbf{g}_{\boldsymbol{\mu}}^{new} &= (W_\mu + \lambda \mathbf{K})^{-1}\Big(W_\mu \mathbf{g}_{\boldsymbol{\mu}}^{old} + u_\mu - W_{\mu\sigma}(\mathbf{g}_{\boldsymbol{\sigma}}^{new} - \mathbf{g}_{\boldsymbol{\sigma}}^{old}) - W_{\mu\xi}(\xi^{new} - \xi^{old})\Big) \\
\mathbf{g}_{\boldsymbol{\sigma}}^{new} &= (W_\sigma + \alpha \mathbf{K})^{-1}\Big(W_\sigma \mathbf{g}_{\boldsymbol{\sigma}}^{old} + u_\sigma - W_{\sigma\mu}(\mathbf{g}_{\boldsymbol{\mu}}^{new} - \mathbf{g}_{\boldsymbol{\mu}}^{old}) - W_{\sigma\xi}(\xi^{new} - \xi^{old})\Big) \\
\xi^{new} &= \xi^{old} + W_\xi^{-1} u_\xi - W_\xi^{-1} W_{\xi\mu}(\mathbf{g}_{\boldsymbol{\mu}}^{new} - \mathbf{g}_{\boldsymbol{\mu}}^{old}) - W_\xi^{-1} W_{\xi\sigma}(\mathbf{g}_{\boldsymbol{\sigma}}^{new} - \mathbf{g}_{\boldsymbol{\sigma}}^{old}).
\end{aligned}$$

These equations can then be manipulated further so that they are in a form similar to the solution for a weighted cubic smoothing spline as discussed in Chapter 3. Thus

$$\mathbf{g}_{\boldsymbol{\mu}}^{new} = S_1(y_1 - W_\mu^{-1} W_{\mu\sigma} \mathbf{g}_{\boldsymbol{\sigma}}^{new} - W_\mu^{-1} W_{\mu\xi} \xi^{new}) \tag{4.6}$$

$$\mathbf{g}_{\boldsymbol{\sigma}}^{new} = S_2(y_2 - W_\sigma^{-1} W_{\sigma\mu} \mathbf{g}_{\boldsymbol{\mu}}^{new} - W_\sigma^{-1} W_{\sigma\xi} \xi^{new}) \tag{4.7}$$

$$\xi^{new} = y_3 - W_\xi^{-1} W_{\xi\mu} \mathbf{g}_{\boldsymbol{\mu}}^{new} - W_\xi^{-1} W_{\xi\sigma} \mathbf{g}_{\boldsymbol{\sigma}}^{new} \tag{4.8}$$

where

$$\begin{aligned}
S_1 &= (W_\mu + \lambda \mathbf{K})^{-1} W_\mu \\
S_2 &= (W_\sigma + \alpha \mathbf{K})^{-1} W_\sigma
\end{aligned}$$

are hat matrices and

$$\begin{aligned} y_1 &= \mathbf{g}_{\boldsymbol{\mu}}^{old} + W_\mu^{-1}u_\mu + W_\mu^{-1}W_{\mu\sigma}\mathbf{g}_{\boldsymbol{\sigma}}^{old} + W_\mu^{-1}W_{\mu\xi}\xi^{old} \\ y_2 &= \mathbf{g}_{\boldsymbol{\sigma}}^{old} + W_\sigma^{-1}u_\sigma + W_\sigma^{-1}W_{\sigma\mu}\mathbf{g}_{\boldsymbol{\mu}}^{old} + W_\sigma^{-1}W_{\sigma\xi}\xi^{old} \\ y_3 &= \xi^{old} + W_\xi^{-1}u_\xi + W_\xi^{-1}W_{\xi\mu}\mathbf{g}_{\boldsymbol{\mu}}^{old} + W_\xi^{-1}W_{\xi\sigma}\mathbf{g}_{\boldsymbol{\sigma}}^{old}. \end{aligned}$$

The degrees of freedom for each of the updating equations $\mathbf{g}_{\boldsymbol{\mu}}^{new}$ and $\mathbf{g}_{\boldsymbol{\sigma}}^{new}$ are given by the trace of their hat matrix, that is $tr(S_1)$ and $tr(S_2)$ respectively. Due to the constraints of the GEV distribution and the distribution for the r largest order statistics with $\xi \neq 0$, an additional check is required to ensure that

$$1 + \xi\left(\frac{x_i^{(k)} - g_\mu(t_i)}{g_\sigma(t_i)}\right) > 0 \tag{4.9}$$

for $i = 1, 2, \ldots, N$ and $k = 1, 2, \ldots, r$. These constraints should be checked at each iteration in the Fisher scoring algorithm.

4.3.2 Fisher Scoring when $\xi = 0$

When $\xi = 0$, the Fisher scoring algorithm simplifies to the expression given by Rosen and Cohen (1996), so that

$$\begin{pmatrix} W_\mu + \lambda\mathbf{K} & W_{\mu\sigma} \\ W_{\sigma\mu} & W_\sigma + \alpha\mathbf{K} \end{pmatrix}\begin{pmatrix} \mathbf{g}_{\boldsymbol{\mu}}^{new} - \mathbf{g}_{\boldsymbol{\mu}}^{old} \\ \mathbf{g}_{\boldsymbol{\sigma}}^{new} - \mathbf{g}_{\boldsymbol{\sigma}}^{old} \end{pmatrix} = \begin{pmatrix} u_\mu - \lambda\mathbf{K}\mathbf{g}_{\boldsymbol{\mu}}^{old} \\ u_\sigma - \alpha\mathbf{K}\mathbf{g}_{\boldsymbol{\sigma}}^{old} \end{pmatrix}$$

where

$$\begin{aligned} u_\mu &= \frac{\partial l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, 0; r)}{\partial \mathbf{g}_{\boldsymbol{\mu}}} & u_\sigma &= \frac{\partial l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, 0; r)}{\partial \mathbf{g}_{\boldsymbol{\sigma}}} \\ W_\mu &= E\left(-\frac{\partial^2 l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, 0; r)}{\partial \mathbf{g}_{\boldsymbol{\mu}}\partial \mathbf{g}_{\boldsymbol{\mu}}^T}\right) & W_\sigma &= E\left(-\frac{\partial^2 l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, 0; r)}{\partial \mathbf{g}_{\boldsymbol{\sigma}}\partial \mathbf{g}_{\boldsymbol{\sigma}}^T}\right) \\ W_{\mu\sigma} &= E\left(-\frac{\partial^2 l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, 0; r)}{\partial \mathbf{g}_{\boldsymbol{\mu}}\partial \mathbf{g}_{\boldsymbol{\sigma}}^T}\right) & W_{\sigma\mu} &= E\left(-\frac{\partial^2 l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, 0; r)}{\partial \mathbf{g}_{\boldsymbol{\sigma}}\partial \mathbf{g}_{\boldsymbol{\mu}}^T}\right) \end{aligned}$$

are all evaluated at $\mathbf{g}_{\boldsymbol{\mu}}^{old}$ and $\mathbf{g}_{\boldsymbol{\sigma}}^{old}$. Therefore, when $\xi = 0$ the Fisher scoring algorithm only involves $2N$ equations and the updating trial estimates $\mathbf{g}_{\boldsymbol{\mu}}^{new}$ and $\mathbf{g}_{\boldsymbol{\sigma}}^{new}$ are given by

$$\begin{aligned} \mathbf{g}_{\boldsymbol{\mu}}^{new} &= S_1(y_1 - W_\mu^{-1}W_{\mu\sigma}\mathbf{g}_{\boldsymbol{\sigma}}^{new}) \\ \mathbf{g}_{\boldsymbol{\sigma}}^{new} &= S_2(y_2 - W_\sigma^{-1}W_{\sigma\mu}\mathbf{g}_{\boldsymbol{\mu}}^{new}) \end{aligned}$$

where

$$S_1 = (W_\mu + \lambda \mathbf{K})^{-1} W_\mu$$
$$S_2 = (W_\sigma + \alpha \mathbf{K})^{-1} W_\sigma$$

are hat matrices and

$$y_1 = \mathbf{g}_{\boldsymbol{\mu}}^{old} + W_\mu^{-1} u_\mu + W_\mu^{-1} W_{\mu\sigma} \mathbf{g}_{\boldsymbol{\sigma}}^{old}$$
$$y_2 = \mathbf{g}_{\boldsymbol{\sigma}}^{old} + W_\sigma^{-1} u_\sigma + W_\sigma^{-1} W_{\sigma\mu} \mathbf{g}_{\boldsymbol{\mu}}^{old}.$$

4.3.3 Elements of the Fisher Scoring Algorithm

Smith (1986) and Tawn (1988) presented the second derivatives of the log-likelihood function and elements of the expected information matrix for the distribution of the r largest order statistics, when $\xi = 0$ and $\xi \neq 0$ respectively. However, this was done for the application of parametric models and a different definition was used for specifying the shape parameter. Rosen and Cohen (1996) presented the elements of the Fisher scoring algorithm when $\xi = 0$ and $r > 1$. Although Pauli and Coles (2001) applied the non-stationary GEV distribution and distribution for the r largest order statistics when $\xi \neq 0$, mathematical details for the elements of the Fisher scoring algorithm were not provided.

The first and second derivatives of the log-likelihood function $l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r)$ with respect to $\mathbf{g}_{\boldsymbol{\mu}}$, $\mathbf{g}_{\boldsymbol{\sigma}}$ and ξ can be derived from first principles. As discussed in Chapter 2, when deriving expressions for the elements of the expected information matrix, that is expectations of the negative second derivatives of the log-likelihood $l(\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}, \xi; r)$, when $\xi \neq 0$ the general expectations provided by Tawn (1988)

$$E\big[Y\big] = \frac{(-\xi)^{c-m}}{\Gamma(j)} \sum_{p=0}^{m} (-1)^p \binom{m}{p} \Gamma^{(c)}(j + b\,\xi - p\,\xi + 1)$$

where $Y = \left(Z^{(k)}\right)^m \left(1 + \xi Z^{(k)}\right)^{-\left(\frac{1}{\xi} + b\right)} \left(\ln\left(1 + \xi Z^{(k)}\right)\right)^c$ and $Z^{(k)} = \frac{X^{(k)} - \mu}{\sigma}$ is used. When $\xi = 0$, the general expectations provided by Smith (1986) gives the expectation

$$E\left[\left(Z^{(k)}\right)^m \exp\left(-\rho Z^{(k)}\right)\right] = \frac{(-1)^m \Gamma^{(m)}(k + \rho)}{\Gamma(k)} \tag{4.10}$$

where $Z^{(k)} = \frac{X^{(k)}-\mu}{\sigma}$ is used. Furthermore, in deriving the expectations, various properties and relationships for the Gamma and Digamma function were used repeatedly. These can be found in Appendix A.

For completeness and due to modifications to the various models, explicit expressions for the elements of the Fisher scoring algorithm were derived from first principles. A total of eight different models were considered, that is using various combinations with $r = 1$ or $r > 1$, $\xi \neq 0$ or $\xi = 0$ and keeping the scale parameter constant $\sigma(t) = \sigma$ or allowing the scale parameter to vary with time. Full details regarding the explicit expression for the derivatives of the log-likelihood and elements of the expected information matrix for the various models can be found in Appendix B, C, D, E, F, G, H and I.

4.3.4 Practical Implementation of the Fisher Scoring Algorithm

The Fisher Scoring algorithm is implemented using a back-fitting algorithm that involves two loops, an inner and an outer loop, and additional checks to determine that the constraints of the GEV distribution or the distribution for the r largest order statistics with $\xi \neq 0$ have been satisfied. The process starts by choosing initial estimates for $\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}$ and ξ, which are checked against the constraints given by equation (4.9) for $i = 1, 2, \ldots, N$. If the constraints are satisfied, the penalized log-likelihood given by (4.5) is calculated at the initial estimates and the algorithm moves onto the inner loop; otherwise new initial estimates for $\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}$ and ξ must be chosen.

The inner loop involves an iterative process starting with the initial estimates for $\mathbf{g}_{\boldsymbol{\mu}}^{old}$, $\mathbf{g}_{\boldsymbol{\sigma}}^{old}$ and ξ^{old} to calculate $\mathbf{g}_{\boldsymbol{\mu}}^{new}$ in equation (4.6), then, using (4.6) and the initial estimates for $\mathbf{g}_{\boldsymbol{\mu}}^{old}$, $\mathbf{g}_{\boldsymbol{\sigma}}^{old}$ and ξ^{old} to calculate $\mathbf{g}_{\boldsymbol{\sigma}}^{new}$ in equation (4.7) and finally, using (4.6), (4.7) and the initial estimates for $\mathbf{g}_{\boldsymbol{\mu}}^{old}$, $\mathbf{g}_{\boldsymbol{\sigma}}^{old}$ and ξ^{old} to calculate ξ^{new} in equation (4.8). The process is then repeated using the new estimates as the initial estimates and the cycling between the three updating equations (4.6), (4.7) and (4.8) continues until $\sum_{i=1}^{N} |g_{\sigma}^{new}(t_i) - g_{\sigma}(t_i)| < 0.01$ when the inner loop is said to have converged.

After each iteration in the inner loop, the constraints need to be checked. If any of the terms given by equation (4.9) for $i = 1, 2, \ldots, N$ are less than zero,

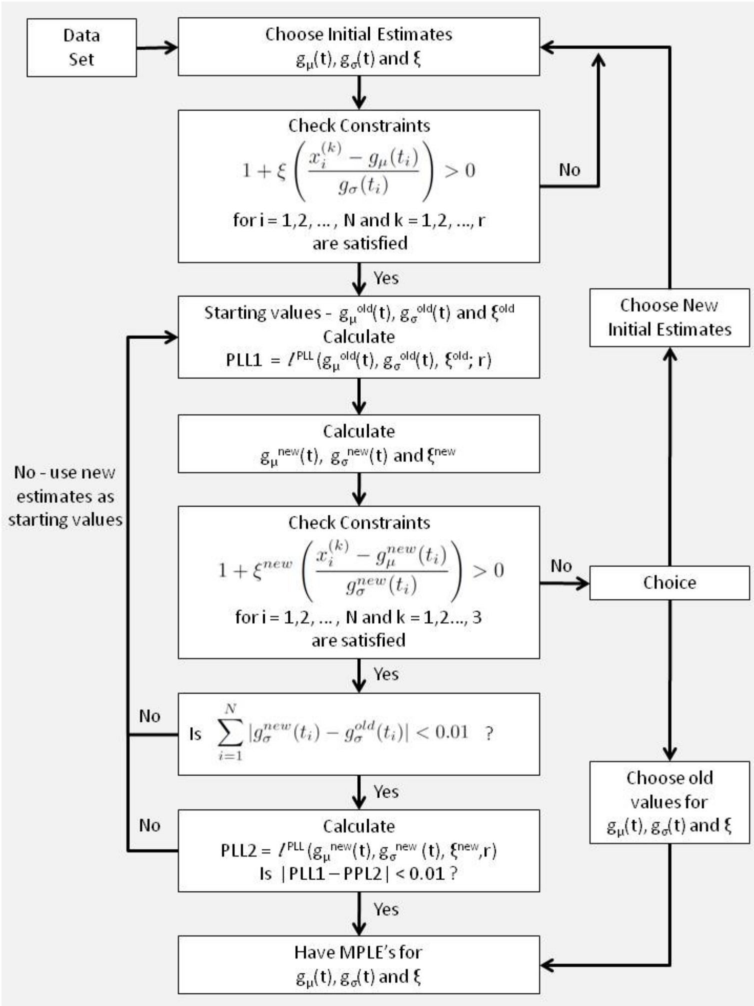

Figure 4.1: Flow chart for determine MPLE's using the Fisher scoring algorithm

a choice must be made, either the process must be started again using new starting values for $\mathbf{g}_{\boldsymbol{\mu}}, \mathbf{g}_{\boldsymbol{\sigma}}$ and ξ or, alternatively, the parameters chosen for the MPLE's are those that last satisfied the constraints.

Once the inner loop converges, a new value for the penalized log-likelihood can be calculated at the values of $\mathbf{g}_{\boldsymbol{\mu}}^{new}$, $\mathbf{g}_{\boldsymbol{\sigma}}^{new}$ and ξ^{new}, that is the new estimates from the final iteration of the inner loop. If the absolute difference between the new value for the penalized log-likelihood and the old value is more than 0.01, the algorithm then returns to the inner loop to obtain new estimates. The process of moving between the outer and inner loop continues until the absolute difference between the new value for the penalized log-likelihood and the old value is less than 0.01, when the outer loop is said to have converged. Once the outer loop converges, estimates from the final iteration are used for the MPLE's.

The back-fitting algorithm and appropriate checks are summarized as a flow chart in Figure 4.1.

4.4 Choosing the Smoothing Parameter

Different approaches to selecting the optimal smoothing parameter, within the context of smoothing NCS, were discussed in Section 3.4. The approach used for choosing the smoothing parameters α and λ, when the penalized log-likelihood approached is used to smooth extremes, is cross-validation.

Returning to first principles, that is the general setting, the 'leave-one-out' cross-validation score function is given by

$$CV(\alpha,\lambda) = \frac{1}{n}\sum_{j=1}^{n}(x_j - \hat{f}^{(-j)}(t_j))^2$$

where x_j is the j^{th} observation, $\hat{f}^{(-j)}(t_j)$ is the predicted value of the j^{th} observation and the fitted curve $\hat{f}^{(-j)}(t)$ is estimated by removing the j^{th} observation from the dataset.

When attempting to smooth extremes using the non-stationary GEV distribution or the non-stationary distribution for the r largest order statistics, the j^{th} observation is removed from the dataset and the MPLE's $\hat{g}_\mu^{(-j)}(t)$, $\hat{g}_\sigma^{(-j)}(t)$ and $\hat{\xi}^{(-j)}$ are calculated by maximizing equation (4.5) using the remaining data. The predicated value $\hat{f}^{(-j)}(t_j)$ for the j^{th} observation, assuming that it is the k^{th} largest observation observed at time point t_i, is then estimated using the expected value

$$E\left[X_i^{(k)}\right] = \hat{g}_\mu^{(-i)}(t_i) - \frac{\hat{g}_\sigma^{(-i)}(t_i)}{\hat{\xi}^{(-i)}}\left(1 - \frac{\Gamma(k-\hat{\xi}^{(-i)})}{\Gamma(k)}\right).$$

The cross-validation score function can therefore be written as

$$CV(\alpha,\lambda) = \frac{1}{Nr}\sum_{k=1}^{r}\sum_{i=1}^{N}\left(x_i^{(k)} - E\left[X_i^{(k)}\right]\right)^2$$

The optimal smoothing parameters values for α and λ are those that minimize $CV(\alpha,\lambda)$.

Chapter 5

The Simulation Study

The aim of this mini-dissertation is to investigate the performance of the roughness penalty approach in modelling non-stationary extremes. Chapter 4 laid down the mathematical foundation for smoothing non-stationary extremes by fitting smoothed natural cubic splines (NCS's) to the location and scale parameters of the GEV distribution and the distribution for the r largest order statistics. The aim of this chapter is to assess the performance of this approach using simulated data and to identify the effects of changing certain parameters. Matlab version 7.8.0.347 (R2009a) was used to simulate the data, fit the various models to the simulated data sets and to generate the results.

The structure of the chapter is as follows. In Section 1, the objectives of the simulation study are outlined. Then, a detailed outline of how extreme random variates are simulated is given in Section 2. The results from the full simulation with $\xi = 0$ and $\sigma(t) = \sigma$ are presented in Section 3. Four individual simulations are examined in greater detail in Section 4 in order to identify the effects that the smoothing parameters have on the maximum penalized likelihood estimates (MPLE's) for the location, scale and shape parameters. Finally, conclusions from the simulation study are presented in Section 5.

5.1 Objectives of the Simulation Study

By using simulated data, the examples in this chapter serve the purpose of investigating how the extreme value theory methodology is implemented. This has the advantage of focusing on the performance of the approach rather than the conclusions drawn from a particular application to real data. The main objectives of the simulation study are as follows:

1. To determine if it is always possible to identify an optimal smoothing parameter when using cross-validation.

2. To evaluate the effects of changing the sample size n and the number of extremes r.

3. To consider the properties of the MPLE's for the location, scale and shape parameters ($\hat{g}_\mu(t)$, $\hat{g}_\sigma(t)$ and $\hat{\xi}$) and to compare the estimates to their true values, $\mu(t)$, $\sigma(t)$ and ξ respectively.

4. To identify the effects of changing the smoothing parameters α and λ on the MPLE's for the location, scale and shape parameters, for the non-stationary GEV distribution and the non-stationary distribution for the r largest order statistics.

The first three objectives are addressed by performing a full simulation on the non-stationary Gumbel distribution and the distribution for the r largest order statistics with $\xi = 0$ and a constant scale parameter $\sigma(t) = \sigma$, by repeatedly simulating data sets for specified configurations of n and r. It was not feasible to conduct the full simulation for the scenarios when $\xi \neq 0$ or $\sigma(t) = g_\sigma(t)$ due to the computation time required to select the optimal smoothing parameters. The fourth objective is addressed by considering four individual simulations. The first two individual simulations consider the case when $\xi = 0$ and $r = 5$, one with a constant scale parameter and the second with a non-constant scale parameter. The final two individual simulations consider the case when $\xi \neq 0$ and $r = 5$, one with a constant scale parameter and the second with a non-constant scale parameter.

5.2 Simulating Extremes

Part of the simulation study requires the efficient generation of extreme random variates. The aim of this section is to outline the procedures used to simulate both maxima and the r largest extreme random variates from the appropriate distributions.

5.2.1 Simulating Maxima

Random variates from the GEV distribution can be generated by making use of the probability integral transform, sometimes referred to as the inverse probability transform method. An extreme random variate x from the GEV distribution, whose cumulative distribution function $G(x)$ is given by equation (2.8), can be generated using the following two steps:

1. Generate a value u from the standard uniform distribution $U(0,1)$.
2. Identify the value x which solves the equation $G(x) = u$, that is find the value x such that $x = G^{-1}(u)$ where G^{-1} is the inverse of G.

The inverse of the GEV distribution function is obtained by inverting equation (2.5) to give

$$x = \mu - \frac{\sigma}{\xi} + \frac{\sigma}{\xi}(-\ln u)^{-\xi}. \tag{5.1}$$

In the case when $\xi \to 0$, the inverse of the Gumbel distribution function is obtained by inverting equation (2.8) to give

$$x = \mu - \sigma \ln(-\ln u). \tag{5.2}$$

5.2.2 Simulating the r Largest Extremes

The inverse probability transform method can theoretically be used to simulate the r largest extreme random variates by generating the first component, for example the maximum or rth largest extreme, from its marginal distribution, then generating the second from its distribution conditional on the first, and so on. However, a relatively simpler and more efficient method involves simulating the r largest standard extreme random variates with $\xi = 0$, using results from David and Nagaraja (2003), and then applying the appropriate transform to obtain the r largest extreme random variates for general μ, σ and ξ.

Results from David and Nagaraja (2003)

As discussed in Chapter 2, David and Nagaraja (2003) showed that the limiting joint distribution function for the r largest standard extremes with $\xi = 0$, whose limiting asymptotic joint density function is given by

$$f(z^{(1)}, \ldots, z^{(r)}) = \exp\Big(-\exp\big(-z^{(r)}\big)\Big) \prod_{k=1}^{r} \exp\big(-z^{(k)}\big), \tag{5.3}$$

coincides with the joint distribution of

$$\lambda^{-1}(Y_1),\ \lambda^{-1}(Y_1+Y_2),\ \ldots,\ \lambda^{-1}(Y_1+Y_2+\ldots+Y_r) \tag{5.4}$$

where Y_i are independent exponential random variates with mean 1 and $\lambda^{-1}(y) = -\ln(y)$. The procedure used to simulate the r largest standard extreme random variates with $\xi = 0$ involves the following steps:

1. Generate r random numbers $u_1, u_2, \ldots, u_r$ from the uniform distribution $U(0,1)$ and let $y_j = -\ln u_j$ for $j = 1, 2, \ldots, r$. Note that each y_i is an independent and identically distributed exponential random variate with mean 1.

2. Define $m_k = \sum_{j=1}^{k} y_j$. Then the j^{th} largest standard extreme random variate with $\xi = 0$ is calculated as $z^{(j)} = -\ln(m_j)$ for $j = 1, 2, \ldots, r$.

Transforms

When simulating extremes, it is useful to exploit the relationships between the different extreme value distributions. Let $Z^{(k)}$ be a kth largest standard extreme random variable with $\xi = 0$. The joint density function of $(Z^{(1)}, \ldots, Z^{(r)})$ is therefore given by (5.3). Then, by applying the transforms

$$X^{(k)} = \mu + \sigma Z^{(k)} \quad \text{for } k = 1, \ldots, r$$

the Jacobian is given by σ^{-r}. It follows that the joint density function of $(X^{(1)}, \ldots, X^{(r)})$ is identical to that of the r largest order statistics with $\xi = 0$, location parameter μ and scale parameter σ. Furthermore for $\xi \neq 0$, by applying the transforms

$$W^{(k)} = \mu - \frac{\sigma}{\xi} + \frac{\sigma}{\xi} e^{\xi Z^{(k)}}, \quad \text{for } k = 1, \ldots, r$$

the Jacobian is given by $\displaystyle\prod_{k=1}^{r} \frac{1}{\sigma}\Big[1 + \xi\Big(\frac{w^{(k)} - \mu}{\sigma}\Big)\Big]^{-1}$ and the joint density function of $(W^{(1)}, \ldots, W^{(r)})$ is identical to that of the r largest order statistics with location parameter μ, scale parameter σ and shape parameter $\xi \neq 0$.

Therefore, by applying the appropriate transforms to $z^{(j)}$, the j^{th} largest standard extreme random variate with $\xi = 0$, it follows that

$$x^{(j)} = \mu - \frac{\sigma}{\xi} + \frac{\sigma}{\xi} e^{\xi z^{(j)}} \tag{5.5}$$

will be the j^{th} largest extreme random variate with $\xi \neq 0$, and

$$x^{(j)} = \mu + \sigma z^{(j)} \tag{5.6}$$

will be the j^{th} largest extreme random variate with $\xi = 0$. When $r = 1$, equations (5.5) and (5.6) are equivalent to equations (5.1) and (5.2) respectively.

5.3 Full Simulation with $\xi = 0$ and $\sigma(t) = \sigma$

A series of sample data sets from the non-stationary Gumbel maxima model and the r largest order statistic model with $\xi = 0$, with a non-constant location parameter

$$\mu(t) = 10t + 15\sin(0.4\pi t)$$

and with a constant scale parameter $\sigma(t) = 5$ for $t \in [0, 10]$, were generated. A total of 25 different configurations, in terms of samples size n and number r of extremes observed at each time point, were considered. For each configuration of n and r, 100 data set simulations were performed. For $r = 1, 2, 5$ and 10 the different sample sizes included were $n = 10, 15, 20, 30, 50$ and 100. In addition, for $r = 1$ the sample size $n = 200$ was included. The sample size of $n = 200$ for $r = 2, 5$ and 10 and sample sizes of $n > 200$ were not considered because the cross-validation process became too time-consuming.

5.3.1 Procedures

The procedure for the simulation study can be broken down into four different stages. This includes simulating the sample data sets, estimating the MPLE's for a given smoothing parameter λ, estimating the optimal smoothing parameter $\hat{\lambda}$ and comparing the MPLE's with their true parameter values.

1. **Simulating Sample Data Sets**

   For a given configuration of n and r, the interval $[0, 10]$ is divided into n equally spaced points t_i so that $0 = t_1 < t_2 < \ldots < t_{n-1} < t_n = 10$. Each point t_i can be considered as a reference point in time, for example a day, month or year, in which the r largest extremes are observed or, in this case, simulated.

To simulate the r largest extremes at each point t_i, the appropriate values for μ and σ are substituted into equation (5.6). For this scenario, the j^{th} largest extreme random variate with $\xi = 0$ is therefore determined by

$$x_{t_i}^{(j)} = 10t_i + 15\sin(0.4\pi t_i) - 5\ln\left(-\sum_{k=1}^{j}\ln u_k\right) \tag{5.7}$$

where $u_1, u_2, \ldots, u_j$ are independent standard uniform random variates, for $j = 1, 2, \ldots, r$.

2. Parameter Estimation

With only the location parameter being smoothed, the penalized log-likelihood is given by

$$l^{\text{PLL}}(\mathbf{g}_{\boldsymbol{\mu}}, \sigma, 0; r) = l(\mathbf{g}_{\boldsymbol{\mu}}, \sigma, 0; r) - \frac{\lambda}{2}\mathbf{g}_{\mu}^{T}\mathbf{K}\mathbf{g}_{\mu} \tag{5.8}$$

where $\mathbf{K} = \mathbf{Q}\mathbf{R}^{-1}\mathbf{Q}^{\mathbf{T}}$ and $\mathbf{g}_{\mu}^{T} = (g_{\mu}(t_1), g_{\mu}(t_2), \ldots, g_{\mu}(t_n))$.

As described in Chapter 4, to obtain the MPLE $\hat{g}_{\mu}(t)$ for the location parameter $\mu(t)$ and the MPLE $\hat{\sigma}$ for the scale parameter σ, equation (5.8) is maximized by implementing a Fisher scoring algorithm, with the mathematical details provided in Appendix C when $r = 1$ and Appendix E when $r > 1$. As discussed in Section 4.2, the MPLE $\hat{g}_{\mu}(t)$ will be a unique NCS that interpolates $\hat{g}_{\mu}(t_i)$ for $i = 1, 2, \ldots, n$, with knots at the points t_i.

