Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Scuola di Dottorato in
Scienze Statistiche
Ciclo XX

Computational methods for complex problems in extreme value theory

Simone Padoan

Direttore: Prof. A. Salvan

Supervisore: Prof. S. Coles

Co-supervisori: Prof. M. Wand, Prof. F. Pauli, Dott. S. Sisson

31 Gennaio 2008

Contents

List of Figures

List of Tables

Acknowledgments

I am especially grateful to my advisor Prof. Stuart Coles for introducing me to extreme value theory, for his guidance and complete support and friendship.

I would like to express my deep thanks to Dr. Scott Sisson for his hospitality during my time spent in the University of New South Wales, his valuable input and support. I would also like to express my gratitude to Prof. Matt Wand for introducing me to the semiparametric regression and for his generous contribution to my research, and friendship. I would like to thank also Dr. Francesco Pauli, Dr. Nicola Sartori and Dr. Cristiano Varin for their useful suggestions and discussions during my research. I am also indebted to Prof. Anthony Davison for his precious advices. Many thanks also to Prof. Richard Smith for providing me with the USA rainfall dataset.

I am deeply indebted to Padova University, the PhD director Prof. Alessandra Salvan and my PhD Professors. Last, but not least I would like to give my special thanks to my family and Angie for all their support and affection.

Introduzione

Gli eventi estremi fanno parte della vita reale, in particolare quelli riguardanti l'ambiente dove viviamo, possono avere un grande impatto sulla nostra vita di ogni giorno. Degli esempi sono le innondazioni, gli uragani, i terremoti ed altre calamità. Questi fenomeni capitano di rado ma le loro consequenze possono avere un impatto drammatico sulle persone, le città e le abitazioni. In campo scientifico un notevole impegno è dedicato allo studio delle cause e delle conseguenze di eventi catastrofici, con l'obiettivo di predirne il loro verificarsi. Inoltre, un ampio sforzo è indirizzato per comprendere il possibile collegamento di questi fenomeni con il riscaldamento della terra e il cambiamento climatico. Conseguentemente, di recente l'analisi dei valore estremi sta acquisendo una notevole importanza ed un numero sempre crescente di ricercatori si stanno interessando a questa area di ricerca e in particolar modo gli statistici e i probabilistici.

La modellazione statistica dei valori estremi ha inizio circa a metà degli anni ottanta. L'analisi statistica degli eventi estremi ci aiuta a capirne l'intensità e a predirne la frequenza con cui questi fenomeni si verificano. Se queste informazioni sono disponibili, allora alcune misure preventive possono essere adottate per mitigarne gli effetti.

L'obiettivo di questa tesi e di fornire metodi statistici per la stima di eventi estremi per due particolari processi: sequenze non stazionarie univariate di valori estremi e sequenze stazionarie di estremi in ambito spaziale. In entrambi i casi gli aspetti statistici consistono nell'inferenza parametrica e non parametrica dei modelli usati e nell'adattamento dei modelli ai processi ambientali, con particolare attenzione rispetto le temperature massime e ai livelli massimi delle piogge rilevati in diversi siti.

La distribuzione dei valori estremi è ampiamente adottata nell' analisi degli eventi rari. Il suo utilizzo e motivato da risultati asintotici che si basano su una classe di processi stazionari ragionevolmente ampia. Un modo di procedere che tiene conto della non stazionarietà è quello di considerare come modello di base la distribuzione generalizzata dei valori estremi, ma definendo per i parametri del modello delle strutture di dipendenza con molteplici covariate (modelli di regressione). Tradizionalmente si sono utilizzati modelli di regressione parametrici ma recentemente l'interesse si è spostato verso l'alternativa offerta dai modelli non o semi parametrici. Degli esempi sono gli articoli di Davison e Ramesh (2002) e Chavez-Demoulin e Davison (2005), i quali hanno dimostrato l'utilità dell'approccio non parametrico per alcuni tipi di modelli dei valori estremi. In particolare, i primi autori hanno utilizzato la verosimiglianza locale mentre i secondi le spline di lisciamento. Nonostante ciò, la letteratura sui modelli di lisciamento dei valori estremi è ancora scarsa e nella sua infanzia. In questa tesi si propone, come possibile alternative ai metodi già adottati, l'utilizzo di spline di lisciamento ma secondo la formulazione che si basa sul paradigma dei modelli misti.

Una conseguenza di questo approccio, consiste nel fatto che per costruzione i parametri di lisciamento delle spline corrispondono alle componenti di varianza del modello. Così i metodi basati sulla verosimiglianza o le tecniche Bayesiane possono essere applicate per l'inferenza, la valutazione e l'adeguamento del modello (per esempio Ruppert, Wand e Carroll, 2003). Inizialmente si considera il caso più semplice, dove è trattata una sola struttura di dipendenza da covariate, rigurdanti il parametro di posizione. È stato sviluppato così, un metodo di stima basato sull'approccio della verosimiglianza, il quale ne risulta una versione estesa per il modello che incorpora anche gli effetti casuali. L'approccio che fa uso dei modelli misti è ben conosciuto nel contesto dei modelli lineari e lineari generalizzati, ma è nuovo nell'ambito della distribuzione generalizzata dei valori estremi. Comunque questo metodo comporta la necessità di risolvere degli integrali multivariati che sono analiticamente intrattabili. Ne consegue così che la formulazione classica della funzione di verosimiglianza è compromessa. Allora consideriamo due possibili opzioni: in un caso il metodo di stima si basa su una approssimazione della funzione di verosimiglianza ottenuta tramite il metodo di Laplace, in un altro caso il metodo di stima si basa su una approssimazione simile alla funzione di verosimiglianza penalizzata. Entrambi i metodi sono utilizzati e messi a confronto. Un' attraente caratteristica dovuta dalla formulazione delle spline di lisciamento tramite i modelli misti è che non sono necessarie ulteriori procedure per la stima dei parametri di lisciamento, come ad esempio il metodo della validazione incrociata. Invece, questi ultimi possono facilmente venire stimati tramite l'utilizzo di un'approssimazione della funzione di verosimiglianza. Questo primo modello ha una adeguata applicabilità ma rappresenta anche un punto di partenza per ulteriori estensioni come quello al caso della componente di scala.

Mentre nel caso univariato la teoria e l'analisi statistica dei valori estremi e ben sviluppata, nel caso multivariato ci sono molte meno linee guida. Questo può essere un problema perchè in molti processi ambientali come per esempio le alluvioni, il naturale dominio è quello spaziale. In ambito spaziale, l'analogo delle sequenze stazionarie univariate e multivariate di valori estremi sono costituiti dai processi massimamente stabili (per esempio de Haan e Pickands, 1986; Resnick, 1987). Questi sono stati sviluppati da de Haan (1984) ed hanno simili risultati teorici della distribuzione per i valori estremi ma estesi al dominio spaziale. I processi massimamente stabili, forniscono un utile approccio perchè permettono di modellare gli estremi incorporando nel modello le dipendenze temporali ma anche quelle spaziali. Dal punto di vista statistico una classe parametrica di processi massimamente stabili assieme con un semplice metodo di stima sono illustrati da Smith (1990). Ulteriori metodi statistici ed altre analisi di dati sono stati discussi nello stesso ambito da Coles (1993) e Coles e Tawn (1996). Dato che non è possibile derivare l' espressione analitica della funzione di densità, i metodi inferenziali basati sulla funzione di verosimiglianza non sono facili da applicare. Comunque alcuni stimatori non parametrici sono stati proposti da de Haan e Pereira (2006).

In questa tesi consideriamo due differenti approcci inferenziali. Il primo si basa sulla funzione di verosimiglianza composita che fornisce un'approssimazione della funzione di verosimiglianza (Linsday, 1988). Dimostriamo come i metodi di stima, basati sulla funzione di verosimiglianza a coppie, forniscono uno strumento flessibile per la stima anche nel contesto spaziale e che i risultati ottenuti sono sensati e ragionevoli.

Infine, illustriamo un alternativo metodo di stima basato sul approccio Bayesiano. Un modo per superare le difficoltà indotte dall'intrattabilità della funzione di verosimiglianza e fornito dai metodi computazionali conosciuti come ABC (Approximate Bayesian Computation). Questi metodi possono essere onerosi dal punto di vista computazionale ma in alcuni casi forniscono ragionevoli risultati inferenziali. Indaghiamo cosi l'applicabilità di questi metodi nel contesto degli estremi spaziali.

Introduction

Rare events are part of the real world but inevitably environmental extreme events may have a massive impact on everyday life. We are familiar, for example, with the consequences and damage caused by hurricanes and floods etc. Consequently, there is considerable attention in studying, understanding and predicting the nature of such phenomena and the problems caused by them, not least because of the possible link between extreme climate events and global warming or climate change. Thus the study of extreme events has become ever more important, both in terms of probabilistic and statistical research.

Statistical modelling of extreme values has flourished since about the mid-1980s. Such analysis, for instance, can help by estimating both the rate and magnitude of rare events, so that precautionary measures can be taken to prevent catastrophic phenomena, plan for their impact and mitigate their effects.

This thesis aims to provide statistical modelling and methods for making inferences about extreme events for two types of process. First, non-stationary univariate processes; second, spatial stationary processes. In each case the statistical aspects focus on model fitting and parameter estimation with applications to the modelling of environmental processes including, in particular, nonstationary extreme temperature series and spatially recorded rainfall measures.

The Generalized Extreme Value distribution (GEV) is widely adopted model for extremal events in the univariate context. It's motivation derives from asymptotic arguments that are based on reasonably wide classes of stationary processes. For modelling extremes of nonstationary sequences it is commonplace to still use the GEV as a basic model, but to handle the issue of nonstationarity by regression modelling of the GEV parameters. Traditionally this has been done using parametric models (Coles 2001, chapter 6), but there has been considerable recent interest in the possibility of nonparametric or semiparametric modelling of extreme value model parameters. For example, Davison and Ramesh (2002) and Chavez-Demoulin and Davison (2005) have demonstrated the usefulness of nonparametric regression, or smoothing, for certain types of extreme value models. The former used a local likelihood approach, while the latter used smoothing splines. Nevertheless, the literature on smoothing in extremal models remains scarce and in its infancy. As a novel alternative, this thesis proposes the use of mixed model-based splines for extremal models. A compelling feature of this approach is that the smoothing parameters correspond to variance components, so maximum likelihood or Bayesian techniques can be applied for model fitting, assessment and inference (e.g. Ruppert,Wand and Carroll, 2003). We start with the simplest case, developing nonparametric estimation for a smoothly varying location parameter within the GEV model. The approach is effectively maximum likelihood for an expanded version of the model that includes random effects. This approach is well used and developed when data are normally distributed or have a distribution within the exponential family, but is novel for extremes. However, the inclusion of the random effects leads to analytically intractable K-dimensional integrals for the likelihood formulation. Two different options are considered: one an approximation to the likelihood based on Laplace's approximation; the other based on a further approximation that is closer in spirit to a penalized likelihood function. The two methods are compared and contrasted by means of a simulation study.

An attractive feature of this general approach to nonparametric smoothing is that the extent of smoothing - in our case expressed as variance components - are estimated as part of the inference procedure. Consequently, there is no need for secondary procedures such as cross validation to determine smoothing parameters. We provide a quick and simple way of model fitting. This has uses in its own right, but also provides good starting values for more complex models, enabling the GEV scale parameter to also be covariate dependent, for example.

Whilst the theory and statistical practice of univariate extremes is well developed, there is much less guidance for the modelling of spatial extremes. This creates problems because many environmental processes - such as rainfall - have a natural spatial domain. The spatial analogue of univariate or multivariate extreme value models is the class of max-stable processes. (e.g de Haan and Pickands, 1986; Resnick, 1987). Max-stable processes were first developed by de Haan (1984) and have a similar asymptotic motivation, but expanded to a spatial domain, as the GEV distribution in the univariate case. They provide a general and useful approach to model extremal processes incorporating temporal or, more commonly, spatial dependence. On the statistical side, a parametric class of max-stable processes, together with a simple approach for inference, is provided by Smith (1990). Statistical methods for max-stable processes and data analysis of practical problems are discussed further by Coles (1993) and Coles and Tawn (1996). However, likelihood methods for such models are complicated by the intractability of density functions in all but the most trivial cases, although some alternative nonparametric estimators have been proposed by de Haan and Pereira (2006).

In this thesis we consider two different approaches. The first is *composite* (or *pseudo*)-likelihood which serves as a surrogate of the full likelihood (Linsday, 1988). We demonstrate that the *composite* likelihood procedure performs reasonably well and provide a flexible framework for inference.

Finally, we also explore the possibility of a Bayesian analysis of max-stable processes. This is obviously complicated by the intractability of the likelihood function, so we turn to a class of recently developed procedures referred to as ABC (Approximate Bayesian Computation). These methods are computationally intensive, and not easy to apply for highly structured problems such as max-stable process. Nonetheless, we show the method to have value and that reasonable inference can be obtained in some cases.

Part I

Smoothing extremes

Chapter 1

Extreme value theory

The extreme type theorems are important to the study of extreme value theory. In the literature, Fisher and Tippett (1928) were the first to discover the these and later their results were proved by Gnedenko (1943). Galambos (1987), Resnick (1987) are interesting reference books on the technical aspect. Coles (1990) gives a detailed introduction of statistical aspects, with emphasis on maximum likelihood methods in parameter estimation. Essentially, the extreme type theorems establish that for a sequence of i.i.d. random variables with suitable normalizing constants, the limiting distribution of maximum statistics, if it exists, follows one of three types of extreme value distributions.

1.1 Basic results

Let $\{X_i\}_{i\geq 1}$ be a sequence of independent and identically distributed random variables with marginal distribution function F, that is each X_i is distributed according to F. Denote the maximum of n consecutive elements of the sequence by $M_n = \max X_1, \ldots, X_n$. As n increases, M_n approaches the upper end-point, $w = \sup\{y : F(y) < 1\}$, of F and the limiting distribution of M_n is a point mass at w. A normalization is required to obtain a non-degenerate limit. As in the central limit theorem, a linear normalization is traditional and the limit, as n approaches infinity, is sought for the distribution function

$$P\left(\frac{M_n - b_n}{an} \leq y\right) = F^n(a_n y + b_n)$$

for sequences of constants $a_n > 0$ and $b_n \in \mathbb{R}$. The following theorem, due to Fisher and Tippett (1928), characterizes all of the possible limit distributions. Denote weak convergence by $\underrightarrow{w}$.

Theorem. *If there exist sequences of constants $an > 0$ and $b_n \in \mathbb{R}$ such that*

$$P\left(\frac{M_n - b_n}{an} \leq y\right) \underrightarrow{w} F(y) \quad \text{as} \quad n \to \infty$$

for a non-degenerate distribution function F, then F is a generalized extreme value distribution function,

$$F(y; \mu, \psi, \xi) = \exp\left[-\left\{1 + \xi\left(\frac{y-\mu}{\psi}\right)\right\}_+^{-1/\xi}\right], \quad -\infty < \mu,\, \xi < \infty, \quad \psi > 0,$$

defined on $\{y : 1 + \xi(y - \mu)/\psi > 0\}$ *where* $y_+ = \max(0, y)$*. For* $\xi = 0$*,* F *is defined by the limit as* $\xi \to 0$*.*

The generalized extreme value distribution function comprises three subclasses:

$$\begin{array}{lll} I & \exp\left[-\exp\left\{-\left(\frac{y-\beta}{\alpha}\right)\right\}\right] & \text{for} \quad y \in \mathbb{R}, \\ II & \exp\left\{-\left(\frac{y-\beta}{\alpha}\right)^{-\gamma}\right\} & \text{for} \quad y > \beta, \\ III & \exp\left[-\left\{-\left(\frac{y-\beta}{\alpha}\right)^{\gamma}\right\}\right] & \text{for} \quad y < \beta, \end{array}$$

where $\alpha > 0$ and $\gamma > 0$. These sub-classes correspond to the generalized extreme value distribution with $\xi = 0$, $\xi > 0$ and $\xi < 0$, and are known by the names I Gumbel, II Fréchet and III Weibull. "Standard" versions of these distributions refer to the special case $\alpha = 1$, $\beta = 0$ and $\gamma = 1$.

The approximation $P(M_n \leq y) \approx F\{(y - b_n)/a_n\}$ for large n motivates the generalized extreme value distribution as a model for maxima of blocks with large but finite lengths since the normalizing constants can be assimilated into the location and scale parameters μ and ψ.

The corresponding (probability) density function has expression

$$f(y) = \psi^{-1}\left\{1 + \xi\left(\frac{y-\mu}{\psi}\right)\right\}^{-1/\xi-1} \exp\left[-\left\{1 + \xi\left(\frac{y-\mu}{\psi}\right)\right\}^{-1/\xi}\right], \qquad (1.1)$$

provided that $\{y : 1 + \xi(y - \mu)/\psi > 0\}$

Chapter 2

Introduction to semiparametric regression

Splines continue to play a central role in nonparametric and semiparametric regression modelling. Recent descriptions include Ruppert, Wand and Carroll (2003) and Denison, Holmes, Mallick and Smith (2002). In these references, smooth functional relationships are fitted using a large basis of spline functions subject to penalization. Up until the mid-1990s most literature on spline-based nonparametric regression was focused on smoothing splines, and their multivariate extension thin plate splines, where the penalty takes a specific form and the number of basis functions roughly equals the sample size (e.g. Wahba, 1990; Green and Silverman, 1994). However, in recent years, there has been a lot of research on more general spline/penalty strategies, most of which use considerably fewer basis functions. The main forces include:

- more complicated models, often with several smooth functions;
- larger data sets, where smoothing and thin plate splines become computationally intractable,
- mixed model and Bayesian representations of smoothers that lend themselves to the use of established software, such as `BUGS`, `lme`() in `R` and `PROC MIXED` in `SAS`; provided the number of basis functions is relatively low.

Ruppert, Wand and Carroll (2003) summarize and provide access to many of these developments. The term *penalized splines* has emerged as a descriptor for general spline fitting subject to penalties. Therefore sometimes the term *smoothing splines* will be used imprecisely, but with the same general idea in mind. Penalized splines have been applied successfully with linear and generalized linear models. Only recently have spline models been explored in more complex settings such as extreme value models (e.g. Pauli and Coles, 2001; Chavez-Demoulin and Davison, 2005; Yee and Stephenson 2007). In this chapter we will briefly introduce penalized spline regression, penalized spline representations under a mixed model approach and illustrate the application to the linear and generalized linear models. In the next Chapter the penalized spline mixed model representation for sample extremes will be discussed.

2.1 Penalized spline regression

The term *nonparametric regression* is often referred to as the problem of estimating an unspecified "smooth" function f from a scatterplot (x_i, y_i), $i = 1, \ldots, n$. Many different approaches for this general objective of smoothing a scatterplot exist. Here we focus on the *penalized splines* method which has the attractiveness of being a relatively straightforward extension of linear regression modelling. A comprehensive exposition to this topic with many useful references is given by Green and Silverman (1994).

Consider the simplest ordinary nonparametric regression setting

$$y_i = f(x_i) + \varepsilon_i, \quad 1 \leq i \leq n, \tag{2.1}$$

where $(x_i, y_i) \in \mathbb{R}^2$. Suppose that the ε_i are random variables with $\mathbb{E}(\varepsilon_i) = 0$ and variance σ_ε^2 and that an estimate of $f(x) = \mathbb{E}(y|x)$, the corresponding underlying trend, is required over the interval $[a,\, b]$ containing the x_i's. f is assumed to be a generic "smooth" function, for example a polynomial of some order p, with which different types of nonlinear structures can be accommodated. In order to fit any complicated structure of f by, for example the method called penalized splines, the function must be represented defining a nonparametric regression such as a spline model.

For an integer $K \leq n$ let $\kappa_1, \ldots, \kappa_K$ be a sequence of knots such that

$$a = \kappa_1 < \kappa_2 < \ldots < \kappa_j < \ldots < \kappa_{K-1} < \kappa_K = b$$

and let $(x - \kappa_1)_+, \ldots, (x - \kappa_K)_+$ be the *linear spline basis functions* defined by these knots, where $(x)_+ = \max(0, x)$. Then the simplest spline model is a *linear* spline model defined as

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k (x - \kappa_K)_+$$

where β_0, β_1 and u_k for $i = 1, \ldots, K$ are coefficients. The result is a piecewise linear function obtained as a linear combination of the linear spline basis functions $1, x, (x - \kappa_1)_+, \ldots, (x - \kappa_K)_+$. Alternative spline basis functions are, for example the truncated power basis of degree p, the radial basis function, B-spline, etc. An introduction to these basic alternatives are discussed in detail in Ruppert, Wand and Carroll (2003, p. 67–74). A more convenient formulation of a spline models is given by the general class

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k z_k(x) \tag{2.2}$$

where $z_1, \ldots, z_k$ are spline basis functions which must be suitably specified. The benefit of using sophisticated basis functions might be substantial. For a deeper discussion on spline basis functions, penalized splines and direct generalization of smoothing splines such as the O'Sullivan penalized splines, we refer to Green and Silverman (1994, p.12), Welham, Cullis, Kenward and Thompson (2007) and Wand and Ormerod (2007).

Lastly, we mention knot choice. There are sophisticated algorithms that use the data to choose K, the number of knots. Some examples are discussed in Ruppert, Wand and Carroll (2003, p. 127–128). A common default in the penalized spline literature is $K = \min(n_U/4, 35)$, where n_U is the number of unique x_i's. Given K, the distribution of the knots may also have some effect on the results. One strategy is to use

$$\kappa_k = \left(\frac{k+1}{K+2}\right) \text{th sample quantile of the unique } x_i$$

while an alternative recommend by Eilers and Marx (1996) is to use equally-spaced knots. In Figure 2.1 we have illustrated an example of nonparametric regression by using fossil data (Ruppert, Wand and Carroll, 2003 p. 129). The data consists of 106 measurements of ratios of strontium isotopes found in fossil shells and their age. We can see how it is possible to handle any complex type of structure by simply adding more linear spline basis functions with form $(x - \kappa)_+$. It is clear that the fit improves with a larger set of knots. However, an excess of knots can cause too much flexibility or overfitting. To avoid this problem attention must be paid when determining the number of knots, as they have consequences on the roughness of the fit. One way to overcome this problem is to select the number and the location of the knots by the simple methods listed above, and then fit the model by using the technique known as *penalized spline regression*. In other words we suppose to use a large number of knots K, but then

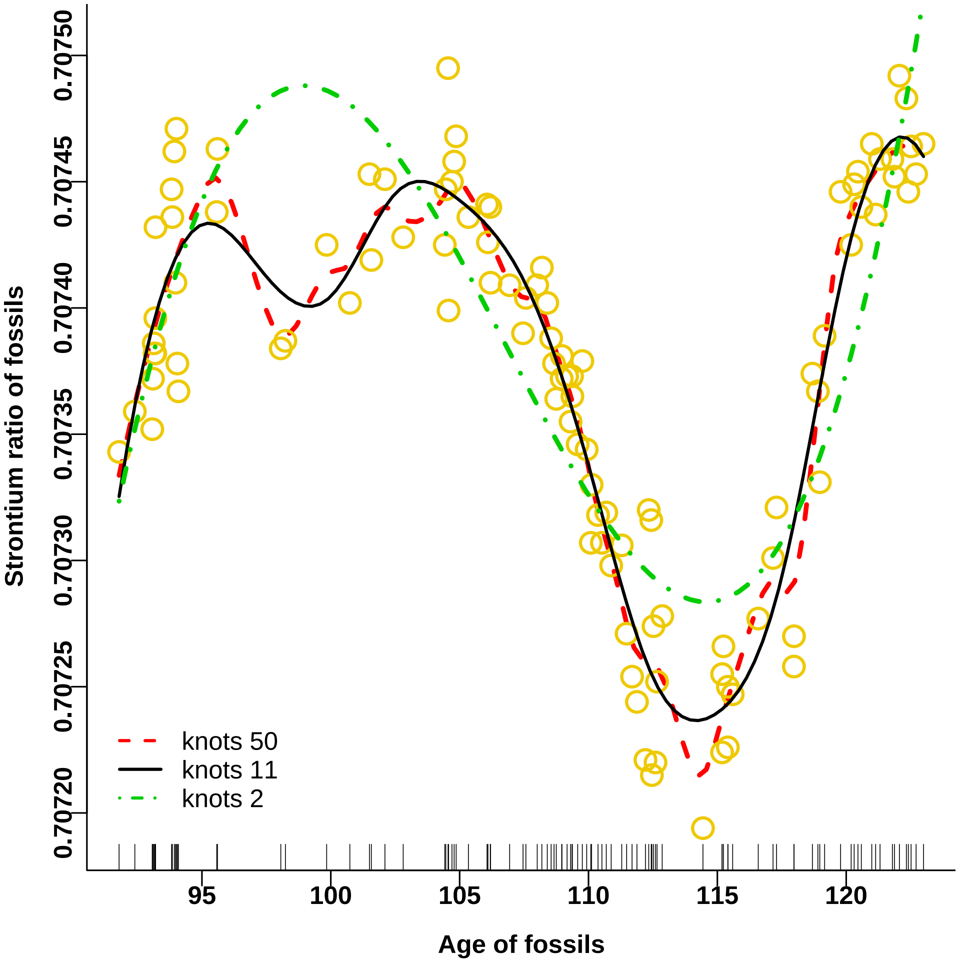

Figure 2.1: *Fossil data: ratios of strontium isotopes found in fossil shells versus the age of the fossils. The model fitting is reported by using 50 (dotted red line), 11 (solid black line) and 2 (broken green line) knots.*

we constrain the coefficients of the spline basis functions which depend by the knots, so to reduce their influence. This is done in the following way.

The ordinary least-squares criterion with form

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\nu}}, \quad \text{where} \quad \hat{\boldsymbol{\nu}} \text{ minimizes } \|\mathbf{y} - \mathbf{X}\boldsymbol{\nu}\|^2,$$

with $\boldsymbol{\nu} = [\beta_0, \beta_1, u_1, \ldots, u_K]^T$, without any constraint on $\boldsymbol{\nu}$ often leads to overfitting. This problem can be avoided by imposing a constraint on $\boldsymbol{\nu}$ such as $\sum_{k=1}^{K} u_k < B$ for a smoother fit to the scatterplot. In this case the least squares minimization problem can be written as

$$\text{minimize } \|\mathbf{y} - \mathbf{X}\boldsymbol{\nu}\|^2 \quad \text{subject to} \quad \boldsymbol{\nu}^T\mathbf{D}\boldsymbol{\nu} \leq B,$$

where

$$\mathbf{D} = \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times K} \\ \mathbf{0}_{K\times 2} & \mathbf{I}_{K\times K} \end{bmatrix}.$$

The amount of smoothness is controlled by B, and does not depend on the number or placement of knots. The solution of the constrained optimization problem has connections with ridge regression (Hoerl and Kennard, 1970). Solution requires a Lagrange multiplier argument, i.e. for some λ, choose $\boldsymbol{\nu}$ to minimize

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\nu}\|^2 + \lambda^2\boldsymbol{\nu}^T\mathbf{D}\boldsymbol{\nu}. \tag{2.3}$$

This has the solution

$$\hat{\boldsymbol{\nu}}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda^2)^{-1}\mathbf{X}^T\mathbf{y}.$$

The term $\lambda^2\boldsymbol{\nu}^T\mathbf{D}\boldsymbol{\nu}$ is called the *roughness penalty*. It penalizes fits that are too rough thus yielding a smoother result. The quantity λ is the *smoothing* parameter which controls the amount of smoothing. The case $\lambda = 0$ corresponds to the unconstrained case. By contrast when $\lambda > 0$ we have downweighted the influence of the knots, so the fit is a little less rough.

In Figure 2.2 the regression analysis of Fossil data is again reported. This time we illustrate the model fitting by setting different values of the smooth parameter. We can see that the more the smooth parameter's value is increased, the smoother the model fitting.

Instead, in Figure 2.3 the model fitting is reported by using different spline basis functions. In particular we considered the truncated spline basis (top-left panel), the quadratic spline basis (top-right panel) and the cubic spline basis (bottom-left panel). As we can see the application of different basis for the fossil data does not have much influence on the fitting. In the bottom-right panel we have reported, along with the fit, the variability bands. For a discussion on the variability bands we refer to Ruppert, Wand and Carroll (2003, p. 133–137).

The curve estimation method described in this section is based on the roughness penalty approach that has other complementary arguments, note in literature such as *smoothing splines* (Wahba, 1990; Green and Silverman, 1994). Note that penalized splines are a more general approach than smoothing splines so that the latter can also be implemented by the former. Connections between the two approaches are well established, see Ruppert, Wand and Carroll (2003; p. 74–75) and Green and Silverman (1994).

So far the penalized spline regression has been described without need to focus on the random structure of the variables contemplated. For instance, if we

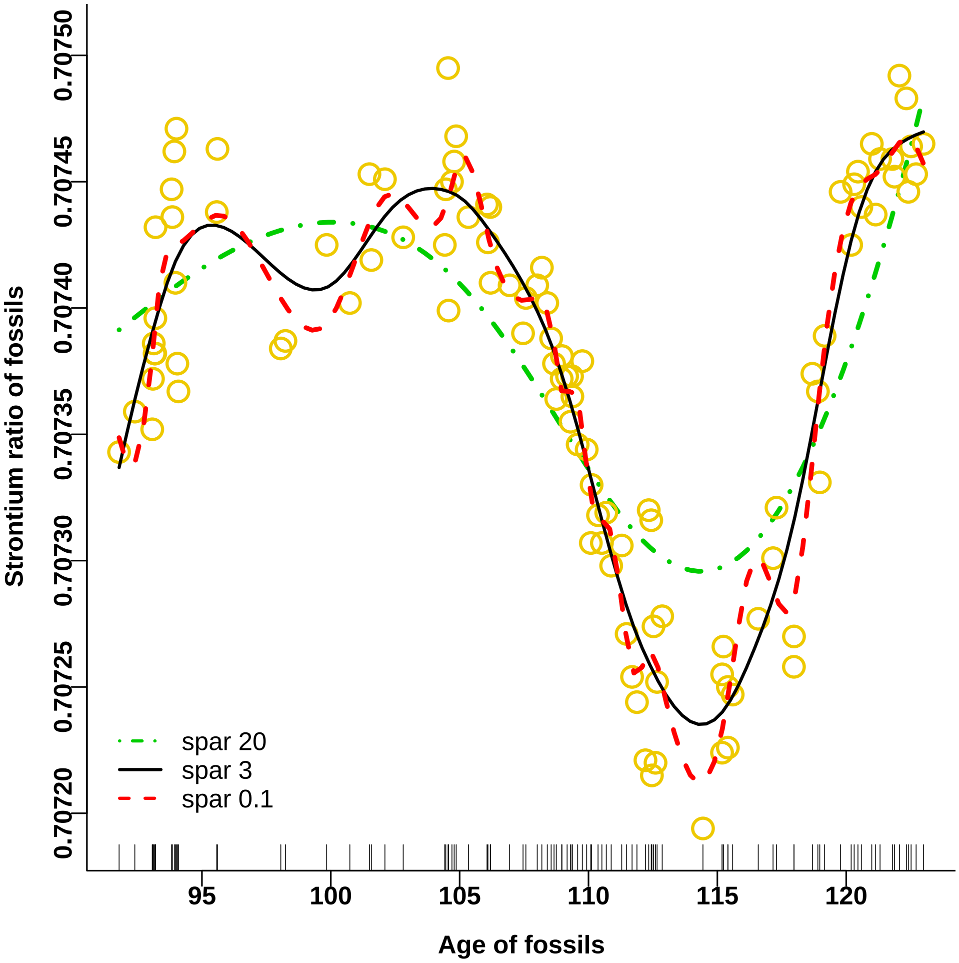

Figure 2.2: *Fossil data: ratios of strontium isotopes found in fossil shells versus the age of the fossils. The model fitting is reported by using smooth parameters equal to 20 (dotted green line), 3 (solid black line) and 0.1 (broken red line) knots.*

suppose that the y_i are normally distributed, then the minimization of the sum of the penalized squared criterion is equivalent to the maximization of the *penalized log-likelihood* (e.g Green, 1987). The roughness penalty in this context is given by the smoothing splines penalty, which is essentially the integral of the second derivative of the smooth function f. Roughly speaking, the log likelihood term is penalized by the amount of nonlinearity. But the assumption of normality of the errors in real data analysis are scarcely satisfied. However, the penalized log-likelihood framework can be used for more general models allowing the treatment of responses with density other than normal, see Davison (2003, p. 535–539).

2.2 Mixed Model Formulation

Wand (2003) shows how an estimate of f by penalized splines can be written as the *best linear unbiased predictor* (BLUP) of a mixed model, Robinson (1991). Assume the basic nonparametric regression (2.1) where f is modelled by a linear spline model. Let

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 & \beta_1 \end{pmatrix}^T, \quad \mathbf{u} = \begin{pmatrix} u_1 & \dots & u_K \end{pmatrix}^T$$

be respectively the coefficients of the linear part and the truncated line functions of the linear spline model. These are associated to the design matrices

$$\mathbf{X} = [1 \quad x_i]_{1\le i\le n}, \qquad \mathbf{Z}[(x_i - \kappa_k)_+]_{\substack{1\le i\le n \\ 1\le k\le K}}. \tag{2.4}$$

Then the penalized least square criterion (2.3) for fitting a nonparametric regression can be equivalently reformulated as

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \underset{\boldsymbol{\beta},\,\mathbf{u}}{\text{argmin}}(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\|^2 + \lambda^2\|\mathbf{u}\|^2).$$

Thus if the coefficients $\mathbf{u}$ are treated as random with $\text{Cov}(\mathbf{u}) = \sigma_u^2$ then the minimization criterion corresponds to the BLUP of the standard "general" linear mixed model with the following identity $\lambda^2 = \sigma_\varepsilon^2/\sigma_u^2$. The minimization of $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\|^2$ subject to the penalty $\lambda^2\|\mathbf{u}\|^2$ (that is imposing a restriction on the distribution of $\mathbf{u}$) defines a penalized least squares criterion as well. Broadly a regression spline problem can be formulated by the following general linear mixed model representation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \varepsilon, \quad \begin{bmatrix} \mathbf{u} \\ \varepsilon \end{bmatrix} \sim \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_u^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_\varepsilon^2\mathbf{I} \end{bmatrix} \right).$$

While estimation of $\boldsymbol{\beta}$ and prediction of $\mathbf{u}$ can be done without the Gaussianity assumption by BLUP (e.g Robinson, 1991), if the normality hypothesis for the response and the random effects are taken into account, then the variance components can be estimated by standard likelihood methods. So, given that the variance components correspond in someway to the smoothing parameters, then data can support us in selecting, from the likelihood approach, the right amount of smoothing. The quantity $\|\mathbf{u}\|^2$ in the penalty term of the minimization criterion derives from the normality assumptions of $\mathbf{u}$. Under the Gaussian model, inference may be based on the marginal normal density of $\mathbf{y}$, which gives the log likelihood for $(\boldsymbol{\beta}, \sigma_\varepsilon^2, \sigma_u^2)$ with form

$$\ell(\boldsymbol{\beta}, \sigma_\varepsilon^2, \sigma_u^2) = -\frac{1}{2}\{n\log(2\pi) + \log|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\},$$

where $\mathbf{V} = \mathbf{Z}\,\sigma_u^2\mathbf{I}\,\mathbf{Z} + \sigma_\varepsilon^2\mathbf{I}$. For known $(\sigma_\varepsilon^2, \sigma_u^2)$ the likelihood maximum estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}.$$

In principle likelihood inference for σ_u^2 and σ_ε^2 may be obtained via maximization of the profile log likelihood

$$\ell_P(\sigma_\varepsilon^2, \sigma_u^2) = -\frac{1}{2}\big\{\log|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) + n\log(2\pi)\big\}.$$

However, the maximum likelihood estimators (obtained by maximization of the profile log likelihood) may have large downward bias because no adjustment is made for the degrees of freedom lost in estimating the vector $\boldsymbol{\beta}$. Therefore an adjustment is required. One can be provided by maximizing the modified log likelihood

$$\ell_R(\sigma_\varepsilon^2, \sigma_u^2) = \ell_P(\sigma_u^2, \sigma_\varepsilon^2) - \frac{1}{2}\log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|.$$

In this way the corresponding maximum likelihood estimator of the variance components are divided for the right degrees of freedom $n-p$, Davison (2003, p. 457–458, 657–659). This log likelihood version is known as *restricted* log likelihood, and the associated estimator as *restricted* maximum likelihood estimator (REML). The "modified" term above means that the log likelihood profile has been modified, rather than the modified likelihood of Barndorff-Nielse (1983). Restricted log-likelihood turns out to be equivalent to the use of the marginal likelihood corresponding to the marginal density, rather than the full density of the data (e.g. Barndorff-Neilsen and Cox, 1994).

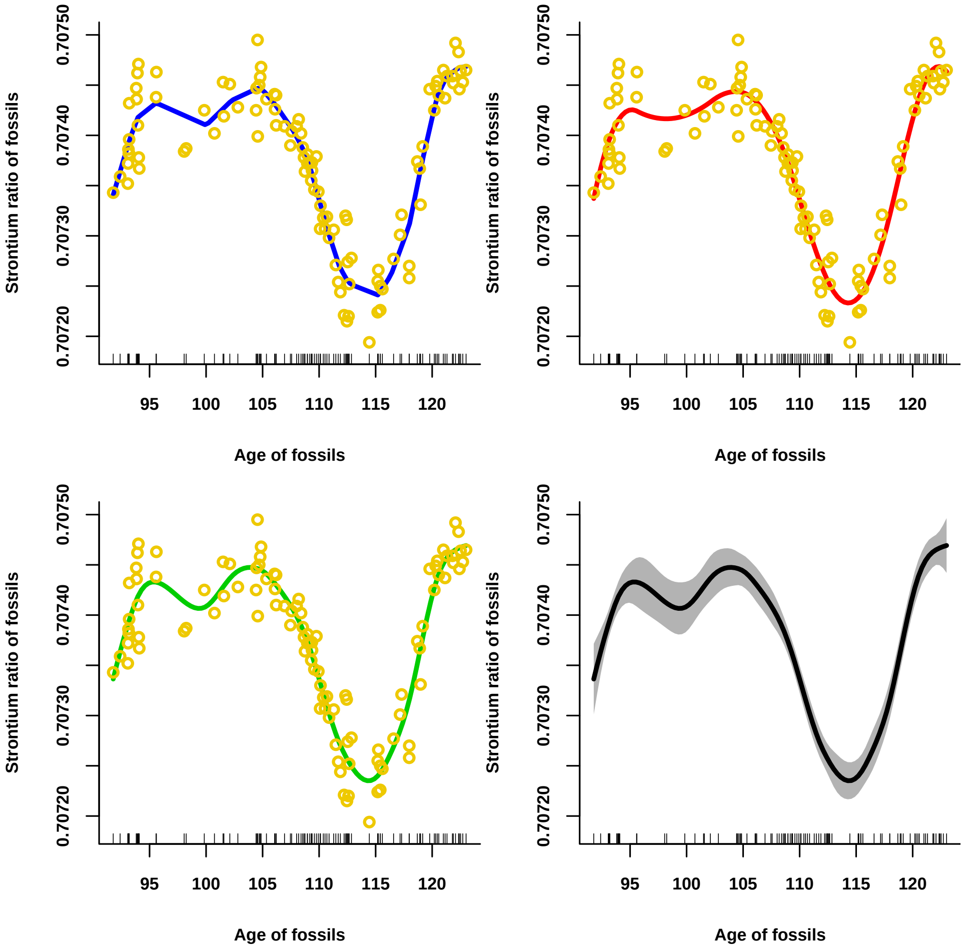

Figure 2.3: *Fossil data: ratios of strontium isotopes found in fossil shells versus the age of the fossils. The model fitting is reported by using different spline basis functions: truncated (top-left), quadratic (top-right) and cubic (bottom-left). In the bottom-right panel the fit (using the cubic basis function) along with its variability bands is reported.*

For known $\boldsymbol{\beta}$, σ_u^2 and σ_ε^2, the random effects are predicted by using the *best prediction* $\hat{\mathbf{u}} = \mathbb{E}(\mathbf{u}|\mathbf{y})$ which results in

$$\hat{\mathbf{u}} = \sigma_u^2 \mathbf{I}\,\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}),$$

where for the practice determination of $\mathbf{u}$ the estimates of $(\boldsymbol{\beta}, \sigma_u^2, \sigma_\varepsilon^2)$ will be used. The Maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ and the best predictor $\hat{\mathbf{u}}$ for given σ_u^2 and σ_ε^2 are equivalent to the solutions obtained solving the penalized least squares problem, leading to

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \mathbf{C}(\mathbf{C}^T\mathbf{C} + \lambda^2\mathbf{D})^{-1}\mathbf{C}^T\mathbf{y}$$

where $\mathbf{C} = [\mathbf{X}\ \mathbf{Z}]$ and $\mathbf{D} = \text{diag}(0, 0, 1, \ldots, 1)$.

2.3 Generalized Semiparametric Regression

Generalized linear models (GLMs) allow the handling of non-Gaussian response variables in parametric regression problems. Similarly GLMs may be fruitfully utilized in nonparametric regression using the penalized spline framework, when the responses are evidently nonnormally distributed. In particular, with general responses the penalized spline regressions, handled using the mixed model representation, need to be redefined. For this reason we evoke the *generalized linear mixed model* (Ruppert, Wand and Carroll, 2003, p. 203–206) paradigm in order to specify properly the nonparametric regression. This change involves challenging efforts by means of overcoming the computational difficulties resulting in the fitting for the more complex model.

Specifically, suppose the response vector $\mathbf{y}$ has distribution belonging to the exponential family with density function

$$f(\mathbf{y};\, \boldsymbol{\beta}) = \exp\left(\frac{\mathbf{y}^T g(\mathbf{X}\boldsymbol{\beta}) - \mathbf{1}^T b\{g(\mathbf{X}\boldsymbol{\beta})\}}{\phi} + \mathbf{1}^T c(\mathbf{y}, \phi)\right),$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients related with the design $\mathbf{X}$ of form (2.4) which is associated with the predictor variables x_i. $b(\cdot)$ is a function of the natural parameter $\eta_i = g(\mathbf{x}_i^T\boldsymbol{\beta})$, and $c(\cdot)$ is a function of $\mathbf{y}$ and the dispersion parameter ϕ. These include assuming several forms corresponding to the different models: Poisson, Bernoulli, etc. Finally the linear predictor $\mathbf{x}_i^T\boldsymbol{\beta}$ is related to the mean of the response by the relation $g^{-1}\{\mathbb{E}(y)\} = \mathbf{x}_i^T\boldsymbol{\beta}$, where g^{-1} is the link function, McCullagh and Nelder (1989).

The generalized linear mixed model representation of the penalized spline regression is a natural extension of the model structure illustrated in the previous section for the generalized responses. However the more complex distributional form of the responses involve complications in statistical modelling. It is worth remembering that penalized log likelihood is a well known inference method widely utilized in smoothing regressions with Gaussian and non-Gaussian data.

Essentially, under the likelihood umbrella, the estimation criterion consists of maximizing the log likelihood but also penalizing it with a term that takes the nonlinearity amount into account. Note that this corresponds to the minimization of the penalized sum of the squares, only when we treat normal responses. The key point is that, in order to fit a spline regression model, the mixed model formulation can provide a similar criteria to that of the penalized log likelihood.

Precisely, assume the linear spline model (2.2) with basis functions (2.4) and coefficients $(\boldsymbol{\beta}, \mathbf{u})$. Let the response vector $\mathbf{y}$ be a member of the exponential family with density

$$f(\mathbf{y};\, \boldsymbol{\beta}) = \exp\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T b\{\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}\} + \mathbf{1}^T c(\mathbf{y})\},$$

where the dispersion parameter is assumed known, for example $\phi = 1$ and the link function g is equal to the identity. With this specified structural design, the solution to the penalized linear spline problem is given by

$$\left[\begin{array}{c}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}\end{array}\right] = \underset{\boldsymbol{\beta},\,\mathbf{u}}{\mathrm{argmax}}\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) + \sigma_u^{-2}\|\mathbf{u}\|^2\},$$

Ruppert, Wand and Carroll (2003, p. 215–216). Assume also that the coefficients $\mathbf{u}$ are random variables normally distributed, $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2\mathbf{I})$ with zero mean and covariance matrix $\sigma_u^2\mathbf{I}$. Then, combining the probability density functions of the exponential family and the random effects yields the likelihood for $(\boldsymbol{\beta},\, \sigma_u^2)$

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta},\, \sigma_u^2) = f(\mathbf{y};\, \boldsymbol{\beta}, \sigma_u^2) &= \int_{\mathbb{R}^k} f(\mathbf{y}|\mathbf{u}) f(\mathbf{u})\, d\mathbf{u}\\ &\propto |\sigma_u^2\mathbf{I}|^{-1/2}\int_{\mathbb{R}^k}\exp\left\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) - \frac{1}{2}\mathbf{u}^T\,\sigma_u^{-2}\mathbf{I}\,\mathbf{u}\right\} d\mathbf{u}.\end{aligned} \tag{2.5}$$

We refer to (2.5) as the *integrated* likelihood. To perform maximum likelihood estimation, we need to be able to compute this likelihood by integrating out the random effects. Whereas the prediction of the random effects can be obtained by the best predictor $\hat{\mathbf{u}} = \mathbb{E}(\mathbf{u}|\mathbf{y})$ (their predictions will depend on the likelihood estimates). Note that $\mathbf{u}$ are random quantities (not parameters) so maximum likelihood cannot be used to estimate (or predict) them. However, the estimation and prediction are compromised by intractable high dimensional integrals. In fact with non linear models, such non linearity prevents integration over the random effects, so that it is not usually possible to find the exact likelihood. No closed form solution is available either for the best predictor. Various strategies have been proposed by different authors (e.g. Lindstrom and Bates, 1990). Most of them replace the integration by joint maximization over the parameters and random effects, and then use linearization for the conditional modes of the random effects to estimate the variance components.

More precisely, laplace's method (e.g. Severini, 2005, p.276) can be applied to the integrals over $\mathbf{u}$, expanding the integral argument as a quadratic Taylor expansion about its maximum, see appendix A.1 for details. This, or other methods are approximate, either based on theoretical approximations or numerical approximations of integrals, see Wolfinger and O"Connell (1993). Essentially, the goodness of the approximation depends on the order used in the Taylor expansion. In some cases high-order expansion could be required in order to obtain adequate inference results.

Application of Laplace's method leads to the approximate (integrated) log likelihood

$$\ell_{\text{INT}}(\boldsymbol{\beta},\, \sigma_u^2) \simeq \mathbf{y}^T(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\hat{\mathbf{u}}) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\hat{\mathbf{u}}) - \frac{1}{2}\hat{\mathbf{u}}^T\,\sigma_u^{-2}\mathbf{I}\,\hat{\mathbf{u}} - \frac{1}{2}\log|\mathcal{I}_{\mathbf{uu}}(\widehat{\mathbf{u}}_{\boldsymbol{\beta},\sigma^2})|,$$

where the quantity $\mathcal{I}_{\mathbf{uu}}(\widehat{\mathbf{u}}_{\boldsymbol{\beta},\sigma^2}) = \mathbf{I} + \sigma_u^2\mathbf{I}\mathbf{Z}^T\,\mathrm{diag}\{b''(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\hat{\mathbf{u}})\}\mathbf{Z}$ - for σ_u^2 fixed - is the observed information matrix of the sum between the exponential and the normal probability densities respect to $\mathbf{u}$. A model feature of spline regression under the generalized mixed model representation is that a penalized log likelihood function arises as an approximation of the integrated log likelihood, $\ell_{\text{INT}}(\boldsymbol{\beta},\, \sigma_u^2)$.

More precisely, excluding the quantity $-\frac{1}{2}\log|\mathcal{I}_{\mathbf{uu}}(\widehat{\mathbf{u}}_{\boldsymbol{\beta},\sigma_u^2})|$ in the integrated log likelihood we acquire a similar expression of the penalized log likelihood for generalized linear models. The key idea of the penalized log likelihood governed by the mixed model paradigm is to treat the fixed and random effects $(\boldsymbol{\beta}, \mathbf{u})$ as coefficients, but penalizing the random effects according to the restriction $\mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$. It turns out that the fitting and inference could be, by some means, based on the penalized log likelihood. In fact for given σ_u^2, Breslow and Clayton (1993) argued how for ease of fitting that the quantity $\mathcal{I}_{\mathbf{u}\widehat{\mathbf{u}}}(\widehat{\mathbf{u}}_{\boldsymbol{\beta},\sigma^2})$ can be (somehow) neglected in order to estimate the coefficients $\boldsymbol{\beta}$. But then for given σ_u^2 the penalized log likelihood

$$\ell_{\text{PL}}(\boldsymbol{\beta},\, \sigma_u^2) = \mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \frac{1}{2}\mathbf{u}^T\, \sigma_u^{-2}\mathbf{I}\,\mathbf{u}.$$

can be used in order to base the inference procedure for $(\boldsymbol{\beta}, \mathbf{u})$. By maximization of $\ell_{\text{PL}}(\boldsymbol{\beta},\, \sigma_u^2)$ respect with the fixed and random effects (treated as parameters) leads to the approximate likelihood estimates $(\widehat{\boldsymbol{\beta}},\, \widehat{\mathbf{u}})$. Estimates can be provided, for instance as solution of the score equations by a Fisher scoring algorithm (Green, 1989). Maximization of the penalized log likelihood corresponds essentially to maximization of the joint "likelihood" of the observed data and random effects simultaneously, as proposed by Harville and Mee (1984). The amount of smoothing of the penalized spline with the mixed model approach corresponds to the variance component, and the fit depends on the current estimates of σ_u^2. So given $(\widehat{\boldsymbol{\beta}},\, \widehat{\mathbf{u}})$, maximum likelihood estimation of σ_u^2 involves maximizing the profile log likelihood. There is no closed form solution, so it has to be done numerically. However, a common approach is to alternatively apply the restricted log likelihood, although its use is controversially justified. We refer to Breslow and Clayton (1993), see also Ruppert, Wand & Carroll (2003, p. 205–206) for details. They provide a description of an iterative scheme in order to obtain the model fitting and to estimate variance components.

It is also worth remembering that Lee and Nelder (1996) introduced an alternative approach which does not require integration when $f(\mathbf{y}|\mathbf{u})$ is a member of the GLMs. They propose to base inference on what they call the *h-likelihood* which is essentially the product between the likelihood of the observed data and the density function of the random effects. It turns out that taking the logarithm of that product, again, leads to an expression similar to what here we have called penalized log likelihood. However, also using the h-likelihood, the generalized linear models case involves computational challenges for fitting, respect the easiest linear models case. Lee, Nelder and Pawitan (2006) also based their model fitting and inference in many cases on a stepwise iterative algorithm similar to that of Breslow and Clayton (1993). Therefore it seems widely accepted in literature that the model fitting and inference can be based on stepwise procedures.

Chapter 3

Mixed model-based additive models for sample extremes

3.1 Introduction

Extreme value models have asymptotic theoretical results that makes them particulary well suited for statistical applications that focus on extreme events. Coles (2001) provides a comprehensive introduction to this topic. Assume to observe a sequence of a process measurement, such as daily temperature, hourly rainfall levels, etc. and suppose that we are interested in the maximum of the process over a period of time of observation, for example the annual maxima. Then the generalized extreme value (GEV) distribution has emerged as the most common family for modelling such data.

Here briefly, let $Z_1, \ldots, Z_n$ be an independent and identically distributed (i.i.d.) set of random variables and let $M_n = \max(Z_1, \ldots, Z_n)$ denote the sample maximum. Then, the limiting distribution as $n \to \infty$ of $(M_n - a_n)/b_n$ (if such a sequences of constants $\{b_n > 0\}$ and $\{a_n\}$ exist) must be a member of the generalized extreme value family of distributions (e.g. von Mises, 1954; Jenkinson, 1955). A random variable Y has a GEV distribution, denoted by $Y \sim$ GEV(μ, ψ, ξ) if its cumulative distribution function is given by:

$$F(y; \mu, \psi, \xi) \exp\left[-\left\{1 + \xi\left(\frac{y-\mu}{\psi}\right)\right\}_{+}^{-1/\xi}\right], \quad -\infty < \mu,\, \xi < \infty, \quad \psi > 0,$$

where $x_+ = \max(0, x)$ and μ, ψ and ξ are respectively location, scale and shape parameters. The GEV distribution may be divided into the following three subfamilies: Fréchet distribution (Fischer-Tippett type III) for $\xi > 0$, Weibull distribution (Fischer-Tippett type II) for $\xi < 0$ and Gumbel-type distribution (Fischer-Tippett type I) when $\xi \to 0$; see Fisher and Tippett (1928). For a deeper discussion with a review of theoretical results see Chapters 3 and 4 of Coles (2001).

Although it is possible to study the asymptotic of maxima of processes with specified forms of non–stationarity, the results are generally too specific to be of use in modelling data for which the form of non-stationarity is unknown. Because of the generality with which such effects may arise in practice, there is little point in the theoretical study of non–stationary or covariate-dependent extremes. Instead, these aspects are best addressed from a purely statistical viewpoint, trying to model changes in the marginal behavior of extremes rather than appealing to any additional asymptotic theory. Thus it is usual to model non-stationarity of extremes directly through the parameters of the standard models. For example, to allow for a linear trend in the underlying level of extremal behavior.

Suppose we observe n sample maxima $y_1, \ldots, y_n$ with corresponding covariate vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$. General GEV regression models (e.g. Chapter 6 Coles, 2001) take the form

$$y_i|\mathbf{x}_i \sim \text{GEV}(\mu(\mathbf{x}_i), \psi(\mathbf{x}_i), \xi(\mathbf{x}_i)), \tag{3.1}$$

where, for example, $\mu(\mathbf{x}_i) = g((\mathbf{X}\boldsymbol{\beta})_i)$, g is a link function, $\boldsymbol{\beta}$ is a vector of regression coefficients and $\mathbf{X}$ is a design matrix associated with the $\mathbf{x}_i$'s. Similar structures may be imposed upon $\psi(\mathbf{x}_i)$ and $\xi(\mathbf{x}_i)$. For non-stationary sequences of extremes a pragmatic approach is to use the standard extreme value distribution as a basic template and then enhance it by statistical modelling of the model parameters. The result is an attractive parametric family where non-stationary is expressed in terms of extreme value parameters. An immediate advantage of this approach is that the regression coefficients can be estimated via maximum likelihood which is easily adaptable to changes in model structures. Davison and Ramesh (2000) and Hall and Tajvidi (2000) argue that parametric models for (3.1) can be too restrictive, and have advocated non-parametric approaches. Pauli and Coles (2001) and Chavez-Demoulin and Davison (2005) have also demonstrated the usefulness of non-parametric regression, or smoothing in extreme value contexts. The first of these papers used a Bayesian approach, while the second used the classic based-likelihood approach. Chavez-Demoulin and Davison (2005) also treated the additive model extension, where the effect of several covariates can be considered simultaneously and flexibly.

The aim of this chapter is to explore an alternative approach to additive model fitting and inference for sample extreme responses. It is based on the mixed model/splines paradigm that has achieved a great deal of success in other contexts during the last decade. Ruppert, Wand and Carroll (2003) provide a summary of this general approach. A compelling feature of this approach is that the smoothing parameters correspond to variance components, so maximum likelihood or Bayesian techniques can be applied for model fitting, assessment and inference. Complications such as spatial or temporal correlation, missing data and measurement error are more easily incorporated.

3.2 Model structures

Ruppert, Wand and Carroll (2003) and Wand (2003) have discussed how penalized splines can be carried out in a mixed model framework for Gaussian and exponential family models. Here, we focus on the generalized extreme value models.

Let $y_1, \ldots, y_n$ be n observed sample with associated explanatory variables x_i. Assuming that the location parameter in the GEV distribution is smooth on an interval $[a, b]$, then the simplest time-nonhomogeneous spline mixed model is given by

$$y_i \sim \text{GEV}(\mu(x_i), \psi, \xi) \quad -\infty < \mu(x_i), \xi < \infty, \; \psi > 0 \quad x_i \in \mathbb{R}. \tag{3.2}$$

Mixed model-based penalized spline models for μ take the general form

$$\mu(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k z_k(x); \quad u_1, \ldots, u_k \quad \text{i.i.d.} \quad N(0, \sigma^2),$$

where $z_1, \ldots, z_K$ is an appropriate set of spline basis functions. The simplest version is $z_k(x) = (x - \kappa_k)_+$, where $\kappa_1, \ldots, \kappa_K$ is a dense set of knots within the range of the x_i's. More sophisticated basis functions are also recommended for consideration. See for example, Welham, Cullis, Kenward and Thompson (2006) and Wand and Ormerod (2007). The latter reference describes the z_k corresponding to the R function `smooth.spline()`. The choice of K has a secondary effect and, for many signals, about 20 knots are sufficient.

Let $\mathbf{y} = (y_1, \ldots, y_n)$ and define the design matrix

$$\mathbf{X} = [1 \quad x_i]_{1 \le i \le n}, \qquad \mathbf{Z}[z_k(x_i)]_{\substack{1 \le i \le n \\ 1 \le k \le K}}$$

associated with fixed $\boldsymbol{\beta} = [\beta_0 \ \beta_1]^T$ and random effects $\mathbf{u} = [u_1 \ldots u_K]^T$. Given $\mathbf{u}$, the y_i are conditionally independent with distribution,

$$y_i | \mathbf{u} \sim \text{GEV}(\boldsymbol{\mu}_i, \psi, \xi),$$

where the linear predictor $\boldsymbol{\eta}_i(\mathbf{X}\beta + \mathbf{Z}\mathbf{u})_i$ is related to $\boldsymbol{\mu}_i$ by the link function $g(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i$. Note that $\boldsymbol{\mu}$ is related to the conditional mean of $\mathbf{y}$ given $\mathbf{u}$ via

$$\text{E}(\mathbf{y}|\mathbf{u}) = \begin{cases} \boldsymbol{\mu} + \mathbf{1}\psi\{\Gamma(1-\xi) - 1\}/\xi, & \text{for} \quad \xi \neq 0 \\ \boldsymbol{\mu} + \mathbf{1}\psi\gamma, & \text{for} \quad \xi = 0 \end{cases}$$

where $\mathbf{1}$ is a vector of n one, $\gamma = 0.57721566\cdots$ is the Euler's constant and Γ is the Gamma function.

Let $\mathbf{C} = [\mathbf{X} \,|\, \mathbf{Z}]$ be the matrix obtained combining the columns of design matrices $\mathbf{X}$ and $\mathbf{Z}$, and with vector $\boldsymbol{\nu}^T = [\boldsymbol{\beta}^T \ \mathbf{u}^T]$ the $K + 2$ coefficients of fixed and random effects. With this notation the conditional probability density function of $y_i|\mathbf{u}$ and the probability density function of $\mathbf{u}$ random effects, have the expressions

$$f(\mathbf{u}; \sigma^2) = (2\pi)^{-K/2}(\sigma^2)^{-K} \exp\left(-\frac{\|\mathbf{u}\|^2}{2\sigma^2}\right) \quad \text{and}$$

$$f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}, \psi, \xi) = \prod_{i=1}^{n} \frac{1}{\psi} \left\{1 + \xi\left(\frac{(\mathbf{y} - \mathbf{C}\boldsymbol{\nu})_i}{\psi}\right)\right\}^{-\frac{1}{\xi}-1} \exp\left[-\left\{1 + \xi\left(\frac{(\mathbf{y} - \mathbf{C}\boldsymbol{\nu})_i}{\psi}\right)\right\}^{-\frac{1}{\xi}}\right].$$

The norm for fitting (3.2) is estimation of the parameters via maximization of the likelihood:

$$\mathcal{L}(\boldsymbol{\beta}, \psi, \xi, \sigma^2) = f(\mathbf{y}; \boldsymbol{\beta}, \psi, \xi, \sigma^2) = \int_{\mathbb{R}^K} f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}, \psi, \xi) f(\mathbf{u}; \sigma^2)\, d\mathbf{u}$$

and prediction of the random effects via the best predictor $\widehat{\mathbf{u}} = E(\mathbf{u}|\mathbf{y})$. We term the likelihood above as *integrated* likelihood (see Chapter 2, Section 2.3). However, both are hindered by intractable integrals. Instead, we appeal to the approximate (integrated) log likelihood and the ideas of the penalized log likelihood.

Integrated likelihood $\mathcal{L}(\boldsymbol{\beta}, \psi, \xi, \sigma^2)$ for the parameters set $(\boldsymbol{\beta}, \psi, \xi, \sigma^2)$ derives from integrating out the random effects $\mathbf{u}$. The analytical solution of that integral is not easy to derive straightforwardly, but an approximation can be reached by application of Laplace's method to $\mathcal{L}(\boldsymbol{\beta}, \psi, \xi, \sigma^2)$, see appendix A.3 for details. From the penalized (quasi) likelihood (e.g. Green, 1987; Breslow and Clayton, 1993) inference procedures, based on approximate maximum likelihood, can be developed for generalized linear mixed models and also for curves estimation.

The penalized log likelihood is widely used in scatterplot smoothing with normal and generalized responses but it has also been applied in the extreme value context by Pauli and Coles (2001). Application of Laplace's method to $\mathcal{L}(\boldsymbol{\beta}, \psi, \xi, \sigma^2)$, and then taking the logarithm we derive the following approximate log likelihood

$$\begin{aligned}\ell_{\text{INT}}(\boldsymbol{\beta}, \psi, \xi, \sigma^2) &= \log\{f(\mathbf{y}; \boldsymbol{\beta}, \psi, \xi, \sigma^2)\} \\ &\simeq \log\{f(\mathbf{y}|\widehat{\mathbf{u}}; \boldsymbol{\beta}, \psi, \xi)\} + \log\{f(\widehat{\mathbf{u}}; \sigma^2)\} - \tfrac{1}{2}\log|\mathcal{I}_{\mathbf{uu}}(\widehat{\mathbf{u}}, \psi, \xi, \sigma^2)|,\end{aligned}$$

where $\mathcal{I}_{\mathbf{uu}}(\widehat{\mathbf{u}}; \boldsymbol{\beta}, \psi, \xi, \sigma^2)$ - for $(\boldsymbol{\beta}, \psi, \xi, \sigma^2)$ fixed - is the information matrix of the sum between the log of the conditional distribution $f(\mathbf{y}|\mathbf{u})$, and the log of the random effects distribution $f(\mathbf{u})$, respect with $\mathbf{u}$. The details and the analytical expression of the observed information matrix is given in the appendix A.3.