3. Smoothing Parameter Selection

The smoothing parameter λ has a direct effect on both the MPLE's $\hat{g}_{\mu}(t)$ and $\hat{\sigma}$. It is therefore important to be able to determine the value of the smoothing parameter that has the "best fit". 50 different values for the smoothing parameter λ were considered on the interval $[0.0005, 0.025]$, where the upper and lower end point of the interval were selected based on preliminary results. Then for each smoothing parameter λ the cross-validation score function $CV(\lambda)$ is calculated.

Given that the cross-validation score function is calculated for 50 different smoothing parameters, by construction there will always be a minimum, that is, $CV(\lambda)$ can be ranked from smallest to largest. However, that minimum may not be unique and secondly, the smoothing parameter λ

giving rise to that minimum may give results for the MPLE's $\hat{\mu}$ and $\hat{\sigma}$ that are not sensible. Therefore, for each dataset it was possible to determine a smoothing parameter λ_{min} that minimized the cross-validation score function $CV(\lambda)$. Using λ_{min} it was then possible to compare the MPLE $\hat{\sigma}$ to the true scale parameter $\sigma = 5$ to determine if the results made sense. The following decision rule was implemented to determine if an optimal smoothing parameter could be determined. If $\hat{\sigma} \geq 1$, then λ_{min} was accepted as being the optimal smoothing parameter, that is $\hat{\lambda} = \lambda_{min}$, and if $\hat{\sigma} < 1$, the results were deemed not to be meaningful and it was concluded that no optimal smoothing parameter could be found.

Due to the fact that there is no guarantee that an optimal smoothing parameter can be found, an additional smoothing parameter was recorded. This was the smoothing parameter $\lambda_{\sigma=5}$ that gave the closest MPLE $\hat{\sigma}$ to the true scale parameter $\sigma = 5$. Note that in practice this is not possible as the true scale parameter is unknown.

4. **Assessment of Parameter Estimates**

One of the advantages of using simulated data is that the MPLE's $\hat{g}_\mu(t)$ and $\hat{\sigma}$ can be directly compared to the true location parameter $\mu(t) = 10t + 15\sin(0.4\pi t)$ and scale parameter $\sigma(t) = 5$ respectively.

The mean summed squared error (MSSE), introduced by Ruppert *et al.* (2003), is used to compare the MPLE for the location parameter $g_\mu(t)$ to the true location parameter $\mu(t)$, for a given smoothing parameter λ. The MSSE considers the squared error between the true and estimated functions for the location parameter at each knot t_i (Ruppert *et al.*, 2003) so that

$$MSSE_\mu(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{g}_\mu(t_i) - \mu(t_i)\right]^2. \tag{5.9}$$

When comparing $MSSE_\mu(\lambda)$ for various smoothing parameters, a smaller value for $MSSE_\mu(\lambda)$ would indicate a better fit relative to the true location parameter.

With the use of a constant scale parameter $\sigma = 5$ in the simulation study, a direct comparison can be made between the MPLE $\hat{\sigma}$ and σ.

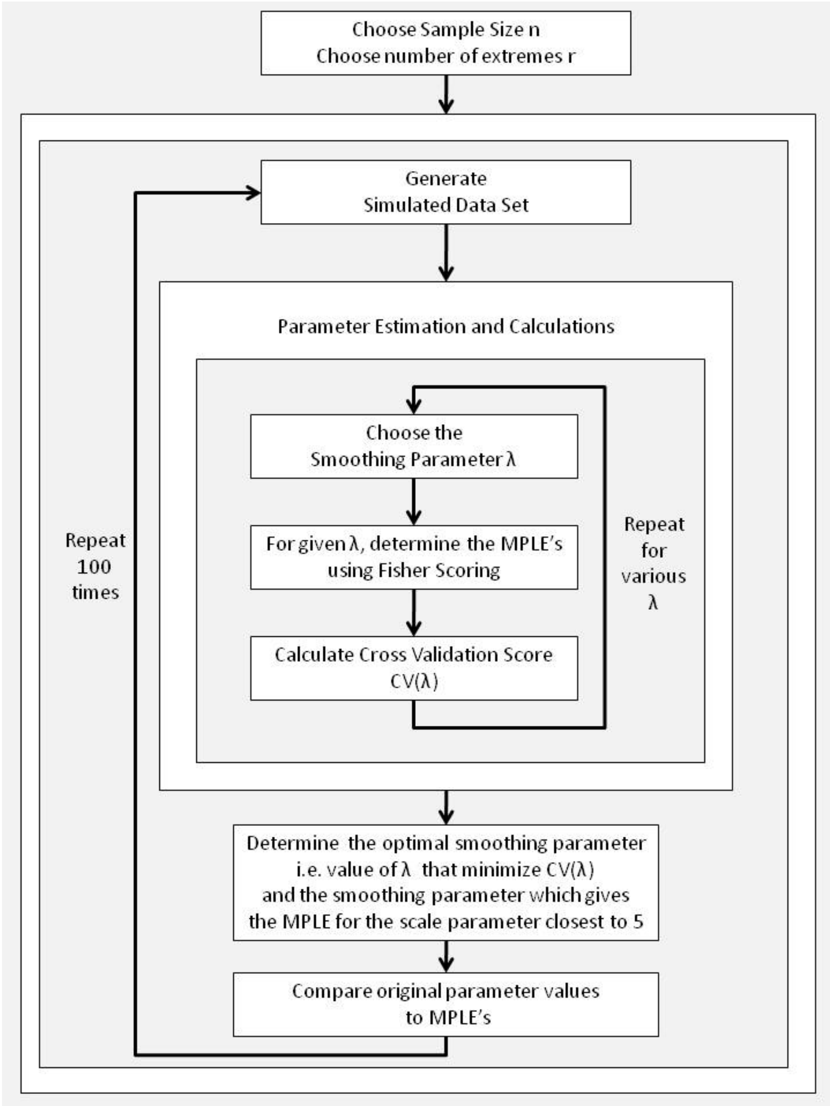

Figure 5.1: Flowchart of the procedure followed for the simulation study with $\xi = 0$

An outline of the procedures followed in the simulation study with $\xi = 0$ is summarized in the flow chart in Figure 5.1.

5.3.2 Results

Within the context of smoothing extremes, there has been certain criticism regarding the use of cross-validation as an automatic selection criterion for the smoothing parameter. Chavez-Demoulin and Davison (2005) describe the use of cross-validation to select the smoothing parameter as "computationally demanding", Chavez-Demoulin and Embrechts (2004) describes the process as becoming "computationally costly when the size of the data increases" and Pauli and Coles (2001) explain that in practice they found it "difficult to implement within the extreme value context because of the heavy computational burden involved". Furthermore, there is no evidence that it is always possible to determine an optimal smoothing parameter.

Over the last few years, modern computational power has substantially increased the speed at which calculations can be made, thus making it easier to implement cross-validation techniques. Consequently, for every simulated data set, the time taken to run the cross-validation process was recorded and also whether or not it was possible to determine an optimal smoothing parameter. A summary of the cross-validation results, from the simulation study with $\xi = 0$, when $r = 1$ is given in Table 5.1 and when $r > 1$ in Table 5.2.

From the results in Tables 5.1 and 5.2, the following observations can be made:

- There is an increase in the proportion of data sets for which an optimal smoothing parameter can be found as the sample size n increases. It appears that with a sample size of $n = 50$, it is almost always possible to obtain an optimal smoothing parameter. Surprisingly however, the proportion of data sets for which an optimal smoothing parameter exists decreases as the number of extremes r increases.

- The average time taken to run the cross-validation process, in order to determine the optimal smoothing parameter, increases rapidly with an increase in the sample size n. This confirms the statement made by Chavez-Demoulin and Embrechts (2004) that the cross-validation process becomes "computationally costly when the size of the data increases". This is particularly evident with a sample size of $n = 200$.

- For sample sizes of $n \leq 50$, there appears to be no direct relationship between the average time to run the cross-validation process and the number of extremes r. However, for the sample sizes $n = 50$ and $n = 100$, increasing the number of extremes r leads to a substantial increase in the average time taken to run the cross-validation process.

For each of the 2500 data sets, recalling that there were 25 different configurations of n and r with 100 data sets simulated for each configuration, the MPLE's $\hat{g}_\mu(t)$ and $\hat{\sigma}$ where calculated for each of the 50 different smoothing parameters selected. If it was possible to determine an optimal smoothing parameter $\hat{\lambda}$, the MPLE's $\hat{g}_\mu(t)$ and $\hat{\sigma}$ were recorded for both the smoothing parameters $\hat{\lambda}$ and $\lambda_{\sigma=5}$, that is for the optimal smoothing parameter and smoothing parameter that gave the closest MPLE $\hat{\sigma}$ to the true scale parameter $\sigma = 5$. Furthermore, to make a comparison of the performance of the MPLE $\hat{g}_\mu(t)$ for both the smoothing parameters $\hat{\lambda}$ and $\lambda_{\sigma=5}$, the MSSE is calculated for each smoothing parameter using equation (5.9), that is $MSSE_\mu(\hat{\lambda})$ and $MSSE_\mu(\lambda_{\sigma=5})$. However, if it was not possible to determine an optimal smoothing parameter $\hat{\lambda}$, the MSSE was not recorded. A summary of these results when $r = 1$ is given in Table 5.3 and when $r > 1$ in Table 5.4.

From the results in Tables 5.3 and 5.4, the following observations can be made:

- When using the optimal smoothing parameter $\hat{\lambda}$, increasing the sample size n and number of extremes r provides a relatively better MPLE $\hat{\sigma}$ compared to the true scale parameter $\sigma = 5$. Furthermore, an increase in n and r results in a smaller standard deviation of the MPLE $\hat{\sigma}$, suggesting less variation in the estimates.

- The average $MSSE_\mu(\hat{\lambda})$ decreases with an increase in sample size n and number of extremes r. This suggests that when using the optimal smoothing parameter $\hat{\lambda}$, relatively better estimates are obtained for the MPLE $\hat{g}_\mu(t)$ by increasing n and r.

- It is also interesting to note that the optimal smoothing parameter $\hat{\lambda}$ appears to provided on average better estimates for the location parameter $\mu(t)$ than the smoothing parameter $\lambda_{\sigma=5}$, with $MSSE_\mu(\hat{\lambda}) < MSSE_\mu(\lambda_{\sigma=5})$.

	Sample Size n						
	10	15	20	30	50	100	200
Proportion where optimal smoothing parameter λ found							
	21%	57%	81%	96%	100%	100%	100%
Time taken to run cross-validation							
Minimum (seconds)	8.12	8.49	12.48	16.60	44.07	216.77	2039.71
Maximum (seconds)	26.63	29.94	52.63	108.82	326.36	5016.35	4713.05
Average (seconds)	13.81	18.00	24.30	47.50	98.39	353.50	2865.32
Standard Deviation	3.82	5.00	8.55	23.05	61.05	471.26	529.38
Optimal Estimate for Smoothing parameter $\hat{\lambda}$							
Average	0.0112	0.0073	0.0068	0.0065	0.0063	0.0076	0.0101
Standard Deviation	0.0069	0.0029	0.0029	0.0025	0.0028	0.0036	0.0048

Table 5.1: Summary of the cross-validation results from the simulation study with $r = 1$

		Sample Size n					
		10	15	20	30	50	100
Proportion where optimal smoothing parameter λ found							
	r = 2	20%	76%	89%	97%	100%	100%
	r = 5	5%	69%	85%	95%	98%	100%
	r = 10	1%	57%	80%	92%	96%	100%
Time taken to run cross-validation							
Minimum (seconds)	r = 2	9.78	15.50	26.15	40.75	81.91	300.32
	r = 5	10.08	19.03	28.15	48.51	163.85	564.82
	r = 10	13.29	25.53	39.80	79.19	262.40	973.02
Maximum (seconds)	r = 2	30.47	65.75	102.48	94.71	194.43	437.17
	r = 5	30.70	38.03	43.99	63.91	195.49	1062.00
	r = 10	18.03	30.86	46.37	114.05	302.35	1590.37
Average (seconds)	r = 2	18.04	33.88	46.61	58.41	117.50	355.22
	r = 5	18.44	23.75	31.64	55.42	175.27	663.14
	r = 10	15.34	27.94	43.00	85.98	278.79	1050.04
Standard Deviation	r = 2	5.12	11.89	17.16	13.62	17.80	27.69
	r = 5	4.12	4.09	2.04	2.73	5.64	100.62
	r = 10	0.96	1.05	1.29	4.93	7.82	93.38
Optimal Estimate for Smoothing parameter $\hat{\lambda}$							
Average	r = 2	0.0052	0.0060	0.0070	0.0083	0.0082	0.0120
	r = 5	0.0019	0.0058	0.0082	0.0098	0.0128	0.0166
	r = 10	0.0025	0.0067	0.0102	0.0137	0.0162	0.0197
Standard Deviation	r = 2	0.0027	0.0029	0.0027	0.0036	0.0035	0.0048
	r = 5	0.0004	0.0026	0.0035	0.0047	0.0064	0.0067
	r = 10	0.0000	0.0047	0.0049	0.0060	0.0062	0.0067

Table 5.2: Summary of the cross-validation results from the simulation study with $r > 1$

	Sample Size n						
	10	15	20	30	50	100	200
MPLE for Scale Parameter $\hat{\sigma}$							
Minimum	2.0972	2.0431	2.2369	2.3645	2.3680	3.7358	4.1076
Maximum	9.8508	6.5757	6.7966	6.9636	6.0935	5.4358	5.4458
Average	5.4355	3.6273	3.5060	3.9863	4.1619	4.6329	4.7926
Standard Deviation	2.1341	1.1971	1.0556	0.8275	0.6238	0.3953	0.2928
Average MSSE for Location Parameter $\mu(t)$							
$MSSE_\mu(\hat{\lambda})$	52.9983	23.9652	15.9412	10.2474	5.4795	2.9319	1.4938
$MSSE_\mu(\lambda_{\sigma=5})$	50.8774	33.0345	22.5185	13.9483	8.5545	4.0328	1.7297
Proportion where $MSSE_\mu(\hat{\lambda}) < MSSE_\mu(\lambda_{\sigma=5})$							
	0.5238	0.7895	0.7412	0.7396	0.7900	0.7900	0.7200

Table 5.3: Summary of the MPLE results from the simulation study with $r = 1$

		Sample Size n					
		10	15	20	30	50	100
MPLE for Scale Parameter $\hat{\sigma}$							
Minimum	r = 2	2.0297	2.0538	2.2379	2.5545	3.4526	4.0383
	r = 5	3.4448	3.1212	2.9302	3.4136	3.8388	4.3510
	r = 10	4.1042	3.9056	3.9065	3.9493	4.3170	4.5259
Maximum	r = 2	6.3601	6.6235	6.7784	6.4136	5.8270	5.9505
	r = 5	5.7836	5.6359	5.5563	6.0567	5.7474	5.4327
	r = 10	4.1042	5.5525	5.7729	5.4463	5.4574	5.4194
Average	r = 2	3.3387	3.5985	3.9711	4.2464	4.5310	4.7659
	r = 5	4.3412	4.2992	4.3723	4.5992	4.8330	4.8680
	r = 10	4.1042	4.5985	4.6712	4.7472	4.8569	4.9188
Standard Deviation	r = 2	1.1093	0.8565	0.8605	0.6568	0.4728	0.3437
	r = 5	0.8674	0.5962	0.4717	0.4491	0.3361	0.2059
	r = 10	0.0000	0.4019	0.3811	0.2720	0.2188	0.1685
Average MSSE for Location Parameter $\mu(t)$							
$MSSE_\mu(\hat{\lambda})$	r = 2	17.8471	9.2860	7.5781	4.9150	2.8728	1.5588
	r = 5	9.2623	5.1233	4.0223	2.5210	1.4536	0.7589
	r = 10	7.8089	3.3317	2.7490	1.7476	0.9788	0.5110
$MSSE_\mu(\lambda_{\sigma=5})$	r = 2	28.5712	17.8450	13.3973	7.3920	3.8938	1.8721
	r = 5	11.0742	5.7108	4.4562	2.7000	1.4923	0.7695
	r = 10	5.0809	3.0003	2.6496	1.6542	0.9499	0.5257
Proportion where $MSSE_\mu(\hat{\lambda}) < MSSE_\mu(\lambda_{\sigma=5})$							
	r = 2	0.8500	0.8158	0.8202	0.8660	0.8100	0.7900
	r = 5	0.8000	0.5797	0.6941	0.7158	0.5918	0.5500
	r = 10	0.0000	0.4737	0.4875	0.4130	0.4583	0.3700

Table 5.4: Summary of the MPLE results from the simulation study with $r > 1$

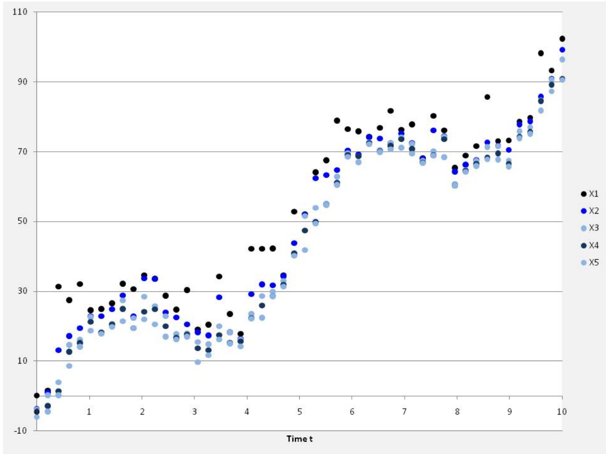

Figure 5.2: Simulated sample data set 1 with $\mu(t) = 10t+15\sin(0.4\pi t)$, $\sigma(t) = 5$ and $\xi = 0$

5.4 Individual Simulations

The aim of this section is to identify the effects of changing the smoothing parameters α and λ on the MPLE's for the location, scale and shape parameters. This is achieved by taking a closer look at individual data sets simulated from the non-stationary distributions for the r largest order statistics. When $\xi = 0$ and $\xi \neq 0$, two individual simulations were considered. The first simulated data set relates to the scenario where only the location parameter $\mu(t)$ is smoothed with a NCS and the second simulated data set relates to the smoothing of both the location and scale parameters $\mu(t)$ and $\sigma(t)$ with NCS's. In order to try to remove the effects that different sample sizes n, numbers of extremes r and the time interval $[0, T]$ have on parameter estimates, these factors where kept constant for all the individual simulations considered, with $n = 50$, $r = 5$ and a time interval $[0, 10]$.

5.4.1 Case of $\xi = 0$

Simulated data set with σ constant

The first simulated data set is taken directly from the simulation study described earlier with $\xi = 0$, $\mu(t) = 10t + 15\sin(0.4\pi t)$ and $\sigma(t) = 5$. For each of the fifty equally spaced time points t_i, five realizations are drawn from the non-stationary distribution for the 5 largest order statistics with $\xi = 0$, such that the j^{th}

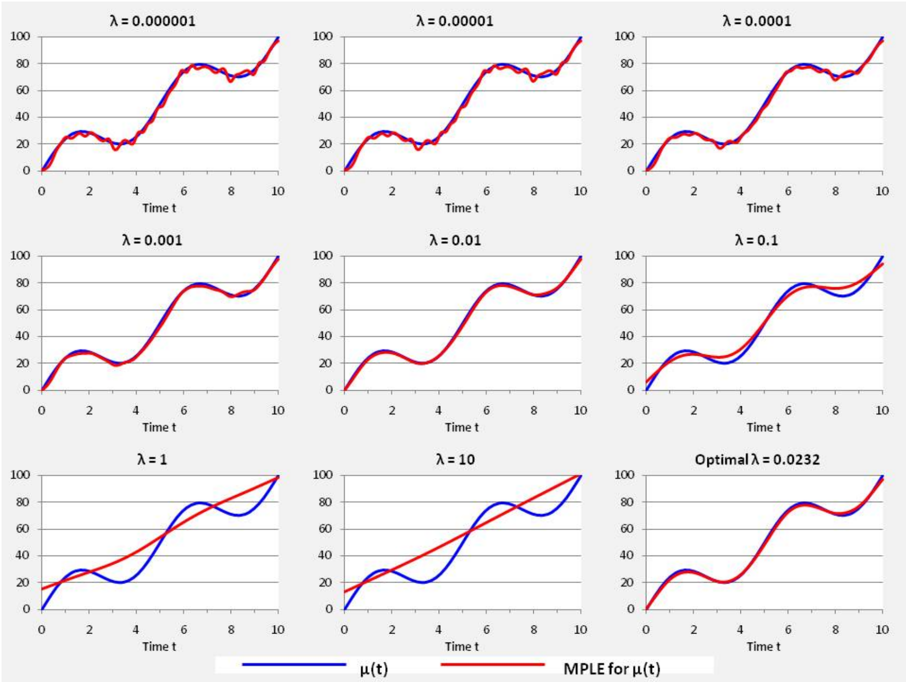

Figure 5.3: MPLE $\hat{g}_{\mu(t)}$ for various smoothing parameters for the simulated sample data set 1

largest extreme random variate is given by equation (5.7) for $j = 1, 2, \ldots, 5$. The simulated data set, labelled simulated sample data set 1, is illustrated in Figure 5.2.

The MPLE's $\hat{g}_{\mu}(t)$ and $\hat{\sigma}$ where calculated for nine smoothing parameters, including the optimal smoothing parameter $\hat{\lambda}$ selected using the cross validation criterion, by maximizing the penalized likelihood (5.8). The MPLE $\hat{\sigma}$ and the $MSSE_{\mu}(\lambda)$ obtained for the various smoothing parameters λ are shown in Table 5.5. Its evident that the size of the smoothing parameter λ also has a direct effect on the MPLE estimate for the scale parameter $\hat{\sigma}$. When λ is small, $\hat{\sigma}$ is relatively smaller in comparison to the original σ and as λ increases in size, so does the value for $\hat{\sigma}$.

Figure 5.3 represents graphically the MPLE $\hat{g}_{\mu}(t)$ compared to the original location parameter $\mu(t)$, for the various smoothing parameters λ, showing that as λ increases in size, the MPLE $\hat{g}_{\mu}(t)$ gets relatively smoother until it eventually shows no curvature and becomes a straight line.

Smoothing Parameter λ	MPLE $\hat{\sigma}$	$MSSE_{\mu}(\lambda)$
0.000001	4.028	5.774
0.00001	4.032	5.426
0.0001	4.157	3.757
0.001	4.477	1.738
0.01	4.721	1.015
0.1	6.062	13.241
1	9.935	100.097
10	10.933	130.521
$\hat{\lambda} = 0.0232$	4.834	1.424

Table 5.5: Results for the MPLE for $\hat{\sigma}$ and the MSSE for $\mu(t)$ for the simulated sample data set 1

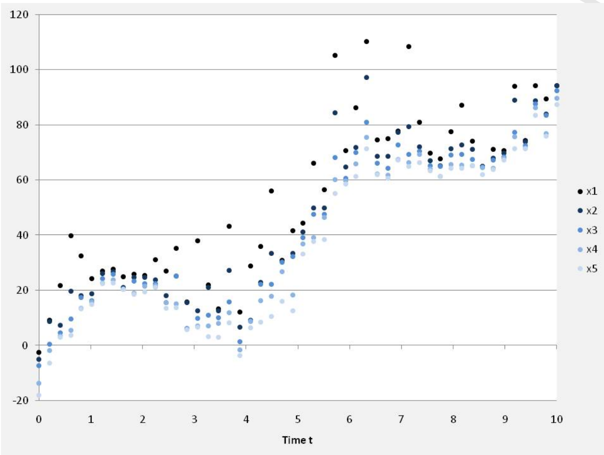

Figure 5.4: Simulated sample data set 2 with $\mu(t) = 10t + 15\sin(0.4\pi t)$, $\sigma(t) = 10 - 5\sin(0.3\pi t)$ and $\xi = 0$

Simulated data set with $\sigma(t)$ varying

The second individual simulation with $\xi = 0$ investigates the performance of the model when both the location parameter $\mu(t)$ and the scale parameter $\sigma(t)$ are smoothed with NCS's $g_{\mu}(t)$ and $g_{\sigma}(t)$ respectively.

The data is simulated from the non-stationary distribution for the 5 largest order statistics with $\xi = 0$ and

$$\mu(t) = 10t + 15\sin(0.4\pi t)$$

$$\sigma(t) = 10 - 5\sin(0.3\pi t)$$

Smoothing	Smoothing Parameter α						
Parameter λ	0.00001	0.0001	0.001	0.01	0.1	1	100
0.00001	29.046	27.437	22.150	19.254	21.483	31.791	44.887
0.0001	17.470	16.261	13.499	11.796	14.025	24.699	39.036
0.001	10.751	9.045	7.219	6.909	9.200	20.333	36.361
0.01	9.545	8.079	8.124	7.063	5.148	15.278	32.373
0.1	62.915	65.491	69.547	66.413	45.352	64.003	64.374

Table 5.6: Results for the MSSE for $\mu(t)$ for the simulated sample data set 2

so that at time t_i the j^{th} largest extreme random variate with $\xi = 0$ is given by

$$x^{(j)} = 10t + 15\sin(0.4\pi t) + [10 - 5\sin(0.3\pi t)]\ln\left(-\sum_{k=1}^{j}\ln u_k\right)$$

where $u_1, u_2, \ldots, u_j$ are independent standard uniform random variates for $j = 1, 2, \ldots, 5$. The simulated data set, labelled simulated sample data set 2, is illustrated in Figure 5.4.

The MPLE's $\hat{g}_\mu(t)$ and $\hat{g}_\sigma(t)$ are determined by maximizing the penalized log-likelihood

$$l^{\text{PLL}}(\mathbf{g}_{\boldsymbol{\mu}}, \sigma, 0; r) = l(\mathbf{g}_{\boldsymbol{\mu}}, \sigma, 0; r) - \frac{\lambda}{2}\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu - \frac{\alpha}{2}\mathbf{g}_\sigma^T\mathbf{K}\mathbf{g}_\sigma$$

by implementing a Fisher scoring algorithm as discussed in Chapter 4, with the mathematical details provided in Appendix D. The results for the MPLE's are depicted graphically in Figures 5.5 and 5.6 for various smoothing parameters λ and α. As expected, small values of λ and α result in MPLE's that fit the data more closely and show relatively more variation, whereas large values of λ and α result in 'smoother' curve estimates for the MPLE's. In addition, the MPLE's are compared to the original location and scale parameters by considering the MSSE as shown in Tables 5.6 and 5.7. The results show that the smoothing parameter λ effects both MPLE's $\hat{g}_\mu(t)$ and $\hat{g}_\sigma(t)$, and similarly for the smoothing parameter α. It is also interesting to note how the MSSE changes by fixing one smoothing parameter and varying the other smoothing parameter.

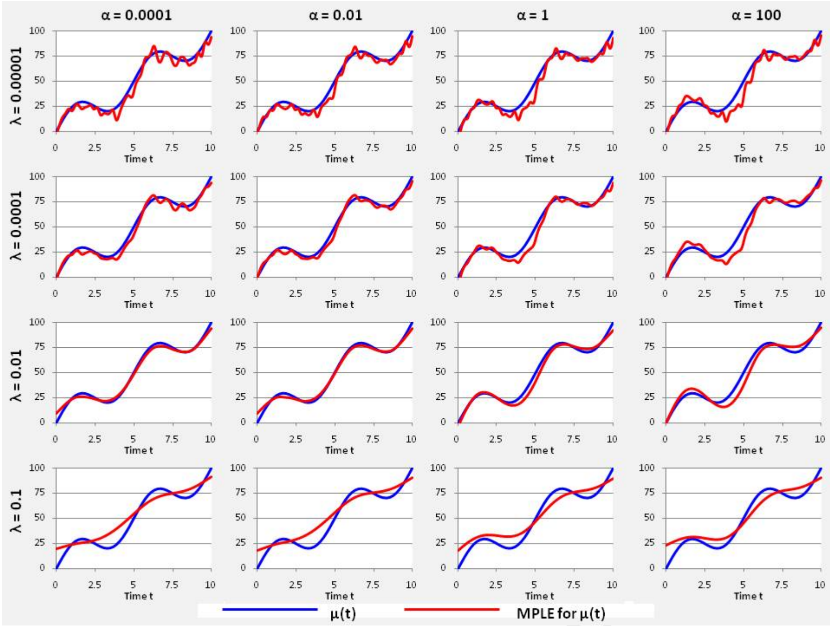

Figure 5.5: MPLE $\hat{g}_{\mu(t)}$ for various smoothing parameters for the simulated sample data set 2

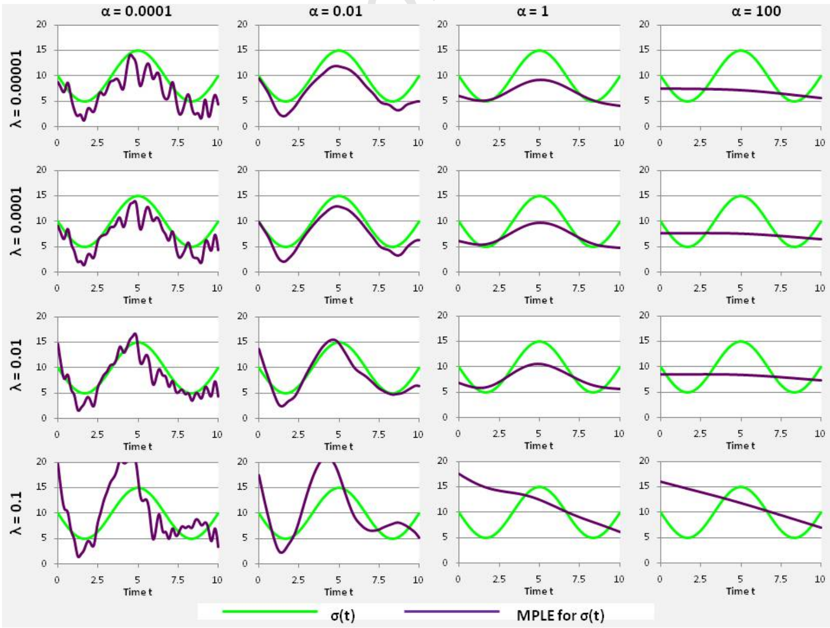

Figure 5.6: MPLE $\hat{g}_{\sigma(t)}$ for various smoothing parameters for the simulated sample data set 2

Smoothing	Smoothing Parameter α						
Parameter λ	0.00001	0.0001	0.001	0.01	0.1	1	100
0.00001	10.778	8.694	6.018	5.161	6.131	10.172	15.025
0.0001	7.783	6.141	3.889	3.132	3.987	7.947	12.972
0.001	6.445	4.797	2.827	2.078	2.620	6.401	11.697
0.01	6.448	5.180	3.959	2.777	1.928	5.406	11.128
0.1	17.590	17.709	17.820	14.932	6.018	27.154	25.140

Table 5.7: Results for the MSSE for $\sigma(t)$ for the simulated sample data set 2

5.4.2 Case of $\xi \neq 0$

Simulated data set with σ constant

The first simulation with $\xi \neq 0$ considers the case when the scale parameter $\sigma(t)$ is kept constant and only the location parameter $\mu(t)$ is smoothed with a NCS $g_\mu(t)$. The data set is simulated from the non-stationary distribution for the 5 largest order statistics with

$$\begin{aligned} \mu(t) &= 10 + 5\sin(0.4t\pi), \\ \sigma(t) &= 3, \\ \text{and} \qquad \xi &= 0.1. \end{aligned}$$

The j^{th} largest extreme random variate with $\xi \neq 0$ is obtained by substituting the appropriate values for μ, σ and ξ into Equation (5.5). Therefore, at the time point t_i the j^{th} largest extreme random variate is given by

$$x_t^{(j)} = 5\sin(0.4t\pi) - 20 + 30\exp\left(-0.1\ln\left(-\sum_{k=1}^{j}\ln u_k\right)\right)$$

for $j = 1, 2, \ldots, 5$. The data set, labelled simulated sample data set 3, is illustrated in Figure 5.7.