Now omitting the term $-\frac{1}{2}\log|\mathcal{I}_{\mathbf{uu}}(\widehat{\mathbf{u}}, \boldsymbol{\beta}, \psi, \xi, \sigma^2)|$ in the approximate log likelihood $\ell_{\text{INT}}(\boldsymbol{\beta}, \psi, \xi, \sigma^2)$, we get the penalized log likelihood expression

$$\begin{aligned}\ell_{\text{PL}}(\boldsymbol{\nu}, \psi, \xi, \sigma^2) &= \log\{f(\mathbf{y}|\widehat{\mathbf{u}}; \boldsymbol{\beta}, \psi, \xi)\} + \log\{f(\widehat{\mathbf{u}}; \sigma^2)\} \\ &= -n\log(\psi) - \frac{1+\xi}{\xi}\mathbf{1}^T \log\left\{\mathbf{1} + \xi\left(\frac{\mathbf{y} - \mathbf{C}\boldsymbol{\nu}|_{\mathbf{u}=\widehat{\mathbf{u}}}}{\psi}\right)\right\} \\ &\quad - \mathbf{1}^T\left\{\mathbf{1} + \xi\left(\frac{\mathbf{y} - \mathbf{C}\boldsymbol{\nu}|_{\mathbf{u}=\widehat{\mathbf{u}}}}{\psi}\right)\right\}^{-\frac{1}{\xi}} - \frac{K}{2}\log(\sigma^2) - \frac{\|\widehat{\mathbf{u}}\|^2}{2\sigma^2},\end{aligned} \tag{3.3}$$

where $\mathbf{C}\boldsymbol{\nu}|_{\mathbf{u}=\widehat{\mathbf{u}}}$ means that $\mathbf{C}\boldsymbol{\nu}$ is computed for $\mathbf{u} = \widehat{\mathbf{u}}$, and $\mathbf{C}\boldsymbol{\nu} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$.

Formula (3.3) is similar to the formulation used by Pauli and Coles (2001). They approached the model fitting and variance components estimation treating them as two divided problems. As it is commonly used in nonparametric literature, they assessed the variance structure by cross validation. Alternatively, in our setting we will see that the variance components can be estimated as model parameters. Note that this is an important feature, given that the variance components in practice correspond to the the smoothing parameters. So, by this approach the model fitting and inference can be carried out through likelihood methods alone.

Analogously to Section 2.3, for fixed σ^2 the penalized log likelihood (3.3), derived from the approximate integrated log likelihood $\ell_{\text{INT}}(\boldsymbol{\beta}, \psi, \xi, \sigma^2)$, can provide a remedy for the model fitting. In fact, the strategy under the mixed model approach is to treat the fixed and random effects $\boldsymbol{\nu} = (\boldsymbol{\beta}, \mathbf{u})$ as coefficients in the penalized log likelihood (3.3), but to penalize the $\mathbf{u}$ according to the restriction $\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Thus the maximization of (3.3) respect with $\boldsymbol{\nu}$ furnishes an estimate of the regression parameters. This and the accompanying methods for inference of the scale, shape and the variance components are described in the next section.

3.2.1 Estimation procedure

With the generalized extreme value model we need to provide the scale and shape parameter estimates in addition to the model fitting, this is because ψ and ξ are unknown quantities. For fixed σ^2, estimates of $(\boldsymbol{\nu}, \psi, \xi)$ can potentially be provided by means of a Newton-Raphson algorithm step applied to (3.3). Given those current estimates $(\widehat{\boldsymbol{\nu}}, \widehat{\psi}, \widehat{\xi})$, maximization of the profile penalized log likelihood respect with σ^2 yields an estimate of the variance component. Then we iterate these two stages until convergence of the parameters. However, the addition of scale and shape parameters jointly with regression coefficients hinders the application of this method by providing misleading estimates. We found by simulation exercises that the estimate results were dependent on particular starting values, and that the iterative scheme may be very numerically unstable. This is because the Newton-Raphson seems to be an unreliable method in order to simultaneously obtain the estimates of scale, shape and regression parameters.

Alternatively, dividing the estimation of $(\boldsymbol{\nu}, \psi, \xi)$ into two separate stages yields a better performing method. In conclusion, the stepwise algorithm that allows us to estimate the regression coefficients, the model parameters and the variance components is provided by the following iterative scheme:

Iterative Scheme: Fitting and inference in GEV spline Mixed Model

1. Set starting values: $\hat{\boldsymbol{\nu}}, \hat{\psi}, \hat{\xi}, \hat{\sigma}^2$.

2. Update $\hat{\boldsymbol{\nu}}$ by maximizing the penalized log likelihood $\ell_{\text{PL}}(\boldsymbol{\nu}, \hat{\psi}, \hat{\xi}, \hat{\sigma}^2)$.

3. Update $(\hat{\psi}, \hat{\xi})$ by maximizing the log likelihood $\ell_{\text{M}}(\psi, \xi, \hat{\sigma}^2)$.

4. Update $\hat{\sigma}^2$ by maximizing the log likelihood $\ell_{\text{M}}(\sigma^2)$.

5. Repeat steps 2–4 until convergence.

At the first stage the maximum penalized likelihood estimate of $\boldsymbol{\nu}$ can be obtained by the Newton-Raphson method but substituting the observed information with the Fisher information (Prescott and Walden, 1980). The Newton-Raphson updating step is given by

$$\widehat{\boldsymbol{\nu}}^{(i+1)} = \widehat{\boldsymbol{\nu}}^{(i)} + \mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\widehat{\boldsymbol{\nu}}^{(i)}_{\psi,\xi,\sigma^2}, \psi, \xi, \sigma^2)^{-1}\, \mathsf{D}_{\boldsymbol{\nu}}\ell_{\text{PL}}(\widehat{\boldsymbol{\nu}}^{(i)}, \psi, \xi, \sigma^2),$$

where $\mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\widehat{\boldsymbol{\nu}}_{\psi,\xi,\sigma^2})$ is the observed information matrix and $\mathsf{D}_{\boldsymbol{\nu}}\ell_{\text{PL}}(\boldsymbol{\nu}, \psi, \xi, \sigma^2)$ is the gradient for $\boldsymbol{\nu}$ for fixed (ψ, ξ, σ^2) at the corresponding maximum penalized log likelihood estimate $\widehat{\boldsymbol{\nu}}^{(i)}_{\psi,\xi,\sigma^2}$. An explicit expression for $\mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\widehat{\boldsymbol{\nu}}_{\psi,\xi,\sigma^2})$ its expectation and $\mathsf{D}_{\boldsymbol{\nu}}\ell_{\text{PL}}(\boldsymbol{\nu}, \psi, \xi, \sigma^2)$ are given in Appendix A.4.

At the second stage, estimates of $(\psi,\ \xi)$ can be obtained by maximization of the (modified) log likelihood,

$$\ell_{\text{M}}(\psi, \xi, \sigma^2) = \ell_{\text{PL}}(\widehat{\boldsymbol{\nu}}_{\psi,\xi,\sigma^2}, \psi, \xi, \sigma^2) - \tfrac{1}{2}\log|\mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\widehat{\boldsymbol{\nu}}_{\psi,\xi,\sigma^2}, \psi, \xi, \sigma^2)|.$$

Note that the modified term is not referring to the modified likelihood of Barndorff-Nielsen (1983). At the last stage, estimates of (σ^2) can be obtained by maximization of the log likelihood,

$$\begin{aligned}\ell_{\text{M}}(\sigma^2) \;\; = \;\; & \ell_{\text{PL}}(\widehat{\boldsymbol{\nu}}_{\psi,\xi,\sigma^2}, (\widehat{\psi}, \widehat{\xi})_{\sigma^2}, \sigma^2) - \tfrac{1}{2}\log|\mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\widehat{\boldsymbol{\nu}}_{(\widehat{\psi},\widehat{\xi})_{\sigma^2},\sigma^2}, (\widehat{\psi}, \widehat{\xi})_{\sigma^2}, \sigma^2)| \\ & -\tfrac{1}{2}\log|\mathcal{I}_{(\psi,\xi)(\psi,\xi)}((\widehat{\psi}, \widehat{\xi})_{\sigma^2}, \sigma^2)|\end{aligned}$$

where $\mathcal{I}_{(\psi,\xi)(\psi,\xi)}((\widehat{\psi}, \widehat{\xi})_{\sigma^2}, \sigma^2)$ is the observed information matrix for (ψ, ξ) for fixed σ^2 at the corresponding maximum log likelihood estimate $(\widehat{\psi}, \widehat{\xi})_{\sigma^2}$. An explicit expression for $\mathcal{I}_{(\psi,\xi)(\psi,\xi)}((\widehat{\psi}, \widehat{\xi})_{\sigma^2}, \sigma^2)$ is given in Appendix A.3.

Maximization of the log likelihoods in the second and third steps can be obtained by using the quasi-Newton numerical maximization routines (e.g. Broyden, 1967).

The use of the restricted likelihood method requires that the model parameters have to be orthogonal, for example in the case of the two parameters (λ, ψ). The adjustment quantity is justified because in the log likelihood an estimate $\widehat{\lambda}_\psi$ appears rather than λ, see Davison (2003, p. 657) for details. The adjustment is a penalization of the log likelihood that depends on the information available from λ. Large size of λ involves stronger penalization of the log likelihood than when it is small. In our case the regression coefficients are asymptomatically independent of the ψ and ξ parameters, Tawn (1988). Instead, the orthogonality condition does not hold exactly for the parameters (ψ, ξ) and σ^2. We found that the adjustments in the iterative scheme provide an improvement on the estimate results. This is especially the case for the variance components which can be substantially smaller, involving the hinderance of the model fitting and parameter estimates without the adjustments. However, the performance of this approximate method is tested by means of a simulation study that is illustrated in Section 3.4.

In practice we have also found that an easy alternative to model fitting and inference is provide by directly using the approximate log likelihood $\ell_{\text{INT}}(\boldsymbol{\beta}, \psi, \xi, \sigma^2)$. In fact, application of the quasi-Newton numerical maximization routines (e.g. Broyden, 1967) to $\ell_{\text{INT}}(\boldsymbol{\beta}, \psi, \xi, \sigma^2)$ boils down to adequate estimate results. The implementation is easily managed and the likelihood evaluation is feasible even if it may be computationally demanding. Variance estimates can be estimated consistently using the Jacobian matrix computed at the maximum and obtained from the numerical maximization routine.

In the simulation study of Section 3.4 we will illustrate the performance of the estimates obtained with the method based on $\ell_{\text{INT}}(\boldsymbol{\beta}, \psi, \xi, \sigma^2)$. A comparison between the two methods is also outlined.

Finally, it is important to remember that asymptotic likelihood results for the GEV distribution are subject to restrictions, Smith (1985).

3.2.2 Variability bands

Generally, calculating variability bands in function estimations consists of adding and subtracting from the estimated function two times its estimated standard error, Bowman and Azzalini (1997, p.75–76). They are considered as approximate pointwise confident intervals, Hastie and Tibshirani (1990).

Once an estimate of the regression coefficients $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{u}})$ has been obtained, the fit is given by:

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{u}} = \mathbf{C}\widehat{\boldsymbol{\nu}}.$$

A naïve expression for variability bands at x_i is given by:

$$(\mathbf{C}\widehat{\boldsymbol{\nu}})_i \pm z_{\alpha/2}\sqrt{\{\mathbf{C}\,\widehat{\text{Cov}(\widehat{\boldsymbol{\nu}}|\mathbf{u})}\mathbf{C}^T\}_{ii}}.$$

where $z_{\alpha/2}$ is the quantile of level $1-\alpha$ of $N(0,1)$, $\widehat{\text{Cov}(\widehat{\boldsymbol{\nu}}|\mathbf{u})}$ is the estimated covariance matrix of $\widehat{\boldsymbol{\nu}}$ given $\mathbf{u}$. Note that these are not $100(1-\alpha)\%$ pointwise confident intervals, although we assume that they provide an approximation.

3.2.3 Additive models extension

We now consider the extension where several variates may impact on the sample extremes $y_1, \ldots, y_n$. If $\mathbf{x}_i$ is d-variate then

$$\mu(\mathbf{x}_i) = f_1(x_{i1}) + \ldots + f_d(x_{id})$$

defines a general additive model for μ. Here the f_j are general smooth functions. The mixed model-based penalized splines of Section 3.2 can accommodate this extension by setting

$$\mathbf{X} = [1\ \mathbf{x}_{i1} \cdots\ \mathbf{x}_{i1}]_{1\leq i\leq n} \quad \mathbf{Z} = [\underset{1\leq k\leq K_1}{z_k(x_{i1})} \ldots \underset{1\leq k\leq K_d}{z_k(x_{id})}]_{1\leq i\leq n},$$

associated with $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_d]^T$, $\mathbf{u} = [u_1, .., u_{K_1}, .., u_1, .., u_{K_d}]^T$. Also,

$$\text{Cov}(\mathbf{u}) = \mathbf{G}_{\boldsymbol{\sigma}^2}\, \text{blockdiag}(\sigma_1^2 \mathbf{I}_{K_1}, \ldots, \sigma_d^2 \mathbf{I}_{K_d}),$$

with K_j the number of spline basis functions used for f_j. The fitting procedure described in the previous section is basically the same, but with longer $\boldsymbol{\beta}$ and $\mathbf{u}$ vectors, and σ^2 is replaced by the vector $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_d^2)$.

3.3 Illustrative examples

In this section some case studies of nonparametric regression for generalized extreme value responses are presented. For simplicity, the simulations are based on extreme values that are obtained as GEV distribution realizations rather than computing the block maxima over n units of observations. However, this does not substantially change the results of the simulation study.

For each scatterplot smoothing we illustrate the graphic regression fitting, reporting the model parameter estimates and the variance component selected. The common simulation design for the examples are:

- Synthetic observations are drawn from $y_i \sim$ GEV$(\mu(x_i),\ \psi,\ \xi)$ with sample of length $n = 250$ and $x_i \sim$ U$(0, 1)$.

- $K = 20$ knots for the predictor, with $\kappa_k = \frac{k+1}{K+2}$th sample quantile of unique predictor values.

- Radial cubic basis function modelling of a function f entails putting $\mu(x) = \beta_0 + \beta_1 x + \mathbf{Z}\mathbf{u}$, where

$$\mathbf{Z}[\underset{1\leq k\leq K}{|x_i - \kappa_k|^3}\ \underset{1\leq k,k'\leq K}{|\kappa_k - \kappa_{k'}|^3}]^{-1/2}_{1\leq i\leq n},$$

  and for $k = 1, \ldots, K$, $\{\kappa_k\}$ is a sequence of knots, $|x_i - \kappa_k|^3$ are basis functions and $k \leq k'$.

- $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I})$.

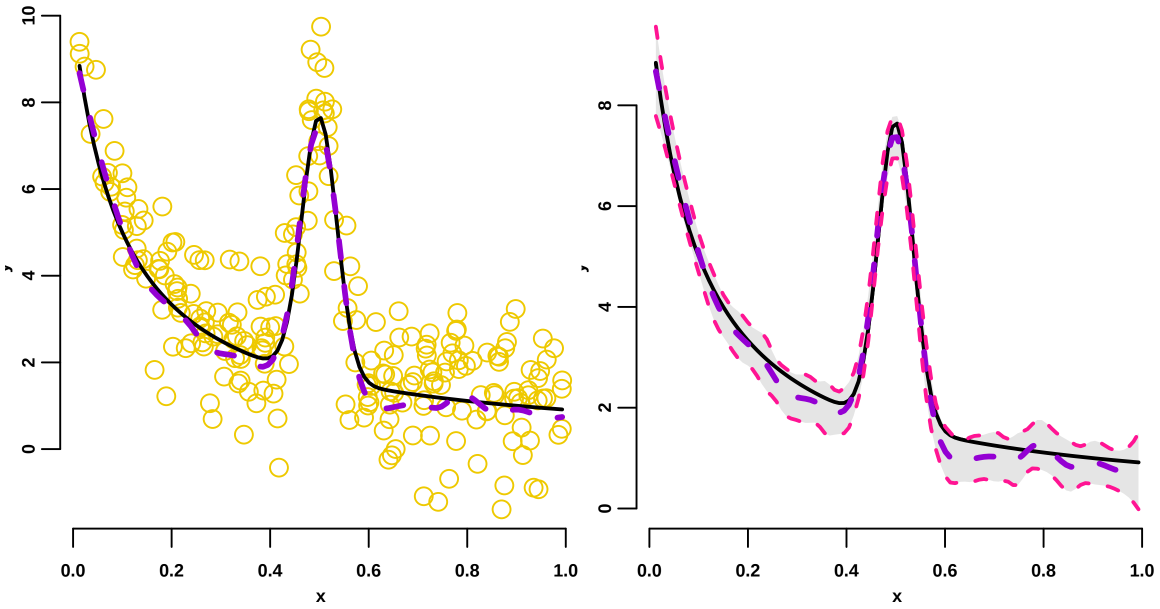

Figure 3.1: *Simulation example: in the left panel the circles indicate the GEV realizations. The location parameter is set up with the bump function. The true signal is the solid (black) line and its estimate with the broken (violet) line. In the right panel the true signal, its estimate and the* 95% *variability bands are illustrated.*

Example 1. **Bump function**: *we considered the regression function suggested by Ruppert, Wand and Carroll (2003, p. 128) with expression*

$$\mu(x) = \frac{1}{0.1 + x} + 8\exp\{-400(x - 0.5)^2\} \quad \text{with } x \in (0, 1).$$

Model parameters have been set as $\psi = 1$ and $\xi = -0.4$ so that the synthetic data was generated from a Weibull model. The left panel of Figure 3.1 shows the results for the simulation exercise. The black solid line is the regression function and the dotted line is the model fitting. In the right panel the real function and its estimate accompanied by the variability bands are reported. From Figure 3.1 it appears that the signal has been closely reproduced corresponding to the estimated smoothing parameter $\widehat{\sigma}_u = 70(12)$. The model parameter estimates are obtained as $\widehat{\psi} = \texttt{1.126(0.056)}$ and $\widehat{\xi} = \texttt{-0.405(0.035)}$.

Example 2. **Trigonometric functions**: *the second case study considers two trigonometric functions with forms*

$$\mu(x) = \sin 4\pi x \quad \text{and} \quad \mu'(x) = \exp\{\sin(4\pi x) + x\} \quad \text{with } x \in (0, 1).$$

Model parameters have been set to $\psi = 0.4$ and $\xi = 0.6$ in the first case, and $\psi' = 1$ and $\xi' = 0.4$ in the second. Figure 3.2 shows the results of the second simulation exercise. We can see from the left panels how the true signals (black solid lines) are adequately replicated from its estimates (broken violet lines). For the first function the smoothing parameter estimate is $\widehat{\sigma}_u = 12(3)$ and for the second is $\widehat{\sigma}'_u = 25(5)$. Between parenthesis the standard deviations are reported. In the right panels the regression estimates accompanied with the variability bands are illustrated. The model parameter estimates are respectively $\widehat{\psi} = \texttt{0.393(0.018)}$,

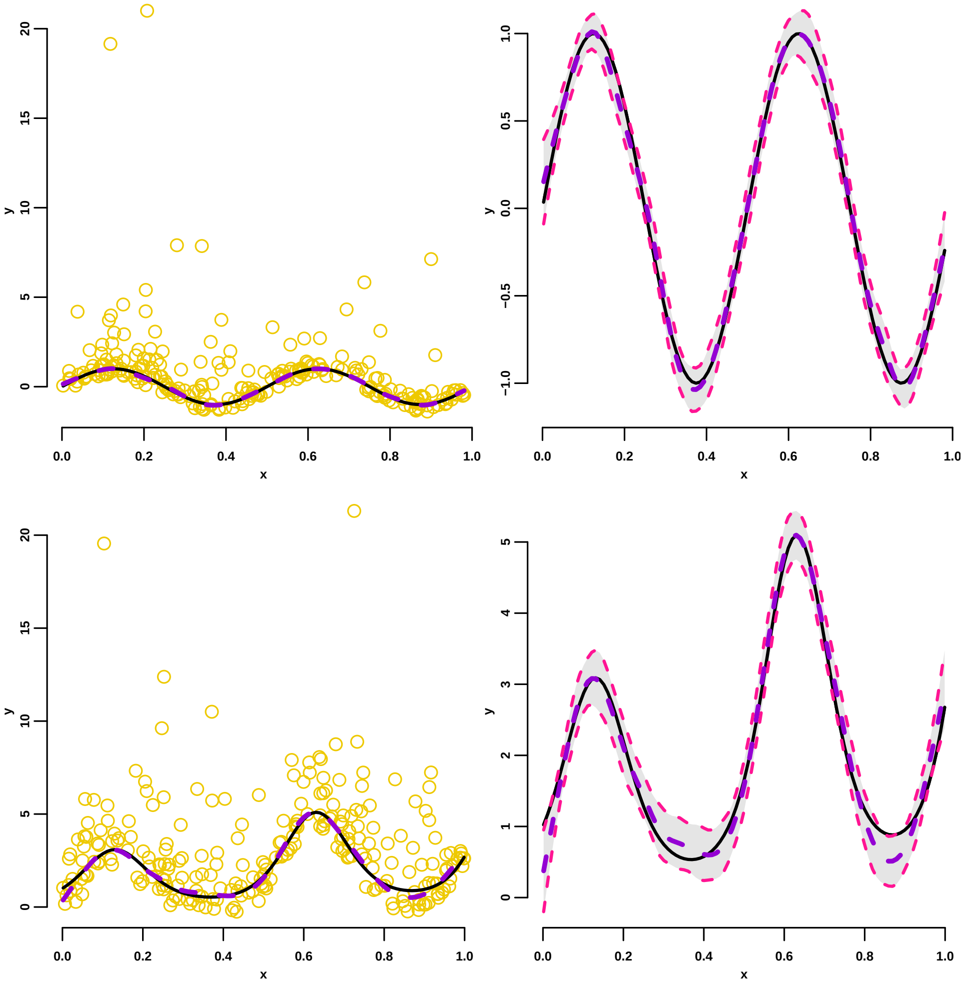

Figure 3.2: *Simulation example: in the left panels the circles indicate the GEV realizations. The location parameters are set up with the trigonometric functions. The true signals are the solid (black) lines and their estimates with the broken (violet) lines. In the right panels the true signals, their estimates and the* 95% *variability bands are illustrated.*

$\widehat{\xi} =$ 0.575(0.064) for the first case and $\widehat{\psi}' =$ 0.964(0.046), $\widehat{\xi}' =$ 0.394(0.061) for the second.

Example 3. **Additive components**: *the last study exposes the case of additive predictors. Consider the following functions*:

$$f_1(x) = \sin(\pi x) \quad \text{and} \quad f_2(x) = \sin(2\pi x) \quad \text{with } x \in (0,1),$$

so that $\mu(x) = f_1(x) + f_2(x)$. Synthetic data are generated with model parameters $\psi = 0.6$ and $\xi = 0.4$. Figure 3.3 shows the true signals f_1 and f_2 respectively from left to right panels with (black) solid lines. The function estimates are reported with the middle broken lines. The outer broken lines and grayed areas represent the 95 % variability bands. Even in this example the signals are evidently replicated and the respective estimated smoothing parameters are $\widehat{\sigma}_{u1} = 1.052(0.213)$ and $\widehat{\sigma}_{u2} = 3.445(0.983)$. The model parameter estimates and their standard errors resulted $\widehat{\psi} =$ 0.614(0.029) and $\widehat{\xi} =$ 0.414(0.059).

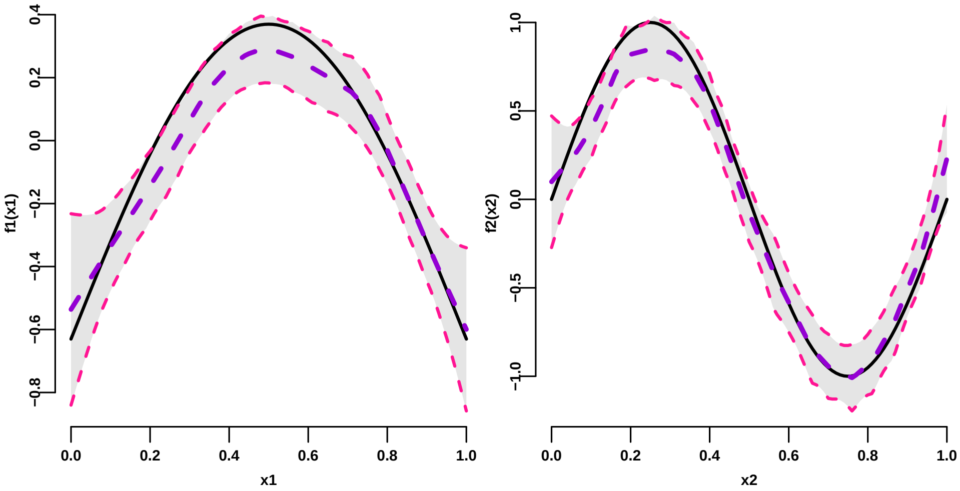

Figure 3.3: *Simulation example: left and right panels show the estimates of the additive function components. The location parameters are set up with the functions above reported. The true signals are the solid (black) lines and their estimates with the broken (violet) lines. In both panels the true signals, their estimates and the* 95% *variability bands are illustrated.*

3.4 Simulation Study

We investigated the performance of the mixed model-based for extremes with a simulation study. Let

$$\mu(x) = 2x + \cos(4\pi x) \qquad 0 \leq x \leq 1.$$

Data was generated in two steps. Firstly a sample $x_1, \ldots, x_n$ was drawn from a uniform distribution on $(0, 1)$. Secondly, given the x_i's, *n* realizations were drawn according to $y_i \sim \text{GEV}(\mu(x_i), \psi, \xi)$. The shape parameter ξ was set to -0.4, 0 and 0.4 corresponding to the three different types of GEV distributions. Also, different values of the scale parameters were considered. We performed 500 data replications for each configuration. In each case estimation was performed using the likelihood-based algorithm of Section 3.2.1 and the approximate log likelihood $\ell_{\text{INT}}(\boldsymbol{\beta}, \psi, \xi, \sigma^2)$.

Results for estimation of the scalar parameters are summarized in Table 3.1. The estimates seem reasonably accurate for all three-type distributions even though the estimation methods involves approximated likelihood functions. With the Fréchet distribution we observe a bias on the shape parameter for sample size equal to 100 with the method based on the penalized log likelihood. This bias effect is a consequence of the approximated likelihood functions. A bias source can arise from an inadequacy of Laplace's approximation integral when used with the heavy tail distribution (Fréchet). The simulation results indicate a considerable reduction of the estimates bias when the method based on the approximate $\ell_{\text{INT}}(\boldsymbol{\beta}, \psi, \xi, \sigma^2)$ is used. However, for larger sample size the variability of estimates distribution decrease gradually in accordance with the standard asymptotic likelihood estimate theory.

Figure 1 conveys the performance of the function estimation component. The estimates appear corresponding to the 10th, 50th and 90th percentiles of the replica-

Distribution	method	n	$\hat{\psi}$	$\hat{\xi}$	$\hat{\sigma}$
Fréchet	penalized	500	0.611 (0.029)	0.398 (0.040)	9.6 (1.0)
-	-	100	0.645 (0.073)	0.365 (0.099)	9.1 (2.1)
-	integrated	500	0.597 (0.029)	0.402 (0.043)	8.8 (0.6)
-	-	100	0.590 (0.070)	0.411 (0.124)	8.6 (1.3)
		true	*0.6*	*0.4*	
Gumbel	penalized	500	0.601 (0.023)	-0.005 (0.032)	9.7 (0.7)
-	-	100	0.607 (0.056)	-0.019 (0.080)	9.9 (1.6)
-	integrated	500	0.598 (0.023)	0.001 (0.031)	8.8 (0.8)
-	-	100	0.604 (0.055)	-0.014 (0.088)	8.3 (1.5)
		true	*0.6*	*0*	
Weibull	penalized	500	0.601 (0.021)	-0.396 (0.027)	9.5 (0.7)
-		100	0.591 (0.042)	-0.377 (0.066)	9.8 (1.5)
-	integrated	500	0.601 (0.024)	-0.405 (0.029)	8.4 (1.4)
-	-	100	0.588 (0.049)	-0.389 (0.063)	8.4 (1.6)
		true	*0.6*	*-0.4*	

Table 3.1: *Smoothing and nuisance parameters estimates of three-types: the first column indicates the distribution take into account, the second indicates which method is used for the estimation, the third report the data sample sizes while the third column indicates the sample size. From columns 4–6 GEV scale, shape and variance components estimates are given. Standard errors are in brackets.*

tion-wise deviance measures; given by

$$D(\boldsymbol{\mu}; \hat{\boldsymbol{\mu}}) = 2\{\ell_{\boldsymbol{\mu}}(\boldsymbol{\mu}) - \ell_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}})\},$$

where $\ell_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}})$ is the log likelihood computed for $y_i = \hat{\mu}_i$. For the larger sample sizes the fitted curves are approximately matching the true curves for all three percentiles and distributions. With sample size 100 the results are still acceptable. In addition, we note for the Fréchet case how a lack of accuracy for the fitted curve obtained corresponding with the 90th percentile is expected due to the nature of the heavy-tailed distribution.

3.5 Real data analysis

In this section we consider some analysis with real dataset. England and Switzerland temperatures have been the focus of the study. We explore the temperatures behavior over time and asses the relationship along with some covariates.