The MPLE's $\hat{g}_\mu(t)$, $\hat{\sigma}$ and $\hat{\xi}$ are determined by maximizing the penalized log-likelihood

$$l^{\text{PLL}}(\mathbf{g}_{\boldsymbol{\mu}}, \sigma, \xi; r) = l(\mathbf{g}_{\boldsymbol{\mu}}, \sigma, \xi; r) - \frac{\lambda}{2}\mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu$$

by implementing a Fisher scoring algorithm as discussed in Chapter 4, with the mathematical details provided in Appendix H.

The MPLE's, $\hat{\sigma}$ and $\hat{\xi}$, and the $MSSE_\mu(\lambda)$ obtained for various smoothing parameters λ are shown in Table 5.8. The results show that the smoothing

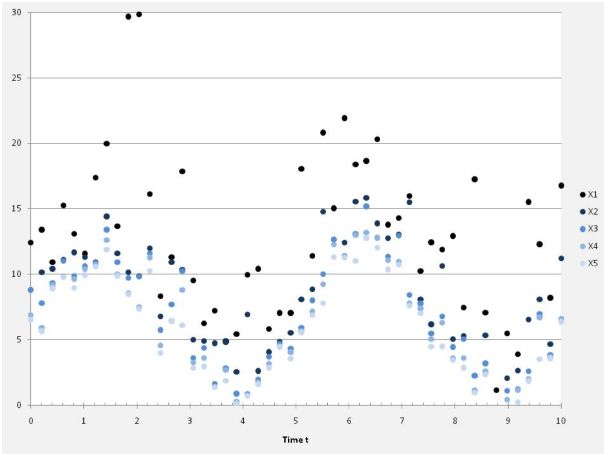

Figure 5.7: Simulated sample data set 3 with $\mu(t) = 10 + 5\sin(0.4t\pi)$, $\sigma(t) = 3$ and $\xi = 0.1$

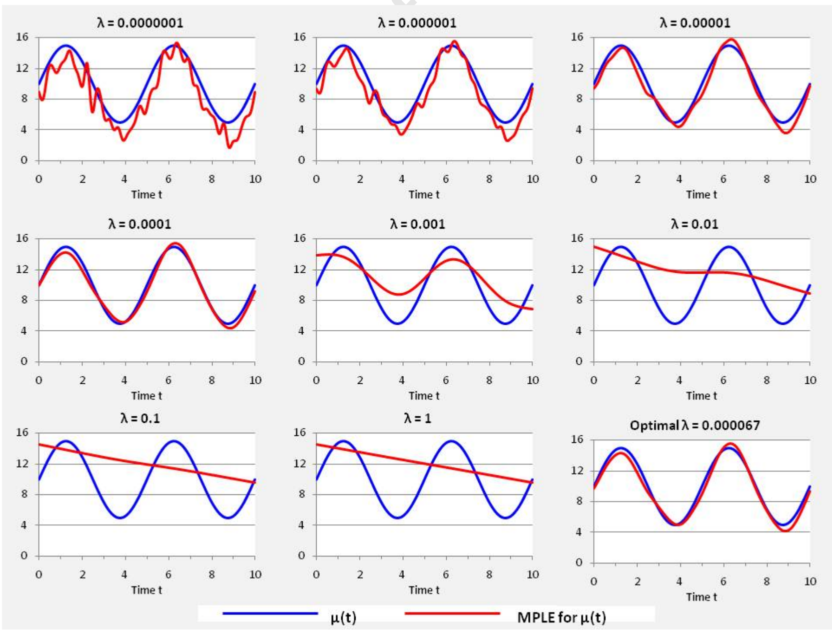

Figure 5.8: Comparison of the MPLE $g_\mu(t)$ with the original $\mu(t)$ for different smoothing parameters λ for simulated sample data set 3

Smoothing Parameter λ	MPLE $\hat{\sigma}$	MPLE $\hat{\xi}$	$MSSE_{\mu}(\lambda)$
0.0000001	2.8332	0.8468	5.4307
0.000001	2.9768	0.5694	3.8025
0.00001	2.9831	0.3163	2.5908
0.0001	2.9670	0.2141	2.3015
0.001	3.5947	0.0178	4.8904
0.01	4.0773	−0.0661	16.7375
0.1	4.1814	−0.0838	16.8432
1	4.2024	−0.0858	16.8125
$\hat{\lambda} = 0.000067$	2.9628	0.2350	2.3550

Table 5.8: Scale and shape parameter estimates for various smoothing parameters λ for simulated sample data set 3

parameter λ has a direct effect on all the MPLE's. Increasing λ increases the value of $\hat{\sigma}$ and decreases the value of $\hat{\xi}$. As before, as λ increases in size, the MPLE $\hat{g}_{\mu}(t)$ gets relatively smoother until it eventually shows no curvature and becomes a straight line. The effect on the MPLE $\hat{g}_{\mu}(t)$ is shown graphically in Figure 5.8.

Simulated data set with $\sigma(t)$ varying

The second simulation with $\xi \neq 0$ investigates the performance of the model when both the location parameter $\mu(t)$ and the scale parameter $\sigma(t)$ are smoothed with NCS's $g_{\mu}(t)$ and $g_{\sigma}(t)$ respectively. The data set is simulated from a non-stationary distribution for the 5 largest order statistics with

$$\begin{aligned} \mu(t) &= 10 + 5\sin(0.4\pi t), \\ \sigma(t) &= 2 + \sin(0.3\pi t), \\ \text{and} \qquad \xi &= -0.1. \end{aligned}$$

The j^{th} largest extreme random variate is obtained by substituting the appropriate values for μ, σ and ξ into Equation (5.5). Therefore, at the time point t_i the j^{th} largest extreme random variate is given by

$$x_{t_i}^{(j)} = 30+5\sin(0.4\pi t)+10\sin(0.3\pi t)+(20-10\sin(0.3\pi t))\exp\left(0.1\ln\left(-\sum_{k=1}^{j}\ln u_k\right)\right)$$

for $j = 1, 2, \ldots, r$. The data set, labelled simulated sample data set 4, is illustrated in Figure 5.9.

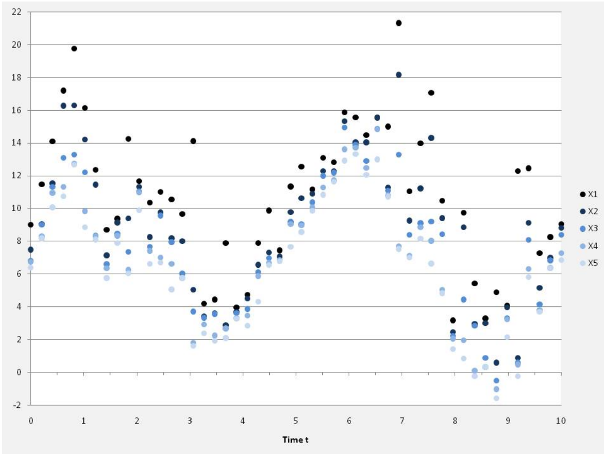

Figure 5.9: Simulated sample data set 4 with $\mu(t) = 10 + 5\sin(0.4\pi t)$, $\sigma(t) = 2 + \sin(0.3\pi t)$ and $\xi = -0.1$

Smoothing Parameter λ	Smoothing Parameter α			
	0.0000001	0.00001	0.001	0.1
0.0000001	-0.3905	0.2086	0.4217	0.5642
0.00001	0.1696	-0.0212	-0.0076	0.0844
0.001	0.8392	0.2560	-0.1335	-0.1996
0.1	0.8557	0.6348	-0.0129	-0.2382

Table 5.9: Results for the MPLE for $\hat{\xi}$ for various smoothing parameters for simulated sample data set 4

The MPLE's $\hat{g}_\mu(t)$, $\hat{g}_\sigma(t)$ and $\hat{\xi}$ are determined by maximizing the penalized log-likelihood

$$l^{\text{PLL}}(\mathbf{g}_{\boldsymbol{\mu}}, \sigma, \xi; r) = l(\mathbf{g}_{\boldsymbol{\mu}}, \sigma, \xi; r) - \frac{\lambda}{2}\mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu - \frac{\alpha}{2}\mathbf{g}_\sigma^T \mathbf{K} \mathbf{g}_\sigma$$

by implementing a Fisher scoring algorithm, with the mathematical details provided in Appendix I. The results for the MPLE $\hat{\xi}$ are shown in Table 5.9 and the results for the MPLE's $\hat{g}_\mu(t)$ and $\hat{g}_\sigma(t)$ are depicted graphically for various smoothing parameters λ and α in Figures 5.10 and 5.11 . In addition, the the MPLE's $\hat{g}_\mu(t)$ and $\hat{g}_\sigma(t)$ are compared to the original location and scale parameters, $\mu(t)$ and $\sigma(t)$ respectively, by considering the MSSE, and the results are shown in Tables 5.10 and 5.11. The results show that the smoothing parameter λ and α have a direct effect on the MPLE's $\hat{g}_\mu(t)$, $\hat{g}_\sigma(t)$ and $\hat{\xi}$.

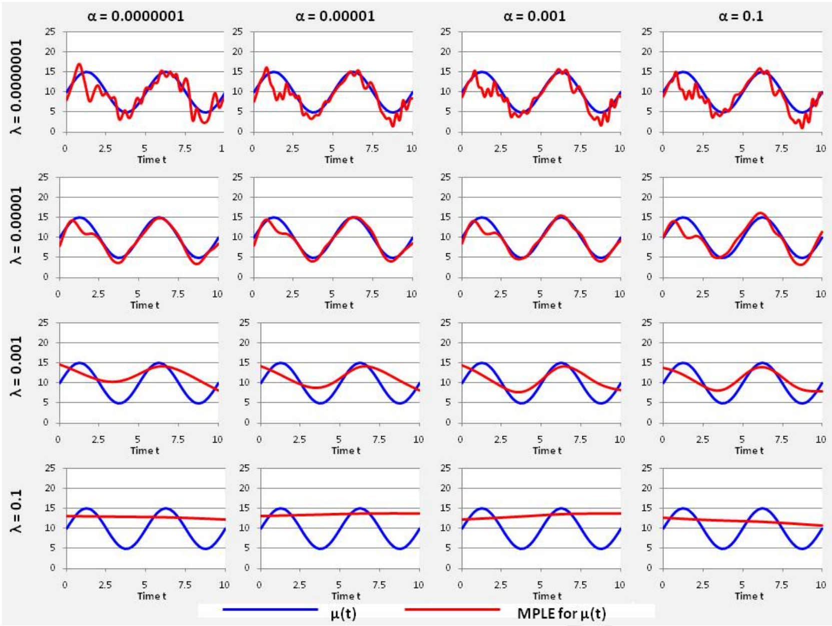

Figure 5.10: Comparison of the MPLE $g_{\mu}(t)$ with the original $\mu(t)$ for different smoothing parameters λ and α for simulated sample data set 4

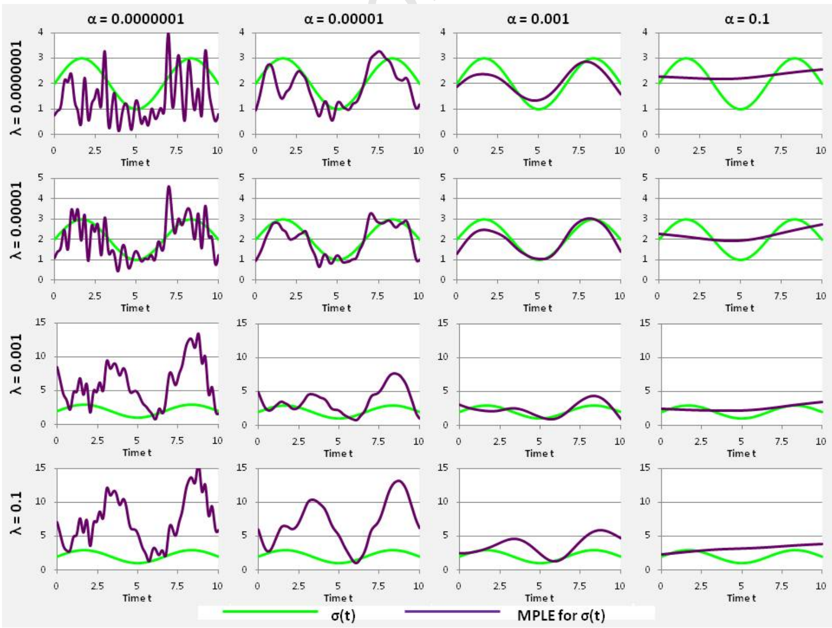

Figure 5.11: Comparison of the MPLE $g_{\sigma}(t)$ with the original $\sigma(t)$ for different smoothing parameters λ and α for simulated sample data set 4

Smoothing Parameter λ	Smoothing Parameter α			
	0.0000001	0.00001	0.001	0.1
0.0000001	4.3177	3.8527	3.6632	4.7008
0.00001	1.4610	1.2649	1.2659	2.8270
0.001	11.0161	8.8425	6.1522	4.8217
0.1	19.5857	26.3851	23.7894	14.2191

Table 5.10: Results for the MSSE for $\mu(t)$ for simulated sample data set 4

Smoothing Parameter λ	Smoothing Parameter α			
	0.0000001	0.00001	0.001	0.1
0.0000001	1.3715	0.4008	0.1268	0.4093
0.00001	0.4868	0.2722	0.1223	0.3324
0.001	22.8231	4.5841	0.5171	0.5037
0.1	36.0295	32.0291	3.3456	1.6606

Table 5.11: Results for the MSSE for $\sigma(t)$ for simulated sample data set 4

5.5 Conclusions from the Simulation Study

The results and examples from the simulation study have firstly served the purpose of demonstrating how the roughness penalty approach is implemented when modelling non-stationary extremes and secondly, have identified the effects of changing certain parameters such as the sample size n, the numbers of extremes r and the smoothing parameter on the MPLE's.

From the full simulation, with $\xi = 0$ and $\sigma(t) = \sigma$, the results show that it is not always possible to identify an optimal smoothing parameter when using cross-validation and that the time taken to run the cross-validation process, in order to determine the optimal smoothing parameter, increases rapidly with an increase in the sample size n. In terms of performance, using the optimal smoothing parameter leads to a lower MSSE on average compared to other smoothing parameters. Furthermore, by increasing the sample size n and the number of extremes r, it was evident that relatively better estimates are obtained for the MPLE's.

When considering the results from the individual simulations, it is evident that the smoothing parameters have a direct effect on the MPLE's. As expected, small values for the smoothing parameter resulted in a very close fit to the underlying data and increasing the value for the smoothing parameter leads to a relatively 'smoother' estimate, with the estimate approaching a straight line as the smoothing parameters tend to infinity. It is however interesting to observe the effect the smoothing parameters have on the MPLE's for constant parameters. For the case when only the location parameter $\mu(t)$ is being smoothed, increasing the smoothing parameter increases the value of the MPLE for the scale parameter $\hat{\sigma}$ and decreases the value of the MPLE for the shape parameter $\hat{\xi}$.

Chapter 6

Application of Smoothing Extremes to Real Data

In this chapter, the smoothing of non-stationary extremes is illustrated on two real data sets. The data consist of motor and of property reinsurance claims incurred by the Swiss Reinsurance Company Ltd (Swiss Re) in selected African countries. I am extremely grateful to Melissa Leicester for putting me in contact with Rudolf Senn, who so kindly provided the data. For reasons of confidentiality, these data sets are not publicly available.

The structure of the chapter is as follows. A brief overview regarding the procedures used for exploring and analysing the extreme values for a real data set are outlined in Section 1. The analysis of the motor reinsurance claims, which involves modelling the sample block maxima with the use of the non-stationary GEV distribution, is presented in Section 2. Finally, in Section 3 an analysis of the property reinsurance claims, which were modelled using the non-stationary distribution for the r largest order statistics with $\xi = 0$, is presented.

6.1 Exploring and Analysing Real Data Sets

In the previous chapter, the modelling of non-stationary extremes from simulated data was considered. The advantages of using simulated data is that there is no need to block the data and extract the r largest order statistics, there is an obvious choice for which model to select and, in addition, which starting values to use for the estimates in the Fisher scoring algorithm. However, when using real world data these choices and decisions become less obvious and it is therefore important to consider what factors will affect these decisions. Therefore, before describing the results from the analysis of the two real data sets, the basic procedures that were used in the process are outlined in this section. These include the procedures for checking the original data, blocking the data, selecting the model and estimating the MPLE's.

Procedure 1: Checking the original data

The first step in analysing a data set is to check the data for consistency and to try to identify whether or not there are any abnormal points present. Abnormal data points could arise due to data being incorrectly recorded and should be checked. After cleaning the data, it is useful to provide a basic summary of the underlying data by considering some descriptive statistics and visual displays, for example the mean, the variance, a histogram and a time plot. This should provide some insight into the prominent features of the data. Finally, one should also determine whether or not the data needs to be transformed. Transformations will usually be applied to reduce the spread or to improve interpretability.

Procedure 2: Blocking the data

As described in Chapter 2, the data must be blocked and the r largest observations within each block extracted. Practical consideration often lead to selecting the blocks based on reference points in time, for example choosing blocks based on a time interval of a month or a year. It is important to recall that when choosing the size of the blocks and number of order statistics r used for each block, there is a bias-variance trade-off. Furthermore, when using time intervals for the block size, its important to check that there are enough observations within each block for the asymptotic assumptions to still hold.

Procedure 3: Selecting the model

The primary factor affecting the choice of model will depend on the purpose of the investigation and the ultimate goal of the analysis. Secondly, the physical data and the form that the data take will affect the decisions made. For argument's sake, if the data consist only of the annual maxima, it will not be possible to model the r largest extremes in each year. Therefore, it is important to explore the blocked data in order to gain more information to help make

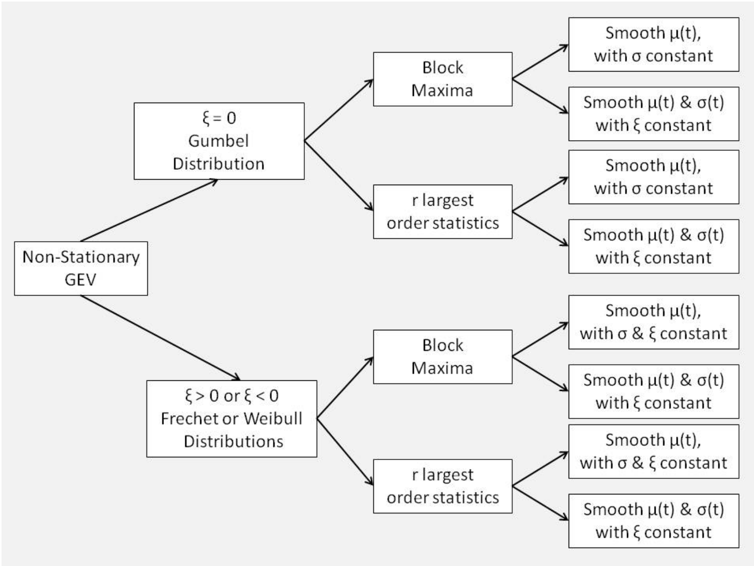

Figure 6.1: Flow chart of choosing the model

more informed decisions. The initial extreme value analysis may consist of fitting the stationary GEV distribution or the stationary distribution for the r largest order statistics to the blocked data and considering the relevant model diagnostics including quantile, probability and return level plots. If there is any evidence of a trend, the assumptions of stationarity can then be relaxed.

Once the decision to model the extremes using a non-stationary distribution is made, three further decisions need to be made before the final model is selected. Firstly, the modeller must decide whether the shape parameter ξ is set to zero or allowed to vary and secondly, whether to model only the maxima or the r largest extremes. Finally, the decision regarding which parameters to be smoothed must be made. Figure 6.1 provides a summary of the final three decisions which may lead to one of eight possible models being selected.

Procedure 4: Estimating the MPLE's

Once a model has been chosen, the MPLE's are obtained by implementing a Fisher scoring algorithm to maximize the penalized log-likelihood, as discussed in Section 4.3.4. Given that the main aim of this type of analysis is exploratory, it is useful to consider various smoothing parameters as well as the optimal smoothing parameter when calculating the MPLE's.

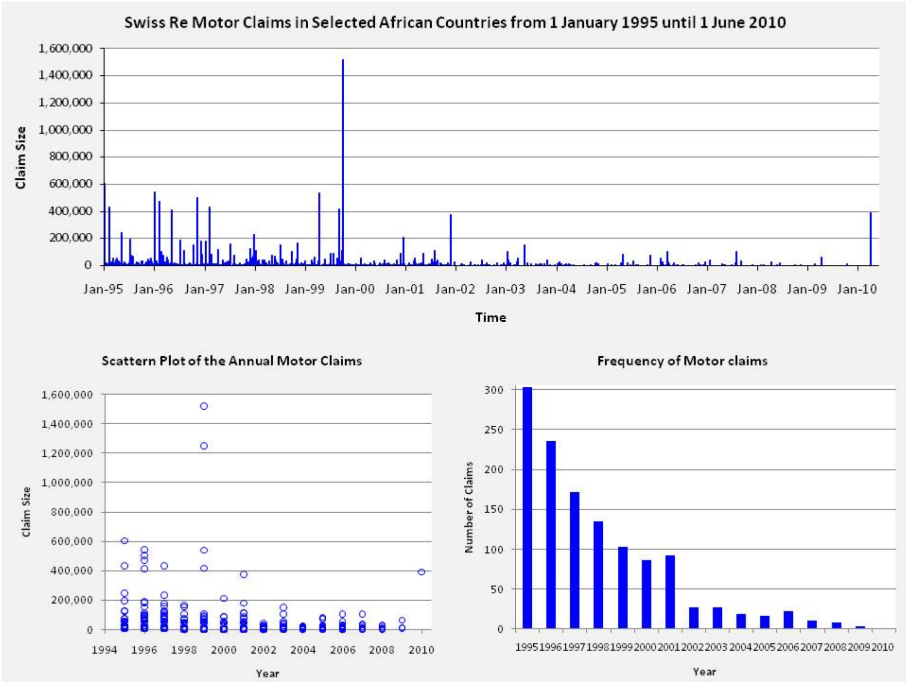

Figure 6.2: Time plot of the motor claims recorded from the 1 January 1995 till 1 June 2010, a scatter plot of motor claims grouped by year and a bar plot for the frequency of the annual motor claims

6.2 Motor Reinsurance Claims

The Swiss Re motor data set contains 1262 reinsurance claims which are at least as large as 1000 CHF (Swiss Francs), recorded from 1 January 1995 to 31 May 2010. The minimum of 1000 CHF was used to remove small claims arising from currency conversions, clean ups, booking corrections and migrations. The data were gathered from selected African countries stemming from claims incurred by Swiss Re. Therefore the data set did not include claims for which Swiss Re advised on a precautionary basis, thereby incurring no cost, but for which the direct insurer might still have paid.

6.2.1 Preliminary Analysis

The time plot for all the motor claims is given in Figure 6.2 along with a scatter plot of the annual claims and a bar plot for the frequency of annual claims. A first glance at Figure 6.2 reveals that the frequency of claims is declining

Statistic	All Claims
Sample size n	1 262
Minimum	1 000.6084
Lower Quartile	1 976.3254
Median	4 233.0975
Upper Quartile	12 052.2829
Maximum	1 522 691.5512
99% Quantile	392 249.8318
Mean	18 994.6464
Standard Deviation	73 963.1940
Skewness	12.5801
Kurtosis	208.3737

Table 6.1: Some descriptive statistics for the motor claims data

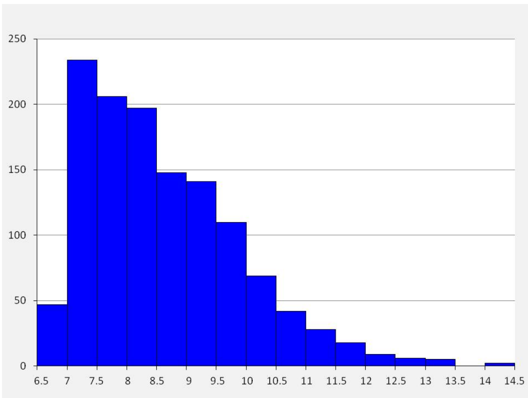

Figure 6.3: Histogram of the log-transformed motor claims

over time and furthermore, it appears that the trend in the size of the claims is declining over time, with the exception of the two notably large claims that occurred in 1999. Some descriptive statistics for the motor claims data are presented in Table 6.1 and a histogram of the log-claims is shown in Figure 6.3. These reveal that the data are heavy-tailed and skewed to the right.

Three decisions were made when modelling the motor claims data. Firstly, claims that occurred in 2010 were included in the analysis due to the large claim that occurred on 1 May 2010. Secondly, due to the size and variability of the motor claims, the data were log-transformed and the log-claim data modelled. Thirdly, due to the decreasing frequency of the motor claims over time, the data was blocked using time periods of one year with only the largest claim in each

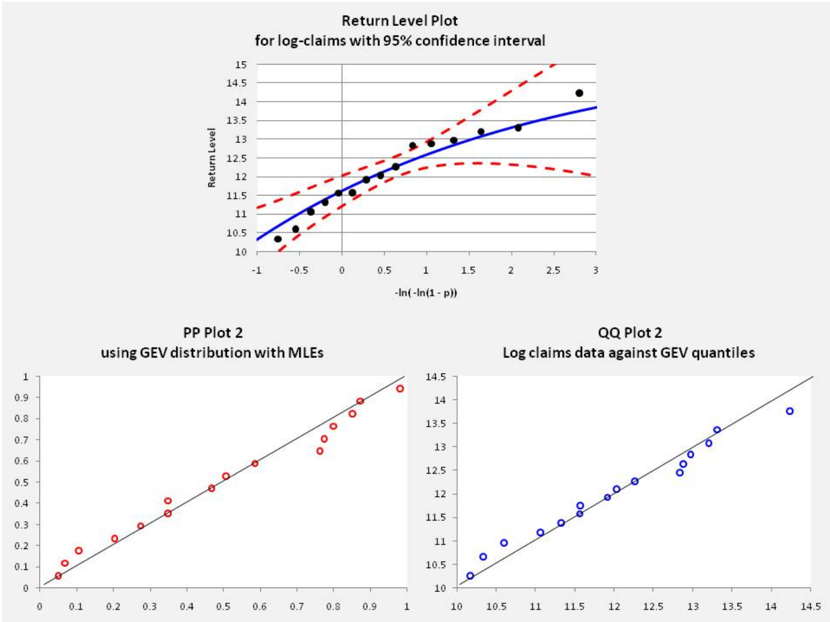

Figure 6.4: Probability, quantile and return-level plots for the annual maximum log-claim

year being extracted.