3.5.1 Application to English temperature data

In this section we consider the maximum Central England Temperature (CET). The dataset consists of daily maximum temperatures representative of a roughly triangular area of the United Kingdom enclosed by Lancashire, London and Bristol recorded from 1878 to 2006. The analysis focuses on the trend of the annual maxima of the temperatures. So in this case the blocks of maxima correspond to a time period of one year so that the equal length sequences of daily observations

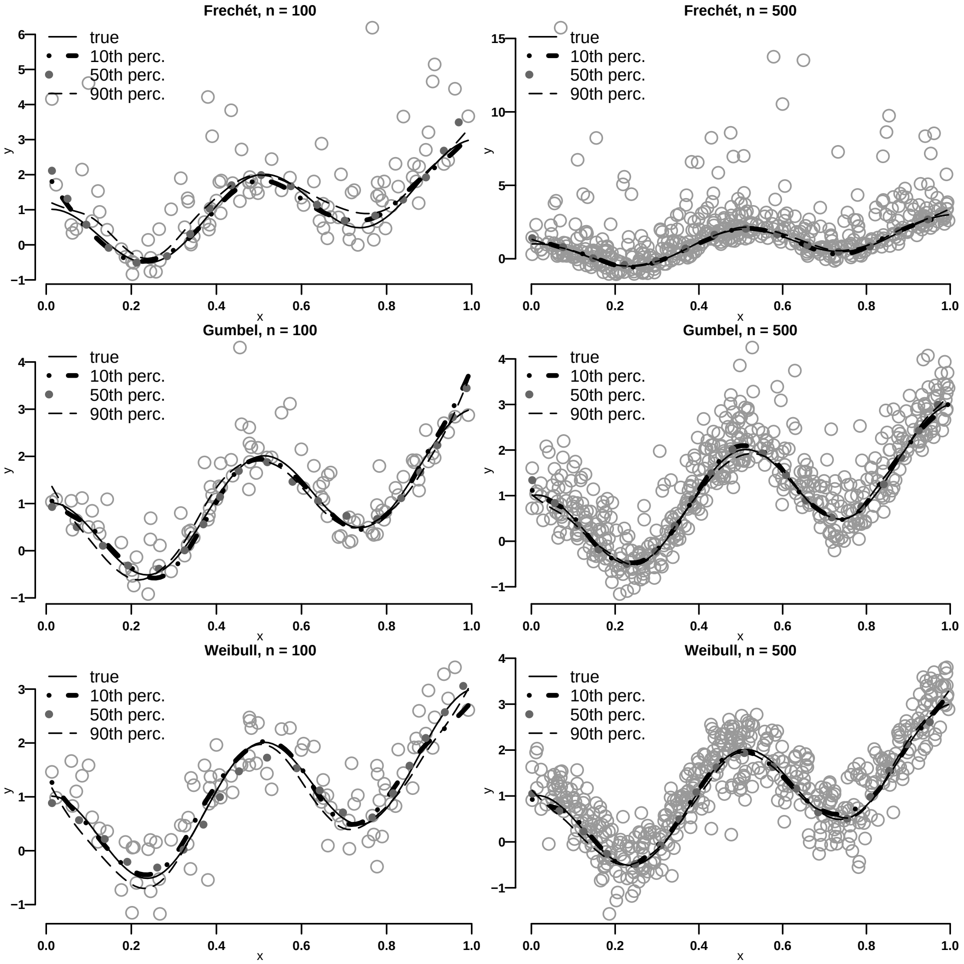

Figure 3.4: *Fitting of the location parameter: the smoothing function μ and its fitting for the three-type distributions and different sample sizes are plotted on the panels.*

in a year are used to compose them. The annual mean of the North Atlantic Oscillation and Southern Oscillation Index are also considered. The North Atlantic Oscillation index measures the difference of mean atmospheric sea-level pressures near the Azores and near Iceland. The Southern Oscillation Index measures the difference of mean atmospheric sea-level pressures near Tahiti and Darwin, Australia. All three daily and monthly series are available on the web respectively from: `http://hadobs.metoffice.com/hadcet/index.html` and `http://www.cru.uea.ac.uk /cru/data/pci.htm`. A large amount of literature has established that the North Atlantic Oscillation has climatic effects on European and North American winters and Southern Oscillation on Australia's climate. From this, we were curious to see if these two environmental processes could have an effect also for the annual maxima temperature. The aim of this analysis is not to provide an exhaustive investigation of the temperature behaviour, but rather to illustrate how our tools can be used to assess dependence of the extreme on covariates. The additive regression model is composed from the maxima temperature which play the role of the response variable, whereas time, North Atlantic Oscillation and southern Oscillation Index are predictors. The relationship is shown in Figure 3.5. The link function is set to be the identity. Firstly, we have fitted the full spline model obtaining the following estimation results: $\hat{\psi} = 1.33\,(0.094)$ and $\hat{\xi} = -0.12\,(0.064)$. Standard errors are reported in parentheses, based on the approximate Fisher information matrix. The variance components estimates of time, North Atlantic Oscillation and Southern Oscillation Index are, respectively: $\hat{\sigma}_T^2 = 0.050\,(0.0465)$, $\hat{\sigma}_N^2 = 0.001\,(0.0015)$ and $\hat{\sigma}_S^2 = 0.001\,(0.0015)$. In Figure 3.5 we plot the fitted maxima trends against the three covariates (from top-left to bottom-left panel). The shaded regions are variability bands; see the Appendix for details. In top-left panel we have the regression of extreme temperatures versus the time. The trend of annual maxima of temperatures is increasing over the whole period. From 1978 to 2006 the range of the maxima temperature trend is about 2 degrees Celsius. Initially the trend has increased from nearly 20 degrees to 21 degrees in about 1991. In around the last 15 years the trend has increased again of the same amount arriving at nearly 22 degrees. Thus it seems from this brief analysis that in recent years the maxima temperature trend has been more rapidly increasing. Differently, the relationship between the annual maxima of temperatures, North Atlantic Oscillation and southern Oscillation Index are only apparent. In fact, the slope coefficients estimates of North Atlantic and Southern Oscillation Index are: $\hat{\beta}_N = -0.10(0.261)$ and $\hat{\beta}_S = -0.128(0.174)$. From this we can conclude that there is no evidence of linear trend between the annual maxima of temperatures and those two predictors. Then, we have investigated a second spline model where we considered the time dependence but we do not take into account the North Atlantic and Southern Oscillation Index covariates. The estimation results of this second model are: $\hat{\psi} = 1.32\,(0.092)$ and $\hat{\xi} = -0.11\,(0.065)$. The estimate of the variance component for the time covariate is resulted: $\hat{\sigma}_T^2 = 0.055(0.049)$, which, for the time covariate, are fairly the same of the previous model. The model fitting is illustrated in Figure 3.6. Finally, we considered a third model defining only a linear trend for the time covariate and without taking into account the smoothing part.

At this stage in order to assess the adequacy of the parametric model we can test the null of the variance components, Ruppert, Wand and Carroll (2003, p. 146). In other words we performed the hypothesis test, $H_0 : \sigma_T^2 = 0$ against $H_1 : \sigma_T^2 > 0$. From the hypothesis test results we can assume that if variance components are zero then a parametric model should be preferred (linear in this case) to the nonparametric alternative. Note that this is not a trivial problem because under the null hypothesis the variance components are on the boundary of their parameter space. For example the standard test performed within the likelihood ratio paradigm does not have the usual chi-squared distribution but rather a mixture of chi-square, Selft and Liang (1987). Moreover, under the mixed model paradigm the random effects induce a dependence factor (e.g. Miller, 1977) and the mixture asymptotic distribution does not hold for penalized spline models, Crainiceanu and Ruppert (2002). In order to test the null hypothesis we used the likelihood ratio test, but for the reasons just discussed we determine the null distribution of the likelihood ratio test statistic by a simulation-based alternative. The critical value of the test has been obtained by Monte Carlo simulation. More precisely, for the data vector $\mathbf{y}$ and the variance component $\theta = \sigma_T^2$, the ratio test statistic is

$$\mathrm{LRT}(\mathbf{y}) = 2\{\ell(\widehat{\theta}; \mathbf{y}) - \ell(\widehat{\theta}_0; \mathbf{y})\}, \tag{3.4}$$

where $\widehat{\theta}_0$ maximizes the penalized log likelihood under the null hypothesis that the variance components could be removed from the model, and $\widehat{\theta}$ under the alternative. We compute the statistic (3.4) with the observed data, we say $\mathrm{LRT}(\mathbf{y}^{obs})$. Fixing the model parameter equal to $\widehat{\theta}_0^{obs}$, the maximum likelihood estimates ob-

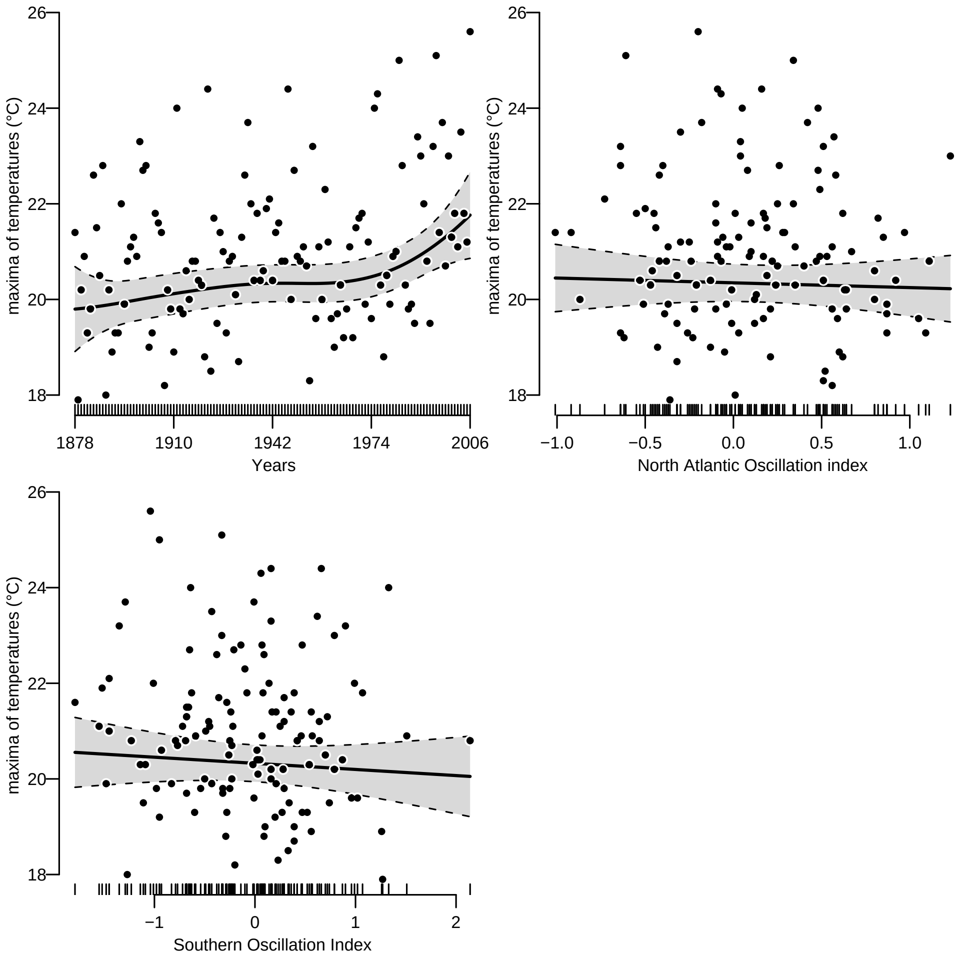

Figure 3.5: *Central England Temperatures example. The three panels are from the top-left to the bottom-left panels, respectively, annual maxima of temperatures versus time, North Atlantic Oscillation and Southern Oscillation Index. Continuous lines express the fitting trend and the shaded regions are variability bands.*

tained under null hypothesis with the observed data, we simulate $M = 10000$ synthetic data from the spline mixed-model for extremes (under the null hypothesis it consists of a GEV model with linear trend). Then for each simulated data we estimate the smoothing, the GEV and the dispersion parameters according to the models under the null and the alternative hypotheses and so we compute the test statistic LRT($\mathbf{y}^{sim}$) by using (3.4). In this way we obtain a sequence of values that simulate the distribution of the likelihood ratio test under the null hypothesis. Finally the p-value of the test is the proportion of simulated values LRT($\mathbf{y}^{sim}$) that exceed the statistic computed with the real data. In other words

$$p\text{-value} = \frac{\sum_{m=1}^{M} I\{\text{LRT}(\mathbf{y}^{sim}) > \text{LRT}(\mathbf{y}^{obs})\}}{M},$$

where $I\{B\}$ is the indicator function of the set B. Using this simulation-based method we found a p-value = 0.0002. We conclude that the null hypothesis of linearity (H_0) should be rejected, given that the observed statistic LRT($\mathbf{y}^{obs}$) is in the upper tail of the null simulated distribution.

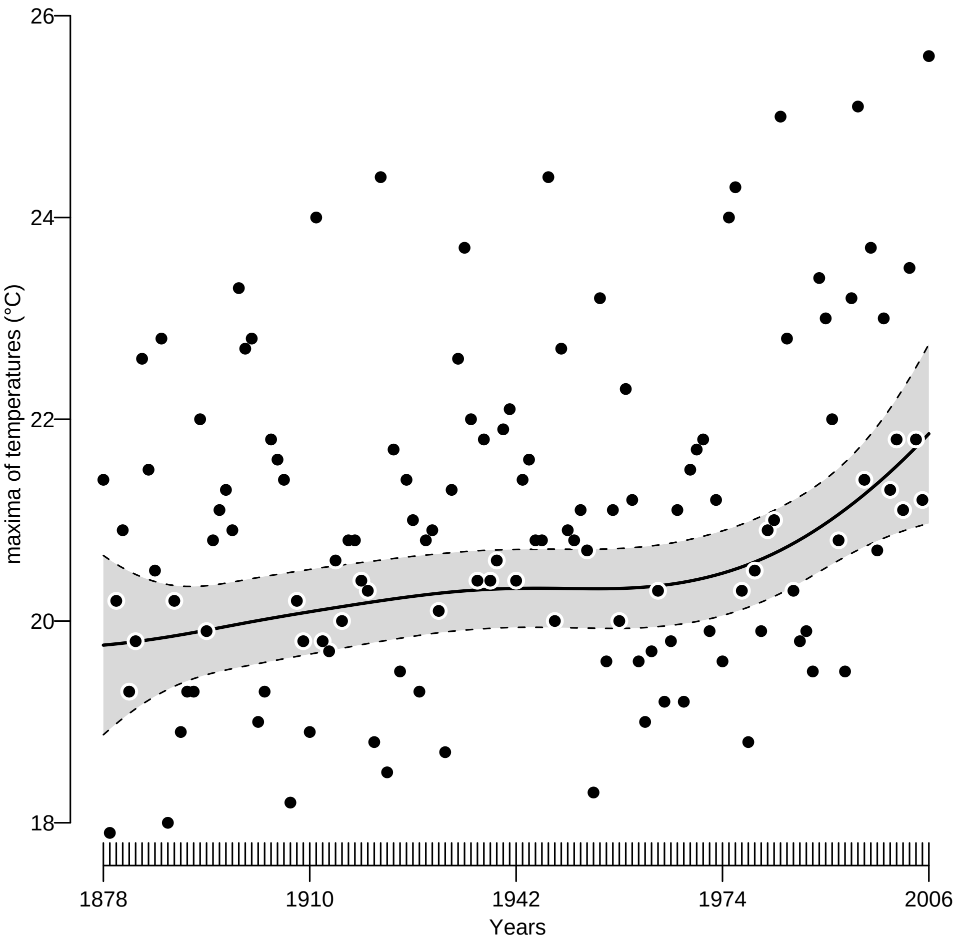

Figure 3.6: *Central England Temperatures example. The panel shows the annual maxima of temperatures versus the time. Continuous lines express the fitting trend and the shaded regions are variability bands.*

3.5.2 Application to Swiss temperature data

In the second case we studied the maximum temperatures of Switzerland. The data sequences consists of daily maximum temperatures in degrees Celsius recorded in the Zuerich city from 1901 to 2006. The daily precipitation amount in meters was relieved in the same city. These data series and others are available from the website of the European Climate Assessment & Dataset (ECA&D) project at url: `http://eca.knmi.nl/`. This meteorological institute provide some indices for monitoring and analysing changes in climate extremes, as well as the daily dataset needed to calculate these indices.

In principle the aim of the analysis has been to investigate the presence of a temporal trend on the maxima temperatures. The non-stationarity hypothesis of the extreme temperatures seems acceptable and supported from the pattern of the real data, see left panel of Figure 3.7. Establishing if the trend is real or only apparent and describing the correct pattern of the process is not so immediate and simple. This is because often the assumptions at the base of the parametric regression model might be too restrictive in order to describe the trend of environmental processes. For this reason we approach the problem through semiparametric regression adopting the extreme value model as a basic template. This methodology can provides, rather than an exhaustive investigation of the temperature behaviour, an appropriate model in order to assess the dependence of the extremes on covariates. An explorative graphic analysis confirmed also the presence of a dependence between the maxima temperatures and the total annual amount of precipitation recorded (large rainfall episodes have a decreasing temperature effect), see right panel of Figure 3.7. Assuming the dependence between the these two processes seems realistic as much as the time-nonhomogeneous assumption. Thus we manage both the time and rainfall dependence in a unified framework. In particular we specify an additive model for sample extremes where the maxima temperatures play the role of the response variable, whereas the time and the precipitation amounts are the predictors. Variations through time and precipitation amounts are modeled as penalized splines in the location parameter of the appropriate extreme value model, instead by contrast the shape and scale parameters are assumed constant. In particular we have used radial basis functions with 20 knots. The link function is set to be the identity, so that the setup is the same as the one described in the numerical example section. The estimates and standard error results of scale and shape parameters are: $\hat{\psi} = 1.564\,(0.125)$ and $\hat{\xi} = -0.216\,(0.079)$. These indicate mild data variability. Standard errors are reported in parentheses, based on the approximate Fisher information matrix. The variance components estimates of time and annual precipitation amounts are, respectively: $\hat{\sigma}^2_T = 0.82\,(0.54)$, $\hat{\sigma}^2_P = 0.28\,(0.18)$.

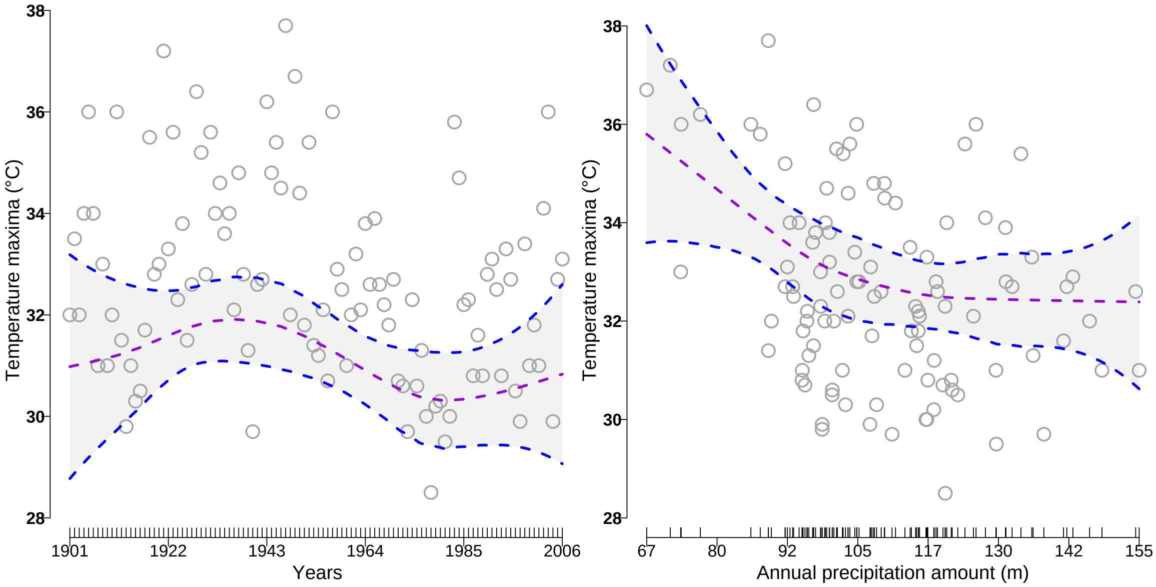

Figure 3.7: *Switzerland Temperatures example. The circles into the two panels are, respectively, maxima annual temperatures versus time and maxima annual temperatures versus total annual amount of precipitation. Continuous lines illustrate the fitting trend and the shaded regions and broken lines are variability bands.*

In Figure 3.7 we plot the fitted maxima trends against the two covariates (time in the left panel, precipitation amount in the right panel). The shaded regions are variability bands; see the Section 3.2.2 for details. The left panel suggests variations of the trend maxima in time, similar to a trigonometric function. The right panel suggests a decreasing variation of the trend maxima with the increasing of the precipitation amount as expected. The relationship does not seem to be linear evident, the maxima temperature trend decreases with the increases of the precipitation amount as expected, but for larger values of the latter the trend decrease with smaller rates.

We have assessed the adequacy of the parametric model by testing the null of the variance components as in the previous section. In other words we performed the hypothesis test, $H_0 : \sigma_T = \sigma_P = 0$ versus $H_1 : \sigma_T > 0, \sigma_P > 0$. The critical value of the test has been obtained by Monte Carlo simulation. More precisely, for the data vector $\mathbf{y}$ and parameter vector $\boldsymbol{\sigma} = (\sigma_T, \sigma_P)$, the ratio test statistic is

$$\mathrm{LRT}(\mathbf{y}) = 2\{\ell(\widehat{\boldsymbol{\sigma}}; \mathbf{y}) - \ell(\widehat{\boldsymbol{\sigma}}_0; \mathbf{y})\}, \tag{3.5}$$

where $\widehat{\boldsymbol{\sigma}}_0$ maximizes the penalized log likelihood under the null hypothesis that the variance components could be removed from the model, and $\widehat{\boldsymbol{\sigma}}$ under the alternative. We compute the statistic (3.5) with the observed data, we say $\mathrm{LRT}(\mathbf{y}^{obs})$. Fixing the model parameter equal to $\widehat{\boldsymbol{\sigma}}_0^{obs}$, the maximum likelihood estimates obtained under null hypothesis with the observed data, we simulate $M = 1000$ synthetic data from the spline mixed-model for extremes (under the null hypothesis it consists of a GEV model with linear trend). Then for each simulated data we estimate the smoothing, the GEV and the dispersion parameters according to the models under the null and the alternative hypotheses and so we compute the test statistic $\mathrm{LRT}(\mathbf{y}^{sim})$ by using (3.5). In this way we obtain a sequence of values that simulate the distribution of the likelihood ratio test under the null hypothesis. Finally the p-value of the test is the proportion of simulated values $\mathrm{LRT}(\mathbf{y}^{sim})$ that exceed the statistic computed with the real data. In other words

$$p\text{-value} = \frac{\sum_{m=1}^{M} I\{\mathrm{LRT}(\mathbf{y}^{sim}) > \mathrm{LRT}(\mathbf{y}^{obs})\}}{M},$$

where $I\{B\}$ is the indicator function of the set B. Using this simulation-based method we found a p-value = 0.027. We conclude that the null hypothesis of linearity (H_0) should be rejected, given that the observed statistic $\mathrm{LRT}(\mathbf{y}^{obs})$ is in the upper tail of the null simulated distribution.

Chapter 4

Model extension to the scale function

4.1 Introduction

In the previous chapter we discussed the statistical modelling and inference of nonparametric regression for GEV models. Basically we focused on the estimation of regression functions for the location model parameter as a function of predictors $\mathbf{x}_i$. The main properties of the antecedent framework are: the regression functions have been modeled by linear or nonlinear regression splines, errors have been assumed to have a GEV distribution, and scale and shape parameters have been considered unchanged respect with the predictors. The model description is shortened and synthesized by saying

$$y_i \sim \mathrm{GEV}(\mu(\mathbf{x}_i), \psi, \xi) \quad -\infty < \mu(\mathbf{x}_i),\, \xi < \infty, \quad \psi > 0; \quad \mathbf{x}_i \in \mathbb{R}^d,$$

where $\mu(\mathbf{x}_i)$ is modeled by a spline model. It seems plausible that trends appear in real data accompanied with scale variations and assuming, however the distribution, inalterability. Usually this remark is supported by the real data pattern. This is apparently the case with the data illustrated in Figure (3.7). In other words, we can still suppose the GEV distribution as the appropriate model, but assuming its location and scale ψ changes as the predictor changes. The GEV shape model parameter is difficult to estimate precisely, and also the regularity conditions of the maximum likelihood estimator and the obtainable estimates are restricted as they depend on the shape parameter value assumed, Smith (1985). So, models that alow the shape parameter to be modeled as a smooth function of predictors could be unreliable in the inference stage. Nonetheless, some authors such Yee and Stephenson (2007) have explored the alternative of modelling all three GEV parameters using spline models. However, here we have not yet considered the opportunity to also model the shape parameter. Instead, we focus on the resulting model formed by taking into account location and scale parameters which are modeled by nonparametric regressions.

4.2 Extension to scale parameter

The model extension is formulated as follows. We describe initially the univariate predictor case. Let $y_1, \ldots, y_n$ be n observed extreme values associated with explanatory variables x_i, where $i = 1, \ldots, n$ and $n \in \mathbb{N}$. As in Chapter 3, Section 3.2, we assumed that the block maxima $Y_1, \ldots, Y_n$ are independent variables from a GEV distribution. Where i's random variable Y_i represents the maximum of a process (daily temperatures, etc.) over m time units of observations. So that, for instance if m is the number of days in a year we deal with the annual maxima. Assume also that the location μ and scale ψ parameters of the GEV distribution are undefined smooth functions on an interval $[a, b]$, where the x's are defined. Then the time-nonhomogeneous spline mixed model, or more generally we say the spline mixed model for extreme values is given by

$$y_i \sim \text{GEV}(\mu(x_i), \psi(x_i), \xi) \quad -\infty < \mu(x_i),\, \xi < \infty,\ \psi(x_i) > 0;\ x_i \in \mathbb{R}. \tag{4.1}$$

Where the mixed model-based penalized spline models for the location parameter μ is defined by

$$\mu(x) = \alpha_0 + \alpha_1 x + \sum_{k=1}^{K_\mu} u_k z_k(x); \quad u_1, \ldots, u_k \quad \text{i.i.d.} \quad N(0, \sigma_u^2) \tag{4.2}$$

and for the scale parameter ψ by

$$\psi(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K_\psi} v_k z_k(x); \quad v_1, \ldots, v_k \quad \text{i.i.d.} \quad N(0, \sigma_v^2). \tag{4.3}$$

where $z_1, \ldots, z_K$ is an appropriate set of spline basis functions that depend on a dense set of knots $\kappa_1, \ldots, \kappa_K$ within the range of the x_i's. Note that the spline basis functions used for $\mu(x)$ could be different from that used for $\psi(x)$, similarly for the number of knots that we have opportunely indicated by K_μ and K_ψ.

More precisely, let $\mathbf{y} = (y_1, \ldots, y_n)$ be the response vector for which we define the design matrices

$$\mathbf{X} = [1 \quad x_i]_{1\le i\le n}, \qquad \mathbf{Z}_\mu [\, z_k(x_i)\,]_{\substack{1\le i\le n \\ 1\le k\le K_\mu}} \qquad \mathbf{Z}_\psi [\, z_k(x_i)\,]_{\substack{1\le i\le n \\ 1\le k\le K_\psi}}$$

associated with fixed $\boldsymbol{\alpha} = [\alpha_0\ \alpha_1]^T$ and random effects $\mathbf{u} = [u_1 \ldots u_{K_\mu}]^T$ for the location function, and fixed $\boldsymbol{\beta} = [\beta_0\ \beta_1]^T$ and random effects $\mathbf{v} = [v_1 \ldots v_{K_\psi}]^T$ for the scale function. Given $(\mathbf{u}, \mathbf{v})$, the y_i are conditionally independent with distribution,

$$y_i|(\mathbf{u}, \mathbf{v}) \sim \text{GEV}(\boldsymbol{\mu}_i, \boldsymbol{\psi}_i, \xi),$$

where the linear predictor $\boldsymbol{\eta}_i(\mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}_\mu \mathbf{u})_i$ is related to $\boldsymbol{\mu}_i$ by the link function $g(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i$ and the linear predictor $\boldsymbol{\gamma}_i(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_\psi \mathbf{v})_i$ is related to $\boldsymbol{\psi}_i$ by the link function $h(\boldsymbol{\psi}_i) = \boldsymbol{\gamma}_i$. Note that $\boldsymbol{\mu}$ and $\boldsymbol{\psi}$ are related to the conditional mean and the variance of $\mathbf{y}$ given $(\mathbf{u}, \mathbf{v})$ by the relations

$$\text{E}(\mathbf{y}|\mathbf{u}) = \begin{cases} \boldsymbol{\mu} + \mathbf{1}\psi\{\Gamma(1-\xi) - 1\}/\xi, & \text{for} \quad \xi \neq 0 \\ \boldsymbol{\mu} + \mathbf{1}\psi\gamma, & \text{for} \quad \xi = 0 \end{cases}$$

and

$$\text{V}(\mathbf{y}|\mathbf{u}, \mathbf{v}) = \begin{cases} \boldsymbol{\psi}\mathbf{1}(\Gamma(1-2\xi) - \Gamma(1-\xi)^2)/\xi & \text{for} \quad \xi \neq 0 \\ \boldsymbol{\psi}^2\pi^2/6 & \text{for} \quad \xi = 0, \end{cases}$$

where Γ is the Gamma function.

Note that both $(\mathbf{u}, \mathbf{v})$ have been assumed normally distributed so we have the doubled random effects set given by

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} \sim \text{N}\Bigg(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_u^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_v^2 \mathbf{I} \end{bmatrix} \Bigg).$$

Let $\boldsymbol{\nu}^T = [\boldsymbol{\alpha}^T\, \mathbf{u}^T]$ be the $K_\mu + 2$ vector of fixed and random effects associated with the design matrix $\mathbf{C}_\mu = [\mathbf{X}|\mathbf{Z}_\mu]$ of the location function, and $\boldsymbol{\tau}^T = [\boldsymbol{\beta}^T\, \mathbf{v}^T]$ be the $K_\psi + 2$ vector of fixed and random effects associated with the design matrix $\mathbf{C}_\psi = [\mathbf{X}|\mathbf{Z}_\psi]$ of the scale function. Suppose that the link functions g and h are respectively the identity and the natural logarithm. Considering the previous specifications and denoting with $f(y_i|(\mathbf{u},\mathbf{v}))$ the GEV conditional density of $y_i|(\mathbf{u},\mathbf{v})$ and with $f(\mathbf{u},\mathbf{v})$ the multivariate unconditional density of the random effects $(\mathbf{u},\mathbf{v})$, then those density functions have expressions

$$f(\mathbf{u},\mathbf{v};\sigma_u^2,\sigma_v^2) = (2\pi)^{-(K_\mu+K_\psi)/2}(\sigma_u)^{-K_\mu}(\sigma_v)^{-K_\psi}\exp\left\{-\frac{1}{2}\left(\frac{\|\mathbf{u}\|^2}{\sigma_u^2}+\frac{\|\mathbf{v}\|^2}{\sigma_v^2}\right)\right\} \quad \text{and}$$

$$\begin{aligned} f(\mathbf{y}|\mathbf{u},\mathbf{v};\boldsymbol{\alpha},\boldsymbol{\beta},\xi) &= \prod_{i=1}^{n}\frac{1}{\exp(\mathbf{C}_\psi\boldsymbol{\tau})_i}\left\{1+\xi\left(\frac{(\mathbf{y}-\mathbf{C}_\mu\boldsymbol{\nu})_i}{\exp(\mathbf{C}_\psi\boldsymbol{\tau})_i}\right)\right\}^{-\frac{1}{\xi}-1} \\ &\quad \exp\left[-\left\{1+\xi\left(\frac{(\mathbf{y}-\mathbf{C}_\mu\boldsymbol{\nu})_i}{\exp(\mathbf{C}_\psi\boldsymbol{\tau})_i}\right)\right\}^{-\frac{1}{\xi}}\right]. \end{aligned}$$

The norm for fitting such a model is given by the estimation of the model parameters via maximization of the likelihood,

$$\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\beta},\xi,\sigma_u^2,\sigma_v^2) = \int_{\mathbb{R}^{K_\mu}}\int_{\mathbb{R}^{K_\psi}} f(\mathbf{y}|(\mathbf{u},\mathbf{v}))f(\mathbf{u},\mathbf{v})\,d\mathbf{u}\,d\mathbf{v},$$

and prediction of the random effects $(\mathbf{u},\mathbf{v})$ via the best predictor $(\widehat{\mathbf{u}},\widehat{\mathbf{v}}) = E((\mathbf{u},\mathbf{v})|\mathbf{y})$. However, as we have already discussed in Chapter 3, Section 3.2 both the likelihood function and the best predictor's analytical expressions can not be easily determined due to the intractable high dimensional integrals. We found that the solution previously proposed in order to provide the model fitting and the parameter estimates (3, Section 3.2–3.2.1), performs poorly considering the scale extension. This result may be due to the scarce approximation that Laplace's method provided for the likelihood function. Although, we have not yet conducted an in depth study on this issue. Further studies should be undertaken in order to conclude if model fitting and model assessments can be provided based on the likelihood approach. Nonetheless, the model fitting and inference for the model (4.1) can be performed by alternatively using the Bayesian approach, in particular via the application of Markov Chain Monte Carlo methods. In the next section we will discuss the details for the model fitting and inference of the spline mixed model for extremes.