Initially, an assumption of stationarity is made and the annual maxima of log-claims are modelled as independent observations from the GEV distribution. Maximization of the GEV log-likelihood function, given by equation (2.12), leads to maximum likelihood estimates $(\tilde{\mu}, \tilde{\sigma}, \tilde{\xi}) = (11.6234, 1.1175, -0.2926)$ with an approximate variance-covariance matrix of

$$V = \begin{bmatrix} 0.1005 & 0.0068 & -0.0289 \\ 0.0068 & 0.0552 & -0.0309 \\ -0.0289 & -0.0309 & 0.0428 \end{bmatrix}.$$

As discussed in section 2.4.1, with $\hat{\xi} > -0.5$ the maximum likelihood estimators will have the usual asymptotic properties. Therefore, an approximate 95% confidence interval for the location parameter μ is given by $11.6234 \pm 1.96\sqrt{0.1005} = [11.0021, 12.2448]$. Similarly, an approximate 95% confidence interval for scale parameter $\sigma = [0.6569, 1.5781]$ and the shape parameter $\xi = [-0.6982, 0.1131]$.

Probability, quantile and return-level plots were used to assess the fit of the GEV distribution and are shown in Figure 6.4. With a negative maximum likelihood estimate for the shape parameter ξ, the return-level plot is concave and

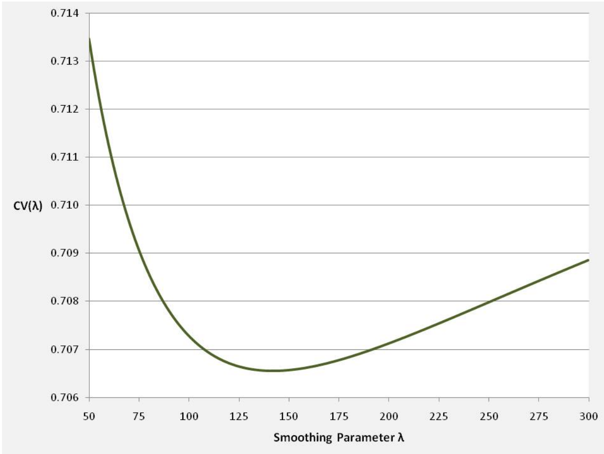

Figure 6.5: Cross-Validation Score Function for the annual maximum log-claim data

only one observation lies outside the 95% confidence interval. The probability and quantile plots are near linear, suggesting a reasonable fit. Therefore, the only clear evidence of a time trend in the data comes from the scatter plot of the annual claims in Figure 6.2.

6.2.2 Smoothing the Sample Maxima

The analysis now involves modelling the annual maxima of the log-claims in a non-stationary environment, where the location parameter is allowed to vary over time whilst keeping the scale and shape parameters constant. To allow for the smoothing of the extremes over time, the sample maxima are modelled with the use of the non-stationary GEV block maxima model by fitting a smoothed NCS to the location parameter. For a given smoothing parameter λ, the penalized log-likelihood function

$$l^{PLL}(\mathbf{g}_\mu, \sigma, \xi; 1) = l(\mathbf{g}_\mu, \sigma, \xi; 1) - \frac{1}{2}\lambda \mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu$$

is maximized using a Fisher scoring algorithm.

The first smoothing parameter considered was the optimal smoothing parameter $\hat{\lambda}$ determined by minimizing the cross-validation score function, which is illustrated in Figure 6.5. Using the optimal smoothing parameter $\hat{\lambda} = 142$ resulted in a MPLE $\hat{\sigma} = 1.2365$ for the scale parameter and $\hat{\xi} = -0.3481$ for the shape parameter, with the MPLE for the location parameter $\hat{g}_\mu(t)$ shown in Figure 6.6. In addition, Figure 6.6 illustrates the expected log-claim and the 10, 20 and 100 year return periods. At each time point t, the expected log-claim $E[X_t]$ is estimated by substituting the MPLE's into

$$E[X_t] = g_\mu(t) - \frac{\sigma}{\xi} + \frac{\sigma}{\xi}\,\Gamma(1-\xi)$$

and the estimate for the $1/p$ year return period, corresponding to the MPLE for the extreme quantile $z_{p,t}$, is calculated by substituting the MPLE's into

$$\hat{z}_{p,t} = g_\mu(t) - \frac{\sigma}{\xi}\left[1 - \{-\ln(1-p)\}^{-\xi}\right].$$

It is clear from Figure 6.6 that the MPLE $\hat{g}_\mu(t)$ has a linear decreasing trend from 1995 to 2003, after which the trend levels off. This suggests that the maximum motor claims incurred by Swiss Re have been decreasing over time. It is also interesting to note that in 1999 and 2010, the maximum claims exceed the 10 year return period were almost exactly 10 years apart. However, the 20 year return period has not been exceeded over the period of investigation, suggesting that in the next couple of years Swiss Re may expect a large claim to exceed these levels.

Having modelled the log-transformed motor claims, the results depicted in Figure 6.6 are transformed by taking exponents to see the effects on the original motor claim data, with these results being illustrated in Figure 6.7.

In addition to the optimal smoothing parameter, other smoothing parameters were considered for exploratory purposes. For illustration, a second smaller smoothing parameter of $\lambda = 10$ is considered here. The aim of using the smaller smoothing parameter $\lambda = 10$ is to allow for a “closer fit” to the underlying data and to explore potentially different trends in the MPLE for the location parameter. The results are shown in Figures 6.8 and 6.9 with MPLE's $\hat{\sigma} = 1.2417$ and $\hat{\xi} = -0.3611$. Note that using the smoothing parameter $\lambda = 10$ suggests a slightly different trend for the MPLE $\hat{g}_\mu(t)$. There is still a decreasing trend from 1995 to 2005, but, after 2005 there is an increasing trend that picks up on the large claim in 2010.

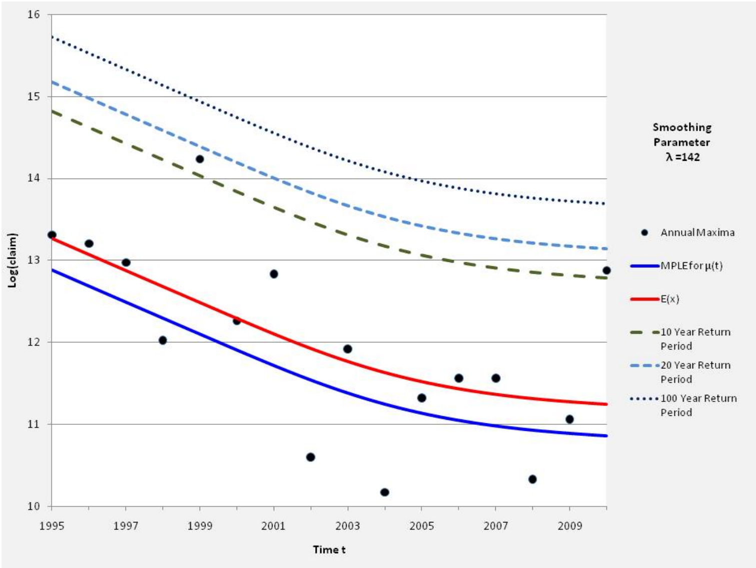

Figure 6.6: Maximum penalized likelihood estimator of $\mu(t)$ and various return periods for the annual motor log-claim data using the optimal smoothing parameter $\hat{\lambda} = 142$

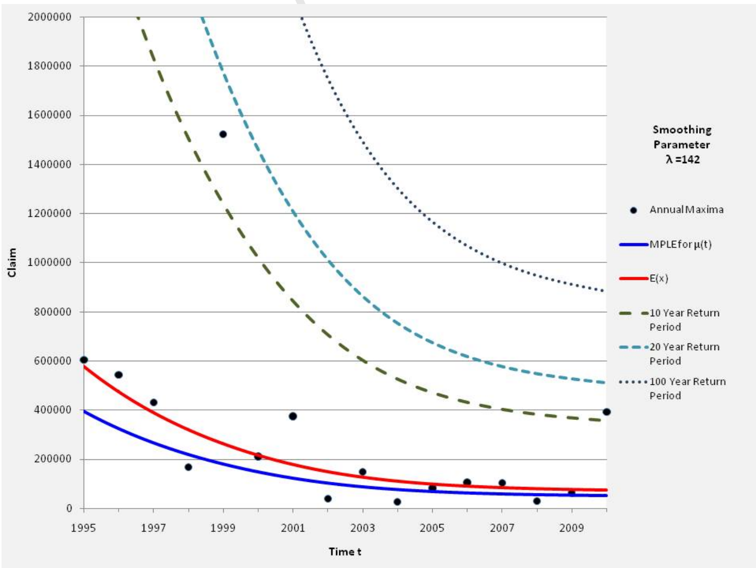

Figure 6.7: Taking exponents to get claim data using the optimal smoothing parameter $\hat{\lambda} = 142$

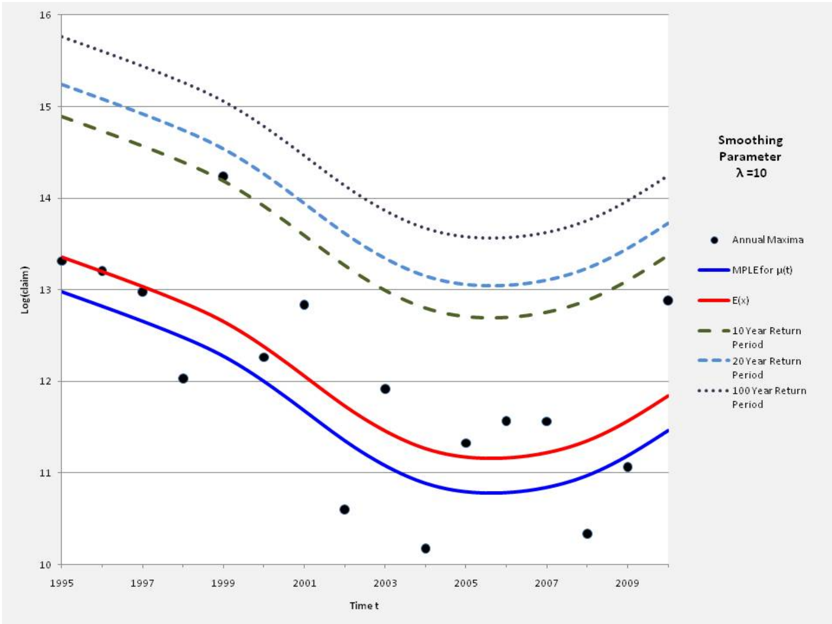

Figure 6.8: Penalized likelihood estimator of $\mu(t)$ and various return periods for the annual motor log-claim data using a smoothing parameter $\lambda = 10$

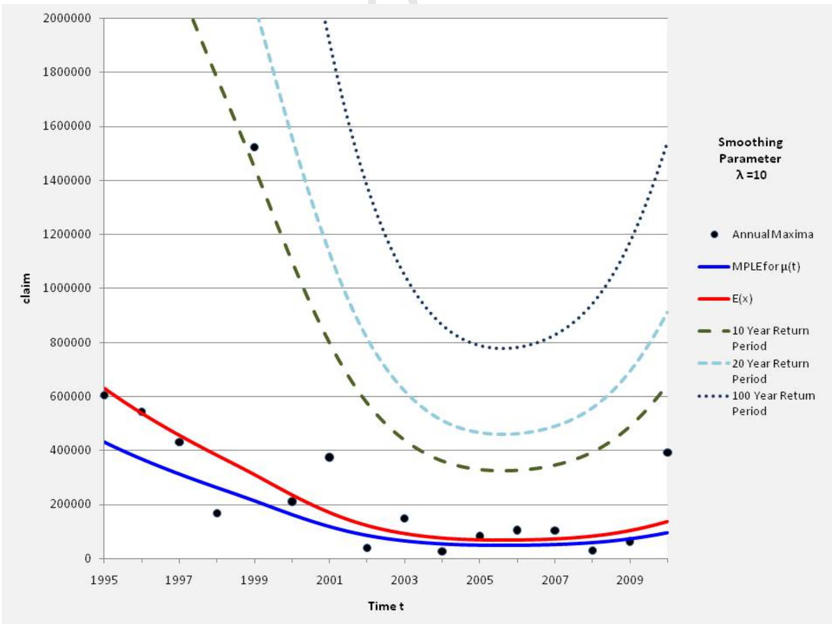

Figure 6.9: Taking exponents to get claim data using smoothing parameter $\lambda = 10$

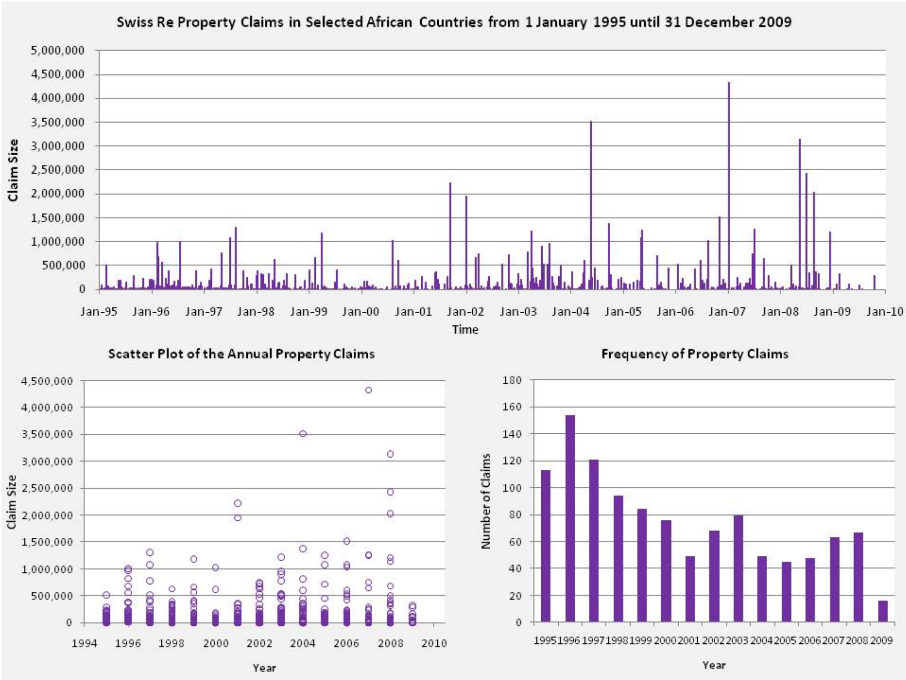

Figure 6.10: Time plot of the property claims recorded from the 1 January 1995 till 31 December 2009, a scatter plot of property claims grouped by year and a bar plot for the frequency of the annual property claims

6.3 Property Reinsurance Claims

The Swiss Re property data set contains 1126 reinsurance claims which are at least as large as 5000 CHF (Swiss Francs), recorded from 1 January 1995 to 31 December 2009. As for the motor reinsurance claims, the minimum of 5000 CHF was used to remove small claims arising from currency conversions, clean ups, booking corrections and migrations. The data were gathered from selected African countries stemming from claims incurred by Swiss Re. Therefore the data set did not include claims for which Swiss Re advised on a precautionary basis, thereby incurring no cost, but for which the direct insurer might still have paid.

6.3.1 Preliminary Analysis

The time plot for all the property claims is given in Figure 6.10 along with a scatter plot of the annual claims and a bar plot for the frequency of annual

Statistic	All Claims
Sample size n	1 126
Minimum	5 008.6927
Lower Quartile	8 618.4991
Median	21 295.0142
Upper Quartile	72 319.8541
Maximum	4 322 725.9491
99% Quantile	1 260 110.1916
Mean	104 390.6513
Standard Deviation	290 535.1141
Skewness	7.49893
Kurtosis	77.2904

Table 6.2: Some descriptive statistics for the property claim data

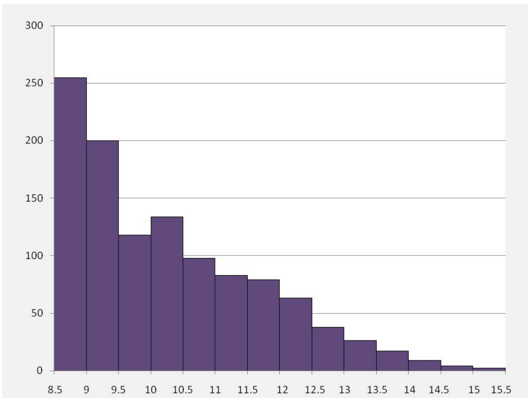

Figure 6.11: Histogram of the log-claims incurred by Swiss Re for property reinsurance in selected African countries

claims. A first glance at Figure 6.10 reveals that the trend in the size of the annual claim maxima is increasing up until 1 January 2007 after which the trend in the annual claim maxima appears to be declining. Furthermore, it is evident that in certain years there is more than one notably large claim and also that the frequency of claims has a cyclical and declining trend over time.

Some descriptive statistics for the property claims data are presented in Table 6.2 and the histogram of the log-claims is shown in Figure 6.11. These reveal that the data are heavy-tailed and skewed to the right.

Three decisions were made when modelling the property claims data. Firstly, due to the size and variability of the property claims, the data were log-transformed and the log-claim data modelled. Secondly, the data were blocked using time

Number of Extremes r	Location MLE $\hat{\mu}$	Standard Error for $\hat{\mu}$	Scale MLE $\hat{\sigma}$	Standard Error for $\hat{\sigma}$
1	13.6945	0.1716	0.6644	0.1271
2	13.8891	0.1310	0.7169	0.0874
3	14.0205	0.1396	0.9350	0.0825
4	14.0902	0.1216	0.9403	0.0651
5	14.1350	0.1087	0.9398	0.0537
6	14.2256	0.1063	1.0065	0.0492
7	14.4516	0.1120	1.1457	0.0491
8	14.9421	0.1318	1.4415	0.0552
9	14.9461	0.1229	1.4255	0.0495
10	15.1129	0.1217	1.4875	0.0473

Table 6.3: Maximum likelihood estimates and standard errors for selected values of r for the property log-claim data

periods of one year and, given that there are a sufficient number of claims in each year, with only 16 claims in 2009, and due to the fact that the asymptotic results break down with increasing r, the 10 largest claims in each year were extracted. Thirdly, to illustrate the application of the Gumbel distribution to real data, the shape parameter was set to zero, that is $\xi = 0$.

Initially, an assumption of stationarity is made and the property log-claim data were modelled using the distribution for the r largest order statistics with $\xi = 0$ for $r = 1, 2, \ldots, 10$. Maximization of the log-likelihood function, for the stationary distribution of the r largest order statistics with $\xi = 0$, resulted in the maximum likelihood estimates and standard errors as shown in Table 6.3 for various r. As anticipated, the standard errors decrease as the value of r increases, for $r = 1, 2, \ldots, 6$. However, for $r = 7$ and 8 there is an increase in the standard error for the location parameter and an increase for $r = 8$ in the standard error for the scale parameter, which is not expected and might be a result of the asymptotic results breaking down. Furthermore, the parameter estimates increase as r increases and there is no stability in the estimates. Therefore, the decision was taken to only consider up to a maximum of $r = 6$ for the largest claims in each year when analysing the property log-claim data.

The accuracy of the stationary model fit was examined by considering probability and quantile plots. It was evident from the diagnostic plots that there was a clear lack of fit due to a trend. For illustration, probability and quantile

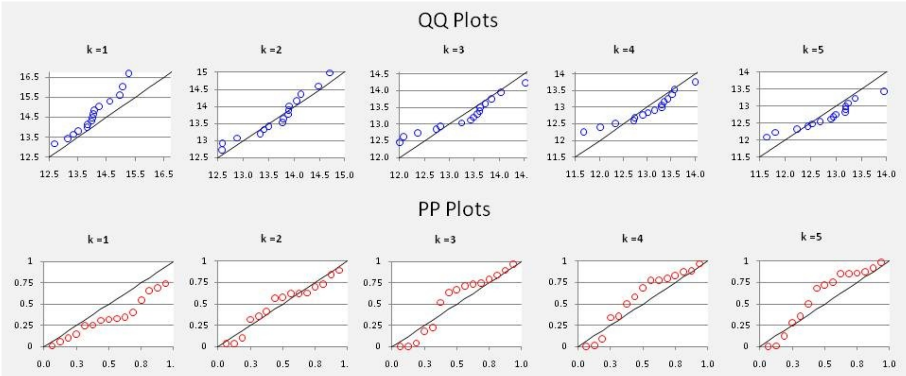

Figure 6.12: Model diagnostics for the property log-claim data on basis of fitted r largest order statistic model with $r = 5$ and $\xi = 0$

plots for the property log-claim data on the basis of the fitted r largest order statistic model with $r = 5$ are shown in Figure 6.12.

6.3.2 Smoothing the Sample Extremes

The analysis now involves modelling the extremes in a non-stationary environment, where the location parameter is allowed to vary over time whilst keeping the scale parameter constant. To allow for the smoothing of the extremes over time, the sample extremes are modelled with the use of the non-stationary distribution for the r largest order statistics with $\xi = 0$ and fitting a smoothed NCS to the location parameter. The process of identifying the parameter estimates is identical to that used in the Gumbel simulation study discussed in Section 5.3.1. For a given smoothing parameter λ, the penalized log-likelihood function

$$l^{PLL}(\mathbf{g}_\mu, \sigma, 0; r) = l(\mathbf{g}_\mu, \sigma, 0; r) - \frac{1}{2}\lambda \mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu$$

is maximized using a Fisher scoring algorithm.

Two sets of smoothing parameters for each value of r were used in the analysis. The first set comprised of the optimal smoothing parameters $\hat{\lambda}$ which varied depending on the size of r and the second set used a constant smoothing parameter $\lambda = 50$ to make a more direct comparison of the MPLE's with the same λ.

The MPLE $\hat{\sigma}$ for the various smoothing parameters are given in Table 6.4. Figure 6.13 displays the MPLE $\hat{g}_\mu(t)$ along with the log-claims for the property

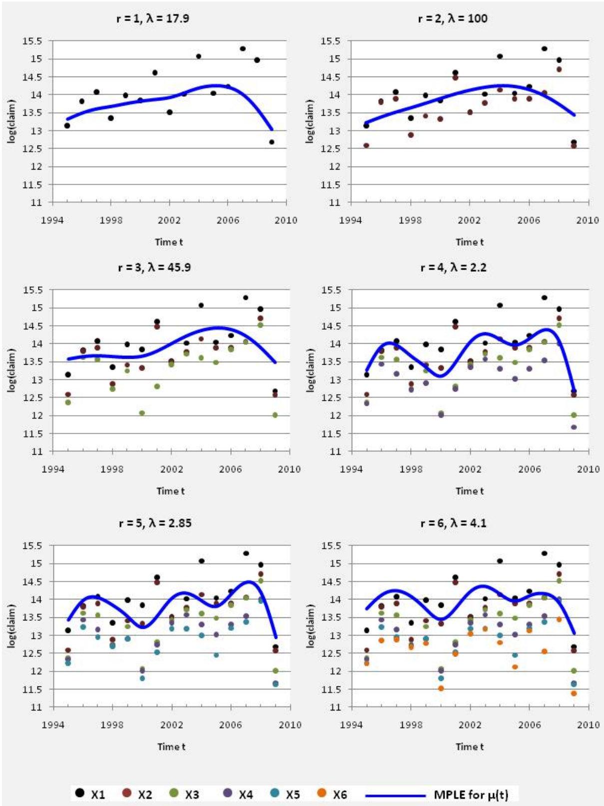

Figure 6.13: Property log-claim data and MPLE for $\mu(t)$ using the optimal smoothing parameter $\hat{\lambda}$ and for various r

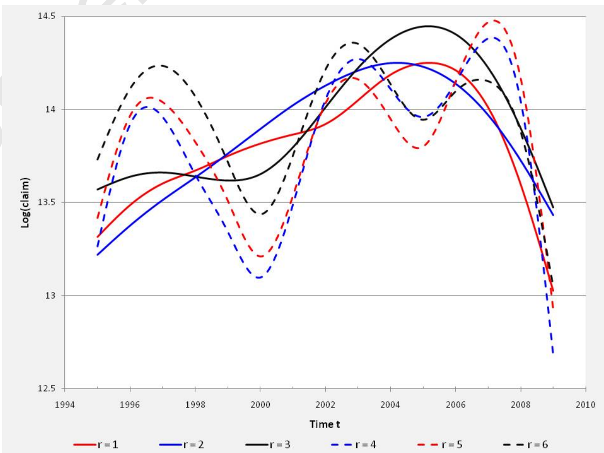

Figure 6.14: MPLE for $\mu(t)$ for various r using the optimal smoothing parameters $\hat{\lambda}$

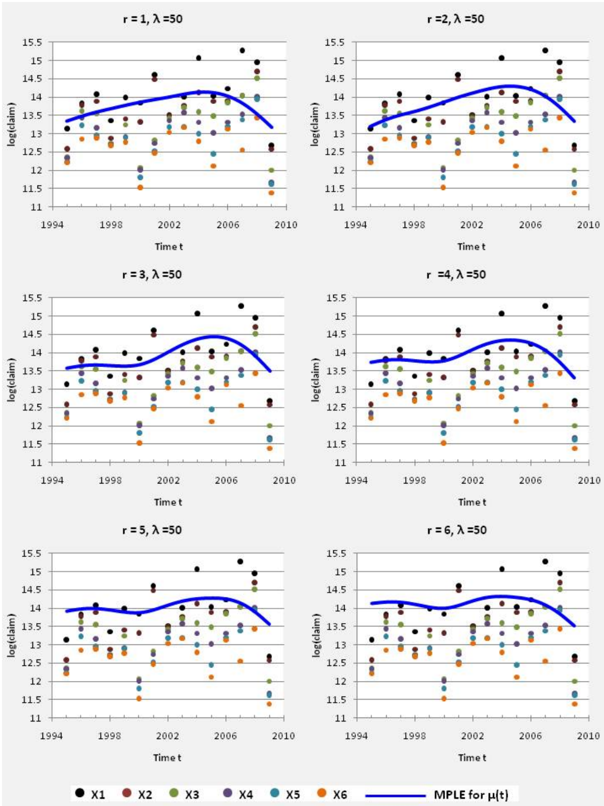

Figure 6.15: Property log-claim data and MPLE for $\mu(t)$ using the smoothing parameter $\lambda = 50$ and for various r

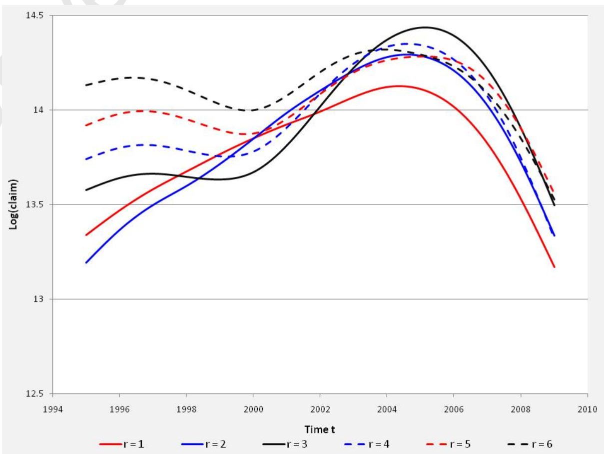

Figure 6.16: MPLE for $\mu(t)$ for various r using a smoothing parameter $\lambda = 50$

Number of Extremes r	MPLE for σ using $\hat{\lambda}$	MPLE for σ using $\lambda = 50$
1	0.4205	0.4351
2	0.5296	0.5071
3	0.7546	0.7609
4	0.6306	0.7745
5	0.6505	0.8390
6	0.7809	0.9116

Table 6.4: Maximum penalized likelihood estimates for σ for selected values of r based on the property log-claim data, using optimal smoothing parameter $\hat{\lambda}$ and $\lambda = 50$

data for different values of r using the optimal smoothing parameters $\hat{\lambda}$. To provide a more direct comparison of the MPLE's $\hat{g}_{\mu}(t)$ for various r, the estimates are combined into one plot and displayed in Figure 6.16. Figures 6.15 and 6.16 display similar plots, except the results are obtained using the constant smoothing parameter $\lambda = 50$.

From the results, it is evident that when $r = 1$ and 2, the extreme property claims are increasing from 1995 up to 2005 after which they begin to decrease. However, when $r \geq 3$ there is evidence of possible cyclical trends in these extremes, which are picked up by including the additional extremes.

Chapter 7

Conclusions

Aims achieved

In this mini-dissertation, the modelling of stationary and non-stationary extremes has been explored. For the modelling of stationary extremes, an introduction to classical extreme value theory, which summarized the main results and, more importantly, drew together a number of key results that are scattered in the literature, was presented. Two distributions were considered, namely the GEV distribution, for modelling block maxima, and the distribution for the r largest order statistics, for modelling the r largest extremes in a block.

For the modelling of non-stationary extremes, a roughness penalty approach, where the location and scale parameters of the GEV distribution and the distribution for the r largest order statistics vary over time and the log-likelihood function is penalized with a roughness penalty, was used. It is known that, when the squared second derivative for the roughness penalty is used, it is optimal to choose the location and scale parameters to be smoothed natural cubic splines. The MPLE's were calculated using a Fisher scoring algorithm and the optimal smoothing parameters were calculated by minimizing the cross-validation score function. The advantage of using the roughness penalty approach is that it provides a flexible framework for exploring smooth trends in sample extremes, with the benefit of balancing the trade-off between 'smoothness' and adherence to the underlying data by simply changing the smoothing parameter.

Simulated data sets where used to investigate the performance of modelling non-stationary extremes, using the roughness penalty approach. A full simulation was conducted to determine if it is always possible to identify an optimal smoothing parameter when using cross-validation, to evaluate the effects of changing the sample size n and the number of extremes r, to consider the properties of the MPLE's for the location, scale and shape parameters and to compare the estimates to their true values. To identify the effects of changing the smoothing parameters on the MPLE's for the location, scale and shape parameters, four individual simulations were considered. The results showed that the smoothing parameters had a direct effect on the MPLE's and that it was not always possible to obtain an optimal smoothing parameter using cross-validation. Furthermore, relatively better parameter estimates were obtained by increasing the sample size n and the number of extremes r, but, this also had a direct effect of increasing the time taken to run the cross-validation process.