Finally note here that the additive models extension has not been discussed, as in Chapter 3, Section 3.2.3. This is because the same rules outlined for the location function can be applied straightforwardly with typographic modifications to the scale function.

4.3 Bayesian Analysis and Markov Chain Monte Carlo

In a mixed model some parameters are treated as random so that in some sense it is like we have specified prior densities for some model parameters. In Chapter 3 we saw that treating the random coefficients as nonrandom unknown quantities (model parameters) we have been able to conduct the model fitting and inference based on the likelihood approach. The mixed model representation of the penalized splines (4.1) consists of defining spline models, specifying the priors on $(\mathbf{u}, \mathbf{v})$ as well as the likelihood $f(\mathbf{y}|\mathbf{u}, \mathbf{v}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_u^2, \sigma_v^2)$. Now assuming that in the spline mixed model framework the fixed effects, the GEV shape parameter and the variance components $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \xi, \sigma_u^2, \sigma_v^2)$, are all random we can fully formulate a Bayesian model.

More precisely, the spline mixed model for extremes (4.1) can be naturally formulated as a hierarchical model under the Bayesian paradigm. The hierarchical Bayes model is constructed by arranging random variables in a hierarchy so that distributions at each level are determined by the random variable of the previous lower levels. In the first stage of the hierarchical model, the GEV distribution for $y_i|(\mathbf{u}, \mathbf{v})$ is set up given the fixed $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and random $(\mathbf{u}, \mathbf{v})$ effects. In the second stage, it is assumed the prior distributions $N((\boldsymbol{a}, \mathbf{b}), \mathbf{G})$ for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and F_ξ for ξ which depend by other extra hyperparameters, and the distribution $N(0, \mathbf{R})$ for $(\mathbf{u}, \mathbf{v})$ that depends on the variance components $\mathbf{R}$. Finally, in the last stage, the prior distribution $F_{\mathbf{R}}$ is assumed for the variance components $\mathbf{R}$. The hierarchical spline model for extremes is synthesized by

1. $y_i \sim \mathrm{GEV}\{(\mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}_\mu \mathbf{u})_i,\ \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_\psi \mathbf{v})_i,\ \xi\}$;

2. $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \sim N((\boldsymbol{a}, \mathbf{b}), \mathbf{G}), \quad (\mathbf{u}, \mathbf{v}) \sim N(0, \mathbf{R})\,\text{and}\,\xi \sim F_\xi$;

3. $\mathbf{R} \sim F_{\mathbf{R}}$,

where it is assumed $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are independent, with $(\boldsymbol{a}, \mathbf{b})$ and $\mathbf{G}$ known. It is assumed also that $\mathbf{u}$ and $\mathbf{v}$ are independent with variance components $\mathbf{R}$ that include the elements (σ_u^2, σ_v^2) but can also include more components in the case where many predictors will impact on the response. Consequently the same extension is also expected for the matrix $\mathbf{G}$ corresponding to $(\boldsymbol{\alpha}, \boldsymbol{\beta})$.

The inference procedure includes the fixed and random effects, and that about the GEV shape parameter and the variance components. First define the likelihood function under the Bayesian framework. Suppose that, given $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v}, \xi, \mathbf{R})$, $\mathbf{y} \sim f(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v}, \xi, \mathbf{R})$ that is the conditional GEV density function. Furthermore, suppose that, given $\mathbf{R}$ the random effects $(\mathbf{u}, \mathbf{v})$ are independent of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and also that $\mathbf{u}$ is independent of $\mathbf{v}$, and $\boldsymbol{\alpha}$ is independent of $\boldsymbol{\beta}$. Then the likelihood function for estimating $\mathbf{R}$ is given by

$$L(\mathbf{R}|\mathbf{y}) = \int\!\!\int\!\!\int f(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v}, \xi, \mathbf{R})\phi(\boldsymbol{\alpha}, \boldsymbol{\beta}|(\boldsymbol{a}, \mathbf{b}), \mathbf{G})\phi(\mathbf{u}, \mathbf{v}|\mathbf{R})\,\pi(\xi)\mathrm{d}\boldsymbol{\alpha}\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{v}\,\mathrm{d}\xi,$$

where the integrals with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\mathbf{u}$ and $\mathbf{v}$ are multivariate. Now if the prior is taken into account, then the posterior for $\mathbf{R}$ can be expressed as

$$f(\mathbf{R}|\mathbf{y}) = \frac{L(\mathbf{R}|\mathbf{y})\,\pi(\mathbf{R})}{\int L(\mathbf{R}|\mathbf{y})\,\pi(\mathbf{R})\,\mathrm{d}\mathbf{R}}, \tag{4.4}$$

where $\pi(\mathbf{R})$ is a prior density function for $\mathbf{R}$. Similar to (4.4) we can obtain the posterior densities $f((\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v})|\mathbf{y})$ for $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v})$ and $f(\xi|\mathbf{y})$ for ξ.

Alternatively, we can consider jointly the full parameter set $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v}, \xi, \mathbf{R})$ and then derive the posterior density, $f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v}, \xi, \mathbf{R}|\mathbf{y})$. So, taking into account the priors densities of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v}, \xi, \mathbf{R})$, their posterior can be formulated as

$$f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v}, \xi, \mathbf{R}|\mathbf{y}) = k\, f(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v}, \xi, \mathbf{R})\, \varphi(\boldsymbol{\alpha}, \boldsymbol{\beta}|(\boldsymbol{a}, \mathbf{b}), \mathbf{G})\, \varphi(\mathbf{u}, \mathbf{v}|\mathbf{R})\, \pi(\mathbf{R})\, \pi(\xi),$$

where

$$\begin{aligned} k &= \Bigg(\int\!\!\int\!\!\int\!\!\int\!\!\int\!\!\int f(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v}, \xi, \mathbf{R})\, \varphi(\boldsymbol{\alpha}, \boldsymbol{\beta}|(\boldsymbol{a}, \mathbf{b}), \mathbf{G})\, \varphi(\mathbf{u}, \mathbf{v}|\mathbf{R}) \\ &\qquad \pi(\mathbf{R})\, \pi(\xi)\, \mathrm{d}\boldsymbol{\alpha}\, \mathrm{d}\boldsymbol{\beta}\, \mathbf{du}\, \mathbf{dv}\, \mathbf{dR}\, \mathrm{d}\xi \Bigg)^{-1}, \end{aligned}$$

φ denotes the multivariate normal density function for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $(\mathbf{u}, \mathbf{v})$ and π denotes the priory density function for $\mathbf{R}$ and ξ (particular forms will be specified later). The computation of the posterior density function can be fairly complicated even for simple models, as for example with linear mixed models. In our case the posterior densities involve integrals that are analytically intractable. For complex models, as in our case, the computation of the posterior density is typically carried out by Markov chain Monte Carlo (MCMC) methods. For example the quantity k is a constant of proportionality, which cannot be computed easily but is still required for inference. The MCMC methods allow sampling from the posterior distribution. The idea is to sample from a chain whose stationary distribution is equal to the posterior. Essentially, the MCMC methods work to divide the model parameters into subsets, and then aim to draw a sample from the conditional distributions given the remaining parameters and data, Zhao, Staudenmayer, Coull and Wand (2006). In our case the parameter set could be broken down into $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v})$, $\mathbf{R}$ and ξ, leading to the conditionals $f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v}|\mathbf{R}, \xi, \mathbf{y})$, $f(\mathbf{R}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v}, \xi, \mathbf{y})$ and $f(\xi|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v}, \mathbf{R}, \mathbf{y})$. By Monte Carlo methods we sample from a density function that is known only up to a constant of proportionality, so that roughly speaking we sample from the posterior density without calculating the proportionality constant (the denominator of the Bayes formula), Tierney (1994). However, to sample from our conditional densities could be somewhat difficult because they are not in any standard family. Then we need to adopt some strategies for attacking the problem in order to provide a suitable solution. Complex algorithms such as Metropolis-Hastings (Hastings 1970), the adaptive rejection sampling (Gilks and Wild, 1992), slice sampling (Besag and Green, 1993) or the Gibbs sampling (Casella and Edward, 1992) can be useful tools in order to complete the sampling scheme.

We do not discuss here any particular algorithm or which could be the more appropriate for our study. For a general discussion we refer to Casella and Edward (1992). Instead, aside discussions about the implementation, we prefer to focus on the model description and data analysis. Then, the simplest approach is to implement the penalized spline mixed model for extremes by using the `OpenBUGS` package (the open source version of `WinBUGS`), which is based on the Gibbs sampling algorithm. Many sophisticated versions are available. In particular we used the `R` interface `BRugs` (see `http://mathstat.helsinki.fi/openbugs/`). Previous works on Bayesian analysis for penalized spline regression by using `WinBUGS` have been explored by Crainiceanu, Ruppert and Wand (2005) and Zhao, Staudenmayer, Coull and Wand (2006).

Now assuming that only a predictor x_i impacts on the response y_i (the simplest case), in order to fit the model (4.1) with `BRugs` we need to specify the prior distributions, the spline basis functions and the number and location of the knots. In particular, we assume independence of the random effects $\mathbf{u}, \mathbf{v}$ so that the matrix $\mathbf{R}$ is diagonal with elements σ_u^2 and σ_v^2. We consider also that the variance components σ_u^2 and σ_v^2 are independent with prior inverse gamma density,

$$\pi(\sigma^2) = \frac{\eta^\lambda}{\Gamma(\lambda)}(\sigma^2)^{-(\lambda+1)} \exp\left(\frac{-\eta}{\sigma^2}\right), \quad \sigma^2 > 0, \quad \eta, \lambda > 0.$$

We take the prior distribution of the fixed effects vector $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ to be of the form $N((\boldsymbol{a}, \mathbf{b}), \mathbf{G})$ for some values $(\boldsymbol{a}, \mathbf{b})$ and covariance matrix $\mathbf{G}$. We have assumed also that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are independent. Finally, we assume an uniform distribution for ξ of the form $U(l, u)$. The hierarchical Bayes model for `OpenBUGS` is described by the following scheme

$$\begin{array}{ll}
\text{1st level} & L_i = \frac{1}{\psi_i}\left\{1 + \xi\left(\frac{y_i-\mu_i}{\psi_i}\right)\right\}^{-\frac{1}{\xi}-1} \exp\left[-\left\{1 + \xi\left(\frac{y_i-\mu_i}{\psi_i}\right)\right\}^{-\frac{1}{\xi}}\right] \\
 & \mu_i = (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i \\
 & \log\psi_i = (\mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{v})_i \\
 & \\
\text{2nd level} & (\boldsymbol{\alpha}, \boldsymbol{\beta}) \sim N((\boldsymbol{a}, \mathbf{b}), \mathbf{G}) \\
 & (\mathbf{u}, \mathbf{v}) \sim N(0, \mathbf{R}) \\
 & \xi \sim U(l, u) \\
 & \\
\text{3rd level} & \sigma_u^2 \sim IG(\lambda_u, \eta_u) \\
 & \sigma_v^2 \sim IG(\lambda_v, \eta_v)
\end{array}$$

At the bottom of the hierarchy are the variance components whose distributions depend on the known hyperparameters. At the next level are the fixed and random effects and the GEV shape parameter, whose distributions depend on the variance components and the hyperparameters. The top level contains the data $\mathbf{y}$.

For the practice implementation, the covariance matrix $\mathbf{G}$ is taken to be diagonal with very large entries, so that each entry of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ corresponds to noninformative priors. Similarly, for ξ the interval (l, u) of the prior uniform distribution is set large enough so that it corresponds to a noninformative prior. Finally, for the scale and shape parameters of the inverse gamma (IG) densities, both take small values such as 0.01 for the same reasons discussed previously. With those values, stable model fitting is also guaranteed, as suggested by the sensitivity investigation of Zhao, Staudenmayer, Coull and Wand (2006).

We used radial cubic basis functions for smooth function components. They have the advantage of requiring a relatively small number of knots in order to obtain a smooth function. Also, they have shown to posses good mixing properties in MCMC analysis (Crainiceanu, Ruppert and Wand, 2005). Consider the spline model (4.2) for the locations parameter and (4.3) for the scale. The radial cubic basis functions (Ruppert, Wand and Carrol, 2003, p. 72) take the generic form

$$\mathbf{Z}[\underset{1\le k\le K}{|x_i - \kappa_k|^3}\; \underset{1\le k,k'\le K}{|\kappa_k - \kappa_{k'}|^3}]^{-1/2}_{1\le i\le n}.$$

where $\{\kappa_k\}$ is a sequence of knots for $k = 1, \ldots, K$, $|x_i - \kappa_k|^3$ are basis functions and $k \le k'$.

Empirically we found that for our model with the MCMC approach a relatively small number of knots is required, we say 10. A greater number do not have an evident impact on the model fitting but involve lower convergence of the chains. For the location of the knots we used the rule reported in Section 2.1. Taking into account the Bayesian approach, we explore the MCMC model fitting

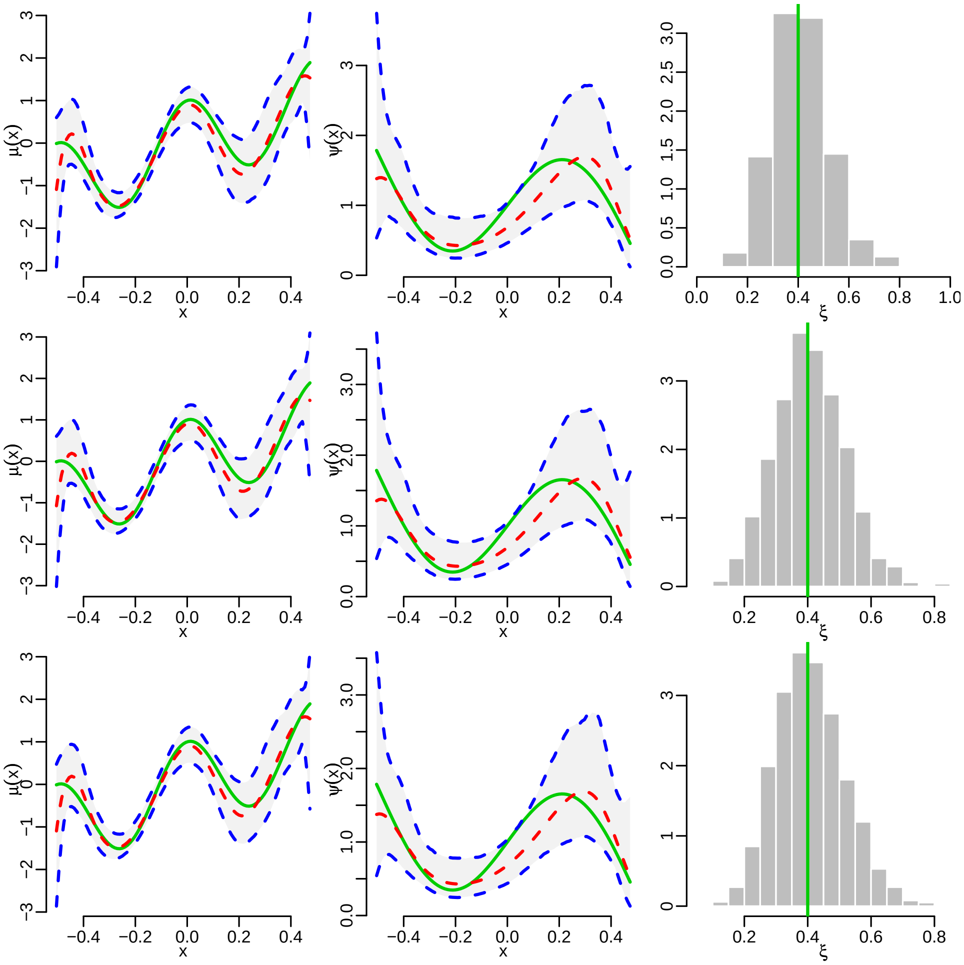

Figure 4.1: *Simulation results: horizontally the plot shows the estimate results of three different chains. The first column shows with the red broken lines the regression function estimates of the location parameter. Instead in the second column the scale parameters are illustrated. The blue broken lines are the 95% credibility intervals, while the true signal is represented with the solid green line. The third column shows the histograms of the posterior density for the shape parameters.*

and inference of the penalized spline mixed model for extremes by some simulation examples. For instance, we have considered the following simulation design. We set up for the location and scale the functions:

$$\mu(x) = 2x + \cos(4\pi x) \quad \text{and} \quad \psi(x) = 1 - 1.5x + \sin(2\pi x), \quad -0.5 \leq x \leq 0.5.$$

We have simulated $n = 100$ values for the covariate x from the uniform distribution, $x_i \sim U(-0.5, 0.5)$. Fixing the shape parameter, $\xi = 0.4$, then we have drawn a sample from a GEV distribution with the above parameters, $y_i \sim \text{GEV}(\mu(x_i), \psi(x_i), \xi)$.

Then we have run a `BRugs` script. We found that a burn in period of length 5000 and keeping 10000 values from the chain with a thinning factor of 5 was sufficient to produce acceptable convergence. A rigorous method that demonstrates the convergence of the chain does not exist. Nonetheless, some diagnosis attempts for the convergence of the chain can be provided. For instance, multiple chains can be run simultaneously from different starting values and then the results of the simulations are compared. In Figure (4.1) the results of three different chains are illustrated in horizontal rows. In the first and second columns the location and scale function's estimates and their 95% credibility intervals are reported. The estimates of the location and scale functions are obtained taking the posterior mean of the regression coefficients using 2000 simulated chain values. In other words, we compute as a point estimate of the true curves: $\widehat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\alpha}} + \mathbf{Z}_{\mu}\widehat{\mathbf{u}}$ and $\widehat{\boldsymbol{\psi}} = \exp(\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}_{\psi}\widehat{\mathbf{v}})$, where the estimates $\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{u}}, \widehat{\mathbf{v}}$ are given by the posterior means. The $(1-\alpha)\%$ credibility intervals are obtained computing the distributions of $\boldsymbol{\mu}$ and $\boldsymbol{\psi}$ with 2000 simulated chain values for the regression coefficients $\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{u}$ and $\mathbf{v}$. Then, for each predictor value we take the quantiles of level α and $1-\alpha$ of the resulting distributions. In the third column histograms of the shape parameter are given. From Figure (4.1) we can see that the three different chains provide, in practice fairly indistinguishable estimate functions and very close posterior distributions to the shape parameter. In Figure (4.2) we have also

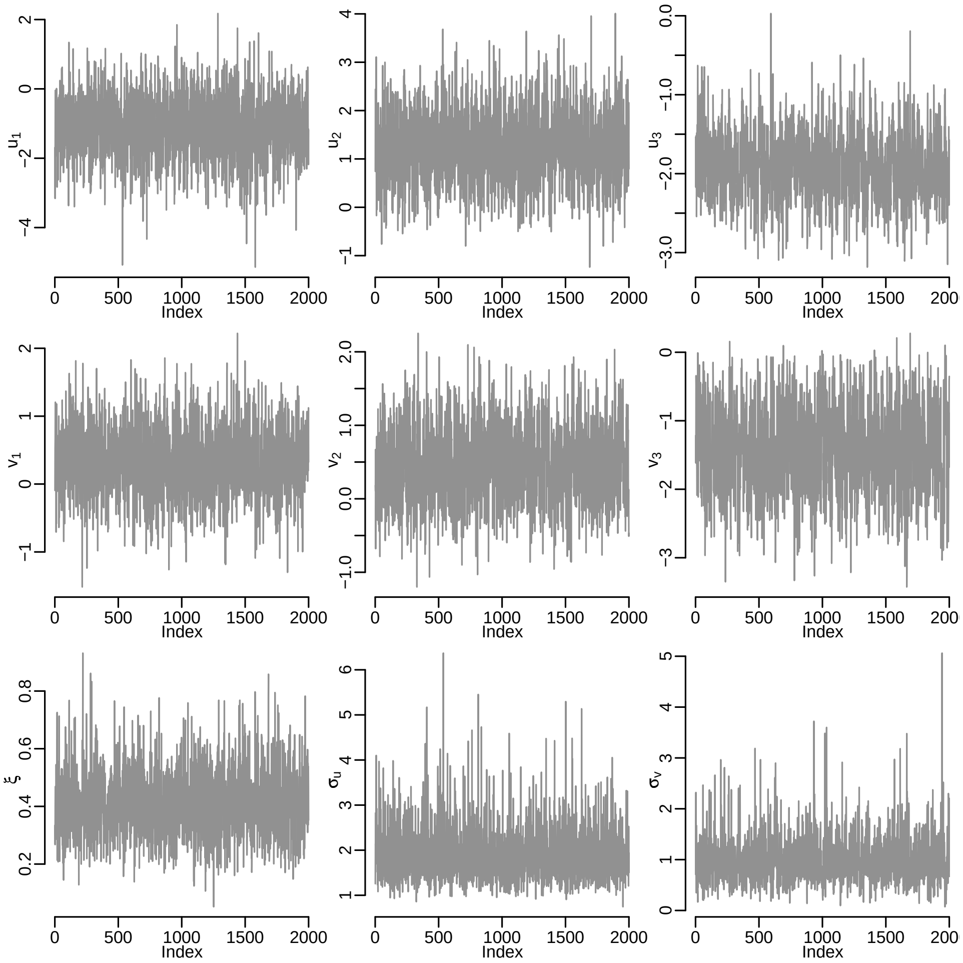

Figure 4.2: *Simulation results: trace plots of three random effects of the regression function for the location (first row) and scale (second row) parameters are illustrated. In the third row, in order from left to right panels, the trace plots of the shape parameter and the two variance components are reported .*

reported some illustrative examples of the chains trace plots after the burn in period. These correspond to a further simulation. In particular we took a sample of three chains of random effects for the location and scale functions, the chains of the shape parameter and the variance components. All these results support convergence of the chains.

In order to illustrate the results of MCMC model fitting we have also performed many different data simulations. The model set up considered is the same of the previous study. In Figure (4.3) we have reported a sample of five simulation results. In the top panels the location function estimate (left) and the scale function estimate (right) are reported. We can see how the location function is well estimated in most of the cases. Instead, the scale function estimate results seem to be less accurate. However, the lacking of the estimates for the scale function was expected due to the nature of the model parameter. In the bottom panels the posterior standard deviations of $\mu(x)$ (right) and $\psi(x)$ are shown. In both cases the hight variance at the boundaries is clear and expected. Moreover, the variance of $\mu(x)$ seems correctly reflect the scale function structure $\psi(x)$ from which it is dependent. Concluding from the data simulations, we have observed

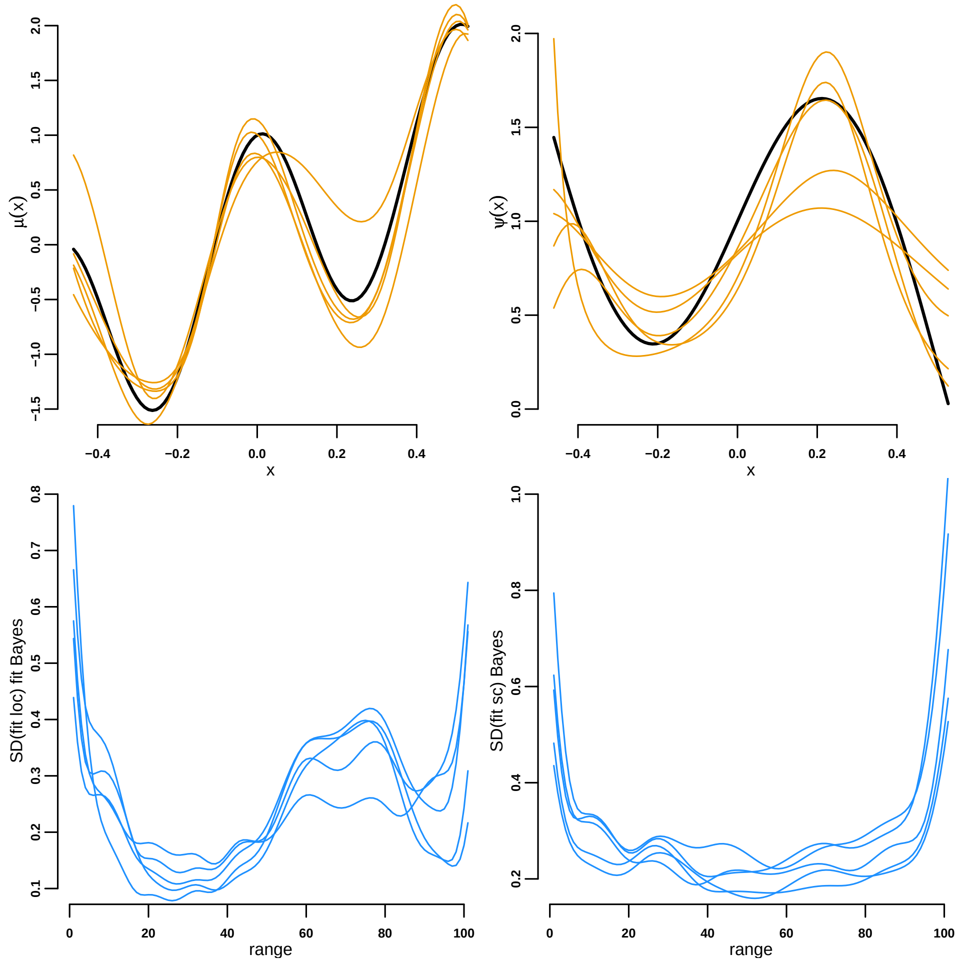

Figure 4.3: *Simulation results: the top panels show the regression function estimates (thin orange lines) of the location (right) and scale (left) parameters estimated by using 5 different simulated dataset. The solid large black lines represent the true signal. The bottom panels show the standard deviations of the two regression functions, left the location and right the scale parameters.*

the satisfactory performance of the Bayesian approach based on MCMC method for the spline mixed model for extremes.

Let us once more consider the Switzerland dataset. We analyzed the annual maxima temperatures with the time and annual amount of precipitation covariates, assuming the model design described in Section 4.2. The handling of multiple covariates in the penalized mixed model framework is allowed by the additive models structure. The Bayesian approach involves specification of prior distributions of all model parameters. Analogously with the simulation exercise, we set the prior distribution of the fixed effects to be $N(\mathbf{0}, \mathbf{G})$ for some diagonal matrix $\mathbf{G}$ with very large entries (noninformative). For the prior of the random effects, zero mean normal distributions are assumed as well, and for the shape parameter we take $U[-3, 3]$. Lastly, the priors of the variance components are $IG(0.01, 0.01)$. The splines design consists of radial cubic basis with 10 knots for each spline.

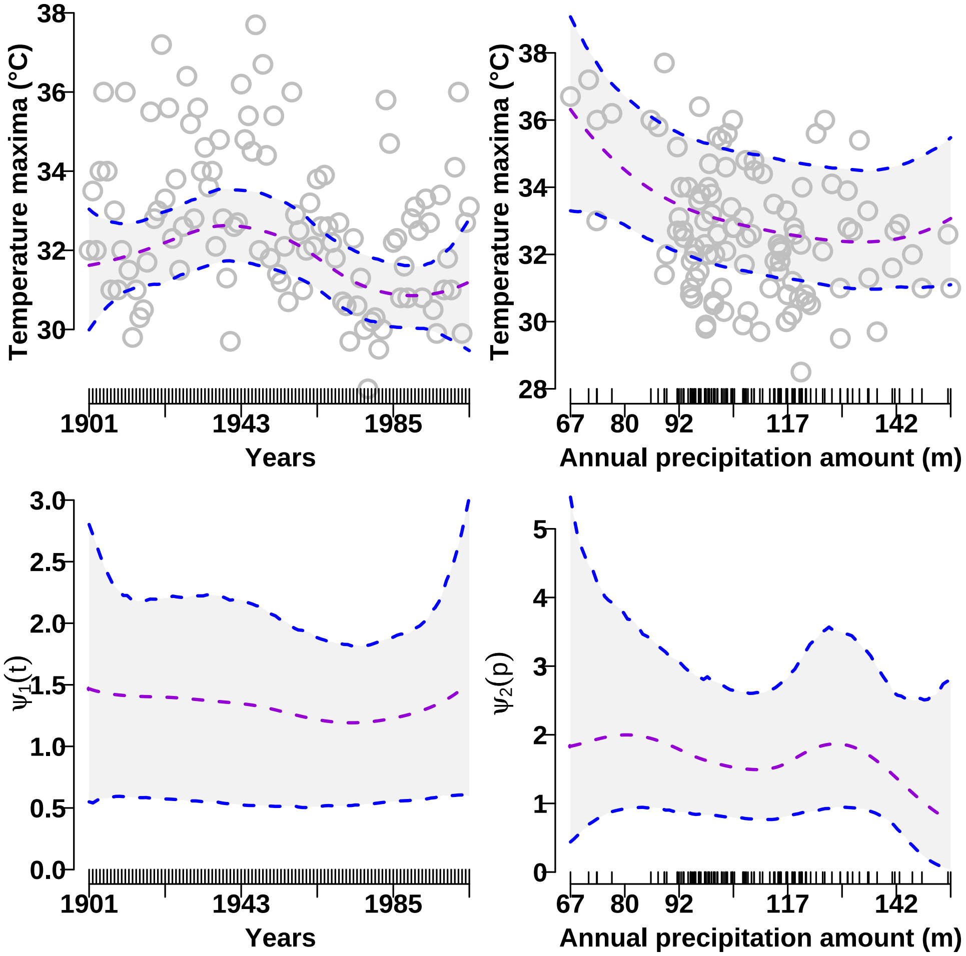

Figure 4.4: *Summary of the fit by using the MCMC method: the panels show the estimates of the location (μ) and scale (ψ) regression (middle broken lines) obtained by the posterior means and the corresponding pointwise 95% credibility intervals (outer broken lines). The top panels display the additive components of μ the bottom those of ψ. The circles depict the temperature maxima vs. the regressors.*

The MCMC implementation of fitting the Bayesian penalized mixed model is led by `BRugs`. The continuous covariate has been standardized (similar to what we have done with the time covariate). We run a `BRugs` script, setting 2000 as the number of iterations with a thinning factor of 5 and a burn in period of length 20000 (after multiple attempts we found that this was sufficient to yield convergence).