Two real data sets have been used to illustrate the roughness penalty approach for modelling of non-stationary extremes, namely motor and property reinsurance claims incurred by the Swiss Reinsurance Company Ltd in selected African countries. The results obtained from analysing the motor reinsurance claims illustrated the flexibility of the approach in identifying trends by simply changing the smoothing parameter. The results obtained from analysing the property reinsurance claims illustrated the advantage of including more extremes when identifying trends. Unfortunately, due to limited information regarding the nature of the real data, it was not possible to infer reasons for the observed trends.

Future research

The following suggestions for future research, which are beyond the scope of this mini-dissertation, can be made. Firstly, to look at calculating confidence bands for the true conditional extreme percentiles when modelling real data. Secondly, to investigate procedures for model-checking when assessing the fit of the non-stationary models. Thirdly, to re-assess the performance of the roughness penalty approach, by re-running the simulation study, by sampling from a number of different distributions that belong to the domains of attraction of the the Gumbel, Fréchet and Weibull distributions. Fourthly, to extend the ideas of modelling non-stationary extremes, using the roughness penalty approach, to peaks-over-threshold and Poisson-based models. Finally, further research needs to be conducted into methods for determining the optimal smoothing parameter using other automatic selection criteria, such as AIC.

Appendices

Appendix A

Gamma and Digamma Functions

The Gamma function is defined as

$$\Gamma(x) = \int_0^{\infty} u^{x-1} e^{-u} \, du$$

Certain relationships that hold with the Gamma function include:

1. $\Gamma(x) = (x-1)!$
2. $x\Gamma(x) = \Gamma(x+1)$
3. $\Gamma(1) = 0! = 1$
4. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$
5. $(1-k)\sum_{j=1}^{r} \frac{\Gamma(j-k)}{\Gamma(j)} = \frac{\Gamma(r+1-k)}{\Gamma(r)}$

The mth derivative of the Gamma function is defined as follows

$$\Gamma^{(m)}(x) = \frac{\partial^m \Gamma(x)}{\partial x^m} = \int_0^{\infty} u^{x-1} e^{-u} (\ln u)^m \, du$$

The Digamma function is defined as follows

$$\Psi(x) = \int_0^{\infty} \left(\frac{e^{-u}}{u} - \frac{e^{-xu}}{1-e^{-u}} \right) du$$

Certain relationships that hold with the Digamma function include:

1. $\Psi(x) = \frac{\partial}{\partial x} \ln \Gamma(x) = \frac{\Gamma^{(1)}(x)}{\Gamma(x)}$
2. $\Psi(x+1) = \Psi(x) + \frac{1}{x}$

3. $\sum_{j=1}^{r} \Psi(j) = r\Big(\Psi(r+1) - 1\Big)$

4. $\frac{\partial \Psi(x+1)}{\partial x} = \Psi'(x+1) = \Psi'(x) - \frac{1}{x^2}$

5. $\frac{\Gamma^{(2)}(r)}{\Gamma(r)} = \Psi'(r) + \Big(\Psi(r)\Big)^2$

6. $\Psi(1) = -\gamma$

7. $\Psi'(1) = \frac{\pi^2}{6}$

8. $\Psi'(\frac{1}{2}) = \frac{\pi^2}{2}$

Appendix B

Gumbel block maxima model with smoothed location and scale parameters

For the non-stationary Gumbel block maxima model with smoothed location and scale parameters, let $\mu = g_\mu(t_i) = g_{\mu i}$ and $\sigma = g_\sigma(t_i) = g_{\sigma i}$, so that log-likelihood function is given by

$$l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; 1) = \sum_{i=1}^{N} \Big\{ -\ln(g_{\sigma i}) - \Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big) - \exp\Big[-\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]\Big\}$$

The penalized log-likelihood is then given by

$$\begin{aligned} l^{PLL}(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; 1) &= l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; 1) - \frac{1}{2}\lambda \int g''_\mu(t)^2 dt - \frac{1}{2}\alpha \int g''_\sigma(t)^2 dt \\ &= l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; 1) - \frac{1}{2}\lambda \mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu - \frac{1}{2}\alpha \mathbf{g}_\sigma^T \mathbf{K} \mathbf{g}_\sigma \end{aligned}$$

The first and second derivatives of the log-likelihood function are given by

$$\begin{aligned} \frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; 1)}{\partial g_{\mu i}} &= \frac{1}{g_{\sigma i}}\Big\{1 - \exp\Big[-\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]\Big\} \\ \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; 1)}{\partial g_{\mu i}^2} &= \frac{-1}{g_{\sigma i}^2} \cdot \exp\Big[-\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big] \\ \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; 1)}{\partial g_{\mu i} \partial g_{\sigma i}} &= -\frac{1}{g_{\sigma i}^2}\Big\{1 - \exp\Big[-\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]\Big\} - \frac{(x_i - g_{\mu i})}{g_{\sigma i}^3} \cdot \exp\Big[-\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big] \\ \frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; 1)}{\partial g_{\sigma i}} &= \frac{1}{g_{\sigma i}}\Big[-1 + \frac{(x_i - g_{\mu i})}{g_{\sigma i}}\Big\{1 - \exp\Big[-\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]\Big\}\Big] \end{aligned}$$

$$\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,0;1)}{\partial g_{\sigma i}^2} = \frac{1}{g_{\sigma i}^2} - \frac{2(x_i-g_{\mu i})}{g_{\sigma i}^3}\Big\{1-\exp\Big[-\Big(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\Big)\Big]\Big\} - \frac{(x_i-g_{\mu i})^2}{g_{\sigma i}^4}\cdot\exp\Big[-\Big(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\Big)\Big]$$

$$\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,0;1)}{\partial g_{\sigma i}\partial g_{\mu i}} = -\frac{1}{g_{\sigma i}^2}\Big\{1-\exp\Big[-\Big(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\Big)\Big]\Big\} - \frac{(x_i-g_{\mu i})}{g_{\sigma i}^3}\cdot\exp\Big[-\Big(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\Big)\Big]$$

$$\forall i\neq j \quad \frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,0;1)}{\partial g_{\mu i}\partial g_{\mu j}}=0\,,\ \frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,0;1)}{\partial g_{\mu i}\partial g_{\sigma j}}=0\,,\ \frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,0;1)}{\partial g_{\sigma i}\partial g_{\sigma j}}=0\,,\ \frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,0;1)}{\partial g_{\sigma i}\partial g_{\mu j}}=0$$

with the first and second derivatives of the roughness penalties given by

$$\frac{\partial(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial\mathbf{g}_\mu} = -\lambda\mathbf{K}\mathbf{g}_\mu \quad\Rightarrow\quad \frac{\partial(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial g_{\mu i}} = -\lambda\big(\mathbf{K}\mathbf{g}_\mu\big)_i$$

$$\frac{\partial^2(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial\mathbf{g}_\mu\partial\mathbf{g}_\mu^T} = -\lambda\mathbf{K} \quad\Rightarrow\quad \frac{\partial^2(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial g_{\mu i}\partial g_{\mu j}} = -\lambda K_{ij}$$

$$\frac{\partial(-\frac{1}{2}\alpha\mathbf{g}_\sigma^T\mathbf{K}\mathbf{g}_\sigma)}{\partial\mathbf{g}_\sigma} = -\alpha\mathbf{K}\mathbf{g}_\sigma \quad\Rightarrow\quad \frac{\partial(-\frac{1}{2}\alpha\mathbf{g}_\sigma^T\mathbf{K}\mathbf{g}_\sigma)}{\partial g_{\sigma i}} = -\alpha\big(\mathbf{K}\mathbf{g}_\sigma\big)_i$$

$$\frac{\partial^2(-\frac{1}{2}\alpha\mathbf{g}_\sigma^T\mathbf{K}\mathbf{g}_\sigma)}{\partial\mathbf{g}_\sigma\partial\mathbf{g}_\sigma^T} = -\alpha\mathbf{K} \quad\Rightarrow\quad \frac{\partial^2(-\frac{1}{2}\alpha\mathbf{g}_\sigma^T\mathbf{K}\mathbf{g}_\sigma)}{\partial g_{\sigma i}\partial g_{\sigma j}} = -\alpha K_{ij}$$

the expectations of the second derivative of the log-likelihood are given by

$$E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,0;1)}{\partial g_{\mu i}^2}\right] = \frac{1}{g_{\sigma i}^2}$$

$$E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,0;1)}{\partial g_{\sigma i}^2}\right] = \frac{1}{g_{\sigma i}^2}\Big(1+\Psi'(2)+(\Psi(2))^2\Big)$$

$$E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,0;1)}{\partial g_{\sigma i}\partial g_{\mu i}}\right] = -\frac{\Psi(1)+1}{g_{\sigma i}^2}$$

$$E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,0;1)}{\partial g_{\mu i}\partial g_{\sigma i}}\right] = -\frac{\Psi(1)+1}{g_{\sigma i}^2}$$

By defining,

$$u_\mu = \frac{\partial l(\mathbf{g}_\mu,\mathbf{g}_\sigma,0;1)}{\partial\mathbf{g}_\mu} \qquad u_\sigma = \frac{\partial l(\mathbf{g}_\mu,\mathbf{g}_\sigma,0;1)}{\partial\mathbf{g}_\sigma}$$

$$W_\mu = E\bigg(-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,0;1)}{\partial\mathbf{g}_\mu\partial\mathbf{g}_\mu^T}\bigg) \qquad W_\sigma = E\bigg(-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,0;1)}{\partial\mathbf{g}_\sigma\partial\mathbf{g}_\sigma^T}\bigg)$$

$$W_{\mu\sigma} = E\bigg(-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,0;1)}{\partial\mathbf{g}_\mu\partial\mathbf{g}_\sigma^T}\bigg) \qquad W_{\sigma\mu} = E\bigg(-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,0;1)}{\partial\mathbf{g}_\sigma\partial\mathbf{g}_\mu^T}\bigg)$$

the Fisher scoring equation, for the non-stationary Gumbel block maxima model with smoothed location and scale parameters, can be written as

$$\begin{pmatrix} W_\mu+\lambda K & W_{\mu\sigma} \\ W_{\sigma\mu} & W_\sigma+\alpha K \end{pmatrix}\begin{pmatrix} g_\mu^{new}-g_\mu \\ g_\sigma^{new}-g_\sigma \end{pmatrix} = \begin{pmatrix} u_\mu-\lambda K g_\mu \\ u_\sigma-\alpha K g_\sigma \end{pmatrix}$$

which implies that

$$(W_\mu + \lambda K)(g_\mu^{new} - g_\mu) + W_{\mu\sigma}(g_\sigma^{new} - g_\sigma) = u_\mu - \lambda K g_\mu$$
$$W_{\sigma\mu}(g_\mu^{new} - g_\mu) + (W_\sigma + \alpha K)(g_\sigma^{new} - g_\sigma) = u_\sigma - \alpha K g_\sigma$$

Re-expressing these equations with only g_μ^{new} and g_σ^{new} on the left hand side,

$$g_\mu^{new} = (W_\mu + \lambda K)^{-1}\Big((W_\mu + \lambda K)g_\mu + u_\mu - \lambda K g_\mu - W_{\mu\sigma}(g_\sigma^{new} - g_\sigma)\Big)$$
$$g_\sigma^{new} = (W_\sigma + \alpha K)^{-1}\Big((W_\sigma + \alpha K)g_\sigma + u_\sigma - \alpha K g_\sigma - W_{\sigma\mu}(g_\mu^{new} - g_\mu)\Big)$$

and by manipulating the equations further,

$$g_\mu^{new} = (W_\mu + \lambda K)^{-1} W_\mu W_\mu^{-1}\Big(W_\mu g_\mu + \lambda K g_\mu + u_\mu - \lambda K g_\mu - W_{\mu\sigma} g_\sigma^{new} + W_{\mu\sigma} g_\sigma\Big)$$
$$g_\sigma^{new} = (W_\sigma + \alpha K)^{-1} W_\sigma W_\sigma^{-1}\Big(W_\sigma g_\sigma + \alpha K g_\sigma + u_\sigma - \alpha K g_\sigma - W_{\sigma\mu} g_\mu^{new} + W_{\sigma\mu} g_\mu\Big)$$

Simplifying further and rearranging the terms, the equations become

$$g_\mu^{new} = (W_\mu + \lambda K)^{-1} W_\mu\Big(g_\mu + W_\mu^{-1} u_\mu + W_\mu^{-1} W_{\mu\sigma} g_\sigma - W_\mu^{-1} W_{\mu\sigma} g_\sigma^{new}\Big)$$
$$g_\sigma^{new} = (W_\sigma + \alpha K)^{-1} W_\sigma\Big(g_\sigma + W_\sigma^{-1} u_\sigma + W_\sigma^{-1} W_{\sigma\mu} g_\mu - W_\sigma^{-1} W_{\sigma\mu} g_\mu^{new}\Big)$$

Which are simplified further by using

$$S_1 = (W_\mu + \lambda K)^{-1} W_\mu \qquad y_1 = g_\mu + W_\mu^{-1} u_\mu + W_\mu^{-1} W_{\mu\sigma} g_\sigma$$
$$S_2 = (W_\sigma + \alpha K)^{-1} W_\sigma \qquad y_2 = g_\sigma + W_\sigma^{-1} u_\sigma + W_\sigma^{-1} W_{\sigma\mu} g_\mu$$

to get

$$g_\mu^{new} = S_1(y_1 - W_\mu^{-1} W_{\mu\sigma} g_\sigma^{new})$$
$$g_\sigma^{new} = S_2(y_2 - W_\sigma^{-1} W_{\sigma\mu} g_\mu^{new})$$

The elements of the Fisher scoring algorithm are given by

$$
\begin{aligned}
u_\mu &= \frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; 1)}{\partial \mathbf{g}_\mu} = \left[\frac{1}{g_{\sigma i}}\Big\{1 - \exp\Big[-\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]\Big\}\right]_i \\
u_\sigma &= \frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; 1)}{\partial \mathbf{g}_\sigma} = \left[\frac{1}{g_{\sigma i}}\Big[-1 + \frac{(x_i - g_{\mu i})}{g_{\sigma i}}\Big\{1 - \exp\Big[-\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]\Big\}\Big]\right]_i \\
W_\mu &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; 1)}{\partial \mathbf{g}_\mu \partial \mathbf{g}_\mu^T}\right) = diag\left(\frac{1}{g_{\sigma i}^2}\right) \\
W_\sigma &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; 1)}{\partial \mathbf{g}_\sigma \partial \mathbf{g}_\sigma^T}\right) = diag\left(\frac{1}{g_{\sigma i}^2}\Big(1 + \Psi'(2) + (\Psi(2))^2\Big)\right) \\
W_{\mu\sigma} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; 1)}{\partial \mathbf{g}_\mu \partial \mathbf{g}_\sigma^T}\right) = diag\left(-\frac{\Psi(1) + 1}{g_{\sigma i}^2}\right) \\
W_{\sigma\mu} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; 1)}{\partial \mathbf{g}_\sigma \partial \mathbf{g}_\mu^T}\right) = diag\left(-\frac{\Psi(1) + 1}{g_{\sigma i}^2}\right)
\end{aligned}
$$

Appendix C

Gumbel block maxima model with constant scale parameter and smoothed location parameter

For the non-stationary Gumbel block maxima model with constant scale parameter and smoothed location parameter, let $\mu = g_\mu(t_i) = g_{\mu i}$, so that log-likelihood function is given by

$$l(\mathbf{g}_\mu, \sigma, 0; 1) = \sum_{i=1}^{N} \Big\{ -\ln(\sigma) - \Big(\frac{x_i - g_{\mu i}}{\sigma}\Big) - \exp\Big[-\Big(\frac{x_i - g_{\mu i}}{\sigma}\Big)\Big]\Big\}$$

and the penalized log-likelihood by

$$\begin{aligned} l^{PLL}(\mathbf{g}_\mu, \sigma, 0; 1) &= l(\mathbf{g}_\mu, \sigma, 0; 1) - \frac{1}{2}\lambda \int g''_\mu(t)^2 dt \\ &= l(\mathbf{g}_\mu, \sigma, 0; 1) - \frac{1}{2}\lambda \mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu \end{aligned}$$

The first and second derivatives of the log-likelihood function are given by

$$\begin{aligned} \frac{\partial l(\mathbf{g}_\mu, \sigma, 0; 1)}{\partial g_{\mu i}} &= \frac{1}{\sigma}\Big\{1 - \exp\Big[-\Big(\frac{x_i - g_{\mu i}}{\sigma}\Big)\Big]\Big\} \\ \frac{\partial^2 l(\mathbf{g}_\mu, \sigma, 0; 1)}{\partial g_{\mu i}^2} &= \frac{-1}{\sigma^2}\exp\Big[-\Big(\frac{x_i - g_{\mu i}}{\sigma}\Big)\Big] \\ \frac{\partial^2 l(\mathbf{g}_\mu, \sigma, 0; 1)}{\partial g_{\mu i}\partial\sigma} &= -\frac{1}{\sigma^2}\Big\{1 + \Big(\frac{x_i - g_{\mu i}}{\sigma}\Big)\exp\Big[-\Big(\frac{x_i - g_{\mu i}}{\sigma}\Big)\Big] - \exp\Big[-\Big(\frac{x_i - g_{\mu i}}{\sigma}\Big)\Big]\Big\} \\ \frac{\partial l(\mathbf{g}_\mu, \sigma, 0; 1)}{\partial\sigma} &= \frac{1}{\sigma}\sum_{i=1}^{N}\Big\{-1 + \frac{x_i - g_{\mu i}}{\sigma}\Big(1 - \exp\Big[-\Big(\frac{x_i - g_{\mu i}}{\sigma}\Big)\Big]\Big)\Big\} \end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,0;1)}{\partial\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{N}\Big\{1-2\frac{(x_i-g_{\mu i})}{\sigma}\Big(1-\exp\Big[-\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]\Big) \\
&\quad -\frac{(x_i-g_{\mu i})^2}{\sigma^2}\exp\Big[-\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]\Big\} \\
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,0;1)}{\partial\sigma\partial g_{\mu i}} &= -\frac{1}{\sigma^2}\Big\{1+\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\exp\Big[-\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]-\exp\Big[-\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]\Big\}
\end{aligned}
$$

$$\forall i\neq j \quad \frac{\partial^2 l(\mathbf{g}_\mu,\sigma,0;1)}{\partial g_{\mu i}\partial g_{\mu j}} = 0$$

and the first and second derivatives of the roughness penalty are given by

$$
\begin{aligned}
\frac{\partial(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial\mathbf{g}_\mu} = -\lambda\mathbf{K}\mathbf{g}_\mu \quad &\Rightarrow \quad \frac{\partial(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial g_{\mu i}} = -\lambda\big(\mathbf{K}\mathbf{g}_\mu\big)_i \\
\frac{\partial^2(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial\mathbf{g}_\mu\partial\mathbf{g}_\mu^T} = -\lambda\mathbf{K} \quad &\Rightarrow \quad \frac{\partial^2(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial g_{\mu i}\partial g_{\mu j}} = -\lambda K_{ij}
\end{aligned}
$$

With the expectations of the second derivative of the log-likelihood function given by

$$
\begin{aligned}
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,0;1)}{\partial g_{\mu i}^2}\right] &= \frac{1}{\sigma^2} \\
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,0;1)}{\partial\sigma^2}\right] &= \frac{N}{\sigma^2}\Big(1+\Psi'(2)+(\Psi(2))^2\Big) \\
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,0;1)}{\partial\sigma\partial g_{\mu i}}\right] &= -\frac{\Psi(1)+1}{\sigma^2} \\
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,0;1)}{\partial g_{\mu i}\partial\sigma}\right] &= -\frac{\Psi(1)+1}{\sigma^2}
\end{aligned}
$$

The equations from the Fisher scoring algorithm imply that

$$
\begin{aligned}
g_\mu^{new} &= S_1(y_1 - W_\mu^{-1}W_{\mu\sigma}\sigma^{new}) \\
\sigma^{new} &= S_2(y_2 - W_\sigma^{-1}W_{\sigma\mu}g_\mu^{new})
\end{aligned}
$$

where

$$
\begin{aligned}
S_1 &= (W_\mu+\lambda K)^{-1}W_\mu \\
S_2 &= (W_\sigma)^{-1}W_\sigma = I \\
y_1 &= g_\mu + W_\mu^{-1}u_\mu + W_\mu^{-1}W_{\mu\sigma}\sigma \\
y_2 &= \sigma + W_\sigma^{-1}u_\sigma + W_\sigma^{-1}W_{\sigma\mu}g_\mu
\end{aligned}
$$

with the elements of the Fisher scoring algorithm given by

$$
\begin{aligned}
u_\mu &= \frac{\partial l(\mathbf{g}_\mu, \sigma, 0; 1)}{\partial \mathbf{g}_\mu} = \left[\frac{1}{\sigma}\Big\{1 - \exp\Big[-\Big(\frac{x_i - g_{\mu i}}{\sigma}\Big)\Big]\Big\}\right]_i \\
u_\sigma &= \frac{\partial l(\mathbf{g}_\mu, \sigma, 0; 1)}{\partial \sigma} = \sum_{i=1}^{N} \frac{1}{\sigma}\Big[-1 + \frac{(x_i - g_{\mu i})}{\sigma}\Big\{1 - \exp\Big[-\Big(\frac{x_i - g_{\mu i}}{\sigma}\Big)\Big]\Big\}\Big] \\
W_\mu &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu, \sigma, 0; 1)}{\partial \mathbf{g}_\mu \partial \mathbf{g}_\mu^T}\right) = diag\left(\frac{1}{\sigma^2}\right) \\
W_\sigma &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu, \sigma, 0; 1)}{\partial \sigma^2}\right) = \left(\frac{N}{\sigma^2}\Big(1 + \Psi'(2) + (\Psi(2))^2\Big)\right) \\
W_{\mu\sigma} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu, \sigma, 0; 1)}{\partial \mathbf{g}_\mu \partial \sigma}\right) = ColumnVector\left(-\frac{\Psi(1) + 1}{\sigma^2}\right) \\
W_{\sigma\mu} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu, \sigma, 0; 1)}{\partial \sigma \partial \mathbf{g}_\mu^T}\right) = RowVector\left(-\frac{\Psi(1) + 1}{\sigma^2}\right)
\end{aligned}
$$

Appendix D

r largest order statistics with $\xi = 0$ and smoothed location and scale parameters

For the non-stationary distribution for the r largest order statistics with $\xi = 0$ and with smoothed location and scale parameters, let $\mu = g_\mu(t_i) = g_{\mu i}$ and $\sigma = g_\sigma(t_i) = g_{\sigma i}$, so that log-likelihood function is given by

$$l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r) = \sum_{i=1}^{N} \Bigg[- r \cdot \ln(g_{\sigma i}) - \exp\Big[- \Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}} \Big) \Big] - \sum_{k=1}^{r} \Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}} \Big) \Bigg]$$

The penalized log-likelihood is then given by

$$\begin{aligned} l^{PLL}(\mathbf{g}_\mu, \sigma, 0; r) &= l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r) - \frac{1}{2}\lambda \int g''_\mu(t)^2 dt - \frac{1}{2}\alpha \int g''_\sigma(t)^2 dt \\ &= l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r) - \frac{1}{2}\lambda \mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu - \frac{1}{2}\alpha \mathbf{g}_\sigma^T \mathbf{K} \mathbf{g}_\sigma \end{aligned}$$

The first and second derivatives of the log-likelihood function are given by

$$\begin{aligned} \frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial g_{\mu i}} &= \frac{1}{g_{\sigma i}} \Big\{ r - \exp\Big[- \Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}} \Big) \Big] \Big\} \\ \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial g_{\mu i}^2} &= \frac{-1}{g_{\sigma i}^2} \cdot \exp\Big[- \Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}} \Big) \Big] \\ \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial g_{\mu i} \partial g_{\sigma j}} &= \frac{-1}{g_{\sigma i}^2} \Big\{ r - \exp\Big[- \Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}} \Big) \Big] \Big\} - \frac{(z_i^{(r)} - g_{\mu i})}{g_{\sigma i}^3} \cdot \exp\Big[- \Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}} \Big) \Big] \\ \frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial g_{\sigma i}} &= -\frac{r}{g_{\sigma i}} - \frac{(z_i^{(r)} - g_{\mu i})}{g_{\sigma i}^2} \cdot \exp\Big[- \Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}} \Big) \Big] + \sum_{k=1}^{r} \Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}^2} \Big) \end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial g_{\sigma i}^2} &= \frac{r}{g_{\sigma i}^2} + \frac{2(z_i^{(r)} - g_{\mu i})}{g_{\sigma i}^3} \cdot \exp\Big[-\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\Big] \\
&\quad - \frac{(z_i^{(r)} - g_{\mu i})^2}{g_{\sigma i}^4} \cdot \exp\Big[-\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\Big] - 2\sum_{k=1}^{r}\Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}^3}\Big) \\
\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0, r)}{\partial g_{\sigma i}\partial g_{\mu i}} &= -\frac{r}{g_{\sigma i}^2} - \frac{(z_i^{(r)} - g_{\mu i})}{g_{\sigma i}^3} \cdot \exp\Big[-\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\Big] + \frac{1}{g_{\sigma i}^2} \cdot \exp\Big[-\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\Big]
\end{aligned}
$$

and

$$
\forall i \neq j \quad \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial g_{\mu i}\partial g_{\mu j}} = 0\ ,\ \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial g_{\mu i}\partial g_{\sigma j}} = 0\ ,\ \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial g_{\sigma i}\partial g_{\sigma j}} = 0\ ,\ \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial g_{\sigma i}\partial g_{\mu j}} = 0
$$

The first and second derivatives of the roughness penalties given by

$$
\begin{aligned}
&\frac{\partial(-\frac{1}{2}\lambda \mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu)}{\partial \mathbf{g}_\mu} = -\lambda \mathbf{K} \mathbf{g}_\mu \quad \Rightarrow \quad \frac{\partial(-\frac{1}{2}\lambda \mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu)}{\partial g_{\mu i}} = -\lambda\big(\mathbf{K}\mathbf{g}_\mu\big)_i \\
&\frac{\partial^2(-\frac{1}{2}\lambda \mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu)}{\partial \mathbf{g}_\mu \partial \mathbf{g}_\mu^T} = -\lambda \mathbf{K} \quad \Rightarrow \quad \frac{\partial^2(-\frac{1}{2}\lambda \mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu)}{\partial g_{\mu i}\partial g_{\mu j}} = -\lambda K_{ij} \\
&\frac{\partial(-\frac{1}{2}\alpha \mathbf{g}_\sigma^T \mathbf{K} \mathbf{g}_\sigma)}{\partial \mathbf{g}_\sigma} = -\alpha \mathbf{K} \mathbf{g}_\sigma \quad \Rightarrow \quad \frac{\partial(-\frac{1}{2}\alpha \mathbf{g}_\sigma^T \mathbf{K} \mathbf{g}_\sigma)}{\partial g_{\sigma i}} = -\alpha\big(\mathbf{K}\mathbf{g}_\sigma\big)_i \\
&\frac{\partial^2(-\frac{1}{2}\alpha \mathbf{g}_\sigma^T \mathbf{K} \mathbf{g}_\sigma)}{\partial \mathbf{g}_\sigma \partial \mathbf{g}_\sigma^T} = -\alpha \mathbf{K} \quad \Rightarrow \quad \frac{\partial^2(-\frac{1}{2}\alpha \mathbf{g}_\sigma^T \mathbf{K} \mathbf{g}_\sigma)}{\partial g_{\sigma i}\partial g_{\sigma j}} = -\alpha K_{ij}
\end{aligned}
$$

The expectations of the second derivative of the log-likelihood are given by

$$
\begin{aligned}
E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial g_{\mu i}^2}\Big] &= \frac{r}{g_{\sigma i}^2} \\
E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial g_{\sigma i}\partial g_{\mu i}}\Big] &= -\frac{1}{g_{\sigma i}^2}(r\Psi(r) + 1) \\
E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial g_{\mu i}\partial g_{\sigma i}}\Big] &= -\frac{1}{g_{\sigma i}^2}(r\Psi(r) + 1) \\
E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial g_{\sigma i}^2}\Big] &= \frac{1}{g_{\sigma i}^2}\left[2(r+1)\Psi(r) + r\Psi'(r) + r\left(\Psi(r)\right)^2 + 2 - r - 2\sum_{k=1}^{r}\Psi(k)\right]
\end{aligned}
$$

The equations from the Fisher scoring algorithm imply that

$$
\begin{aligned}
g_\mu^{new} &= S_1(y_1 - W_\mu^{-1} W_{\mu\sigma} g_\sigma^{new}) \\
g_\sigma^{new} &= S_2(y_2 - W_\sigma^{-1} W_{\sigma\mu} g_\mu^{new})
\end{aligned}
$$

where

$$\begin{aligned}
S_1 &= (W_\mu + \lambda K)^{-1} W_\mu \\
S_2 &= (W_\sigma + \alpha K)^{-1} W_\sigma \\
y_1 &= g_\mu + W_\mu^{-1} u_\mu + W_\mu^{-1} W_{\mu\sigma} g_\sigma \\
y_2 &= g_\sigma + W_\sigma^{-1} u_\sigma + W_\sigma^{-1} W_{\sigma\mu} g_\mu.
\end{aligned}$$

The elements of the Fisher scoring algorithm are given by

$$\begin{aligned}
u_\mu &= \frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial \mathbf{g}_\mu} = \left[\frac{1}{g_{\sigma i}} \Big\{ r - \exp\Big[-\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}} \Big) \Big] \Big\} \right]_i \\
u_\sigma &= \frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial \mathbf{g}_\sigma} = \left[-\frac{r}{g_{\sigma i}} - \frac{(z_i^{(r)} - g_{\mu i})}{g_{\sigma i}^2} \cdot \exp\Big[-\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}} \Big) \Big] + \sum_{k=1}^{r} \Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}^2} \Big) \right]_i \\
W_\mu &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial \mathbf{g}_\mu \partial \mathbf{g}_\mu^T} \right) = diag\left(\frac{r}{g_{\sigma i}^2} \right) \\
W_\sigma &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial \mathbf{g}_\sigma \partial \mathbf{g}_\sigma^T} \right) = diag\left(\frac{1}{g_{\sigma i}^2} \left[2(r+1)\Psi(r) + r\Psi'(r) \right.\right. \\
&\qquad \left.\left. + r\,(\Psi(r))^2 + 2 - r - 2\sum_{k=1}^{r} \Psi(k) \right] \right) \\
W_{\mu\sigma} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial \mathbf{g}_\mu \partial \mathbf{g}_\sigma^T} \right) = diag\left(-\frac{1}{g_{\sigma i}^2} (r\Psi(r) + 1) \right) \\
W_{\sigma\mu} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, 0; r)}{\partial \mathbf{g}_\sigma \partial \mathbf{g}_\mu^T} \right) = diag\left(-\frac{1}{g_{\sigma i}^2} (r\Psi(r) + 1) \right)
\end{aligned}$$

Appendix E

r largest model with $\xi = 0$ and a constant scale parameter and smoothed location parameter

For the non-stationary distribution of the r largest order statistics with $\xi = 0$ and with a constant scale parameter and smoothed location parameter, let $\mu = g_\mu(t_i) = g_{\mu i}$, so that log-likelihood function is given by

$$l(\mathbf{g}_\mu, \sigma, 0; r) = \sum_{i=1}^{N} \left[-r \cdot \ln(\sigma) - \exp\left[-\left(\frac{z_i^{(r)} - g_{\mu i}}{\sigma} \right) \right] - \sum_{k=1}^{r} \left(\frac{z_i^{(k)} - g_{\mu i}}{\sigma} \right) \right]$$

and the penalized log-likelihood by

$$\begin{aligned} l^{PLL}(\mathbf{g}_\mu, \sigma, 0; r) &= l(\mathbf{g}_\mu, \sigma, 0; r) - \frac{1}{2}\lambda \int g''_\mu(t)^2 dt \\ &= l(\mathbf{g}_\mu, \sigma, 0; r) - \frac{1}{2}\lambda \mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu \end{aligned}$$

The first and second derivatives of the log-likelihood function are given by

$$\begin{aligned} \frac{\partial l(\mathbf{g}_\mu, \sigma, 0; r)}{\partial g_{\mu i}} &= \frac{1}{\sigma}\left\{ r - \exp\left[-\left(\frac{z_i^{(r)} - g_{\mu i}}{\sigma} \right) \right] \right\} \\ \frac{\partial^2 l(\mathbf{g}_\mu, \sigma, 0; r)}{\partial g_{\mu i}^2} &= \frac{-1}{\sigma^2} \cdot \exp\left[-\left(\frac{z_i^{(r)} - g_{\mu i}}{\sigma} \right) \right] \\ \frac{\partial^2 l(\mathbf{g}_\mu, \sigma, 0; r)}{\partial g_{\mu i} \partial \sigma} &= \frac{-1}{\sigma^2}\left\{ r - \exp\left[-\left(\frac{z_i^{(r)} - g_{\mu i}}{\sigma} \right) \right] \right\} - \frac{(z_i^{(r)} - g_{\mu i})}{\sigma^3} \cdot \exp\left[-\left(\frac{z_i^{(r)} - g_{\mu i}}{\sigma} \right) \right] \\ \frac{\partial l(\mathbf{g}_\mu, \sigma, 0; r)}{\partial \sigma} &= \sum_{i=1}^{N} \left[-\frac{r}{\sigma} - \frac{(z_i^{(r)} - g_{\mu i})}{\sigma^2} \cdot \exp\left[-\left(\frac{z_i^{(r)} - g_{\mu i}}{\sigma} \right) \right] + \sum_{k=1}^{r} \left(\frac{z_i^{(k)} - g_{\mu i}}{\sigma^2} \right) \right] \end{aligned}$$

$$\begin{aligned}\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,0;r)}{\partial\sigma^2} &= \sum_{i=1}^{N}\Bigg[\frac{r}{\sigma^2}+\frac{2(z_i^{(r)}-g_{\mu i})}{\sigma^3}\cdot\exp\Big[-\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big] \\ &\quad -\frac{(z_i^{(r)}-g_{\mu i})^2}{\sigma^4}\cdot\exp\Big[-\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big]-2\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma^3}\Big)\Bigg] \\ \frac{\partial^2 l(\mathbf{g}_\mu,\sigma,0;r)}{\partial\sigma\partial g_{\mu i}} &= -\frac{r}{\sigma^2}-\frac{(z_i^{(r)}-g_{\mu i})}{\sigma^3}\cdot\exp\Big[-\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big]+\frac{1}{\sigma^2}\cdot\exp\Big[-\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big]\end{aligned}$$

and

$$\forall i\neq j \quad \frac{\partial^2 l(\mathbf{g}_\mu,\sigma,0;r)}{\partial g_{\mu i}\partial g_{\mu j}}=0$$

The first and second derivatives of the roughness penalty are given by

$$\begin{aligned}\frac{\partial(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial\mathbf{g}_\mu} &= -\lambda\mathbf{K}\mathbf{g}_\mu \quad\Rightarrow\quad \frac{\partial(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial g_{\mu i}}=-\lambda\big(\mathbf{K}\mathbf{g}_\mu\big)_i \\ \frac{\partial^2(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial\mathbf{g}_\mu\partial\mathbf{g}_\mu^T} &= -\lambda\mathbf{K} \quad\Rightarrow\quad \frac{\partial^2(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial g_{\mu i}\partial g_{\mu j}}=-\lambda K_{ij}\end{aligned}$$

With the expectations of the second derivative of the log-likelihood function given by

$$\begin{aligned}E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,0;r)}{\partial g_{\mu i}^2}\Big] &= \frac{r}{\sigma^2} \\ E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,0;r)}{\partial\sigma\partial g_{\mu i}}\Big] &= -\frac{1}{\sigma^2}(r\Psi(r)+1) \\ E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,0;r)}{\partial g_{\mu i}\partial\sigma}\Big] &= -\frac{1}{\sigma^2}(r\Psi(r)+1) \\ E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,0;r)}{\partial\sigma^2}\Big] &= \frac{N}{\sigma^2}\Bigg[2(r+1)\Psi(r)+r\Psi'(r)+r\left(\Psi(r)\right)^2+2-r-2\sum_{k=1}^{r}\Psi(k)\Bigg]\end{aligned}$$

The equations from the Fisher scoring algorithm imply that

$$\begin{aligned}g_\mu^{new} &= S_1(y_1-W_\mu^{-1}W_{\mu\sigma}\sigma^{new}) \\ \sigma^{new} &= S_2(y_2-W_\sigma^{-1}W_{\sigma\mu}g_\mu^{new})\end{aligned}$$

where

$$\begin{aligned}S_1 &= (W_\mu+\lambda K)^{-1}W_\mu \\ S_2 &= (W_\sigma)^{-1}W_\sigma=I \\ y_1 &= g_\mu+W_\mu^{-1}u_\mu+W_\mu^{-1}W_{\mu\sigma}\sigma \\ y_2 &= \sigma+W_\sigma^{-1}u_\sigma+W_\sigma^{-1}W_{\sigma\mu}g_\mu\end{aligned}$$

with the elements of the Fisher scoring algorithm given by

$$\begin{aligned}
u_\mu &= \frac{\partial l(\mathbf{g}_\mu, \sigma, 0; r)}{\partial \mathbf{g}_\mu} = \left[\frac{1}{\sigma}\Big\{r - \exp\Big[-\Big(\frac{z_i^{(r)} - g_{\mu i}}{\sigma}\Big)\Big]\Big\}\right]_i \\
u_\sigma &= \frac{\partial l(\mathbf{g}_\mu, \sigma, 0; r)}{\partial \sigma} = \sum_{i=1}^{N}\left[-\frac{r}{\sigma} - \frac{(z_i^{(r)} - g_{\mu i})}{\sigma^2}\cdot \exp\Big[-\Big(\frac{z_i^{(r)} - g_{\mu i}}{\sigma}\Big)\Big] + \sum_{k=1}^{r}\Big(\frac{z_i^{(k)} - g_{\mu i}}{\sigma^2}\Big)\right] \\
W_\mu &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu, \sigma, 0; r)}{\partial \mathbf{g}_\mu \partial \mathbf{g}_\mu^T}\right) = diag\left(\frac{r}{\sigma^2}\right) \\
W_\sigma &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu, \sigma, 0; r)}{\partial \sigma^2}\right) = \left(\frac{N}{\sigma^2}\left[2(r+1)\Psi(r) + r\Psi'(r) + r\left(\Psi(r)\right)^2 + 2 - r - 2\sum_{k=1}^{r}\Psi(k)\right]\right) \\
W_{\mu\sigma} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu, \sigma, 0; r)}{\partial \mathbf{g}_\mu \partial \sigma}\right) = ColumnVector\left(-\frac{(r\Psi(r)+1)}{\sigma^2}\right) \\
W_{\sigma\mu} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu, \sigma, 0; r)}{\partial \sigma \partial \mathbf{g}_\mu^T}\right) = RowVector\left(-\frac{(r\Psi(r)+1)}{\sigma^2}\right)
\end{aligned}$$

Appendix F

GEV block maxima model with smoothed location and scale parameters

For the non-stationary GEV block maxima model with smoothed location and scale parameters, let $\mu = g_\mu(t_i) = g_{\mu i}$ and $\sigma = g_\sigma(t_i) = g_{\sigma i}$, so that log-likelihood function is given by

$$l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1) = -\sum_{i=1}^{N} \left(\left[1 + \xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-\frac{1}{\xi}} + \Big(\frac{1}{\xi} + 1\Big) \ln \left[1 + \xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\right] + \ln(g_{\sigma i}) \right)$$

The penalized log-likelihood is then given by

$$\begin{aligned} l^{PLL}(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1) &= l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1) - \frac{\lambda}{2}\int g''_\mu(t)^2 dt - \frac{\alpha}{2}\int g''_\sigma(t)^2 dt \\ &= l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1) - \frac{\lambda}{2}\mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu - \frac{\alpha}{2}\mathbf{g}_\sigma^T \mathbf{K} \mathbf{g}_\sigma \end{aligned}$$

The first and second derivatives of the log-likelihood function are given by

$$\begin{aligned} \frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial g_{\mu i}} &= -\frac{1}{g_{\sigma i}}\left[1 + \xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-\left(\frac{1}{\xi}+1\right)} + \frac{1}{g_{\sigma i}}\big(1+\xi\big)\left[1 + \xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-1} \\ \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial g_{\mu i}^2} &= -\Big(\frac{1+\xi}{g_{\sigma i}^2}\Big)\left[1 + \xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-\left(\frac{1}{\xi}+2\right)} + \frac{\xi}{g_{\sigma i}^2}\big(1+\xi\big)\left[1 + \xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-2} \end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial g_{\mu i}\partial g_{\sigma i}} &= \frac{1}{g_{\sigma i}^2}\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-\left(\frac{1}{\xi}+1\right)} - \left(\frac{1+\xi}{g_{\sigma i}^2}\right)\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-1} \\
&\quad - \left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\left(\frac{1+\xi}{g_{\sigma i}^2}\right)\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad + \left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\left(\frac{\xi(1+\xi)}{g_{\sigma i}^2}\right)\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-2} \\
\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial g_{\mu i}\partial \xi} &= -\frac{1}{g_{\sigma i}\xi^2}\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-\left(\frac{1}{\xi}+1\right)} \ln\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right] \\
&\quad + \frac{1}{g_{\sigma i}}\left(\frac{1}{\xi}+1\right)\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad + \frac{1}{g_{\sigma i}}\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-1} - \left(\frac{1+\xi}{g_{\sigma i}}\right)\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-2} \\
\frac{\partial l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial \mathbf{g}_{\sigma\mathbf{i}}} &= -\frac{1}{g_{\sigma i}}\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-\left(\frac{1}{\xi}+1\right)} \\
&\quad + \left(\frac{1+\xi}{g_{\sigma i}}\right)\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-1} - \frac{1}{g_{\sigma i}} \\
\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial {\mathbf{g}_{\sigma\mathbf{i}}}^2} &= \frac{2}{g_{\sigma i}^2}\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-\left(\frac{1}{\xi}+1\right)} \\
&\quad - \left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)^2\left(\frac{1+\xi}{\sigma^2}\right)\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad - \left(\frac{2(1+\xi)}{g_{\sigma i}^2}\right)\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\left[1+\xi\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\right]^{-1} \\
&\quad + \left(\frac{\xi\,(1+\xi)}{\sigma^2}\right)\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)^2\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-2} + \frac{1}{g_{\sigma i}^2} \\
\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial g_{\sigma i}\partial g_{\mu i}} &= \frac{1}{g_{\sigma i}^2}\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-\left(\frac{1}{\xi}+1\right)} \\
&\quad - \left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\left(\frac{1+\xi}{g_{\sigma i}^2}\right)\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad - \left(\frac{1+\xi}{g_{\sigma i}^2}\right)\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-1} \\
&\quad + \left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\left(\frac{\xi(1+\xi)}{g_{\sigma i}^2}\right)\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-2} \\
\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial g_{\sigma i}\partial \xi} &= -\left(\frac{1}{g_{\sigma i}\xi^2}\right)\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-\left(\frac{1}{\xi}+1\right)} \ln\left(1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right) \\
&\quad + \frac{1}{g_{\sigma i}}\left(\frac{1}{\xi}+1\right)\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)^2\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad + \frac{1}{g_{\sigma i}}\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\left[1+\xi\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\right]^{-1} \\
&\quad - \left(\frac{1+\xi}{g_{\sigma i}}\right)\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)^2\left[1+\xi\left(\frac{x_i-g_{\mu i}}{g_{\sigma i}}\right)\right]^{-2}
\end{aligned}
$$

$$\begin{aligned}
\frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial \xi} &= -\sum_{i=1}^{N} \frac{1}{\xi^2}\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\frac{1}{\xi}} \ln\Big(1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big) \\
&+\sum_{i=1}^{N} \frac{1}{\xi}\left(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\right)\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \\
&+\sum_{i=1}^{N} \frac{1}{\xi^2} \ln\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big] \\
&-\sum_{i=1}^{N} \Big(\frac{1}{\xi}+1\Big)\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-1} \\
\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial \xi^2} &= -\sum_{i=1}^{N} \frac{1}{\xi^4}\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\frac{1}{\xi}} \left(\ln\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]\right)^2 \\
&+\sum_{i=1}^{N} \frac{2}{\xi^3}\left(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\right)\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \ln\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big] \\
&+\sum_{i=1}^{N} \frac{2}{\xi^3}\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\frac{1}{\xi}} \ln\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big] \\
&-\sum_{i=1}^{N} \frac{2}{\xi^2}\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \\
&-\sum_{i=1}^{N} \left(\frac{1}{\xi^2}+\frac{1}{\xi}\right)\left(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\right)^2\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&-\sum_{i=1}^{N} \frac{2}{\xi^3} \ln\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big] \\
&+\sum_{i=1}^{N} \frac{2}{\xi^2}\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-1} \\
&+\sum_{i=1}^{N} \Big(\frac{1}{\xi}+1\Big)\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)^2\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-2} \\
\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial \xi \partial g_{\mu i}} &= -\frac{1}{g_{\sigma i}\xi^2}\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \ln\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big] \\
&+\frac{1}{\sigma}\left(\frac{1}{\xi}+1\right)\left(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\right)\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&+\frac{1}{g_{\sigma i}}\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-1} - \left(\frac{1+\xi}{g_{\sigma i}}\right)\left(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\right)\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-2} \\
\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial \xi \partial g_{\sigma i}} &= -\left(\frac{1}{g_{\sigma i}\xi^2}\right)\left(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\right)\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \ln\Big(1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big) \\
&+\frac{1}{g_{\sigma i}}\left(\frac{1}{\xi}+1\right)\left(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\right)^2\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&+\frac{1}{g_{\sigma i}}\left(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\right)\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{\sigma}\Big)\Big]^{-1} \\
&-\left(\frac{1+\xi}{g_{\sigma i}}\right)\left(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\right)^2\Big[1+\xi\Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-2}
\end{aligned}$$

and

$$\forall i \neq j \quad \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial g_{\mu i} \partial g_{\mu j}} = 0 \qquad , \qquad \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial g_{\sigma i} \partial g_{\sigma j}} = 0$$

with the first and second derivatives of the roughness penalties given by

$$\begin{aligned}
\frac{\partial(-\frac{1}{2}\lambda \mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu)}{\partial \mathbf{g}_\mu} &= -\lambda \mathbf{K} \mathbf{g}_\mu \quad \Rightarrow \quad \frac{\partial(-\frac{1}{2}\lambda \mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu)}{\partial g_{\mu i}} = -\lambda \big(\mathbf{K} \mathbf{g}_\mu\big)_i \\
\frac{\partial^2(-\frac{1}{2}\lambda \mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu)}{\partial \mathbf{g}_\mu \partial \mathbf{g}_\mu^T} &= -\lambda \mathbf{K} \quad \Rightarrow \quad \frac{\partial^2(-\frac{1}{2}\lambda \mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu)}{\partial g_{\mu i} \partial g_{\mu j}} = -\lambda K_{ij} \\
\frac{\partial(-\frac{1}{2}\alpha \mathbf{g}_\sigma^T \mathbf{K} \mathbf{g}_\sigma)}{\partial \mathbf{g}_\sigma} &= -\alpha \mathbf{K} \mathbf{g}_\sigma \quad \Rightarrow \quad \frac{\partial(-\frac{1}{2}\alpha \mathbf{g}_\sigma^T \mathbf{K} \mathbf{g}_\sigma)}{\partial g_{\sigma i}} = -\alpha \big(\mathbf{K} \mathbf{g}_\sigma\big)_i \\
\frac{\partial^2(-\frac{1}{2}\alpha \mathbf{g}_\sigma^T \mathbf{K} \mathbf{g}_\sigma)}{\partial \mathbf{g}_\sigma \partial \mathbf{g}_\sigma^T} &= -\alpha \mathbf{K} \quad \Rightarrow \quad \frac{\partial^2(-\frac{1}{2}\alpha \mathbf{g}_\sigma^T \mathbf{K} \mathbf{g}_\sigma)}{\partial g_{\sigma i} \partial g_{\sigma j}} = -\alpha K_{ij}
\end{aligned}$$

The expectations of the second derivative of the log-likelihood are given by

$$\begin{aligned}
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial g_{\mu i}^2}\right] &= \left(\frac{1+\xi}{g_{\sigma i}^2}\right)\Gamma(2+2\xi) - \left(\frac{\xi(1+\xi)}{g_{\sigma i}^2}\right)\Gamma(1+2\xi) \\
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial g_{\sigma i} \partial g_{\mu i}}\right] &= \left(\frac{1}{\xi\, g_{\sigma i}^2}\right)\Gamma(2+\xi) - \left(\frac{1+\xi}{\xi\, g_{\sigma i}^2}\right)\Gamma(2+2\xi) + \left(\frac{1+\xi}{g_{\sigma i}^2}\right)\Gamma(1+2\xi) \\
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial \xi \partial g_{\mu i}}\right] &= -\left(\frac{\Psi(2+\xi)}{g_{\sigma i}\xi} + \frac{1+\xi}{g_{\sigma i}\,\xi^2}\right)\Gamma(2+\xi) + \left(\frac{1+\xi}{g_{\sigma i}\,\xi^2}\right)\Gamma(2+2\xi) \\
&\quad + \left(\frac{1}{g_{\sigma i}\,\xi}\right)\Gamma(1+\xi) - \left(\frac{1+\xi}{g_{\sigma i}\,\xi}\right)\Gamma(1+2\xi) \\
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial g_{\sigma i}^2}\right] &= \left(\frac{2}{g_{\sigma i}^2\,\xi}\right)\left[\frac{\Gamma(1+\xi+1)}{\Gamma(1)} - \frac{\Gamma(1+1)}{\Gamma(1)}\right] \\
&\quad + \left(\frac{1+\xi}{g_{\sigma i}^2\,\xi^2}\right)\left[\frac{\Gamma(1+2\xi+1)}{\Gamma(1)} - 2\frac{\Gamma(1+\xi+1)}{\Gamma(1)} + \frac{\Gamma(1+1)}{\Gamma(1)}\right] \\
&\quad - \left(\frac{2(1+\xi)}{g_{\sigma i}^2\,\xi}\right)\left[\frac{\Gamma(1+\xi)}{\Gamma(1)} - \frac{\Gamma(1)}{\Gamma(1)}\right] \\
&\quad - \left(\frac{1+\xi}{g_{\sigma i}^2\,\xi}\right)\left[\frac{\Gamma(1+2\xi)}{\Gamma(1)} - 2\frac{\Gamma(1+\xi)}{\Gamma(1)} + \frac{\Gamma(1)}{\Gamma(1)}\right] - \frac{1}{g_{\sigma i}^2} \\
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial g_{\sigma i}^2}\right] &= -\left(\frac{2}{g_{\sigma i}^2\,\xi^2}\right)\Gamma(2+\xi) + \left(\frac{1+\xi}{g_{\sigma i}^2\,\xi^2}\right)\Gamma(2+2\xi) \\
&\quad - \left(\frac{1+\xi}{g_{\sigma i}^2\,\xi}\right)\Gamma(1+2\xi) - \frac{1}{g_{\sigma i}^2\,\xi^2} \\
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial \xi \partial g_{\sigma i}}\right] &= \left[\frac{\Psi(2+\xi)}{g_{\sigma i}\,\xi^2} + \frac{2(1+\xi)}{g_{\sigma i}\,\xi^3}\right]\Gamma(2+\xi) - \left[\frac{1+\xi}{g_{\sigma i}\,\xi^3}\right]\Gamma(2+2\xi) \\
&\quad - \left[\frac{2+\xi}{g_{\sigma i}\,\xi^2}\right]\Gamma(1+\xi) + \left[\frac{1+\xi}{g_{\sigma i}\,\xi^2}\right]\Gamma(1+2\xi) - \left[\frac{\Psi(2)}{g_{\sigma i}\,\xi^2} + \frac{1}{g_{\sigma i}\,\xi^3}\right]
\end{aligned}$$

$$
\begin{aligned}
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial\xi^2}\right] &= \sum_{i=1}^{N}\Bigg(\left[-\frac{2}{\xi^3}\Psi(2+\xi)-2\left(\frac{1+2\,\xi}{\xi^4}\right)\right]\Gamma(2+\xi)+\left(\frac{1+\xi}{\xi^4}\right)\Gamma(2+2\,\xi)\\
&\quad +2\left(\frac{2+\xi}{\xi^3}\right)\Gamma(1+\xi)-\left(\frac{1+\xi}{\xi^3}\right)\Gamma(1+2\,\xi)+\frac{1}{\xi^2}\left[\Psi'(2)+[\Psi(2)]^2\right]\\
&\quad +2\left(\frac{1+\xi}{\xi^3}\right)\Psi(2)-\left(\frac{2}{\xi^2}\right)\Psi(1)+\frac{1-\xi^2}{\xi^4}\Bigg)\\
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial g_{\mu i}\partial\xi}\right] &= -\left(\frac{\Psi(2+\xi)}{g_{\sigma i}\xi}+\frac{1+\xi}{g_{\sigma i}\,\xi^2}\right)\Gamma(2+\xi)+\left(\frac{1+\xi}{g_{\sigma i}\,\xi^2}\right)\Gamma(2+2\,\xi)\\
&\quad +\left(\frac{1}{g_{\sigma i}\,\xi}\right)\Gamma(1+\xi)-\left(\frac{1+\xi}{g_{\sigma i}\,\xi}\right)\Gamma(1+2\,\xi)\\
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial g_{\sigma i}\partial\xi}\right] &= \left[\frac{\Psi(2+\xi)}{g_{\sigma i}\,\xi^2}+\frac{2(1+\xi)}{g_{\sigma i}\,\xi^3}\right]\Gamma(2+\xi)-\left[\frac{1+\xi}{g_{\sigma i}\,\xi^3}\right]\Gamma(2+2\,\xi)\\
&\quad -\left[\frac{2+\xi}{g_{\sigma i}\,\xi^2}\right]\Gamma(1+\xi)+\left[\frac{1+\xi}{g_{\sigma i}\,\xi^2}\right]\Gamma(1+2\,\xi)-\left[\frac{\Psi(2)}{g_{\sigma i}\,\xi^2}+\frac{1}{g_{\sigma i}\,\xi^3}\right]
\end{aligned}
$$

By defining,

$$
\begin{aligned}
u_\mu &= \frac{\partial l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial\mathbf{g}_\mu} & u_\sigma &= \frac{\partial l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial\mathbf{g}_\sigma}\\
u_\xi &= \frac{\partial l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial\xi} & W_\xi &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial\xi^2}\right)\\
W_\mu &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial\mathbf{g}_\mu\partial\mathbf{g}_\mu^T}\right) & W_\sigma &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial\mathbf{g}_\sigma\partial\mathbf{g}_\sigma^T}\right)\\
W_{\mu\sigma} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial\mathbf{g}_\mu\partial\mathbf{g}_\sigma^T}\right) & W_{\sigma\mu} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial\mathbf{g}_\sigma\partial\mathbf{g}_\mu^T}\right)\\
W_{\mu\xi} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial\mathbf{g}_\mu\partial\xi}\right) & W_{\xi\mu} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial\xi\partial\mathbf{g}_\mu^T}\right)\\
W_{\sigma\xi} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial\mathbf{g}_\sigma\partial\xi}\right) & W_{\xi\sigma} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial\xi\partial\mathbf{g}_\sigma^T}\right)
\end{aligned}
$$

the Fisher scoring equation, for the non-stationary GEV black maxima model with smoothed location and scale parameters, can be written as

$$
\begin{pmatrix} W_\mu+\lambda K & W_{\mu\sigma} & W_{\mu\xi}\\ W_{\sigma\mu} & W_\sigma+\alpha K & W_{\sigma\xi}\\ W_{\xi\mu} & W_{\xi\sigma} & W_\xi \end{pmatrix}
\begin{pmatrix} g_\mu^{new}-g_\mu\\ g_\sigma^{new}-g_\sigma\\ \xi^{new}-\xi \end{pmatrix}
=
\begin{pmatrix} u_\mu-\lambda K g_\mu\\ u_\sigma-\alpha K g_\sigma\\ u_\xi \end{pmatrix}
$$

which implies that

$$g_\mu^{new} = S_1(y_1 - W_\mu^{-1} W_{\mu\sigma} g_\sigma^{new} - W_\mu^{-1} W_{\mu\xi}\xi^{new})$$
$$g_\sigma^{new} = S_2(y_2 - W_\sigma^{-1} W_{\sigma\mu} g_\mu^{new} - W_\sigma^{-1} W_{\sigma\xi}\xi^{new})$$
$$\xi^{new} = S_3(y_3 - W_\xi^{-1} W_{\xi\mu} g_\mu^{new} - W_\xi^{-1} W_{\xi\sigma} g_\sigma^{new})$$

where

$$\begin{aligned}
S_1 &= (W_\mu + \lambda K)^{-1} W_\mu \\
S_2 &= (W_\sigma + \alpha K)^{-1} W_\sigma \\
S_3 &= (W_\xi)^{-1} W_\xi = I \\
y_1 &= g_\mu + W_\mu^{-1} u_\mu + W_\mu^{-1} W_{\mu\sigma} g_\sigma + W_\mu^{-1} W_{\mu\xi}\xi \\
y_2 &= g_\sigma + W_\sigma^{-1} u_\sigma + W_\sigma^{-1} W_{\sigma\mu} g_\mu + W_\sigma^{-1} W_{\sigma\xi}\xi \\
y_3 &= \xi + W_\xi^{-1} u_\xi + W_\xi^{-1} W_{\xi\mu} g_\mu + W_\xi^{-1} W_{\xi\sigma} g_\sigma
\end{aligned}$$