Figure 4.4 shows the estimates and the 95% credibility intervals for the regression functions. These are computed by using the means and the quantiles 0.025 and 0.975 of the posteriors. We see from the top panels that the estimate results are reminiscent of those in Figure 3.7. The posterior mean and deviation standard of the shape parameter and variance components result in: $\widehat{\xi} = -0.211(0.101)$, $\widehat{\sigma}^2_{\mu,T} = 0.63(0.41)$, $\widehat{\sigma}^2_{\psi,T} = 0.20(0.12)$, $\widehat{\sigma}^2_{\mu,P} = 0.18(0.11)$, $\widehat{\sigma}^2_{\psi,P} = 0.24(0.16)$. The estimate of the shape parameter (provided by the posterior mean) also supports the compatibility of these results with those of the analysis illustrated in Section (3.5.2). Figure 4.5 illustrates in the left panel the trace plot (of 2000 iterations) of the shape parameter after the burn in period. This supports the convergence of the chain. The right panel shows the posterior estimate by the kernel density method.

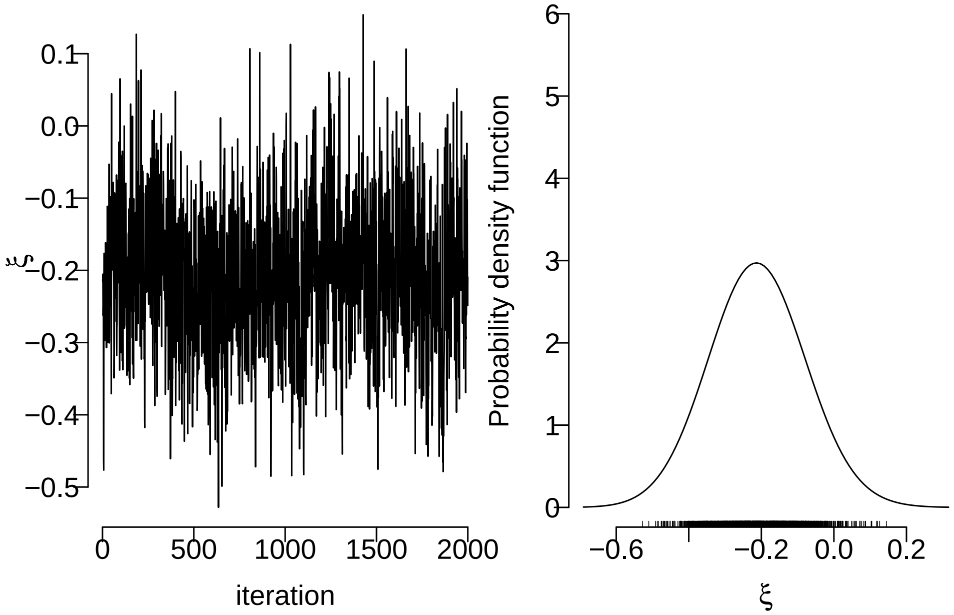

Figure 4.5: *Summary of the fit by using the MCMC method: the left panel shows the trace plot of the shape parameter (2000 iterations). The right panel shows the kernel estimate of the shape parameter posterior density.*

Concluding, as illustrated by this analysis, the MCMC fitting provides compatible results with those previously provided. This is enforced by the similar location curve estimates for both additive components. For the scale parameter the presence of a nonlinear trend seems less evident especially because of the very little change in the posterior mean, and the very large credibility intervals.

The advantage of the Bayesian approach is the versatility of the penalized mixed model. Multiple regressors can be taken into account for the location and scale parameters with a relatively easy MCMC implementation. And the uncertainty in the variance components is more easily assessed respect the likelihood approach.

Part II

Spatial extremes

Chapter 5

Max-stable processes for spatial extremes

5.1 Introduction

The theory of multivariate extreme value distribution is a relatively novel but rapidly growing field. It has been flourishing in the last decades where crucial theoretical bases, as limit theorems, have been founded (e.g Resnick, 1987; Galambos 1987). From the statistical point of view the multivariate extreme value theory provides a suitable probability framework in order to study jointly the patterns of many extremes series. Some application examples are provided in Coles (2001). However, multivariate extreme value distributions have several limitations in spatial context, for instance the dependence structure is not related with the site distances. Moreover with large location numbers, the distribution density function is often intractable, that is we can not easily derive the analytical expression.

Stationary stochastic processes theory such as max-stable processes is relatively similar to the multivariate extreme value theory, de Haan (1984). They provide an infinite dimensional extension of the multivariate extreme value theory, see also de Haan and Pickands (1986). Indeed some multivariate extreme value families can be derived starting from a max-stable formulation (Smith 1990). These processes give a more appropriate theoretical approach in order to model spatial extremes. One of the advantages accomplished from the max-stable processes representation is that the tail dependence among the variables located on the plane decreases monotonically and continuously with the distance. This is a desired property for spatial models. Some applications that illustrate their suitability in extreme value context are given by Smith (1990), Coles and Tawn (1990) and Coles and Tawn (1991).

Spatial extreme models that arise from the max-stable formulation are characterized by having the dependence structure of the random variables involved represented by model parameters. It turns out that the analytical K-dimensional distribution function of these models is not easy to derive for an integer $K > 2$. For this reason, a consolidated inference procedure for the dependent model parameters does not yet exist. A consistent and asymptotically normally distributed estimator of the spatial dependence structure is proposed by de Haan and Pereira (2006). This estimator has nice theoretical properties but we have found that it has poor practical performances in weak spatial dependent cases. Alternatively we propose an inference procedure based on the likelihood approach. In particular, we deal with what is known in literature as the *composite* likelihood estimator, introduced by Linsday (1988).

In the next section we provide a review of the max-stable processes theory. In Section 5.3 we are mainly concerned with the description and discussion of the composite maximum likelihood estimator for spatial extremes. Moreover in Section 5.3.2 we show numerically the performance of the estimator and its behavior for large samples. In section 5.4 we also describe an alternative inference procedure based on the Bayesian approach. Concluding in section 5.5 we illustrate a real data application by using the composite likelihood approach, focusing on rainfall levels recorded in North and South Carolina (USA).

5.2 Definition and modelling

Definition. Let T be an arbitrary space and consider n independent replications of a stochastic process $\{Z'(t)\}_{t\in T}$. Then $\{Z(t)\}_{t\in T}$ is a max-stable process if a suitable sequences of constants $a_n(t) > 0$ and $b_n(t) \in \mathbb{R}$ exist and such that

$$Z(t) = \lim_{n\to\infty} \frac{\max_{i=1}^{n} Z_i'(t) - b_n(t)}{a_n(t)}, \quad t \in T,$$

and provided that the limit exists, Schlater (2003).

In other words $\{Z(t)\}_{t\in T}$ is a max-stable process if $\{\max_{i=1}^{n} Z_i(t) - b_n(t)\}/a_n(t)$ has the same distribution as $\{nZ_1(t)\}_{t\in T}$ where $\{Z_i(t)\}_{t\in T}$ are independent copies of the process and $a_n(t) > 0$ and $b_n(t) \in \mathbb{R}$ are suitable constants, de Haan (1984).

Note that the max-stable formulation can also be seen as an extension of the max-stability property of multivariate extreme value distribution to the continuous processes, see also Resnick (1987). Two properties that follow from the above definition (de Haan 1984; de Haan and Resnick, 1977):

- The one-dimensional marginal distribution function $F_t(z)$ belongs to class of the GEV distribution (three-type), Galambos (1987).
- For any K, the K-dimensional marginal distribution belongs to the class of the multivariate extreme value distributions.

Without loss of generality let us to consider the case when $a_n(t) = n$ and $b_n(t) = 0$ for all t so that the margins are standard Fréchet distributions. This is convenient for the following argumentations.

A max-stable process can be defined by using what is known as its *spectral representation*, de Haan (1984). This definition provides a useful approach in order to obtain models for extreme values. Essentially, the process is defined as a functional of a Poisson process.

In detail, let E be an arbitrary measurable space[1] and $\{X_n, Y_n\}_{n\geq 1}$ be points of a Poisson process $\Pi := \Sigma_n \mathbb{I}(X_n, Y_n)$ on $E \times (0, \infty)$ with mean measure $\mu(dx) \times y^{-2}dy$ for a positive measure μ on E. Consider a measurable function $f(\cdot)$ defined on E for which the following property is valid

$$\int_E f(x,t)\mu(dx) = 1 \quad \forall \quad t \in T.$$

[1] Note that without loss in generality we can assume an Euclidean space.

Then, the stochastic process defined as

$$Z(t) := \max_n \{Y_n f(X_n - t)\}, \quad t \in T, \tag{5.1}$$

forms a max-stable process. The family of random variables $Z(t)$ satisfies the definition listed above. In fact we can show that the joint distribution for any $K = 1, 2, \ldots$ and for $t_1 < \ldots < t_K$ satisfies the property:

$$\begin{aligned}
&\mathbb{P}^m\{Z(t_1) \le mz_1, \ldots, Z(t_k) \le mz_k\} \\
&= \mathbb{P}^m\{Y_n f(X_n - t_k) \le mz_k, n = 1, 2, \ldots; k = 1, \ldots, K\} \\
&= \mathbb{P}^m \left\{ Y_n \le m \min_{k \le K} \frac{z_k}{f(X_n - t_k)}, n = 1, 2, \right\} \\
&= \mathbb{P}^m \left\{ \left(\{X_n, Y_n\} : Y_n > m \min_{k \le K} \frac{z_k}{f(X_n - t_k)} \right) = \emptyset \right\} \\
&= \exp\left\{ -m \int_E \int_{m \min_{k \le K} \frac{z_k}{f(x - t_k)}}^{\infty} \frac{dy}{y^2} \mu(dx) \right\} \\
&= \exp\left\{ - \int_E \max_{k \le K} \left(\frac{f(x - t_k)}{z_k} \right) \mu(dx) \right\} \\
&= \mathbb{P}\{Z(t_1) \le z_1, \ldots, Z(t_k) \le z_k\} \qquad \square
\end{aligned} \tag{5.2}$$

The equivalences obtained in (5.2) arose from the following results. The event in the third row $\{y_n \le m \min_{k \le K} \frac{z_k}{f(x_n - t_k)}\}$ is satisfied if no points of the Poisson process lie in the set $A = (\{X_n, Y_n\} : Y_n > m \min_{k \le K} z_k / f(X_n - t_k))$. Then, $\mathbb{P}\{\Pi(A) = 0\}$, the probability that no points of the Poisson process fall in the set A, is given by a Poisson distribution with mean measure of the set A.

It turns out that for a finite set of indexes $t_1, \ldots, t_K$, the sixth row of (5.2) provides the K-dimensional distribution. It is easy to check from the joint distribution that the univariate marginal Z has Frechét distribution.

The spectral representation of a max-stable process is also appealing because it can be physically interpreted as an extreme environmental process. For instance, in extreme analysis of rainfall and storm events Smith (1990) has formalized the following relations:

- The set E corresponds to the region where the storms are centered (note that in practice $E = \mathbb{R}^2$).
- The measure $\mu(dx)$ describes how the storms are distributed over the region E. The function f defines for the storm with center x_n the diffusion over the region E, and y_n is its intensity.
- Finally, the expression $z(t) = \max_n \{y_n f(x_n - t)\}$ expresses the maximum amount of rainfall in the site t recorded over all independent storms with center x_n and intensity y_n.

Another application of max-stable theory in the environmental processes such as for example, the wind speed is provided by Coles and Walshaw (1994). In this case a max-stable process have been developed on a circular domain that was appropriate for wind directions.

In extreme value theory it is relevant to study the extremal dependence of the random variable $Y \in \mathbb{R}^K$. From extreme value theory literature it is common to consider the joint distribution of the componentwise maximum of n independent replications of the variable $\mathbf{Y}$ (de Haan and Resnick, 1977). Under weak conditions, see Resnick (1987, Ch. 5), for $k = 1, \dots, K$ (with $K \in \mathbb{N}$) there exists a real number ν, with $1 \leq \nu \leq K$, such that the normalized maximum of all the variables converge to the Fréchet distribution with parameter ν. In other words for $n \to \infty$,

$$\mathbb{P}\left\{\max_k \max_{j=1,\dots,n} Y_i^{(j)}/n \leq z\right\} = \mathbb{P}\left\{\max_{j=1,\dots,n} Y_1^{(j)}/n \leq z\right\}^{\nu} = \exp(-\nu/z), \quad \forall z > 0$$

where $Y^{(j)}$ is the jth replicate of the variable Y_k. The quantity ν, is called the *extremal coefficient* and it measures the extremal dependence between $Y_1, \dots, Y_K$. The extremal coefficient represents the effective number of the independent variables in the sequence of K, from which the maximum is taken (Tawn, 1988; Smith, 1990). A more general discussion on the extremal coefficient in the multivariate setting is given in Pickands (1981), see also Schlather and Tawn (2003). However, the main interest here is to study the spatial models for extremes that arise from the max-stable processes and the dependence measure for this class of models. In order, we start considering the extremal dependence for a general stationary process $Z(\mathbf{t})$ with $t \in \mathbb{R}^d$ and unit Fréchet margins. We focus on the extremal dependence for the pairwise structure for a clearer exposition instead of the generic sequence $Z(\mathbf{t}_1), \dots, Z(\mathbf{t}_K)$. The pairwise setting provides sufficient information with a simple structure, so that it can be used instead of the wider approach that involves analytical complications. In principle, we consider the componentwise maximum of n independent replications of $Z(\mathbf{t})$. Under weak conditions (de Haan, 1984; Schlather and Tawn, 2003), there exists a real-valued function $\nu(\cdot)$ so that the normalized maximum asymptotic distribution at the pair (of sites) (i, j) is Fréchet with scale parameter $\nu(\mathbf{t}_i, \mathbf{t}_j)$, where $\mathbf{t}_i, \mathbf{t}_j \in \mathbb{R}^d$. In other words for $n \to \infty$ and $z > 0$ we have

$$\mathbb{P}\left\{\max_{m=1,\dots,n} \max\{Z(\mathbf{t}_i)^{(m)}, Z(\mathbf{t}_j)^{(m)}\}/n \leq z\right\} = \exp(-\nu(\mathbf{t}_i, \mathbf{t}_j)/z).$$

In this setting the term $\nu(\cdot)$ is called the *extremal coefficient function*. A similar definition of the extremal coefficient function arises within the max-stable processes context. In fact we saw before that for a finite set of indexes $k = 1, \dots, K$, the K-dimensional distribution is given by

$$\mathbb{P}\{Z(\mathbf{t}_k) \leq z_k, \text{ for } k = 1, \dots, K\} = \exp\left[-\int_E \max_k \left\{\frac{f(\mathbf{x} - \mathbf{t}_k)}{z_k}\right\} \mu(d\mathbf{x})\right],$$

where now we assume that $\mathbf{t}, \mathbf{x} \in \mathbb{R}^d$. Then for a fixed threshold $z > 0$ it follows that the 2-dimensional distribution function is

$$\mathbb{P}\{Z(t_i) \leq z, Z(t_j) \leq z\} = \exp(-\nu(\mathbf{t}_i, \mathbf{t}_j)/z),$$

where

$$\nu(\mathbf{t}_i, \mathbf{t}_j) = \int_E \max\{f(\mathbf{x} - t_i), f(\mathbf{x} - t_j)\}\mu(d\mathbf{x}),$$

and the indexes $\mathbf{t}_i$ and $\mathbf{t}_j$ are seen as locations. The term $\nu(\cdot)$ is called the extremal coefficient function for the pairwise structure, similarly as we saw before. It follows the property, for $\mathbf{t}_i, \mathbf{t}_j \in \mathbb{R}^d$, that $1 \leq \nu(\mathbf{t}_i, \mathbf{t}_j) \leq 2$. We refer to Schlather and Tawn (2003) for further details.

Note that a further class of max-stable processes derives as a generalization of (5.1) by substituting the function f with an arbitrary random function W, whose positive part $\max\{0, W(\cdot)\}$ is integrable, Schlather (2002).

We focus now on a class of spatial models for extremes that arises from the max-stable processes formulation. In particular we assume that $E = T \subseteq \mathbb{R}^d$, $\mu(d\mathbf{x}) = d\mathbf{x}$ is the Lebesgue measure, and the function *f*, described as the *storm profile* model, is a multivariate normal density function,

$$f(x-t) = (2\pi)^{d/2}|\Sigma|^{-1}\exp\left\{-\frac{1}{2}(x-t)^T\Sigma^{-1}(x-t)\right\},$$

where Σ is a $d \times d$ definite positive covariance matrix. We can not derive the analytical expression of the K-dimensional distribution (5.2) using the storm profile model just reported above. Nonetheless the joint probability can at lest be simulated (Smith, 1990), resolving the integral part as follows:

$$\begin{aligned}
&\int_E \max_k \left\{\frac{f(\mathbf{x}-\mathbf{t}_k)}{z_k}\right\} d\mathbf{x} \\
&= \int_E \sum_k \frac{f(\mathbf{x}-\mathbf{t}_k)}{z_k} I\left\{\frac{f(\mathbf{x}-\mathbf{t}_i)}{z_i} > \max_{j\neq i}\frac{f(\mathbf{x}-\mathbf{t}_j)}{z_j}\right\} d\mathbf{x} \\
&= \sum_k \int_E \frac{f(\mathbf{x})}{z_k} I\left\{\frac{f(\mathbf{x})}{z_i} > \max_{j\neq i}\frac{f(\mathbf{x}-\mathbf{t}_j+\mathbf{t}_i)}{z_j}\right\} d\mathbf{x} \\
&= E\left[\sum_k \frac{1}{z_k} I\left\{\frac{f(\mathbf{X})}{z_i} > \max_{j\neq i}\frac{f(\mathbf{X}-\mathbf{t}_j+\mathbf{t}_i)}{z_j}\right\}\right].
\end{aligned}$$

Note that for practical applications the further simplifications of $E = T = \mathbb{R}^2$ and using the bivariate version of the storm profile model are required. The last equation is useful also in order to derive the the 2-dimensional distribution of the process. In fact, given two locations i, j, with a few steps of algebra it is easy to show that

$$\begin{aligned}
&\mathbb{P}\{Z(\mathbf{t}_i) \leq z_i, Z(\mathbf{t}_j) \leq z_j\} \\
&= \exp\Bigg[-\frac{1}{z_i}\Phi\left(\frac{\theta(\mathbf{h})}{2} + \frac{1}{\theta(\mathbf{h})}\log\frac{z_j}{z_i}\right) - \frac{1}{z_j}\Phi\left(\frac{\theta(\mathbf{h})}{2} + \frac{1}{\theta(\mathbf{h})}\log\frac{z_i}{z_j}\right)\Bigg],
\end{aligned} \tag{5.3}$$

where

$$\theta(\mathbf{h}) = (\mathbf{h}^T\Sigma^{-1}\mathbf{h})^{1/2}, \qquad \Sigma_2^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} \zeta^{-2} & -\frac{\rho}{\gamma\zeta} \\ -\frac{\rho}{\gamma\zeta} & \gamma^{-2} \end{bmatrix},$$

Φ is the standard normal distribution, $\mathbf{h} = \mathbf{t}_i - \mathbf{t}_j$ is the separation between the two sites and $\mathbf{t}_i, \mathbf{t}_j \in \mathbb{R}^2$ are coordinates on the plane. The max-stable process with Gaussian storm profile is known as the *Gaussian extreme value process*, than we refer to (5.3) as the *Gaussian extreme value model*. The quantity $\theta(\mathbf{h})$ is the Mahalanobis distance (which is a distance measure introduced by P. C. Mahalanobis in 1936). The distance is a function of the separation $\mathbf{h}$ and the inverse of covariance matrix Σ_2. The elements of Σ_2 are constants that measure the strength of the tail dependence. Roughly speaking, for small values of the covariance matrix elements we have weak dependence of the extremes and by contrast, large values correspond to strong event dependence. Also as should be required for spatial models, the extreme dependence between the two random variables $Z(\mathbf{t}_i)$

and $Z(\mathbf{t}_j)$ decreases monotonically and continuously with the separations $|\mathbf{h}|$ between $\mathbf{t}_i$ and $\mathbf{t}_j$. Instead, for fixed separation $\mathbf{h}$ between $\mathbf{t}_i$ and $\mathbf{t}_j$ the dependence decreases monotonically as the elements of Σ decrease. More precisely, when $\theta(\mathbf{h}) \to 0$ it represents independence of the extremes, and instead when $\theta(\mathbf{h}) \to \infty$ it means strong dependence.

Observe that other bivariate spatial models can be obtained using different storm profile models, as for example the exponential or t-model (de Haan and Pereira, 2006). Moreover another class of max-stable processes and related spatial models for extremes are discussed by Schlater (2003).

In conclusion, we report few other results related with the extremal dependence structure. From the multivariate extreme value theory we have that the 2-dimensional distribution of the bivariate random variable (Y_i, Y_j) (obtained as the componentwise maximum of n independent replicates) can be factorized as,

$$\mathbb{P}\{Y_i \leq y_i, Y_j \leq y_j\} = \exp\left\{-\left(\frac{1}{y_i} + \frac{1}{y_j}\right) A\left(\frac{y_i}{y_i + y_j}\right)\right\},$$

where $A(\cdot)$ is a function which establishes the dependence of the extremes (Pickands, 1981; Tawn, 1988). It is called the *dependence function*. A connection between the spatial model for extremes given by (5.3) and the dependence function exists. In fact, with few algebraic steps from the 2-dimensional distribution of the Gaussian extreme value process, we can derive the dependence function of the process, that is

$$A(w) = (1-w)\Phi\left(\frac{\theta(\mathbf{h})}{2} + \frac{1}{\theta(\mathbf{h})}\log\frac{1-w}{w}\right) + w\Phi\left(\frac{\theta(\mathbf{h})}{2} + \frac{1}{\theta(\mathbf{h})}\log\frac{w}{1-w}\right),$$

where $w = z_i/(z_i + z_j)$. Moreover, the coefficient $\theta(\mathbf{h})$ of model (5.3) is related with the extremal coefficient $\nu(\mathbf{h})$. Indeed, it is easy to check that the extremal coefficient for the bivariate models is equal to

$$\nu(\mathbf{h}) = 2\Phi(\theta(\mathbf{h})/2).$$

5.2.1 Simulation of a Max-Stable process

Simulating synthetic data from a stochastic process or model has many advantages. For example, collecting data measurements can be expensive and, when involving manual work, lengthly. Then, the simulation of data can be helpful when we desire to asses and validate statistical methods. Moreover, we can assess the fitting of a stochastic process to our data by simulating the process and than comparing the outcomes with observed data. Testing the goodness of fit will depend on our ability to simulate the data from the proposed model. The aim of this section is to describe an efficient method for simulating a max-stable process or random field.

Assume that the random points $\{X_n\}_{n\geq 1}$ of the Poisson process introduced in the previous section, belong to $\mathbb{R}^2$, so that the random field is defined on the plane. Let W be a compact set that we can suppose to be a rectangle or any another compact set with inner points.

To simulate a stationary Poisson point process or random field with mean measure $\mu(W) = \mathbb{E}N(W) = \lambda|W|$ on a bounded set W, where $|\cdot|$ is a Lebesgue measure and λ is the rate per unit, is quite easy. The simulation requires two steps. The first generates the number of points n in W from a random variable with Poisson distribution and parameter $\mu = \lambda|W|$, and the second distributes these uniformly and independently in W. Details of the first step are given in Ripley (1987). Note that Ripley (1987) observed how all random number generators are "defective", where, although most of them will yield a uniform distribution on the intervals, many will yield quite regular patterns on the squares . Ripley (1987) discusses how to use a random number generator for simulations occurring in a spatial context.

Instead, the simulation of stationary max-stable processes or random fields is not as simple because some complications are involved from the process definition (5.1). In fact, the spectral representation of a max-stable process involves the maximum over an infinite sequence of points $\{X_n, Y_n\}_{n\geq 1}$. But in the practice simulations the number of generated points have to be necessarily finite. Nonetheless a simulation of a max-stable process may be carried out under some particular conditions, as Schlater (2002) has demonstrated. In principle, the simulation of a max-stable random field consists of the following essential steps summarized in this iterative scheme.

Iterative Scheme: Simulation of M realizations from a max-stable random field

1. Set a regular or irregular grid of K points in $\mathbb{R}^2$.

2. Define a set $W \subseteq \mathbb{R}^2$ with finite Lebesgue measure $|W|$.

3. Generate a sequence of points $\{x_n\}_{n\geq 1}$ uniformly distributed on W.

4. Simulate the sequence $y_1, \ldots, y_n, n = 1, 2, \ldots$ from unit Fréchet distribution.

5. Compute $Z(t_k) := \max_n\{y_n f(x_n - t_k)\}, \quad k = 1, \ldots, K$.

6. Repeat steps 3–5 for M times.

Now in order to perform a simulation we need to define the sampling set W and the finite number of the points $\{X_n, Y_n\}$. Solutions to these are provided by Lemma 3 and Theorem 4 of Schlater (2002), that we report in the following pages.

***Lemma**: Assume the set W $\subseteq \mathbb{R}^2$ with finite Lebesgue measure $|W|$. Let $\{X_n\}_{n\geq 1}$ be points of a Poisson process on W, and $Y_n = 1/\sum_{i=1}^n S_i$ be a random variable defined assuming that S_i are i.i.d. copies of the exponential variable $S_i \sim Exp(1)$ with $s \geq 0$. Then, the random sequence $\{X_n, |W|\, Y_n\}_{n\geq 1}$ is a Poisson process on $W \times (0, \infty)$ with mean measure $dx \times y^{-2}dy$.*

Proof: The random sequence $S_n = \{\sum_{i=1}^n S_i : n = 1, 2, \ldots\}$ is a homogenous Poisson process on $[0, \infty)$ with mean measure $\mu[0, s) = s$. Consider the transformation $y = 1/s$. Then the transformation of the points of the Poisson process S_n lead to a Poisson process on $(0, \infty]$ with mean measure μ' given for $y > 0$ by

$$\mu'(y, \infty] = \mu\{s \geq 0 : s^{-1} \geq y\} = \mu\{s \geq 0 : s < y^{-1}\} = \mu[0, y^{-1}) = y^{-1}.$$

μ' has density $\alpha(y) = -d/dy\, y^{-1} = y^{-2}$. Considering also the positive constant $|W|$

in the inverse transformation, then the mean measure of $\{X_n, |W|\, Y_n\}_{n\geq 1}$ is given by the product of the measures. So the assertion is demonstrated. □

From the previous lemma we have seen in practice how to build the Poisson process with desired properties on the sampling set W. But it is still unspecified how to get exact simulations of a max-stable process on the finite sampling set W. The next theorem describes this last stage.

Theorem: *Consider the Poisson process Π, $f(\cdot)$ and $Z(t)$ of definition 5.1. Assume a compact set $W \subseteq E$ and that the sequence $\{X_n\}_{n\geq 1}$ is a Poisson process on W_r, defined as $W_r = \bigcup_{t_k \in W} b(t_k, r)$, where $b(\cdot, r)$ is a ball of radius r. Assume also that $f(\cdot)$ has support in $b(\cdot, r)$ for some $r \in (0, \infty)$ and it is bounded by $C \in (0, \infty)$. And finally that $Y_n = 1/\sum_{i=1}^{n} S_i$ where S_i are i.i.d. copies of the exponential variable $S_i \sim$ Exp(1) with $s \geq 0$. Then, on W,*

$$Z'(t_k) = |W_r| \max_n \{Y_n\, f(X_n - t_k)\}, \quad t_k \in W, \tag{5.4}$$

equals in distribution to $Z(t_k)$, and

$$Z'(t_k) = |W_r| \max_n \left[Y_n\, f(X_n - t_k) : n = 1, \ldots, m, \text{ and } m : Y_m C \leq \max_{1\leq n\leq m} \{Y_n f(X_n - t_k)\} \right]$$

is equal to $Z(t_k)$ almost surely.

Proof: Following the previous lemma the process $Z'(t_k)$ is equal in distribution to $Z(t_k)$ on the set W. Instead the second assertion can be derived observing that $f(X_n - t_k) \leq C$ and Y_n is a non-increasing sequence, then it follows that the equality is an immediate consequences. □

This theorem establishes that the transformation 5.4 of the Poisson process defined in the previous lemma is a max-stable process. Moreover, it defines the stopping rule in order to determine the finite number of points that form the process. So the lemma and the theorem together describe the ingredients in order to obtain an exact simulation of a max-stable process.

For max-stable processes with deterministic profile models f, the simulation is relatively easy to implement. For instance, in order to simulate the Gaussian extreme value process the radius r can be selected by the relation $\varphi(r) = \varepsilon$, where φ is the standard normal density and ε is a small tolerance. The constant C depends on the normal density function f. So similarly we can proceed with other storm profile models, as for example the exponential or t-model.

5.3 Inference based on Likelihood approach

We know from the previous section that with particular storm profiles as for example with the bivariate normal density function we can derive the 2-dimensional distribution. The resulting model can be useful for spatial extreme analysis. For example, if data from extreme events are available for some locations spread over a region, then we can estimate, in someway the parameters indicated in (5.3). Once that the model parameters are estimated then the extremal coefficients for all the pairs of locations can be computed, and the tail dependencies can be assessed. Moreover, once the model parameters have been estimated, then by an opportune transformation of (5.3) we will see that the return levels can then be estimated and the rates at which high rainfall levels occur at the sites assessed. In this section we describe an inferential method for spatial extremes, and later we will see an application with real rainfall level data.

We describe the proposed inference methods for spatial extremes focusing on the Gaussian extreme value model. Note that the same arguments can be also extended to other models, we recall that some alternatives are described in Schlather (2002) and de Haan and Pereira (2006). The extremal Gaussian process proposed by Schlather (2002) has been studied but not completed so it is not reported here.