and the elements of the Fisher scoring algorithm are given by

$$\begin{aligned}
u_\mu &= \frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial \mathbf{g}_\mu} = \left(-\frac{1}{g_{\sigma i}} \Big[1 + \xi \Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}} \Big) \Big]^{-\left(\frac{1}{\xi}+1\right)} + \frac{1}{g_{\sigma i}} \big(1+\xi\big) \Big[1 + \xi \Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}} \Big) \Big]^{-1} \right)_i \\
u_\sigma &= \frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial g_\sigma} = \left(-\frac{1}{g_{\sigma i}} \left(\frac{x_i - g_{\mu i}}{g_{\sigma i}} \right) \Big[1 + \xi \Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}} \Big) \Big]^{-\left(\frac{1}{\xi}+1\right)} \right. \\
&\qquad \left. + \left(\frac{1+\xi}{g_{\sigma i}} \right) \left(\frac{x_i - g_{\mu i}}{g_{\sigma i}} \right) \Big[1 + \xi \Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}} \Big) \Big]^{-1} - \frac{1}{g_{\sigma i}} \right)_i \\
u_\xi &= \frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial \xi} = \sum_{i=1}^{N} \Bigg[-\frac{1}{\xi^2} \Big[1 + \xi \Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}} \Big) \Big]^{-\frac{1}{\xi}} \ln \Big(1 + \xi \Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}} \Big) \Big) \\
&\qquad + \frac{1}{\xi} \left(\frac{x_i - g_{\mu i}}{g_{\sigma i}} \right) \Big[1 + \xi \Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}} \Big) \Big]^{-\left(\frac{1}{\xi}+1\right)} \\
&\qquad + \frac{1}{\xi^2} \ln \Big[1 + \xi \Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}} \Big) \Big] \\
&\qquad - \Big(\frac{1}{\xi} + 1 \Big) \Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}} \Big) \Big[1 + \xi \Big(\frac{x_i - g_{\mu i}}{g_{\sigma i}} \Big) \Big]^{-1} \Bigg] \\
W_\mu &= E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial \mathbf{g}_\mu \partial \mathbf{g}_\mu^T} \Big] \\
&= diag\left(\left(\frac{1+\xi}{g_{\sigma i}^2} \right) \Gamma(2 + 2\,\xi) - \left(\frac{\xi\big(1+\xi\big)}{g_{\sigma i}^2} \right) \Gamma(1 + 2\,\xi) \right)
\end{aligned}$$

$$
\begin{aligned}
W_\sigma &= E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial \mathbf{g}_\sigma \partial \mathbf{g}_\sigma^T}\Big] \\
&= diag\left(-\left(\frac{2}{g_{\sigma i}^2\,\xi^2}\right)\Gamma(2+\xi)+\left(\frac{1+\xi}{g_{\sigma i}^2\,\xi^2}\right)\Gamma(2+2\,\xi)-\left(\frac{1+\xi}{g_{\sigma i}^2\,\xi}\right)\Gamma(1+2\,\xi)-\frac{1}{g_{\sigma i}^2\,\xi^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
W_\xi &= E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial \xi^T}\Big] \\
&= N\Bigg(\left[-\frac{2}{\xi^3}\Psi(2+\xi)-2\left(\frac{1+2\,\xi}{\xi^4}\right)\right]\Gamma(2+\xi)+\left(\frac{1+\xi}{\xi^4}\right)\Gamma(2+2\,\xi) \\
&\qquad +2\left(\frac{2+\xi}{\xi^3}\right)\Gamma(1+\xi)-\left(\frac{1+\xi}{\xi^3}\right)\Gamma(1+2\,\xi)+\frac{1}{\xi^2}\Big[\Psi'(2)+[\Psi(2)]^2\Big] \\
&\qquad +2\left(\frac{1+\xi}{\xi^3}\right)\Psi(2)-\left(\frac{2}{\xi^2}\right)\Psi(1)+\frac{1-\xi^2}{\xi^4}\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
W_{\mu\sigma} &= E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial \mathbf{g}_\mu \partial \mathbf{g}_\sigma^T}\Big] \\
&= \text{diag}\left(\left(\frac{1}{\xi\,g_{\sigma i}^2}\right)\Gamma(2+\ \xi)-\left(\frac{1+\xi}{\xi\,g_{\sigma i}^2}\right)\Gamma(2+2\,\xi)+\left(\frac{1+\xi}{g_{\sigma i}^2}\right)\Gamma(1+2\,\xi)\right)_i
\end{aligned}
$$

$$
\begin{aligned}
W_{\sigma\mu} &= E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial \mathbf{g}_\sigma \partial \mathbf{g}_\mu^T}\Big] \\
&= \text{diag}\left(\left(\frac{1}{\xi\,g_{\sigma i}^2}\right)\Gamma(2+\ \xi)-\left(\frac{1+\xi}{\xi\,g_{\sigma i}^2}\right)\Gamma(2+2\,\xi)+\left(\frac{1+\xi}{g_{\sigma i}^2}\right)\Gamma(1+2\,\xi)\right)_i
\end{aligned}
$$

$$
\begin{aligned}
W_{\mu\xi} &= E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial \mathbf{g}_\mu \partial \xi}\Big] \\
&= \text{column vector}\Bigg(-\left(\frac{\Psi(2+\xi)}{g_{\sigma i}\xi}+\frac{1+\xi}{g_{\sigma i}\,\xi^2}\right)\Gamma(2+\xi)+\left(\frac{1+\xi}{g_{\sigma i}\,\xi^2}\right)\Gamma(2+2\,\xi) \\
&\qquad +\left(\frac{1}{g_{\sigma i}\,\xi}\right)\Gamma(1+\xi)-\left(\frac{1+\xi}{g_{\sigma i}\,\xi}\right)\Gamma(1+2\,\xi)\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
W_{\xi\mu} &= E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;1)}{\partial \xi \partial \mathbf{g}_\mu^T}\Big] \\
&= \text{row vector}\Bigg(-\left(\frac{\Psi(2+\xi)}{g_{\sigma i}\xi}+\frac{1+\xi}{g_{\sigma i}\,\xi^2}\right)\Gamma(2+\xi)+\left(\frac{1+\xi}{g_{\sigma i}\,\xi^2}\right)\Gamma(2+2\,\xi) \\
&\qquad +\left(\frac{1}{g_{\sigma i}\,\xi}\right)\Gamma(1+\xi)-\left(\frac{1+\xi}{g_{\sigma i}\,\xi}\right)\Gamma(1+2\,\xi)\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
W_{\sigma\xi} &= E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial \mathbf{g}_\sigma \partial \xi}\Big] \\
&= \text{column vector}\Bigg(\left[\frac{\Psi(2+\xi)}{g_{\sigma i}\,\xi^2} + \frac{2(1+\xi)}{g_{\sigma i}\,\xi^3}\right]\Gamma(2+\xi) - \left[\frac{1+\xi}{g_{\sigma i}\,\xi^3}\right]\Gamma(2+2\,\xi) \\
&\qquad - \left[\frac{2+\xi}{g_{\sigma i}\,\xi^2}\right]\Gamma(1+\xi) + \left[\frac{1+\xi}{g_{\sigma i}\,\xi^2}\right]\Gamma(1+2\,\xi) - \left[\frac{\Psi(2)}{g_{\sigma i}\,\xi^2} + \frac{1}{g_{\sigma i}\,\xi^3}\right] \Bigg)
\end{aligned}
$$

$$
\begin{aligned}
W_{\xi\sigma} &= E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1)}{\partial \xi \partial \mathbf{g}_\sigma^T}\Big] \\
&= \text{row vector}\Bigg(\Bigg(\left[\frac{\Psi(2+\xi)}{g_{\sigma i}\,\xi^2} + \frac{2(1+\xi)}{g_{\sigma i}\,\xi^3}\right]\Gamma(2+\xi) - \left[\frac{1+\xi}{g_{\sigma i}\,\xi^3}\right]\Gamma(2+2\,\xi) \\
&\qquad - \left[\frac{2+\xi}{g_{\sigma i}\,\xi^2}\right]\Gamma(1+\xi) + \left[\frac{1+\xi}{g_{\sigma i}\,\xi^2}\right]\Gamma(1+2\,\xi) - \left[\frac{\Psi(2)}{g_{\sigma i}\,\xi^2} + \frac{1}{g_{\sigma i}\,\xi^3}\right] \Bigg)
\end{aligned}
$$

Appendix G

GEV block maxima model with constant scale parameter and smoothed location parameter

For the non-stationary GEV block maxima model with constant scale parameter and smoothed location parameter, let $\mu = g_\mu(t_i) = g_{\mu i}$, so that log-likelihood function is given by

$$l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; 1) = -\sum_{i=1}^{N}\left(\left[1+\xi\Big(\frac{x_i - g_{\mu i}}{\sigma}\Big)\right]^{-\frac{1}{\xi}} + \Big(\frac{1}{\xi}+1\Big)\ln\left[1+\xi\Big(\frac{x_i - g_{\mu i}}{\sigma}\Big)\right] + \ln(\sigma)\right)$$

and the penalized log-likelihood by

$$\begin{aligned} l^{PLL}(\mathbf{g}_\mu, \sigma, \xi; 1) &= l(\mathbf{g}_\mu, \sigma, \xi, 1) - \frac{1}{2}\lambda \int g''_\mu(t)^2 dt \\ &= l(\mathbf{g}_\mu, \sigma, \xi, 1) - \frac{1}{2}\lambda \mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu \end{aligned}$$

The first and second derivatives of the log-likelihood function are given by

$$\begin{aligned} \frac{\partial l(\mathbf{g}_\mu, \sigma, \xi; 1)}{\partial g_{\mu i}} &= -\frac{1}{\sigma}\left[1+\xi\Big(\frac{x_i - g_{\mu i}}{\sigma}\Big)\right]^{-\left(\frac{1}{\xi}+1\right)} + \frac{1}{\sigma}(1+\xi)\left[1+\xi\Big(\frac{x_i - g_{\mu i}}{\sigma}\Big)\right]^{-1} \\ \frac{\partial^2 l(\mathbf{g}_\mu, \sigma, \xi; 1)}{\partial g_{\mu i}^2} &= -\Big(\frac{1+\xi}{\sigma^2}\Big)\left[1+\xi\Big(\frac{x_i - g_{\mu i}}{\sigma}\Big)\right]^{-\left(\frac{1}{\xi}+2\right)} + \frac{\xi}{\sigma^2}(1+\xi)\left[1+\xi\Big(\frac{x_i - g_{\mu i}}{\sigma}\Big)\right]^{-2} \end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\sigma\partial g_{\mu i}} &= \frac{1}{\sigma^2}\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} - \left(\frac{1+\xi}{\sigma^2}\right)\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-1} \\
&\quad -\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\left(\frac{1+\xi}{\sigma^2}\right)\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad +\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\left(\frac{\xi\big(1+\xi\big)}{\sigma^2}\right)\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-2} \\
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\xi\partial g_{\mu i}} &= -\frac{1}{\sigma\xi^2}\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \ln\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big] \\
&\quad +\frac{1}{\sigma}\left(\frac{1}{\xi}+1\right)\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad +\frac{1}{\sigma}\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-1} - \left(\frac{1+\xi}{\sigma}\right)\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-2} \\
\frac{\partial l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\sigma} &= -\sum_{i=1}^{N}\frac{1}{\sigma}\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \\
&\quad +\sum_{i=1}^{N}\left(\frac{1+\xi}{\sigma}\right)\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-1} \\
&\quad -\sum_{i=1}^{N}\frac{1}{\sigma} \\
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\sigma^2} &= \sum_{i=1}^{N}\frac{2}{\sigma^2}\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \\
&\quad -\sum_{i=1}^{N}\left(\frac{x_i-g_{\mu i}}{\sigma}\right)^2\left(\frac{1+\xi}{\sigma^2}\right)\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad -\sum_{i=1}^{N}\left(\frac{2(1+\xi)}{\sigma^2}\right)\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-1} \\
&\quad +\sum_{i=1}^{N}\left(\frac{\xi\,(1+\xi)}{\sigma^2}\right)\left(\frac{x_i-g_{\mu i}}{\sigma}\right)^2\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-2} \\
&\quad +\sum_{i=1}^{N}\frac{1}{\sigma^2} \\
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial g_{\mu i}\partial\sigma} &= \frac{1}{\sigma^2}\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \\
&\quad -\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\left(\frac{1+\xi}{\sigma^2}\right)\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad -\left(\frac{1+\xi}{\sigma^2}\right)\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-1} \\
&\quad +\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\left(\frac{\xi\big(1+\xi\big)}{\sigma^2}\right)\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-2}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\xi\partial\sigma} &= -\sum_{i=1}^{N}\left(\frac{1}{\sigma\xi^2}\right)\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]^{-\left(\frac{1}{\xi}+1\right)}\ln\Big(1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big)\\
&+\sum_{i=1}^{N}\frac{1}{\sigma}\left(\frac{1}{\xi}+1\right)\left(\frac{x_i-g_{\mu i}}{\sigma}\right)^2\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]^{-\left(\frac{1}{\xi}+2\right)}\\
&+\sum_{i=1}^{N}\frac{1}{\sigma}\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]^{-1}\\
&-\sum_{i=1}^{N}\left(\frac{1+\xi}{\sigma}\right)\left(\frac{x_i-g_{\mu i}}{\sigma}\right)^2\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]^{-2}\\
\frac{\partial l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\xi} &= -\sum_{i=1}^{N}\frac{1}{\xi^2}\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]^{-\frac{1}{\xi}}\ln\Big(1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big)\\
&+\sum_{i=1}^{N}\frac{1}{\xi}\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]^{-\left(\frac{1}{\xi}+1\right)}\\
&+\sum_{i=1}^{N}\frac{1}{\xi^2}\ln\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]\\
&-\sum_{i=1}^{N}\Big(\frac{1}{\xi}+1\Big)\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]^{-1}\\
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\xi^2} &= -\sum_{i=1}^{N}\frac{1}{\xi^4}\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]^{-\frac{1}{\xi}}\left(\ln\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]\right)^2\\
&+\sum_{i=1}^{N}\frac{2}{\xi^3}\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]^{-\left(\frac{1}{\xi}+1\right)}\ln\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]\\
&+\sum_{i=1}^{N}\frac{2}{\xi^3}\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]^{-\frac{1}{\xi}}\ln\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]\\
&-\sum_{i=1}^{N}\frac{2}{\xi^2}\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]^{-\left(\frac{1}{\xi}+1\right)}\\
&-\sum_{i=1}^{N}\left(\frac{1}{\xi^2}+\frac{1}{\xi}\right)\left(\frac{x_i-g_{\mu i}}{\sigma}\right)^2\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]^{-\left(\frac{1}{\xi}+2\right)}\\
&-\sum_{i=1}^{N}\frac{2}{\xi^3}\ln\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]\\
&+\sum_{i=1}^{N}\frac{2}{\xi^2}\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]^{-1}\\
&+\sum_{i=1}^{N}\Big(\frac{1}{\xi}+1\Big)\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)^2\left[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\right]^{-2}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial g_{\mu i}\partial\xi} \quad &= \quad -\frac{1}{\sigma\xi^2}\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)}\ \ln\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big] \\
&\quad +\frac{1}{\sigma}\left(\frac{1}{\xi}+1\right)\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad +\frac{1}{\sigma}\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-1}-\left(\frac{1+\xi}{\sigma}\right)\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-2} \\
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\sigma\partial\xi} \quad &= \quad -\sum_{i=1}^{N}\left(\frac{1}{\sigma\xi^2}\right)\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)}\ \ln\Big(1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big) \\
&\quad +\sum_{i=1}^{N}\frac{1}{\sigma}\left(\frac{1}{\xi}+1\right)\left(\frac{x_i-g_{\mu i}}{\sigma}\right)^2\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad +\sum_{i=1}^{N}\frac{1}{\sigma}\left(\frac{x_i-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-1} \\
&\quad -\sum_{i=1}^{N}\left(\frac{1+\xi}{\sigma}\right)\left(\frac{x_i-g_{\mu i}}{\sigma}\right)^2\Big[1+\xi\Big(\frac{x_i-g_{\mu i}}{\sigma}\Big)\Big]^{-2}
\end{aligned}
$$

and

$$
\forall i\neq j \qquad \frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial g_{\mu i}\partial g_{\mu j}} = 0
$$

The first and second derivatives of the roughness penalty are given by

$$
\begin{aligned}
\frac{\partial(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial\mathbf{g}_\mu} = -\lambda\mathbf{K}\mathbf{g}_\mu \quad &\Rightarrow \quad \frac{\partial(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial g_{\mu i}} = -\lambda\big(\mathbf{K}\mathbf{g}_\mu\big)_i \\
\frac{\partial^2(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial\mathbf{g}_\mu\partial\mathbf{g}_\mu^T} = -\lambda\mathbf{K} \quad &\Rightarrow \quad \frac{\partial^2(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial g_{\mu i}\partial g_{\mu j}} = -\lambda K_{ij}
\end{aligned}
$$

With the expectations of the second derivative of the log-likelihood function given by

$$
\begin{aligned}
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial g_{\mu i}^2}\right] \quad &= \quad \left(\frac{1+\xi}{\sigma^2}\right)\Gamma(2+2\,\xi)-\left(\frac{\xi(1+\xi)}{\sigma^2}\right)\Gamma(1+2\,\xi) \\
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\sigma\partial g_{\mu i}}\right] \quad &= \quad \left(\frac{1}{\xi\,\sigma^2}\right)\Gamma(2+\ \xi)-\left(\frac{1+\xi}{\xi\,\sigma^2}\right)\Gamma(2+2\,\xi)+\left(\frac{1+\xi}{\sigma^2}\right)\Gamma(1+2\,\xi) \\
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\xi\partial g_{\mu i}}\right] \quad &= \quad -\left(\frac{\Psi(2+\xi)}{\sigma\xi}+\frac{1+\xi}{\sigma\,\xi^2}\right)\Gamma(2+\xi)+\left(\frac{1+\xi}{\sigma\,\xi^2}\right)\Gamma(2+2\,\xi) \\
&\quad +\left(\frac{1}{\sigma\,\xi}\right)\Gamma(1+\xi)-\left(\frac{1+\xi}{\sigma\,\xi}\right)\Gamma(1+2\,\xi) \\
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\sigma^2}\right] \quad &= \quad N\Bigg(-\left(\frac{2}{\sigma^2\,\xi^2}\right)\Gamma(2+\xi)+\left(\frac{1+\xi}{\sigma^2\,\xi^2}\right)\Gamma(2+2\,\xi) \\
&\qquad -\left(\frac{1+\xi}{\sigma^2\,\xi}\right)\Gamma(1+2\,\xi)-\frac{1}{\sigma^2\,\xi^2}\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial g_{\mu i}\partial\sigma}\right] &= \left(\frac{1}{\xi\,\sigma^2}\right)\Gamma(2+\,\xi)-\left(\frac{1+\xi}{\xi\,\sigma^2}\right)\Gamma(2+2\,\xi)+\left(\frac{1+\xi}{\sigma^2}\right)\Gamma(1+2\,\xi)\\
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\xi\partial\sigma}\right] &= N\Bigg(\left[\frac{\Psi(2+\xi)}{\sigma\,\xi^2}+\frac{2(1+\xi)}{\sigma\,\xi^3}\right]\Gamma(2+\xi)-\left[\frac{1+\xi}{\sigma\,\xi^3}\right]\Gamma(2+2\,\xi)\\
&\quad -\left[\frac{2+\xi}{\sigma\,\xi^2}\right]\Gamma(1+\xi)+\left[\frac{1+\xi}{\sigma\,\xi^2}\right]\Gamma(1+2\,\xi)-\left[\frac{\Psi(2)}{\sigma\,\xi^2}+\frac{1}{\sigma\,\xi^3}\right]\Bigg)\\
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\xi^2}\right] &= N\Bigg(\left[-\frac{2}{\xi^3}\Psi(2+\xi)-2\left(\frac{1+2\,\xi}{\xi^4}\right)\right]\Gamma(2+\xi)+\left(\frac{1+\xi}{\xi^4}\right)\Gamma(2+2\,\xi)\\
&\quad +2\left(\frac{2+\xi}{\xi^3}\right)\Gamma(1+\xi)-\left(\frac{1+\xi}{\xi^3}\right)\Gamma(1+2\,\xi)+\frac{1}{\xi^2}\left[\Psi'(2)+[\Psi(2)]^2\right]\\
&\quad +2\left(\frac{1+\xi}{\xi^3}\right)\Psi(2)-\left(\frac{2}{\xi^2}\right)\Psi(1)+\frac{1-\xi^2}{\xi^4}\Bigg)\\
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial g_{\mu i}\partial\xi}\right] &= -\left(\frac{\Psi(2+\xi)}{\sigma\xi}+\frac{1+\xi}{\sigma\,\xi^2}\right)\Gamma(2+\xi)+\left(\frac{1+\xi}{\sigma\,\xi^2}\right)\Gamma(2+2\,\xi)\\
&\quad +\left(\frac{1}{\sigma\,\xi}\right)\Gamma(1+\xi)-\left(\frac{1+\xi}{\sigma\,\xi}\right)\Gamma(1+2\,\xi)\\
E\left[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\sigma\partial\xi}\right] &= \sum_{i=1}^{N}\Bigg(\left[\frac{\Psi(2+\xi)}{\sigma\,\xi^2}+\frac{2(1+\xi)}{\sigma\,\xi^3}\right]\Gamma(2+\xi)-\left[\frac{1+\xi}{\sigma\,\xi^3}\right]\Gamma(2+2\,\xi)\\
&\quad -\left[\frac{2+\xi}{\sigma\,\xi^2}\right]\Gamma(1+\xi)+\left[\frac{1+\xi}{\sigma\,\xi^2}\right]\Gamma(1+2\,\xi)-\left[\frac{\Psi(2)}{\sigma\,\xi^2}+\frac{1}{\sigma\,\xi^3}\right]\Bigg)
\end{aligned}
$$

Define

$$
\begin{aligned}
u_\mu &= \frac{\partial l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\mathbf{g}_\mu} & u_\sigma &= \frac{\partial l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\sigma}\\
u_\xi &= \frac{\partial l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\xi} & W_\xi &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\xi^2}\right)\\
W_\mu &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\mathbf{g}_\mu\partial\mathbf{g}_\mu^T}\right) & W_\sigma &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\sigma^2}\right)\\
W_{\mu\sigma} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\mathbf{g}_\mu\partial\sigma}\right) & W_{\sigma\mu} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\sigma\partial\mathbf{g}_\mu^T}\right)\\
W_{\mu\xi} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\mathbf{g}_\mu\partial\xi}\right) & W_{\xi\mu} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\xi\partial\mathbf{g}_\mu^T}\right)\\
W_{\xi\sigma} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\xi\partial\sigma}\right) & W_{\sigma\xi} &= E\left(-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;1)}{\partial\sigma\partial\xi}\right)
\end{aligned}
$$

The Fisher scoring equation can therefore be written as

$$\begin{pmatrix} W_\mu + \lambda K & W_{\mu\sigma} & W_{\mu\xi} \\ W_{\sigma\mu} & W_\sigma & W_{\sigma\xi} \\ W_{\xi\mu} & W_{\xi\sigma} & W_\xi \end{pmatrix} \begin{pmatrix} g_\mu^{new} - g_\mu \\ \sigma^{new} - \sigma \\ \xi^{new} - \xi \end{pmatrix} = \begin{pmatrix} u_\mu - \lambda K g_\mu \\ u_\sigma \\ u_\xi \end{pmatrix}$$

This implies that

$$g_\mu^{new} = S_1(y_1 - W_\mu^{-1} W_{\mu\sigma}\sigma^{new} - W_\mu^{-1} W_{\mu\xi}\xi^{new})$$
$$\sigma^{new} = S_2(y_2 - W_\sigma^{-1} W_{\sigma\mu} g_\mu^{new} - W_\sigma^{-1} W_{\sigma\xi}\xi^{new})$$
$$\xi^{new} = S_3(y_3 - W_\xi^{-1} W_{\xi\mu} g_\mu^{new} - W_\xi^{-1} W_{\xi\sigma}\sigma^{new})$$

where

$$\begin{aligned} S_1 &= (W_\mu + \lambda K)^{-1} W_\mu \\ S_2 &= (W_\sigma)^{-1} W_\sigma = I \\ S_3 &= (W_\xi)^{-1} W_\xi = I \\ y_1 &= g_\mu + W_\mu^{-1} u_\mu + W_\mu^{-1} W_{\mu\sigma}\sigma + W_\mu^{-1} W_{\mu\xi}\xi \\ y_2 &= \sigma + W_\sigma^{-1} u_\sigma + W_\sigma^{-1} W_{\sigma\mu} g_\mu + W_\sigma^{-1} W_{\sigma\xi}\xi \\ y_3 &= \xi + W_\xi^{-1} u_\xi + W_\xi^{-1} W_{\xi\mu} g_\mu + W_\xi^{-1} W_{\xi\sigma}\sigma \end{aligned}$$

Appendix H

r largest order statistics model with $\xi \neq 0$ and smoothed location and scale parameters

For the non-stationary distribution of the r largest order statistics with $\xi \neq 0$ and smoothed location and scale parameters, let $\mu = g_\mu(t_i) = g_{\mu i}$ and $\sigma = g_\sigma(t_i) = g_{\sigma i}$, so that log-likelihood function is given by

$$l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r) = \sum_{i=1}^{N} \left[-\Big[1 + \xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\frac{1}{\xi}} - r\ln(g_{\sigma i}) - \sum_{k=1}^{r}\Big(\frac{1}{\xi} + 1\Big)\ln\Big[1 + \xi\Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}}\Big)\Big]\right]$$

and the penalized log-likelihood by

$$\begin{aligned} l^{PLL}(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r) &= l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r) - \frac{1}{2}\lambda \int g''_\mu(t)^2 dt - \frac{1}{2}\alpha \int g''_\sigma(t)^2 dt \\ &= l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r) - \frac{1}{2}\lambda \mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu - \frac{1}{2}\alpha \mathbf{g}_\sigma^T \mathbf{K} \mathbf{g}_\sigma \end{aligned}$$

The first and second derivatives of the log-likelihood function are given by

$$\begin{aligned} \frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial g_{\mu i}} &= -\frac{1}{g_{\sigma i}}\Big[1 + \xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} + \frac{1+\xi}{g_{\sigma i}}\sum_{k=1}^{r}\Big[1 + \xi\Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-1} \\ \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial g_{\mu i}^2} &= -\frac{1+\xi}{g_{\sigma i}^2}\Big[1 + \xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} + \frac{\xi(1+\xi)}{g_{\sigma i}^2}\sum_{k=1}^{r}\Big[1 + \xi\Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-2} \end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;r)}{\partial g_{\sigma i}\partial g_{\mu i}} &= \frac{1}{g_{\sigma i}^2}\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \\
&\quad -\frac{1+\xi}{g_{\sigma i}^2}\left(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\right)\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad -\frac{(1+\xi)}{g_{\sigma i}^2}\sum_{k=1}^{r}\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-1} \\
&\quad +\frac{\xi(1+\xi)}{g_{\sigma i}^2}\sum_{k=1}^{r}\left(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\right)\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-2} \\
\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;r)}{\partial \xi\partial g_{\mu i}} &= -\frac{1}{g_{\sigma i}\xi^2}\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)}\ln\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big] \\
&\quad +\frac{1}{g_{\sigma i}}\left(\frac{1}{\xi}+1\right)\left(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\right)\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad +\frac{1}{g_{\sigma i}}\sum_{k=1}^{r}\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-1} \\
&\quad -\frac{1+\xi}{g_{\sigma i}}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-2} \\
\frac{\partial l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;r)}{\partial g_{\sigma i}} &= -\frac{1}{g_{\sigma i}}\left(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\right)\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \\
&\quad -\frac{r}{g_{\sigma i}}+\frac{1+\xi}{g_{\sigma i}}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-1} \\
\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;r)}{\partial g_{\sigma i}^2} &= \frac{r}{g_{\sigma i}^2}+\frac{2}{g_{\sigma i}^2}\left(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\right)\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \\
&\quad -\frac{1+\xi}{g_{\sigma i}^2}\left(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\right)^2\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad -\frac{2(1+\xi)}{g_{\sigma i}^2}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-1} \\
&\quad +\frac{\xi(1+\xi)}{g_{\sigma i}^2}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)^2\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-2} \\
\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;r)}{\partial g_{\mu i}\partial g_{\sigma i}} &= \frac{1}{g_{\sigma i}^2}\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)}-\frac{(1+\xi)}{g_{\sigma i}^2}\sum_{k=1}^{r}\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-1} \\
&\quad -\frac{1+\xi}{g_{\sigma i}^2}\left(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\right)\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad +\frac{\xi(1+\xi)}{g_{\sigma i}^2}\sum_{k=1}^{r}\left(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\right)\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-2}
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial \xi \partial g_{\sigma i}} &= -\frac{1}{g_{\sigma i}\xi^2}\left(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\right)\left[1+\xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-\left(\frac{1}{\xi}+1\right)} \ln\left[1+\xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\right] \\
&\quad + \frac{1}{g_{\sigma i}}\left(\frac{1}{\xi}+1\right)\left(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\right)^2\left[1+\xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad + \frac{1}{g_{\sigma i}}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}}\Big)\left[1+\xi\Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-1} \\
&\quad - \frac{1+\xi}{g_{\sigma i}}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}}\Big)^2\left[1+\xi\Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-2} \\
\frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial \xi} &= \sum_{i=1}^{N}\Bigg[-\frac{1}{\xi^2}\left[1+\xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-\frac{1}{\xi}} \ln\left[1+\xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\right] \\
&\quad + \frac{1}{\xi}\left(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\right)\left[1+\xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-\left(\frac{1}{\xi}+1\right)} \\
&\quad + \frac{1}{\xi^2}\sum_{k=1}^{r}\ln\left[1+\xi\Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}}\Big)\right] \\
&\quad - \left(\frac{1}{\xi}+1\right)\sum_{k=1}^{r}\Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}}\Big)\left[1+\xi\Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-1}\Bigg] \\
\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial \xi^2} &= \sum_{i=1}^{N}\Bigg[-\frac{1}{\xi^4}\left[1+\xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-\frac{1}{\xi}}\left(\ln\left[1+\xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\right]\right)^2 \\
&\quad + \frac{2}{\xi^3}\left(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\right)\left[1+\xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-\left(\frac{1}{\xi}+1\right)} \ln\left[1+\xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\right] \\
&\quad - \left(\frac{1}{\xi^2}+\frac{1}{\xi}\right)\left(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\right)^2\left[1+\xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad + \frac{2}{\xi^3}\ln\left[1+\xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\right]\left[1+\xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-\frac{1}{\xi}} \\
&\quad - \frac{2}{\xi^2}\left(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\right)\left[1+\xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-\left(\frac{1}{\xi}+1\right)} \\
&\quad - \frac{2}{\xi^3}\sum_{k=1}^{r}\ln\left[1+\xi\Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}}\Big)\right] \\
&\quad + \frac{2}{\xi^2}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}}\Big)\left[1+\xi\Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-1} \\
&\quad + \left(\frac{1}{\xi}+1\right)\sum_{k=1}^{r}\Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}}\Big)^2\left[1+\xi\Big(\frac{z_i^{(k)} - g_{\mu i}}{g_{\sigma i}}\Big)\right]^{-2}\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial g_{\mu i}\partial \xi} &= -\frac{1}{g_{\sigma i}\xi^2}\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \ln\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big] \\
&+\frac{1}{g_{\sigma i}}\left(\frac{1}{\xi}+1\right)\left(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\right)\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&+\frac{1}{g_{\sigma i}}\sum_{k=1}^{r}\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-1} \\
&-\frac{1+\xi}{g_{\sigma i}}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-2} \\
\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial g_{\sigma i}\partial \xi} &= -\frac{1}{g_{\sigma i}\xi^2}\left(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\right)\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \ln\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big] \\
&+\frac{1}{g_{\sigma i}}\left(\frac{1}{\xi}+1\right)\left(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\right)^2\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&+\frac{1}{g_{\sigma i}}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-1} \\
&-\frac{1+\xi}{g_{\sigma i}}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)^2\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{g_{\sigma i}}\Big)\Big]^{-2}
\end{aligned}
$$

and

$$\forall i \neq j \quad \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial g_{\mu i}\partial g_{\mu j}} = 0 \qquad , \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial g_{\mu i}\partial g_{\sigma j}} = 0$$

with the first and second derivatives of the roughness penalties given by

$$
\begin{aligned}
\frac{\partial(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial \mathbf{g}_\mu} &= -\lambda\mathbf{K}\mathbf{g}_\mu \quad\Rightarrow\quad \frac{\partial(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial g_{\mu i}} = -\lambda\big(\mathbf{K}\mathbf{g}_\mu\big)_i \\
\frac{\partial^2(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial \mathbf{g}_\mu\partial \mathbf{g}_\mu^T} &= -\lambda\mathbf{K} \quad\Rightarrow\quad \frac{\partial^2(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial g_{\mu i}\partial g_{\mu j}} = -\lambda K_{ij} \\
\frac{\partial(-\frac{1}{2}\alpha\mathbf{g}_\sigma^T\mathbf{K}\mathbf{g}_\sigma)}{\partial \mathbf{g}_\sigma} &= -\alpha\mathbf{K}\mathbf{g}_\sigma \quad\Rightarrow\quad \frac{\partial(-\frac{1}{2}\alpha\mathbf{g}_\sigma^T\mathbf{K}\mathbf{g}_\sigma)}{\partial g_{\sigma i}} = -\alpha\big(\mathbf{K}\mathbf{g}_\sigma\big)_i \\
\frac{\partial^2(-\frac{1}{2}\alpha\mathbf{g}_\sigma^T\mathbf{K}\mathbf{g}_\sigma)}{\partial \mathbf{g}_\sigma\partial \mathbf{g}_\sigma^T} &= -\alpha\mathbf{K} \quad\Rightarrow\quad \frac{\partial^2(-\frac{1}{2}\alpha\mathbf{g}_\sigma^T\mathbf{K}\mathbf{g}_\sigma)}{\partial g_{\sigma i}\partial g_{\sigma j}} = -\alpha K_{ij}
\end{aligned}
$$