Second-order partial derivatives of (5.3) yield the 2-dimensional density function

$$
\begin{aligned}
f(z_i,z_j) = \exp\left\{-\frac{1}{z_i}\Phi(w)-\frac{1}{z_j}\Phi(v)\right\}\Bigg[\Bigg\{&-\frac{1}{\theta(\mathbf{h})z_iz_j}\varphi(w)+\frac{1}{z_j^2}\Phi(v)\\
&+\frac{1}{\theta(\mathbf{h})z_j^2}\varphi(v)\Bigg\}\left\{\frac{1}{z_i^2}\Phi(w)+\frac{1}{\theta(\mathbf{h})z_i^2}\varphi(w)-\frac{1}{\theta(\mathbf{h})z_iz_j}\varphi(v)\right\}\\
&+\left\{\frac{\theta(\mathbf{h})-w}{\theta(\mathbf{h})^2z_i^2z_j}\varphi(w)+\frac{\theta(\mathbf{h})-v}{\theta(\mathbf{h})^2z_iz_j^2}\varphi(v)\right\}\Bigg],
\end{aligned}
\tag{5.5}
$$

where φ is the univariate standard normal density function, $w=\theta(\mathbf{h})/2+1/\theta(\mathbf{h})\log(z_j/z_i)$ and $v=\theta(\mathbf{h})/2+1/\theta(\mathbf{h})\log(z_i/z_j)$. For applications the bivariate model represents most of the trivial cases, this is because only two sites $\mathbf{t}_i$ and $\mathbf{t}_j$ are involved. In fact, rainfall analysis often concerns a "large" number of sites, at least greater than 2. Assume that we are interested in studying the rainfall processes over a region where many locations are observed. Then the parameter set (ζ,γ,ρ) can not be estimated by likelihood methods straightforwardly. It can be used only separately for all the marginal events that involve pairs of variables $(Z(\mathbf{t}_i),Z(\mathbf{t}_j))$. Nonetheless, we can still approach the inference problem based on a likelihood approximation. In fact we appeal to a specific class of *pseudo*-likelihood, known in literature as *composite* likelihood (Linsday, 1988). Broadly, the composite likelihood is a likelihood function obtained combining the bivariate likelihood associated with the marginal events (considering pairs of variables), which can be used to provide consistent estimators. In next section we will describe, in detail this alterative approach.

5.3.1 Composite likelihood approach

Consider the Gaussian extreme value model introduced in the previous section and assume, for a finite set of indexes $1,\dots,K$ that the observations $z_{i1},\dots,z_{iK}$ are i.i.d. realizations for $i=1,\dots,n$ of the random variables $Z(\mathbf{t}_1),\dots,Z(\mathbf{t}_K)$ with $\mathbf{t}_1,\dots,\mathbf{t}_K\in\mathbb{R}^2$.

The K-dimensional joint distribution of the process is not easy to derive explicitly when $K>2$. Then the parameters (ζ,γ,ρ) can not be estimated, for example by maximizing the likelihood

$$
\mathcal{L}(\zeta,\gamma,\rho)=\prod_{i=1}^{n}f(z_{i1},\dots,z_{iK};\zeta,\gamma,\rho),
$$

where $f(z_{i1}, \ldots, z_{iK}; \zeta, \gamma, \rho)$ is the i's contribution of the joint density to the likelihood function. In situations where we only know the analytical expression of the marginal or conditional distribution associated to subsets of data, then a likelihood approximation is provided by the composite likelihood, Lindsay (1988), that is defined adding together individual likelihood objects. More precisely, let $\mathcal{F}$ be the parametric statistical model specified by the density function family $\mathcal{F} = \{f(\mathbf{z}; \boldsymbol{\psi}), \mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^K, \psi \in \Psi \subseteq \mathbb{R}^d\}$, and consider the sequence of marginal or conditional events $\{\mathcal{A}_1, \ldots, \mathcal{A}_k\}$ for some $k \in K \subset \mathbb{N}$. Let us not specify any particular sequence of events for the initial description of the general method. Later we will focus on the specific case for the spatial extremes. Broadly, the composite likelihood is defined as

$$\mathcal{L}_C(\boldsymbol{\psi}; \mathbf{z}) = \prod_k \mathcal{L}_k(\boldsymbol{\psi}; \mathbf{z}),$$

where $\mathcal{L}_k(\boldsymbol{\psi}; \mathbf{z}) = f(z_k \in \mathcal{A}_k; \boldsymbol{\psi})$. Following straightforwardly from the previous definition, the composite log likelihood is equal to

$$\ell_C(\boldsymbol{\psi}; \mathbf{z}) = \log \prod_k \mathcal{L}_k(\boldsymbol{\psi}; \mathbf{z}) = \sum_k \ell_k(\boldsymbol{\psi}; \mathbf{z}),$$

where $\ell_k(\boldsymbol{\psi}; \mathbf{z}) = \log \mathcal{L}_k(\boldsymbol{\psi}; \mathbf{z})$. Analogously with the standard likelihood, by differentiation of the the composite log likelihood with respect to parameter vector $\boldsymbol{\psi}$ yields the estimating function

$$U(\boldsymbol{\psi}, \mathbf{Z}) = \frac{\partial}{\partial \boldsymbol{\psi}} \ell_C(\boldsymbol{\psi}; \mathbf{z}) = \sum_k \frac{\partial}{\partial \boldsymbol{\psi}} \ell_k(\boldsymbol{\psi}; \mathbf{z}),$$

which is termed in this context as *composite* score. An estimate of the parameters $\boldsymbol{\psi}$ can be obtained by solving the composite score equation, $U(\boldsymbol{\psi}, \mathbf{Z}) = 0$. The solution of this equation is called the maximum *composite* likelihood estimator, indicated with $\widehat{\psi}_{MCL}$. The key utility of the composite log likelihood is that the composite score equations form an additive estimating function that, can be used to provide consistent parameter estimates where the full likelihood estimator is not available. In other words, because each composite score equation is an unbiased estimating function, then the sum of them is an unbiased estimating function too. Besides, the associated inferential procedures have theoretical properties similar to those of the ordinary likelihood methods under suitable regular conditions (e.g. Lindsay, 1988; Nott and Ryden, 1999) the maximum composite likelihood estimator, $\hat{\psi}_{MCL}$, is consistent and is asymptotically normally distributed, $\hat{\psi}_{MCL} \dot{\sim} N(\psi, I(\psi)^{-1})$, where $I(\psi)^{-1}$ is an approximation of the asymptotic covariance and $I(\psi)$ is known as the *sandwich* information matrix. In particular, let us assume the general setting where we know the density function f that is an approximation of the unknown true density g. If we consider the finite sample version of the score equation $U(\boldsymbol{\psi}, \mathbf{Z}) = 0$ that is $n^{-1} \sum_{i=1}^{n} \partial \log f(\mathbf{z}_i; \widehat{\boldsymbol{\psi}}) / \partial \boldsymbol{\psi} = 0$ where $\widehat{\boldsymbol{\psi}}$ solves the equation, from its Taylor expansion about $\boldsymbol{\psi}_g$, yields

$$\boldsymbol{\psi} \doteq \boldsymbol{\psi}_g + \left\{ -n^{-1} \sum_{i=1}^{n} \frac{\partial^2 \log f(\mathbf{z}_i; \boldsymbol{\psi}_g)}{\partial \boldsymbol{\psi}\, \partial \boldsymbol{\psi}^T} \right\}^{-1} \left\{ -n^{-1} \sum_{i=1}^{n} \frac{\partial \log f(\mathbf{z}_i; \boldsymbol{\psi}_g)}{\partial \boldsymbol{\psi}} \right\}.$$

And a modification of the previous derivation gives

$$\widehat{\boldsymbol{\psi}} \dot{\sim} N(\boldsymbol{\psi}_g, H(\boldsymbol{\psi}_g)^{-1} J(\boldsymbol{\psi}_g) H(\boldsymbol{\psi}_g)^{-1})$$

where $J(\boldsymbol{\psi})$ is the variance of the score function,

$$J(\boldsymbol{\psi}) = n \int \frac{\partial \log f(\mathbf{z};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \frac{\partial \log f(\mathbf{z};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}^T} g(\mathbf{z})\, d\mathbf{z},$$

and $H(\boldsymbol{\psi}_g)$ is the Fisher information,

$$H(\boldsymbol{\psi}_g) = -n \int \frac{\partial^2 \log f(\mathbf{z};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\, \partial \boldsymbol{\psi}^T} g(\mathbf{z})\, d\mathbf{z}.$$

The composite likelihood approach, in the inferential procedure, involves likelihoods (correctly specified) based only on subsets of data (related to marginal or conditional events). Consequently, a sort of model mispecification is introduced with the composite likelihood method.

Given that in practice $g(\mathbf{z})$ is unknown, then quantities $J(\boldsymbol{\psi})$ and $H(\boldsymbol{\psi}_g)$ can be estimated by

$$\hat{J} = \sum_{i=1}^{n} \frac{\partial \log f(\mathbf{z}_i;\widehat{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}} \frac{\partial \log f(\mathbf{z}_i;\widehat{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}^T}, \quad \hat{H} = -\sum_{i=1}^{n} \frac{\partial^2 \log f(\mathbf{z}_i;\widehat{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}\, \partial \boldsymbol{\psi}^T},$$

where the latter is the observed information matrix. As it is expected there is a loss of efficiency, in fact broadly the composite maximum likelihood estimator is not asymptotically efficient (Zhao and Joe, 2005). Some useful results on efficiency issues may be found in Lindsay (1988), although generally a complete analysis is still not provided.

With the Gaussian extreme value model case we consider the bivariate marginal likelihoods related to the bivariate marginal events. In this fashion the composite likelihood is based on a subset of data, and in particular on pairs of observations. This version is known as the *pairwise* likelihood. Consider the sequence of random variables $Z(\mathbf{t}_1), \ldots, Z(\mathbf{t}_K)$ and assume a finite value for K, then the set of all the pairs is $\mathcal{K} = \{K(K-1)/2\}$. So we define the pairwise log-likelihood as

$$\ell_{\mathcal{P}}(\zeta, \gamma, \rho; \mathbf{z}) = \sum_{k \neq j \in \mathcal{K}} \sum_{i=1}^{n} \ell(\zeta, \gamma, \rho; z_{ik}, z_{ij}), \tag{5.6}$$

where $\ell(\zeta, \gamma, \rho; z_{ik}, z_{ij}) = \log f(z_{ik}, z_{ij}; \zeta, \gamma, \rho)$ and $f(z_{ik}, z_{ij}; \zeta, \gamma, \rho)$ is the bivariate density associated to the pair $(z_{\cdot k}, z_{\cdot j})$.

In the Gaussian extreme value model estimates of (ζ, γ, ρ) can not be obtained as the solution of the pairwise score equation, but instead numerical maximization methods are required. Moreover, we saw before that $J(\zeta, \gamma, \rho)$ and $H(\zeta, \gamma, \rho)$ can be estimated by $\hat{J}$ and $\hat{H}$. In practice, in the applications, the estimate of $\hat{H}$ at $(\widehat{\zeta}, \widehat{\gamma}, \widehat{\rho})$ can be given by minus the hessian of the pairwise log likelihood which is easily provided by numerical maximization routines. Instead the quantity $\hat{J}$ can be estimated by the Monte Carlo estimate (Varin, Høst and Skare, 2005),

$$\frac{1}{M} \sum_{m=1}^{M} \frac{\partial \ell_p(\widehat{\zeta}, \widehat{\gamma}, \widehat{\rho}; \mathbf{z}^{(m)})}{\partial \zeta\, \partial \gamma\, \partial \rho} \frac{\partial \ell_p(\widehat{\zeta}, \widehat{\gamma}, \widehat{\rho}; \mathbf{z}^{(m)})}{\partial \zeta\, \partial \gamma\, \partial \rho}^T,$$

where $(\widehat{\zeta}, \widehat{\gamma}, \widehat{\rho})$ is the pairwise maximum likelihood estimate and $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(M)}$ are M data replications obtained from the Gaussian extreme value process with parameters (ζ, γ, ρ) set equal to the pairwise maximum likelihood estimates $(\widehat{\zeta}, \widehat{\gamma}, \widehat{\rho})$. Note that this approach can also be applied to other spatial extreme models proposed, for example by Schlather, (2002) and de Haan and Pereira (2006), where the composite log likelihood setting can be defined using marginal events related to the data.

5.3.2 Unified framework

The composite likelihood may provide a reasonable surrogate to the likelihood for spatial extremes when the latter is not available. The first has important proprieties that make it a useful device for inference also when the likelihood is too computationally intensive. The use of the likelihood for inference has an important advantage in that we can use different model reparametrizations, in order to see that some parametrizations are perhaps more informative for the application. Analogously, this advantage is also valid with the composite likelihood. Specifically, in our setting, by using a particular transformation of the model parameters we can get a more flexible model for the spatial extremes.

We describe the further model extension as follows. Consider the Gaussian extreme value model (5.3) with 2-dimensional density function (5.5). We recall that a compelling feature of the model is that the marginal are unit Frechét with the form $F(z) = \exp(-1/z)$. Suppose now the following transformation $(S(\mathbf{t}_1), S(\mathbf{t}_2)) = g(Z(\mathbf{t}_1), Z(\mathbf{t}_2))$ that is,

$$\begin{aligned} S(\mathbf{t}_1) &= \psi(\mathbf{t}_1)\left\{Z(\mathbf{t}_1)^{\xi(\mathbf{t}_1)} - 1\right\}/\xi(\mathbf{t}_1) + \mu(\mathbf{t}_1) \\ S(\mathbf{t}_2) &= \psi(\mathbf{t}_2)\left\{Z(\mathbf{t}_2)^{\xi(\mathbf{t}_2)} - 1\right\}/\xi(\mathbf{t}_2) + \mu(\mathbf{t}_2), \end{aligned}$$

for some opportune values $-\infty < \mu(\mathbf{t}_i), \xi(\mathbf{t}_i) < \infty$ and $\psi(\mathbf{t}_i) > 0$ with $i = 1, 2$. Note that we focused on the indexes $\mathbf{t}_1$ and $\mathbf{t}_2$ but the discussion is valid for arbitrary indexes $\mathbf{t}_i$ and $\mathbf{t}_j$ belonging to the set of all the pairs. Observe that the location, scale and shape parameters do not have to be necessarily different with $\mathbf{t}_i$. The inverse transformation of $(Z(\mathbf{t}_1), Z(\mathbf{t}_2)) \to (S(\mathbf{t}_1), S(\mathbf{t}_2))$ is

$$\begin{aligned} Z(\mathbf{t}_1) &= \left\{1 + \xi(\mathbf{t}_1)\left(\frac{S(\mathbf{t}_1) - \mu(\mathbf{t}_1)}{\psi(\mathbf{t}_1)}\right)\right\}^{\frac{1}{\xi(\mathbf{t}_1)}} \\ Z(\mathbf{t}_2) &= \left\{1 + \xi(\mathbf{t}_2)\left(\frac{S(\mathbf{t}_2) - \mu(\mathbf{t}_2)}{\psi(\mathbf{t}_2)}\right)\right\}^{\frac{1}{\xi(\mathbf{t}_2)}}, \end{aligned}$$

with jacobian determinant

$$\begin{aligned} |J(s_1, s_2)| &= \frac{1}{\psi(\mathbf{t}_1)\psi(\mathbf{t}_2)}\left\{1 + \xi(\mathbf{t}_1)\left(\frac{s_1 - \mu(\mathbf{t}_1)}{\psi(\mathbf{t}_1)}\right)\right\}^{\frac{1}{\xi(\mathbf{t}_1)} - 1} \\ &\cdot \left\{1 + \xi(\mathbf{t}_2)\left(\frac{s_2 - \mu(\mathbf{t}_2)}{\psi(\mathbf{t}_2)}\right)\right\}^{\frac{1}{\xi(\mathbf{t}_2)} - 1}, \end{aligned}$$

Now if the random vector $\{S(\mathbf{t}_1), S(\mathbf{t}_2)\}$ has GEV marginal distributions and more precisely we say,

$$S(\mathbf{t}_1) \sim \text{GEV}(\mu(\mathbf{t}_1), \psi(\mathbf{t}_1), \xi(\mathbf{t}_1)) \quad S(\mathbf{t}_2) \sim \text{GEV}(\mu(\mathbf{t}_2), \psi(\mathbf{t}_2), \xi(\mathbf{t}_2))$$

then the random variables $Z(\mathbf{t}_1)$ and $Z(\mathbf{t}_2)$ have unit Frechét distribution functions, and the joint distribution function of $(Z(\mathbf{t}_1), Z(\mathbf{t}_2))$ is of form (5.3). Consequentially, from the relationship established with the transformation $(Z(\mathbf{t}_1), Z(\mathbf{t}_2)) \rightarrow (S(\mathbf{t}_1), S(\mathbf{t}_2))$, we have that $(S(\mathbf{t}_1), S(\mathbf{t}_2))$ has the distribution function $F(s_1, s_2) = F(g^{-1}(s_1), g^{-1}(s_1))$ and bivariate density

$$f(s_1, s_2) = f(g^{-1}(s_1), g^{-1}(s_2))|J(s_1, s_2)|. \tag{5.7}$$

Combining (5.7) with (5.5) yields the analytical expression of the bivariate density function of $(S(\mathbf{t}_1), S(\mathbf{t}_2))$ with marginal $S(\mathbf{t}_1)$ and $S(\mathbf{t}_2)$ that have GEV distributions. In other words, the previous transformation produces a bivariate model for spatial extremes but with GEV marginal distributions. With this model parametrization, by using the composite likelihood approach for inference we can not only estimate and asses the tail dependence between the pairs of sites but also asses the rates at which the extreme rainfall levels occur at the sites. The complete set of parameters can be estimated by maximization of the composite likelihood. In practice due to the intractable expression of the bivariate density function, consequently we use the quasi-Newton numerical maximization routines (e.g. Broyden, 1967) in order to maximize the composite log likelihood.

In principle, we could assume marginal GEV distributions with different parameters in each location. For K fixed locations the resulting spatial model has $3K + 3$ parameters that have to be estimated. Although K might be relatively small, we say $K \ll n$ where n is the sample size for each site, the likelihood maximization might be computationally intensive, given that we use numerical maximization routines. Instead, for large K the likelihood maximization turns out to be computationally prohibitive. Alternatively, we can consider a common GEV model with the same parameters (μ, ψ, ξ) for all the sites. However, perhaps in practice few cases can be conformed with these restrictive model assumptions. Thus a solution may be provided by the parsimonious regression model. More precisely, we assume that the marginal have distributions

$$S(\mathbf{t}_k) \sim \text{GEV}(\mu(\mathbf{t}_k), \psi(\mathbf{t}_k), \xi(\mathbf{t}_k)) \quad k = 1, \ldots, K, \tag{5.8}$$

where μ, ψ and ξ are polynomial surfaces of forms

$$f(t_{1,k}, t_{2,k}) = \sum_{i+j \leq p} \beta_{i,j}\, t_{1,k}^i t_{2,k}^j,$$

where $t_{1,k}, t_{2,k} \in \mathbb{R}$ are the plane coordinates of the site $\mathbf{t}_k$ and the quantity p is the order of the surface. Thus there are $(p+1)(p+2)/2$ coefficients. In this way by defining a spatial regression model of degree p we establish a spatial dependence between the GEV parameters. Consequentially, the resulting GEV model has different parameters (μ, ψ, ξ) for each site but only using a relatively small number of regression coefficients. The tail dependence and the regression parameters, $(\zeta, \gamma, \rho, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ represents the regression coefficients for μ, ψ and ξ, can be estimated all together by maximization of the composite likelihood.

In particular the pairwise log likelihood of the previous section is modified including the regression extension as follows

$$\ell_P(\zeta, \gamma, \rho, \boldsymbol{\beta}) = \sum_{j \neq k \in \mathcal{K}} \sum_{i=1}^{n} \log f(s_{i,k}, s_{i,j}; \zeta, \gamma, \rho, \boldsymbol{\beta})$$

where j, k belongs to $\mathcal{K}$, the set index of all the pairs, $\boldsymbol{\beta}$ is the vector of regression coefficients with length that depends by the polynomials degrees of the spatial regressions, and n is the sample size in each location.

5.3.3 Simulation Study

In this section by means of a simulation study we illustrate the use of the composite likelihood method in the spatial extremes context. We conducted a series of simulation exercises using different parameter settings and spatial designs in order to asses the finite sample properties of the composite likelihood estimator.

In particular, the study is led by assuming the Gaussian extreme value model. The model parameters are (ζ, γ, ρ) which define the covariance matrix of a bivariate normal distribution, as it is here reported,

$$\Sigma_2 = \left[\begin{array}{cc} \gamma^2 & \rho\,\gamma\,\zeta \\ \rho\,\gamma\,\zeta & \zeta^2 \end{array} \right].$$

The bivariate normal density (the storm profile model) is related with the Gaussian extreme value model by (5.3). Note that, for fixed sites over a region, the covariance matrix values impact on tail dependence, in determining strong, weak or mild dependence of the extreme events between the locations.

The simulation exercises are performed using an irregular grid of points (locations) over a region of 40×40. In practice the points are randomly located, generating K random values from a uniform variable on an interval $[a, b]$ for each axis, and then these realizations define the plane coordinates of the points. Table 5.1 shows the parameters settings considered in the study. With rows 1–3 we take into account different shapes of the storm profile. Also empirically, for this spatial design, we have found that the settings illustrated in entries 4 and 5 represent strong and weak dependence for most of the pairs, which is of interest in our study.

Parameters settings	γ	ρ	ζ
1. Same strength in both directions	17.32	0	17.32
2. Different strength in both directions	14.14	0	17.32
3. Spatial correlation	14.14	0.61	17.32
4. Strong dependence	44.72	0.61	54.77
5. Weak dependence	4.47	0.61	5.48

Table 5.1: *Parameters configurations for the extreme dependencies: the first column reports the parameter configuration's names. Columns 2–4 report the different parameter values.*

Realizations of simulated data from the model can be performed by implementing the guidelines illustrated in Section 5.3.3. Already implemented software is available in the statistical environment `R`, by the package `RandomFields`, see `http://cran.r-project.org/`.

For each site, n independent replications are generated from the max-stable process and we also fit the model using the same number of observations for each location. The composite log likelihood function has been implemented in `R` and the numerical maximization routine `optim` is used in order to obtain the parameter estimates. The composite log likelihood function, given K sites, is made up of $K(K-1)/2$ distinct log likelihoods. For large K the maximization of the likelihood is computationally demanding, so in some cases it could be necessary to implement in the general-purpose language C the procedures in favor of reducing the computational time.

One simulation exercise illustrates the performance with moderate dataset, $K = 50$ sites and $n = 100$ observations for each site, under different dependence models (listed in Table 5.1). Table (5.2) summarizes the estimation results computed over 500 data replications. From columns 3–4 the averages and the standard deviations are reported for the models. The simulation results indicate good correspondence between the true parameters and the estimates mean and also slightly smaller variances for all four cases. There is no evidence of bias in the estimation of the parameters even in cases of strong and weak dependence. This is good news because other methods, as for example the one proposed by de Haan and Pereira (2006) perform badly in the case of weak dependence.

Models	$\hat{\gamma}$	$\hat{\rho}$	$\hat{\zeta}$
1	17.35(1.28)	0.012(0.117)	17.34(1.42)
true	17.32	0	17.32
2	14.18(1.14)	0.016(0.109)	17.43(1.26)
true	14.14	0	17.32
3	14.15(1.24)	0.609(0.062)	17.44(1.23)
true	14.14	0.61	17.32
4	44.66(3.97)	0.602(0.072)	54.74(5.34)
true	44.71	0.61	54.77
5	4.48(0.21)	0.613(0.038)	5.51(0.22)
true	4.47	0.61	5.48

Table 5.2: *Estimation results: composite likelihood estimates based on 500 simulations of spatial extreme data (100 observations per site) using the Gaussian extreme value family. The first column shows the models contemplated (see Table 5.1). Columns 2–4 show the estimates mean and between parenthesis the standard deviations.*

The second exercise shows the performance in small, moderate and large dataset under the dependence model of row three in Table 5.1. Specifically, we take into account 10, 50 and 100 sites with 100 observations for each . Table 5.3 summarizes the estimation results based on 500 data realizations.

The simulations indicate that some bias occurs, and that the estimate variances are large with small dataset, as for example in the case when $K = 10$ and $n = 10$. With an increasing number of observations per site the estimated parameters have negligible bias and variance (for instance with $n = 200$). However, we can also see that the number of sites does not have much effect on the estimates but a reduction of the variability can be observed. For example, a comparison between 10 and 100 sites with 10 observations in Table 5.3 particulary shows that with the increasing of the number of sites, the estimate variances decrease.

K	n	$\hat{\gamma}$	$\hat{\rho}$	$\hat{\zeta}$
10	*10*	*15.83(4.01)*	*0.559(0.272)*	*17.63(7.00)*
-	*50*	*14.44(1.72)*	*0.820(0.255)*	*17.69(2.01)*
-	**100**	**14.24(1.36)**	**0.613(0.073)**	**17.45(1.53)**
-	*200*	*14.10(0.97)*	*0.612(0.050)*	*17.32(0.98)*
50	*10*	*15.44(3.10)*	*0.610(0.173)*	*18.85(3.81)*
-	*50*	*14.35(1.43)*	*0.612(0.078)*	*17.48(1.78)*
-	**100**	**14.15(1.24)**	**0.609(0.062)**	**17.44(1.23)**
-	*200*	*14.11(0.65)*	*0.612(0.037)*	*17.25(0.86)*
100	*10*	*15.43(2.78)*	*0.603(0.164)*	*18.37(2.99)*
-	*50*	*14.38(1.39)*	*0.608(0.086)*	*17.66(1.69)*
-	**100**	**14.29(1.05)**	**0.624(0.051)**	**17.48(1.25)**
-	*200*	*14.14(0.93)*	*0.603(0.062)*	*17.36(1.23)*
	true	14.14	0.61	17.32

Table 5.3: *Estimation results: composite likelihood estimates based on 500 simulations of spatial extreme data using the Gaussian extreme value model with parameters given by row three of Table 5.1. Columns 1–2 describe the number of sites and observations per site. Columns 3–5 report the estimates mean and between parenthesis the standard deviations.*

5.4 A proposal for Bayesian inference

Recently, in situations where the full maximum likelihood estimator is not available, the Approximate Bayesian Computation (ABC) method has been confirmed as an alternative procedure for inference. Many other applications have since been illustrated for example, Beaumont, Zhang and Bolding (2002), Marjoram, Molitor, Plagnol and Taveré (2003) and in the context of the extreme values by Bortot, Coles and Sisson (2006). Approximate Bayesian Computation methods consist of stochastic simulation algorithms born in a statistical genetics context with the aim of providing an inference technique for the model parameters when the likelihood function is not available. The algorithms can be based on the rejection method Beaumont, Zhang and Bolding (2002) or on Markov chain Monte Carlo (MCMC) techniques Marjoram, Molitor, Plagnol and Taveré (2003). However the common idea is to substitute the likelihood evaluation with the model simulation when it is feasible and not too computationally intensive. The MCMC approaches have demonstrated satisfactory results, see for example Bortot, Coles and Sisson (2006). But the price to pay for avoiding the likelihood evaluation is in many cases long simulation runs. The rejection method approaches (Beaumont *et al*, 2002) can be characterized by fewer low acceptance rates given the global comparison used in the acceptance step. Also, the major difficulty is that the constant c (the likelihood function evaluated at the maximum likelihood estimate), which ensures the total envelope used in the acceptance step, is not easy to derive for most of the non trivial cases, as for example in continuous problems.

The last inconvenience can be overtaken by using an alternative method that avoids the likelihood maximization but only gives a sample that is approximately from the posterior distribution. This is known as Sampling Importance Resampling (SIR). Then we propose here an alternative ABC approach based on the SIR algorithm.

Let $\mathbf{Z}$ be a random vector variable belonging to $\mathcal{Z} \subseteq \mathbb{R}^n$ with probability density function $f(\mathbf{z}|\boldsymbol{\psi})$ parameterized by $\boldsymbol{\psi} \in \Psi \subseteq \mathbb{R}^n$. Denoting with $\pi(\boldsymbol{\psi})$ the prior distribution on $\boldsymbol{\psi}$, then from the Bayesian paradigm the posterior distribution of interest is $f(\boldsymbol{\psi}|\mathbf{z})$, which is given by

$$f(\boldsymbol{\psi}|\mathbf{z}) = \frac{f(\mathbf{z}|\boldsymbol{\psi})\,\pi(\boldsymbol{\psi})}{\int f(\mathbf{z}|\boldsymbol{\psi})\,\pi(\boldsymbol{\psi})d\,\boldsymbol{\psi}}.$$

From the posterior distribution, opportune quantities such as for example the posterior mean, median, etc. can be computed in order to estimate the model parameters. In many situations the posterior density can be hard to derive explicitly. We have already argued that some techniques, as for example the rejection method, can not be easily implemented. The SIR method can represent a suitable alternative in order to get an approximate sample from the posterior.

SIR (Sampling Importance Resampling) is a common method in computational statistics for generating samples from difficult distributions. SIR was first described by Rubin (1987) and it has been used, for example to generate samples from Bayesian posterior distributions Gelman, Carlin, Stern and Rubin (2004) or in sequential importance sampling and particle filtering. Consider temporarily, the case when the random variable Y is discrete. Assume that Y has distribution q. Then SIR can generate samples that approximately have that distribution. To do so, we generate a set of "proposal" samples from a source distribution, p, weight these samples appropriately, then resample these with probability proportional to their weights. Note, we are assuming that the distribution p is defined on the same sample space of q. The SIR algorithm is given by

```
**Algorithm SIR1**: Sampling Importance Resampling's basic algorithm

1. Generate $\{x_1, \ldots, x_N\}$, $N$ proposals from the distribution $p$.

2. Compute the weight $w(x_n) = q(x_n)/p(x_n)$ for $n = 1, \ldots, N$.

3. Draw $\{y_1, \ldots, y_M\}$, $M$ samples $(M \leq N)$ from $\{x_1, \ldots, x_N\}$ with
   replacement and probability proportional to $w(x_m)$ for $m = 1, \ldots, M$.
```

The resulting samples will be approximately distributed according to the unnormalized function $\widehat{q}$. Specifically, $q = \widehat{q}/C$ where C is a normalizing constant. The constant C can be estimated by $C = \sum_{n=1}^{N} w(x_n)$. So, in other words the effect of the resampling step is to take proposals from the distribution p, and "filter" them, so that the samples have a distribution that approximates q. Observe that when the number of proposals increase, we say $N \to \infty$, then the distribution of each sample approaches q. Note that this method is still valid in the continuous case. The algorithm remains the same, but we need to substitute q with the density function $f(y)$ and the proposal distribution p with the density $g(x)$. The samples from $f(y)$ can be obtained by drawing $x_1, \ldots, x_N$ from $g(x)$ and resampling them from the discrete distribution on the set $\{x_1, \ldots, x_N\}$ with probabilities $w(x_i) = f(x_i)/g(x_i)/\sum_{n=1}^{N} f(x_n)/g(x_n)$ for $i = 1, \ldots, N$.