The expectations of the second derivative of the log-likelihood are given by

$$
\begin{aligned}
E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial g_{\mu i}^2}\Big] &= \frac{(1+\xi)^2}{g_{\sigma i}^2\,(1+2\xi)}\,\frac{\Gamma(r+2\xi+1)}{\Gamma(r)} \\
E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial g_{\sigma i}\partial g_{\mu i}}\Big] &= \frac{1}{g_{\sigma i}^2\,\xi\,(1+2\xi)}\left(\frac{\Gamma(r+\xi+1)}{\Gamma(r)}\,(1+2\xi)-\frac{\Gamma(r+2\xi+1)}{\Gamma(r)}\,(1+\xi)^2\right)
\end{aligned}
$$

$$E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;r)}{\partial\xi\partial g_{\mu i}}\Big] = \frac{1}{g_{\sigma i}\,\xi^2\,(1+2\xi)}\left[\frac{\Gamma(r+2\xi+1)}{\Gamma(r)}\,(1+\xi)^2 -(1+2\xi)\,\frac{\Gamma(r+\xi+1)}{\Gamma(r)}\left(\xi\Psi(r+\xi+1)+\frac{1+\xi+\xi^2}{1+\xi}\right)\right]$$

$$E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial g_{\sigma i}^2}\Big] = \frac{1}{g_{\sigma i}^2\,\xi^2\,(1+2\xi)}\left((1+2\xi)\,r-2(1+2\xi)\,\frac{\Gamma(r+\xi+1)}{\Gamma(r)} +(1+\xi)^2\,\frac{\Gamma(r+2\xi+1)}{\Gamma(r)}\right)$$

$$E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;r)}{\partial g_{\mu i}\partial g_{\sigma i}}\Big] = \frac{1}{g_{\sigma i}^2\,\xi\,(1+2\xi)}\left(\frac{\Gamma(r+\xi+1)}{\Gamma(r)}\,(1+2\xi)-\frac{\Gamma(r+2\xi+1)}{\Gamma(r)}\,(1+\xi)^2\right)$$

$$E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;r)}{\partial g_{\mu i}\partial\xi}\Big] = \frac{1}{g_{\sigma i}\,\xi^2\,(1+2\xi)}\left[\frac{\Gamma(r+2\xi+1)}{\Gamma(r)}\,(1+\xi)^2 -(1+2\xi)\,\frac{\Gamma(r+\xi+1)}{\Gamma(r)}\left(\xi\Psi(r+\xi+1)+\frac{1+\xi+\xi^2}{1+\xi}\right)\right]$$

$$E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;r)}{\partial\xi\partial g_{\sigma i}}\Big] = \frac{1}{g_{\sigma i}\,\xi^3\,(1+2\xi)}\Bigg(-r\,(1+2\xi)\,\xi\,\Psi(r+1)-r(1+2\xi) -(1+\xi)^2\,\frac{\Gamma(r+2\xi+1)}{\Gamma(r)} +(1+2\xi)\,\frac{\Gamma(r+\xi+1)}{\Gamma(r)}\left[\xi\,\Psi(r+\xi+1)+\frac{1+(1+\xi)^2}{1+\xi}\right]\Bigg)$$

$$E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;r)}{\partial g_{\sigma i}\partial\xi}\Big] = \frac{1}{g_{\sigma i}\,\xi^3\,(1+2\xi)}\Bigg(-r\,(1+2\xi)\,\xi\,\Psi(r+1)-r(1+2\xi) -(1+\xi)^2\,\frac{\Gamma(r+2\xi+1)}{\Gamma(r)} +(1+2\xi)\,\frac{\Gamma(r+\xi+1)}{\Gamma(r)}\left[\xi\,\Psi(r+\xi+1)+\frac{1+(1+\xi)^2}{1+\xi}\right]\Bigg)$$

$$E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\mathbf{g}_\sigma,\xi;r)}{\partial\xi^2}\Big] = \frac{N}{\xi^4(1+2\xi)}\Bigg(2(1+2\xi)\,\frac{\Gamma(r+\xi+1)}{\Gamma(r)}\left[-\xi\,\Psi(r+\xi+1)-\frac{1+\xi+\xi^2}{1+\xi}\right] +(1+\xi)^2\,\frac{\Gamma(r+2\xi+1)}{\Gamma(r)} +r(1+2\xi)\left[1+2\xi\,\Psi(r+1)+\xi^2\left\{1+\Psi'(r+1)+\Big(\Psi(r+1)\Big)^2\right\}\right]\Bigg)$$

Define

$$u_\mu = \frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial \mathbf{g}_\mu} \qquad u_\sigma = \frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial \mathbf{g}_\sigma}$$

$$u_\xi = \frac{\partial l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi, r)}{\partial \xi} \qquad W_\xi = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi, r)}{\partial \xi^2}\right)$$

$$W_\mu = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial \mathbf{g}_\mu \partial \mathbf{g}_\mu^T}\right) \qquad W_\sigma = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial \mathbf{g}_\sigma \partial \mathbf{g}_\sigma^T}\right)$$

$$W_{\mu\sigma} = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial \mathbf{g}_\mu \partial \mathbf{g}_\sigma^T}\right) \qquad W_{\sigma\mu} = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial \mathbf{g}_\sigma \partial \mathbf{g}_\mu^T}\right)$$

$$W_{\mu\xi} = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial \mathbf{g}_\mu \partial \xi}\right) \qquad W_{\xi\mu} = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial \xi \partial \mathbf{g}_\mu^T}\right)$$

$$W_{\sigma\xi} = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial \mathbf{g}_\sigma \partial \xi}\right) \qquad W_{\xi\sigma} = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \mathbf{g}_\sigma, \xi; r)}{\partial \xi \partial \mathbf{g}_\sigma^T}\right)$$

The Fisher scoring equation, for the non-stationary GEV r largest model with smoothed location and scale parameters and constant shape parameter, can be written as

$$\begin{pmatrix} W_\mu + \lambda K & W_{\mu\sigma} & W_{\mu\xi} \\ W_{\sigma\mu} & W_\sigma + \alpha K & W_{\sigma\xi} \\ W_{\xi\mu} & W_{\xi\sigma} & W_\xi \end{pmatrix} \begin{pmatrix} g_\mu^{new} - g_\mu \\ g_\sigma^{new} - g_\sigma \\ \xi^{new} - \xi \end{pmatrix} = \begin{pmatrix} u_\mu - \lambda K g_\mu \\ u_\sigma - \alpha K g_\sigma \\ u_\xi \end{pmatrix}$$

which implies that

$$g_\mu^{new} = S_1(y_1 - W_\mu^{-1} W_{\mu\sigma} g_\sigma^{new} - W_\mu^{-1} W_{\mu\xi} \xi^{new})$$
$$g_\sigma^{new} = S_2(y_2 - W_\sigma^{-1} W_{\sigma\mu} g_\mu^{new} - W_\sigma^{-1} W_{\sigma\xi} \xi^{new})$$
$$\xi^{new} = S_3(y_3 - W_\xi^{-1} W_{\xi\mu} g_\mu^{new} - W_\xi^{-1} W_{\xi\sigma} g_\sigma^{new})$$

where

$$S_1 = (W_\mu + \lambda K)^{-1} W_\mu$$
$$S_2 = (W_\sigma + \alpha K)^{-1} W_\sigma$$
$$S_3 = (W_\xi)^{-1} W_\xi = I$$
$$y_1 = g_\mu + W_\mu^{-1} u_\mu + W_\mu^{-1} W_{\mu\sigma} g_\sigma + W_\mu^{-1} W_{\mu\xi} \xi$$
$$y_2 = g_\sigma + W_\sigma^{-1} u_\sigma + W_\sigma^{-1} W_{\sigma\mu} g_\mu + W_\sigma^{-1} W_{\sigma\xi} \xi$$
$$y_3 = \xi + W_\xi^{-1} u_\xi + W_\xi^{-1} W_{\xi\mu} g_\mu + W_\xi^{-1} W_{\xi\sigma} g_\sigma$$

Appendix I

r largest order statistics model with $\xi \neq 0$ and a constant scale parameter and smoothed location parameter

For the non-stationary distribution of the r largest order statistics model with $\xi \neq 0$ and with a constant scale and shape parameters and smoothed location parameter, let $\mu = g_\mu(t_i) = g_{\mu i}$, so that log-likelihood function is given by

$$l(\mathbf{g}_\mu, \sigma, \xi; r) = \sum_{i=1}^{N} \left[-\left[1 + \xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{\sigma}\Big)\right]^{-\frac{1}{\xi}} - r\ln(\sigma) - \sum_{k=1}^{r}\Big(\frac{1}{\xi}+1\Big)\ln\left[1 + \xi\Big(\frac{z_i^{(k)} - g_{\mu i}}{\sigma}\Big)\right]\right]$$

and the penalized log-likelihood by

$$\begin{aligned} l^{PLL}(\mathbf{g}_\mu, \sigma, \xi; r) &= l(\mathbf{g}_\mu, \sigma, \xi; r) - \frac{1}{2}\lambda \int g_\mu''(t)^2 dt \\ &= l(\mathbf{g}_\mu, \sigma, \xi; r) - \frac{1}{2}\lambda \mathbf{g}_\mu^T \mathbf{K} \mathbf{g}_\mu \end{aligned}$$

The first and second derivatives of the log-likelihood function are given by

$$\begin{aligned} \frac{\partial l(\mathbf{g}_\mu, \sigma, \xi; r)}{\partial g_{\mu i}} &= -\frac{1}{\sigma}\left[1 + \xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{\sigma}\Big)\right]^{-\left(\frac{1}{\xi}+1\right)} + \frac{1+\xi}{\sigma}\sum_{k=1}^{r}\left[1 + \xi\Big(\frac{z_i^{(k)} - g_{\mu i}}{\sigma}\Big)\right]^{-1} \\ \frac{\partial^2 l(\mathbf{g}_\mu, \sigma, \xi; r)}{\partial g_{\mu i}^2} &= -\frac{1+\xi}{\sigma^2}\left[1 + \xi\Big(\frac{z_i^{(r)} - g_{\mu i}}{\sigma}\Big)\right]^{-\left(\frac{1}{\xi}+2\right)} + \frac{\xi(1+\xi)}{\sigma^2}\sum_{k=1}^{r}\left[1 + \xi\Big(\frac{z_i^{(k)} - g_{\mu i}}{\sigma}\Big)\right]^{-2} \end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial\sigma\partial g_{\mu i}} &= \frac{1}{\sigma^2}\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \\
&-\frac{1+\xi}{\sigma^2}\left(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&-\frac{(1+\xi)}{\sigma^2}\sum_{k=1}^{r}\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big]^{-1} \\
&+\frac{\xi(1+\xi)}{\sigma^2}\sum_{k=1}^{r}\left(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big]^{-2} \\
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial\xi\partial g_{\mu i}} &= -\frac{1}{\sigma\xi^2}\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)}\ln\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big] \\
&+\frac{1}{\sigma}\left(\frac{1}{\xi}+1\right)\left(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&+\frac{1}{\sigma}\sum_{k=1}^{r}\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big]^{-1} \\
&-\frac{1+\xi}{\sigma}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big]^{-2} \\
\frac{\partial l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial\sigma} &= \sum_{i=1}^{N}\Bigg[-\frac{1}{\sigma}\left(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \\
&\qquad -\frac{r}{\sigma}+\frac{1+\xi}{\sigma}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big]^{-1}\Bigg] \\
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial\sigma^2} &= \sum_{i=1}^{N}\Bigg[\frac{r}{\sigma^2}+\frac{2}{\sigma^2}\left(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)} \\
&\qquad -\frac{1+\xi}{\sigma^2}\left(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\right)^2\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&\qquad -\frac{2(1+\xi)}{\sigma^2}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big]^{-1} \\
&\qquad +\frac{\xi(1+\xi)}{\sigma^2}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)^2\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big]^{-2}\Bigg] \\
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial g_{\mu i}\partial\sigma} &= \frac{1}{\sigma^2}\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)}-\frac{(1+\xi)}{\sigma^2}\sum_{k=1}^{r}\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big]^{-1} \\
&-\frac{1+\xi}{\sigma^2}\left(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&+\frac{\xi(1+\xi)}{\sigma^2}\sum_{k=1}^{r}\left(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big]^{-2}
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial\xi\partial\sigma} = \sum_{i=1}^{N}\Bigg[& -\frac{1}{\sigma\xi^2}\left(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\right)\left[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\right]^{-\left(\frac{1}{\xi}+1\right)}\ln\left[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\right] \\
& +\frac{1}{\sigma}\left(\frac{1}{\xi}+1\right)\left(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\right)^2\left[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\right]^{-\left(\frac{1}{\xi}+2\right)} \\
& +\frac{1}{\sigma}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\left[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\right]^{-1} \\
& -\frac{1+\xi}{\sigma}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)^2\left[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\right]^{-2}\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial\xi} = \sum_{i=1}^{N}\Bigg[& -\frac{1}{\xi^2}\left[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\right]^{-\frac{1}{\xi}}\ln\left[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\right] \\
& +\frac{1}{\xi}\left(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\right)\left[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\right]^{-\left(\frac{1}{\xi}+1\right)} \\
& +\frac{1}{\xi^2}\sum_{k=1}^{r}\ln\left[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\right] \\
& -\left(\frac{1}{\xi}+1\right)\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\left[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\right]^{-1}\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial\xi^2} = \sum_{i=1}^{N}\Bigg[& -\frac{1}{\xi^4}\left[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\right]^{-\frac{1}{\xi}}\left(\ln\left[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\right]\right)^2 \\
& +\frac{2}{\xi^3}\left(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\right)\left[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\right]^{-\left(\frac{1}{\xi}+1\right)}\ln\left[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\right] \\
& -\left(\frac{1}{\xi^2}+\frac{1}{\xi}\right)\left(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\right)^2\left[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\right]^{-\left(\frac{1}{\xi}+2\right)} \\
& +\frac{2}{\xi^3}\ln\left[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\right]\left[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\right]^{-\frac{1}{\xi}} \\
& -\frac{2}{\xi^2}\left(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\right)\left[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\right]^{-\left(\frac{1}{\xi}+1\right)} \\
& -\frac{2}{\xi^3}\sum_{k=1}^{r}\ln\left[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\right] \\
& +\frac{2}{\xi^2}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\left[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\right]^{-1} \\
& +\left(\frac{1}{\xi}+1\right)\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)^2\left[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\right]^{-2}\Bigg]
\end{aligned}
$$

$$\begin{aligned}
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial g_{\mu i}\partial\xi} &= -\frac{1}{\sigma\xi^2}\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)}\ \ln\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big] \\
&+\frac{1}{\sigma}\left(\frac{1}{\xi}+1\right)\left(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&+\frac{1}{\sigma}\sum_{k=1}^{r}\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big]^{-1} \\
&-\frac{1+\xi}{\sigma}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big]^{-2} \\
\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial\sigma\partial\xi} &= \sum_{i=1}^{N}\Bigg[-\frac{1}{\sigma\xi^2}\left(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\right)\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+1\right)}\ \ln\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big] \\
&\quad+\frac{1}{\sigma}\left(\frac{1}{\xi}+1\right)\left(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\right)^2\Big[1+\xi\Big(\frac{z_i^{(r)}-g_{\mu i}}{\sigma}\Big)\Big]^{-\left(\frac{1}{\xi}+2\right)} \\
&\quad+\frac{1}{\sigma}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big]^{-1} \\
&\quad-\frac{1+\xi}{\sigma}\sum_{k=1}^{r}\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)^2\Big[1+\xi\Big(\frac{z_i^{(k)}-g_{\mu i}}{\sigma}\Big)\Big]^{-2}\ \Bigg]
\end{aligned}$$

and

$$\forall i\neq j\quad \frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial g_{\mu i}\partial g_{\mu j}}=0$$

The first and second derivatives of the roughness penalty are given by

$$\begin{aligned}
\frac{\partial(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial\mathbf{g}_\mu} &= -\lambda\mathbf{K}\mathbf{g}_\mu \quad\Rightarrow\quad \frac{\partial(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial g_{\mu i}}=-\lambda\big(\mathbf{K}\mathbf{g}_\mu\big)_i \\
\frac{\partial^2(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial\mathbf{g}_\mu\partial\mathbf{g}_\mu^T} &= -\lambda\mathbf{K} \quad\Rightarrow\quad \frac{\partial^2(-\frac{1}{2}\lambda\mathbf{g}_\mu^T\mathbf{K}\mathbf{g}_\mu)}{\partial g_{\mu i}\partial g_{\mu j}}=-\lambda K_{ij}
\end{aligned}$$

The expectations of the second derivative of the log-likelihood function given by

$$\begin{aligned}
E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial g_{\mu i}^2}\Big] &= \frac{(1+\xi)^2}{\sigma^2\,(1+2\xi)}\,\frac{\Gamma(r+2\xi+1)}{\Gamma(r)} \\
E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial\sigma\partial g_{\mu i}}\Big] &= \frac{1}{\sigma^2\,\xi\,(1+2\xi)}\left(\frac{\Gamma(r+\xi+1)}{\Gamma(r)}\,(1+2\xi)-\frac{\Gamma(r+2\xi+1)}{\Gamma(r)}\,(1+\xi)^2\right)
\end{aligned}$$

$$
\begin{aligned}
E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial\xi\partial g_{\mu i}}\Big] &= \frac{1}{\sigma\,\xi^2\,(1+2\xi)}\Bigg[\frac{\Gamma(r+2\xi+1)}{\Gamma(r)}\,(1+\xi)^2 \\
&\quad -(1+2\xi)\,\frac{\Gamma(r+\xi+1)}{\Gamma(r)}\left(\xi\Psi(r+\xi+1)+\frac{1+\xi+\xi^2}{1+\xi}\right)\Bigg] \\
E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial\sigma^2}\Big] &= \frac{N}{\sigma^2\,\xi^2\,(1+2\xi)}\Bigg((1+2\xi)\,r-2(1+2\xi)\,\frac{\Gamma(r+\xi+1)}{\Gamma(r)} \\
&\quad +(1+\xi)^2\,\frac{\Gamma(r+2\xi+1)}{\Gamma(r)}\Bigg) \\
E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial g_{\mu i}\partial\sigma}\Big] &= \frac{1}{\sigma^2\,\xi\,(1+2\xi)}\left(\frac{\Gamma(r+\xi+1)}{\Gamma(r)}\,(1+2\xi)-\frac{\Gamma(r+2\xi+1)}{\Gamma(r)}\,(1+\xi)^2\right) \\
E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial g_{\mu i}\partial\xi}\Big] &= \frac{1}{\sigma\,\xi^2\,(1+2\xi)}\Bigg[\frac{\Gamma(r+2\xi+1)}{\Gamma(r)}\,(1+\xi)^2 \\
&\quad -(1+2\xi)\,\frac{\Gamma(r+\xi+1)}{\Gamma(r)}\left(\xi\Psi(r+\xi+1)+\frac{1+\xi+\xi^2}{1+\xi}\right)\Bigg] \\
E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial\xi\partial\sigma}\Big] &= \frac{N}{\sigma\,\xi^3\,(1+2\xi)}\Bigg(-r\,(1+2\xi)\,\xi\,\Psi(r+1)-r(1+2\xi) \\
&\quad -(1+\xi)^2\,\frac{\Gamma(r+2\xi+1)}{\Gamma(r)} \\
&\quad +(1+2\xi)\,\frac{\Gamma(r+\xi+1)}{\Gamma(r)}\left[\xi\,\Psi(r+\xi+1)+\frac{1+(1+\xi)^2}{1+\xi}\right]\Bigg) \\
E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial\sigma\partial\xi}\Big] &= \frac{N}{\sigma\,\xi^3\,(1+2\xi)}\Bigg(-r\,(1+2\xi)\,\xi\,\Psi(r+1)-r(1+2\xi) \\
&\quad -(1+\xi)^2\,\frac{\Gamma(r+2\xi+1)}{\Gamma(r)} \\
&\quad +(1+2\xi)\,\frac{\Gamma(r+\xi+1)}{\Gamma(r)}\left[\xi\,\Psi(r+\xi+1)+\frac{1+(1+\xi)^2}{1+\xi}\right]\Bigg) \\
E\Big[-\frac{\partial^2 l(\mathbf{g}_\mu,\sigma,\xi;r)}{\partial\xi^2}\Big] &= \frac{N}{\xi^4(1+2\xi)}\Bigg(2(1+2\xi)\,\frac{\Gamma(r+\xi+1)}{\Gamma(r)}\left[-\xi\,\Psi(r+\xi+1)-\frac{1+\xi+\xi^2}{1+\xi}\right] \\
&\quad +(1+\xi)^2\,\frac{\Gamma(r+2\xi+1)}{\Gamma(r)} \\
&\quad +r(1+2\xi)\left[1+2\xi\,\Psi(r+1)+\xi^2\left\{1+\Psi'(r+1)+\Big(\Psi(r+1)\Big)^2\right\}\right]\Bigg)
\end{aligned}
$$

The equations from the Fisher scoring algorithm imply that

$$g_\mu^{new} = S_1(y_1 - W_\mu^{-1}W_{\mu\sigma}\sigma^{new} - W_\mu^{-1}W_{\mu\xi}\xi^{new})$$
$$\sigma^{new} = S_2(y_2 - W_\sigma^{-1}W_{\sigma\mu}g_\mu^{new} - W_\sigma^{-1}W_{\sigma\xi}\xi^{new})$$
$$\xi^{new} = S_3(y_3 - W_\xi^{-1}W_{\xi\mu}g_\mu^{new} - W_\xi^{-1}W_{\xi\sigma}\sigma^{new})$$

where

$$\begin{aligned}
S_1 &= (W_\mu + \lambda K)^{-1}W_\mu \\
S_2 &= (W_\sigma)^{-1}W_\sigma = I \\
S_3 &= (W_\xi)^{-1}W_\xi = I \\
y_1 &= g_\mu + W_\mu^{-1}u_\mu + W_\mu^{-1}W_{\mu\sigma}\sigma + W_\mu^{-1}W_{\mu\xi}\xi \\
y_2 &= \sigma + W_\sigma^{-1}u_\sigma + W_\sigma^{-1}W_{\sigma\mu}g_\mu + W_\sigma^{-1}W_{\sigma\xi}\xi \\
y_3 &= \xi + W_\xi^{-1}u_\xi + W_\xi^{-1}W_{\xi\mu}g_\mu + W_\xi^{-1}W_{\xi\sigma}\sigma
\end{aligned}$$

with

$$u_\mu = \frac{\partial l(\mathbf{g}_\mu, \sigma, \xi; r)}{\partial \mathbf{g}_\mu} \qquad u_\sigma = \frac{\partial l(\mathbf{g}_\mu, \sigma, \xi; r)}{\partial \sigma}$$

$$u_\xi = \frac{\partial l(\mathbf{g}_\mu, \sigma, \xi; r)}{\partial \xi} \qquad W_\xi = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \sigma, \xi; r)}{\partial \xi^2} \right)$$

$$W_\mu = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \sigma, \xi; r)}{\partial \mathbf{g}_\mu \partial \mathbf{g}_\mu^T} \right) \qquad W_\sigma = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \sigma, \xi; r)}{\partial \sigma^2} \right)$$

$$W_{\mu\sigma} = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \sigma, \xi; r)}{\partial \mathbf{g}_\mu \partial \sigma} \right) \qquad W_{\sigma\mu} = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \sigma, \xi; r)}{\partial \sigma \partial \mathbf{g}_\mu^T} \right)$$

$$W_{\mu\xi} = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \sigma, \xi; r)}{\partial \mathbf{g}_\mu \partial \xi} \right) \qquad W_{\xi\mu} = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \sigma, \xi; r)}{\partial \xi \partial \mathbf{g}_\mu^T} \right)$$

$$W_{\xi\sigma} = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \sigma, \xi; r)}{\partial \xi \partial \sigma} \right) \qquad W_{\sigma\xi} = E\left(- \frac{\partial^2 l(\mathbf{g}_\mu, \sigma, \xi; r)}{\partial \sigma \partial \xi} \right)$$

Bibliography

Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models, *Biometrika* **60**: 255–65.

Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2005). *Statistics of Extremes: Theory and Applications*, Wiley Series in Probability and Statistics.

Castillo, E., Hadi, A. S., Balakrishnan, N. and Sarabia, J. M. (2005). *Extreme Value and Related Models wiht Applications in Engineering and Science*, John Wiley & Sons, Inc., Hoboken, New Jersey.

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

Chavez-Demoulin, V. and Embrechts, P. (2004). Smooth extremal models in finance and insurance, *The Journal of Risk and Insurance* **71**: 183–199.

Coles, S. (2001). *An Introduction to Statistical Modelling of Extreme Values*, Springer.

David, H. and Nagaraja, H. N. (2003). *Order Statistics Third Edition*, Wiley Series in Probability and Statistics.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*, Springer.

Embrechts, P., Kluppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Proc. Cambridge Philos. Soc.* **24**: 180–190.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach*, Chapman & Hall.

Laurini, F. and Pauli, F. (2009). Smoothing sample extremes: The mixed model approach, *Computational Statistics adn Data Analysis* **53**: 3842 – 3854.

Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*, Springer-Verlag.

Pauli, F. and Coles, S. (2001). Penalized likelihood inference in extreme value analyses, *Journal of Applied Statistics* **28**: 547–560.

Rosen, O. and Cohen, A. (1996). Extreme percentile regression, *Proceedings of COMPSTAT 94* **70**: 200–214.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge University Press.

Smith, R. (1985). Maximum likelihood estimation in a class of non-regular cases., *Biometrika* **72**: 67–90.

Smith, R. L. (1986). Extreme value theory based on the r largest annual events, *Journal of Hydrology* **86**: 27–43.

Tawn, J. A. (1988). An extreme-value theory model for dependent observations, *Journal of Hydrology* **101**: 227–250.

Wand, M. P. and Padoan, S. A. (2008). Mixed model-based additive models for sample extremes, *Statistics and Probability Letters* **78**: 2850 – 2858.