The SIR method is useful when adopted into a Bayesian context in situations where other methods can not be applied. The method provides an approximate sample from the desired posterior density, as suggested by the following algorithm.

```
Algorithm SIR2: Sampling Importance Resampling's Bayesian algorithm

1. Generate $\{\psi_1, \ldots, \psi_N\}$, $N$ values from the density $g(\psi)$.
2. Compute the weight $w_n = \frac{f(\psi|z)/g(\psi)}{\sum_{n=1}^{N} f(\psi|z)/g(\psi)}$ for $n = 1, \ldots, N$.
3. Draw $\{\psi_1, \ldots, \psi_M\}$, $M$ samples ($M \leq N$) from $\{\psi_1, \ldots, \psi_N\}$ with replacement and probabilities $\{w_1, \ldots, w_N\}$.
```

An approximate sample from the posterior is generated by a weighted resampling. The sample $\{\psi_1, \ldots, \psi_N\}$ is drawn from the density $g(\psi)$ and then we resample from the discrete distribution on the set $\{\psi_1, \ldots, \psi_N\}$ with probabilities w_n. Note that given the relation $f(\psi|z) = kf(z|\psi)\pi(\psi)$ we obtain the following expression for the weights,

$$w_i = \frac{f(z|\psi_i)\pi(\psi_i)/g(\psi_i)}{\sum_{n=1}^{N} f(z|\psi_n)\pi(\psi_n)/g(\psi_n)} \quad \text{for} \quad i = 1, \ldots, N.$$

These weights may be used in the **SIR2** algorithm so that an approximate sample from the posterior can be obtained without necessary knowing $f(\psi|z)$. Now if the sampling density $g(\psi)$ is equal to the prior density $\pi(\psi)$ then the weights assume the form

$$w_i = \frac{f(z|\psi_i)}{\sum_{n=1}^{N} f(z|\psi_n)} \quad \text{for} \quad i = 1, \ldots, N.$$

Example 1: Assume that the random variable Z has Bernoulli distribution with probability ψ, $Z \sim Be(\psi)$. Assume also that the prior for the model parameter is the conjugate beta distribution $\psi \sim B(\alpha, \beta)$.

The conjugate prior distribution has density: $\pi(\psi) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\psi^{\alpha-1}(1-\psi)^{\beta-1}$. The likelihood function has the expression: $f(z|\psi) = \psi^k(1-\psi)^{n-k}$, where $k = \sum_{i=1}^{n} z_i$ and n is the sample size. The posterior density is proportional to, $f(\psi|z) \propto \psi^{k+\alpha-1}(1-\psi)^{n-k+\beta-1}$. This result leads to the posterior $\psi|z$ having distribution $B(\alpha+k, \beta+n-k)$. This example is useful because it allows us to compare the samples obtained from the **SIR2** algorithm with the posterior density of which we know the analytical expression. By means of diagnostic plots and summary statistics we can roughly asses the level of approximation that the SIR method provides.

So we have simulated $z_1, \ldots, z_n$, $n = 100$ synthetic data from a Bernoulli distribution with probability $\psi = 0.3$. Then we have generated approximate samples from the posterior distribution using the **SIR2** algorithm. The prior distribution's set up is $B(5, 5)$ from which we have drawn respectively 500, 1000 and 10000 particles, and resampled 500, 1000 and 2000 values. In Figure 5.1 the results are illustrated. We can see from the histograms that the samples provide a reasonable approximation of the posterior density (solid line). The quantile-quantile plots show the adequacy of the applications for all three cases. However we can also

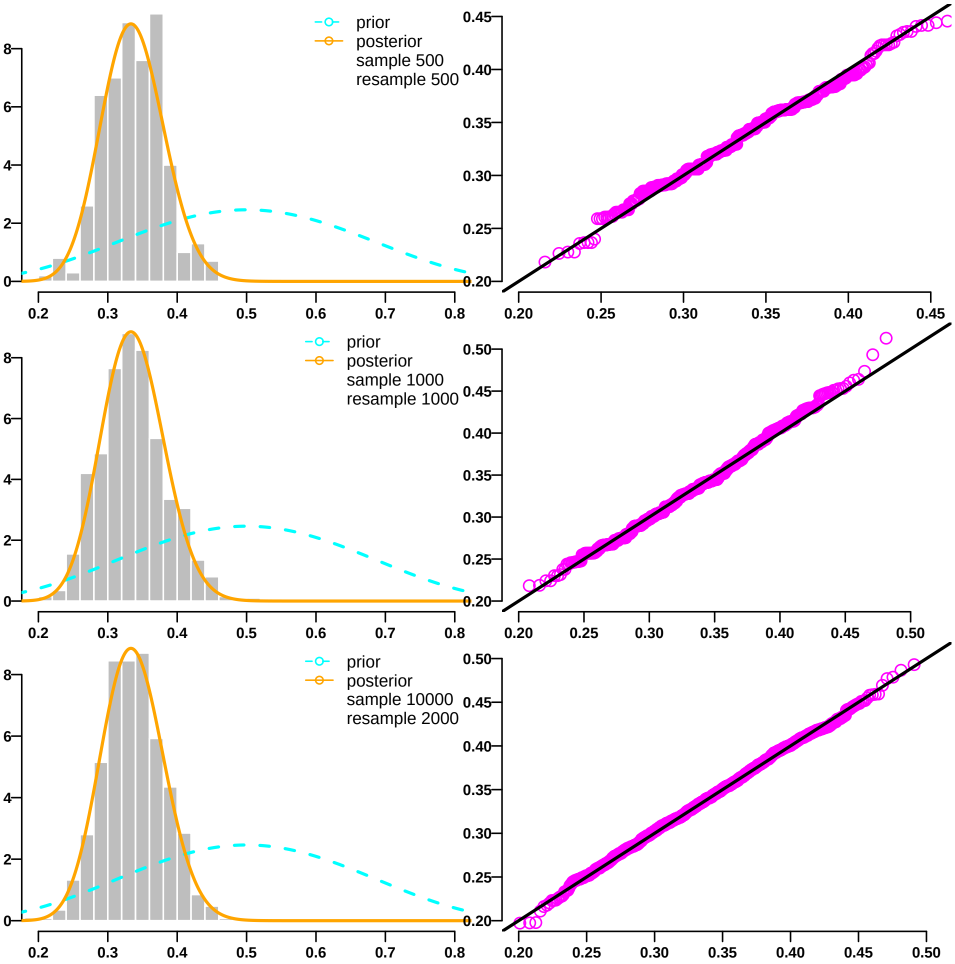

Figure 5.1: *Results example 1: the top to the bottom left panels show the histograms of samples drawn for different sampling and resampling values. The solid lines show the posterior and the broken lines the prior densities. The right panels show the quantile-quantile plots of the true posterior versus the approximate samples.*

see that the approximation improves with the increasing of the sampling and resampling sizes. Moreover, we can also compare the expected value and the variance of the posterior density with the average and variance of the approximate samples, in order to evaluate the accuracy of the approximation. The expected value and the variance of the posterior distribution are:

$$E(\psi|z) = \frac{\alpha + k}{\alpha + \beta + n} \quad \text{and} \quad V(\psi|z) = \frac{(\alpha + k)(\beta + n - k)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}, \tag{5.9}$$

In table 5.4 the averages and the standard deviations of three samples computed with different sampling resampling sizes are reported. In the last row by using the simulated data and formulas (5.9) the mean and the standard deviations of the posterior density are given. In all three cases we can see that the averages and the standard deviations computed from the approximate sample are very close to those of the posterior.

Suppose now that the likelihood evaluation is not available or too time consuming. Observe that for discrete data a sample from the posterior can still be obtained by applying the rejection method under the ABC approach (e.g Marjoram, 2003). The authors also suggested further algorithm modifications in cases

N	M	$\overline{\psi}\|z$	sd$(\psi\|z)$
500	500	0.337	0.043
1000	1000	0.336	0.047
10000	2000	0.337	0.044
		0.336	0.044

Table 5.4: *Summary statistics: columns 1–2 report the sampling and resampling sizes. Columns 3–4 report the means and the standard deviations computed using the approximate samples. The last row gives the mean and the standard deviations of the posterior density with the simulated data.*

when the acceptance rate is too small. However when data are high-dimensional and continuous this approach can be impractical due to the absence of the upper bound of the likelihood. Then one can resort to the ABC version of the approximate SIR method. Essentially, in cases where the underlying stochastic model is easy to simulated and not too computationally demanding the ABC approach to substitute the likelihood evaluation with the model simulation. The method is describe by the following algorithm.

Algorithm SIR3: Sampling Importance Resampling's ABC algorithm

```
1. Generate $\psi_i$, from the prior $\pi(\psi)$.
2. Simulate $z_i^{sim}$ from the model with density $f(z|\psi)$ and particle $\psi_i$,
   and the corresponding statistics $s_i^{sim}$.
3. Compute the distance $d_i = \rho(s_i^{sim}, s_i^{obs})$ and the weight $w_i = K(d_i, \epsilon)I_{(d_i \leq \epsilon)}$.
4. Repeat steps 1–3 until $N$ particles with $w_i > 0$ are obtained.
5. Draw $\{\psi_1, \ldots, \psi_M\}$, $M$ samples $(M \leq N)$ from $\{\psi_1, \ldots, \psi_N\}$ with
   replacement and proportional weights $\{w_1, \ldots, w_N\}$.
```

Essentially, for each generated particle from the prior, data are simulated from the underlying stochastic model. Weights for the samples $\{\psi_1, \ldots, \psi_N\}$ are computed based on a kernel function of the distances between the observed and simulated data using an opportune metric. Then the resulting sample set $\{\psi_1, \ldots, \psi_N\}$ is resampled with proportional weights $\{w_1, \ldots, w_N\}$.

More precisely, the likelihood evaluation is substituted by the comparison between the observed and simulated data. The comparison is done by selecting a suitable metric ρ and an interval ϵ. When $\epsilon \to 0$ the matching of observed and simulated data is required. Instead for $\epsilon \to \infty$ the particles are not "filtered" so that we get the prior density. The observed and simulated data match with frequency proportional to $f(y)$, which can be very low. In order to avoid this a comparison can be done between low-dimensional summaries of the observed and simulated data. The key point is that if the summary statistics S is sufficient then $f(z|\psi) = f(s|\psi)f(y|s)$, where the second term on the right does not depend on ψ. If the reduction of the data dimension is given by sufficient statistics it is without loosing information about ψ. Practically, the advantage of using sufficient statistics is that the matching of observed and simulated data occurs more frequently.

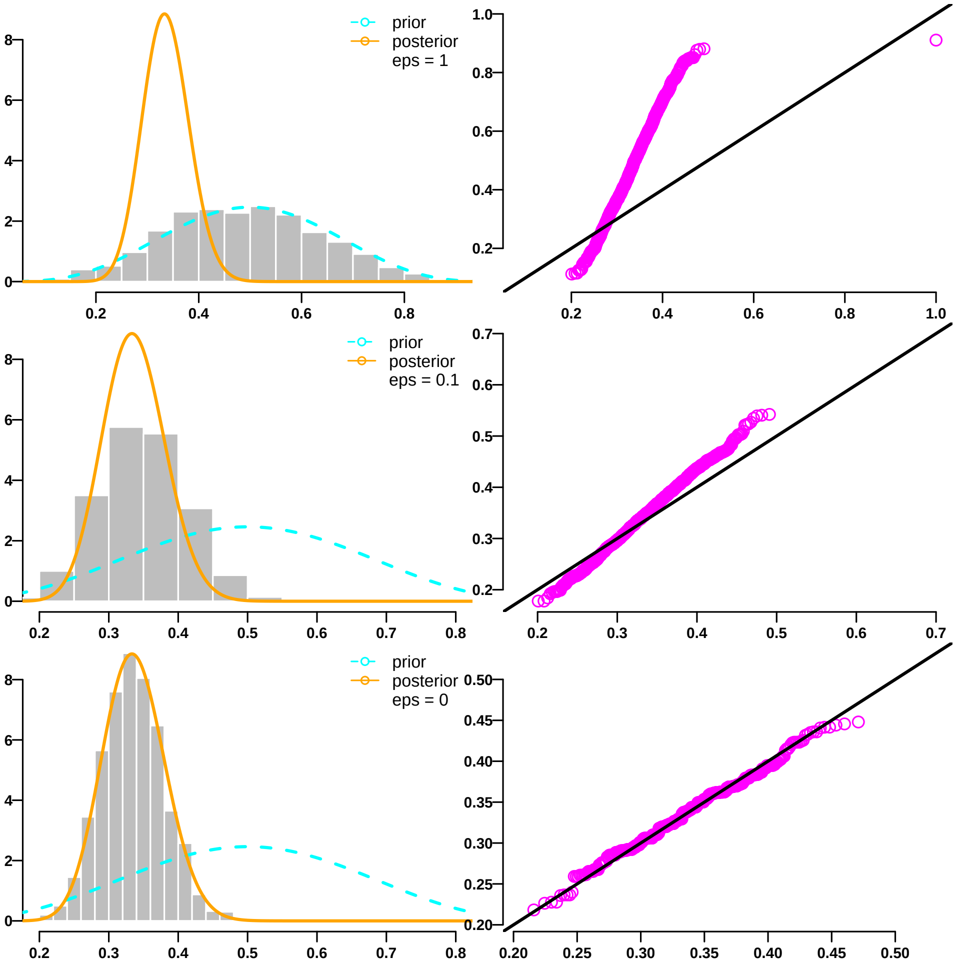

Figure 5.2: *Results example 1: the top to the bottom left panels show the histograms of samples drawn for different tolerances. The solid lines show the posterior and the broken lines the prior densities. The right panels illustrate the quantile-quantile plots between the true posterior versus the approximate samples.*

Finally, the particles that satisfy the criteria, that is the corresponding data are close to the observed for a value less or equal to ϵ, are resampled with proportional weights w_i. These are computed by using a weighting function $K(d_i, \epsilon)I_{(d_i \leq \epsilon)}$, of the distances d_i and interval ϵ. To clarify, the weighting function is composed by a kernel function $K(\cdot, \cdot)$ as for example the Epanechnikov, and an indicator function $I(\cdot)$. So, for $d_i \leq \epsilon$ the respective weighs are positive, and in particular for $d_i \rightarrow 0$ we have $w_i \rightarrow 1$. Instead for $d_i > \epsilon$ the respective weighs are 0. Concluding, the weights are normalized as $w_i = w_i / \sum_{j=1}^{N} w_j$ for $i = 1, \ldots, N$. Observe that with discrete data cases setting $\epsilon \rightarrow 0$ the matching between the observed and simulated data is required. The particles associated to such events are imposed unormalized weighs equal to 1. The algorithm for these situations correspond to that of the rejection method.

Consider once again the illustrative study case given by *example1*. Now, considering the same data of the previous example, we emphasize the approximation level that the approximate samples provide of the posterior density by using the

SIR3 algorithm. We run the **SIR3** algorithm selecting three different tolerances: 1, 0.1 and 0. We fixed the number of acceptances equal to 5000 and then we have resampled 2000 values for each tolerance. The results are illustrated in Figure 5.2. The summary statistics used with the algorithm is the sample mean or success frequencies, $\overline{y} = \sum_{i=1}^{n} y_i/n$. We can see from the histograms that decreasing the tolerance, the approximate samples provide more and more accurate approximation of the posterior density. In fact, it is possible to see from the bottom left and right panels, that reducing the tolerance to zero we reach a high level of accuracy with the approximation. This is also confirmed with the summary statistics results in Table 5.5. With the decreasing of the tolerance values the discrepancy between the mean and standard deviations obtained from the approximate samples and those of the posterior decrease towards zero. In fact this is the case when $\epsilon = 0$.

M	ϵ	$\overline{\psi}\|z$	sd$(\psi\|z)$
2000	1	0.487	0.148
2000	0.1	0.347	0.061
2000	0	0.337	0.044
		0.336	0.044

Table 5.5: *Summary statistics: columns 1–2 report the resampling number and tolerances. Columns 3–4 report the means and the standard deviations computed using the approximate samples. The last row reports the mean and the standard deviations of the posterior density with simulated data.*

Example 2: Assume that the random variable Z has normal distribution with known mean μ and unknown variability σ^2, $Z \sim N(\mu, \sigma^2)$. Assume also that the prior for the model parameter is the conjugate inverse Gamma distribution $\sigma^2 \sim \mathrm{I}\Gamma(\alpha, \beta)$.

The conjugate prior distribution has density: $\pi(\psi) = \frac{\beta^\alpha}{\Gamma(\alpha)}\psi^{-\alpha-1}\exp\left(\frac{-\beta}{\psi}\right)$. The likelihood function has the expression: $L(z|\psi) = (2\pi\psi)^{-n/2}\exp(-k/\psi)$, where $k = \sum_{i=1}^{n}(z_i - \mu)^2/2$ and n is the sample size. Then the posterior distribution is proportional to $f(\psi|z) \propto \psi^{-\alpha-n/2-1}\exp\{-(\beta+k)/\psi\}$. So it turns out that the posterior $\psi|z$ has distribution $\mathrm{I}\Gamma(\alpha + n/2, \beta + k)$. We can test the performance of the SIR algorithms comparing their approximate posterior samples, with the posterior density for which we know the analytical expression. We simulated $z_1, \ldots, z_n$, $n = 100$ synthetic data from a normal distribution with variance $\psi = 9$. Then approximate posterior samples have been obtained by **SIR2** and **SIR3** algorithms drawing $\psi_1, \ldots, \psi_N$, particles from the prior with $\alpha = 0.5$ and $\beta = 0.5$. In particular with the first algorithm we have sampled respectively 500, 1000 and 50000 particles and then resampled 500, 1000 and 2000 values. Instead, with the second algorithm we fixed three tolerances $\epsilon = 10$, $\epsilon = 2$ and $\epsilon = 0.1$. Then we have set up the acceptance number $N = 5000$ and resampled $M = 2000$ values for all three cases. The summary statistic used with **SIR2** is the adjusted sample variance: $s = n\sum(y_i - \overline{y})^2/(n-1)$. We can see from the histograms of Figure (5.3) that even with a continuous variable, the approximate samples from the posterior density (obtained by using the SIR algorithms) provide a reasonable approximation of the latter. In particular with both algorithms, increasing the sampling size for **SIR2** and decreasing the tolerance with **SIR3** we obtain samples from the

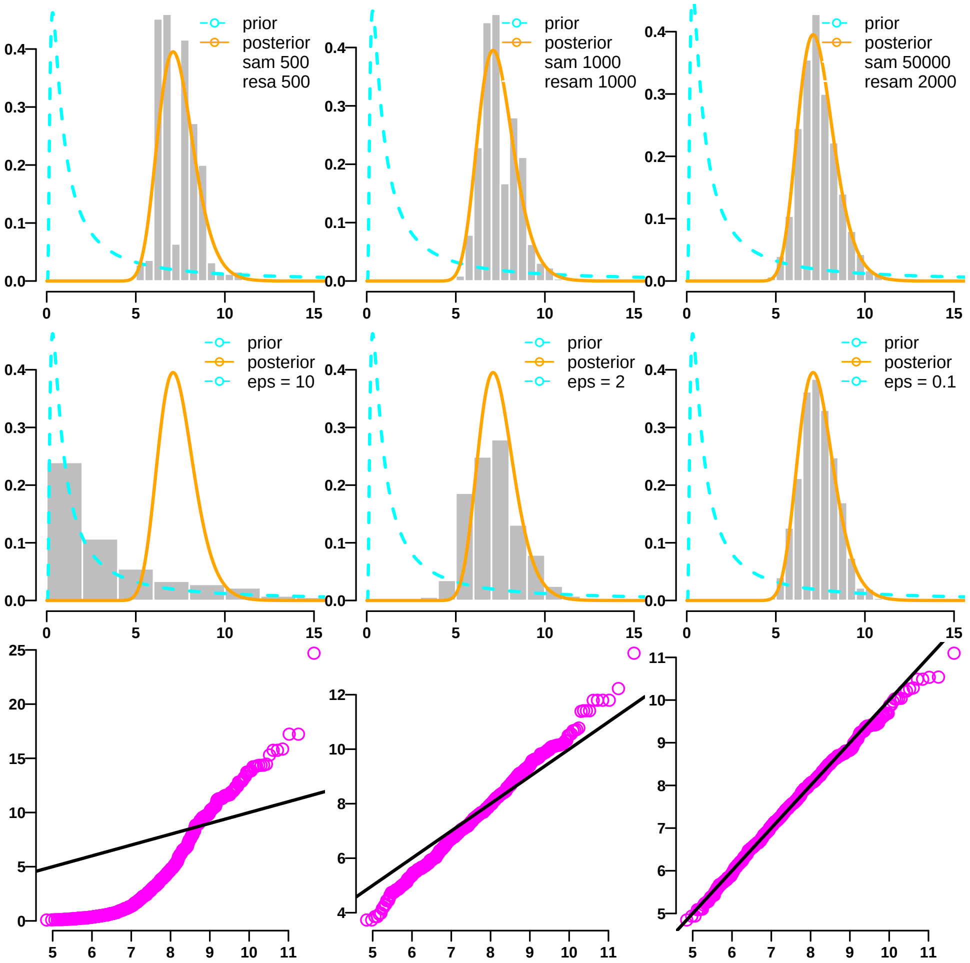

Figure 5.3: *Results example 2: the first row shows the histograms of samples obtained using **SIR2** and the second row using **SIR3**. The solid lines show the posterior and the broken lines the prior densities. The third row illustrates the quantile-quantile plots between the true posterior versus the approximate samples (**SIR3**).*

posterior that provide a better approximation. Figure (5.3) illustrates accurate approximation levels in the panels of the third column, first and second rows. In the third row we have reported the quantile-quantile plots of the posterior density versus the approximate sample obtained with the **SIR3** algorithm. Specifically, the bottom right panel shows that with small tolerance, $\epsilon = 0.1$, the sample provides an accurate approximation of posterior density. Now the expected value and the variance of the posterior distribution are:

$$E(\psi|z) = \frac{\beta + k}{\alpha + n/2 - 1} \quad \text{and} \quad V(\psi|z) = \frac{(\beta + k)^2}{(\alpha + n/2 - 1)^2(\alpha + n/2 - 2)},$$

then we also compared the estimates obtained using the simulated data with the averages and variances obtained from the approximate samples. The results are summarized in Table 5.6. We can see from the third and sixth rows that the differences between the mean and variances computed from the samples and those of the posterior density are close to zero.

We don't posses any rigorous method that demonstrates mathematically the goodness of the approximate sample from the posterior obtained with SIR algorithms, respect with the true posterior. Then ,until here, we have considered

Algorithm	N	M	ϵ	$\overline{\psi}\|z$	sd$(\psi\|z)$
SIR2	500	500	-	7.357	1.054
SIR2	1000	1000	-	7.472	0.992
SIR2	50000	2000	-	7.393	1.077
SIR3	-	2000	10	3.465	3.551
SIR3	-	2000	2	7.170	1.417
SIR3	-	2000	0.1	7.376	1.030
				7.378	1.059

Table 5.6: *Summary statistics: columns 1–3 report the sampling, resampling sizes and tolerances. Columns 4–5 report the means and the standard deviations computed using the approximate samples. The last row reports the mean and the standard deviations of the posterior density with simulated data.*

simple problems where we knew the analytical expression of the likelihood function and posterior density. Therefore for these cases none of the approximation methods are required for inference. But by means of illustrative examples we have compared the performance of the SIR approximate methods with the known posterior density. We have seen with these examples how our proposed method provides a reasonable surrogate to a sample from the posterior density, making the SIR algorithm that we propose useful in problems where the posterior density is not available. For instance in the case of spatial extreme problems introduced in the previous section. We will discuss now with an illustrative example the application of the SIR method in the spatial extreme context. Observe that with the spatial extreme models arising from the max-stable processes, the maximum likelihood method can not be applied for the model inference when we deal with a K-variate distribution with $K > 2$. Then alternative methods such as for example the SIR technique are required. Assume the Gaussian extreme value process with bivariate marginal distribution (5.3). Assume also for simplicity that the covariance matrix Σ_2 of the storm profile has components $\gamma = \zeta$ and $\rho = 0$. We consider in particular the case study given by entry 1 of Table 5.2. In order to conduct the simulation example we need to specify the metric and the summary statistics used in the extreme value setting with the SIR3 algorithm. In particular as summary statistics we used the fitted regression model given by the extremal coefficients versus the Euclidean distances of the pairs. More precisely the regression model is formalized as

$$\nu(\mathbf{h}_k) = g(\|\mathbf{h}_k\|) + \varepsilon_k$$

where $E(\varepsilon_k) = 0$, $V(\varepsilon_k) = \sigma^2$ and

$$g(\|\mathbf{h}_k\|) = \beta_0 + \beta_1\|\mathbf{h}_k\| + \sum_{p=1}^{P} b_p(\|\mathbf{h}_k\| - \kappa_p)_+,$$

with $\|\mathbf{h}_k\|$ the Euclidean distances of the k's pair of locations ($\mathbf{h}_k$ is the vector of plane coordinate differences), $\{\kappa_1, \ldots, \kappa_P\}$ is a set of knots defined on the space of the distances, and b_p are coefficients. In this way, g should be flexible enough to model the relation structure between extremal coefficients and the distances of the pairs, instead of assuming and establishing some particular parametric models. In order to fit such a model we need to somehow provide an estimate of the

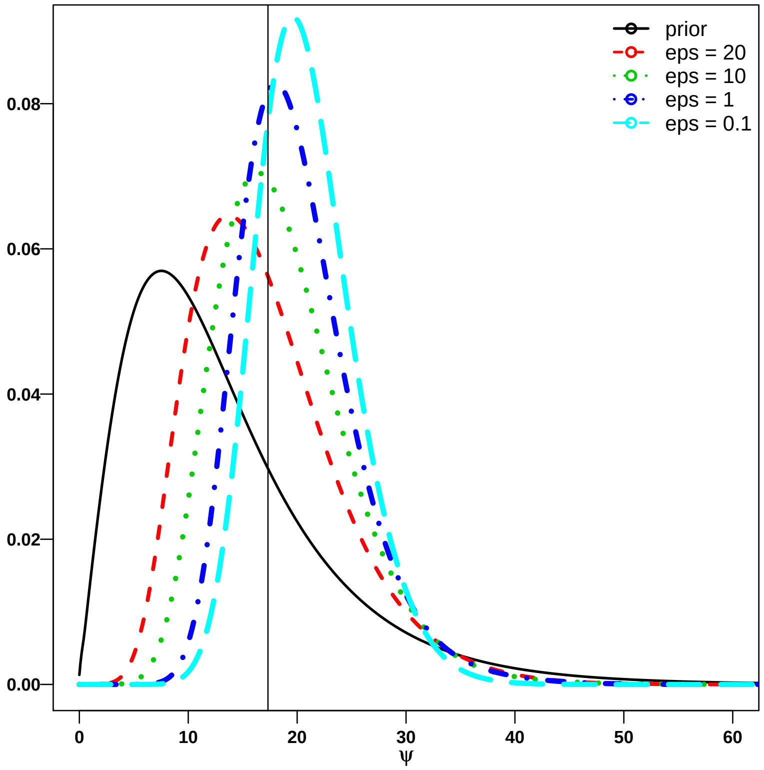

Figure 5.4: *Approximate samples from the posterior density: the plot illustrates the approximate sample from the posterior density obtained with different tolerance levels. The prior density is reported with the solid line. The thin vertical line shows the true parameter value.*

response vector $\boldsymbol{\nu}$. An estimate can be empirically provided in different ways, some methods are proposed for example by Schlater and Tawn (2003). These methods are useful because they provide an estimate of the extremal coefficients independently form the likelihood function of the K-variate model. Finally, the metric used consists of the area between the fitted models (the summary statistics) computed by using the observed and the simulated data. The area is obtained by computing the definite integrals of the two regression curves between zero and the maximum recorded distance, and then taking the absolute value of the integrals difference. An approximate value can be provided numerically by using the `integrate` routine of `R`. With the same spatial configuration used in Section (5.3.3) we have generated a sample $\mathbf{z}^{obs}$ from a max-stable process that we treat as the observed data. Then we used the **SIR3** algorithm in order to obtain an approximate sample from the posterior density. More specifically, particles are drawn from the Wishart prior with elements: $s_1 = 100$, $s_{12} = 0$ and $s_2 = 100$, of the position matrix S_2 and 2 degrees of freedom. Different tolerances have been considered: $\epsilon_1 = 20$, $\epsilon_2 = 10$, $\epsilon_3 = 1$ and $\epsilon_4 = 0.1$, and for each of them at least $N = 1500$ acceptances have been required. Note that for the tolerance ϵ_4, 500000 iterations have been necessary in order to achieve the acceptance number. Then for each of the four cases, samples of $M = 1000$ values are resampled from accepted proposal particles. The resulting approximate samples from the posterior density are illustrated in Figure (5.4). We recall that given the previous assumption about the the covariance matrix of the storm profile, that is $\gamma = \zeta$ and $\rho = 0$, then we deal with the posterior of a single parameter. Note that decreas-

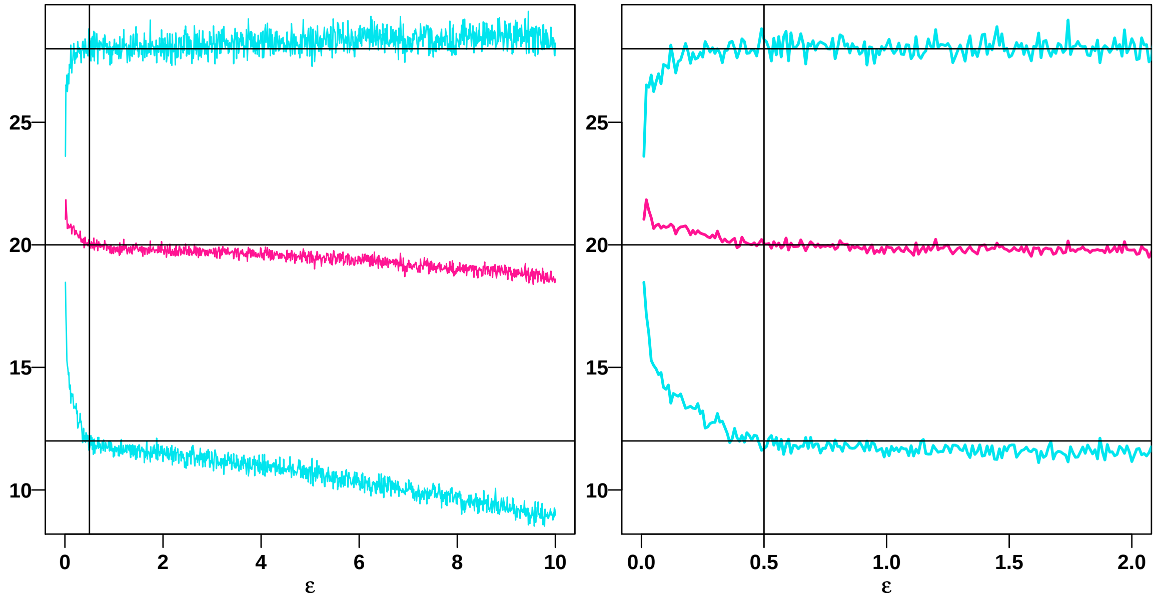

Figure 5.5: *Approximate samples from the posterior density: the plots show the mean and the standard deviations of the approximate samples for different tolerances. The middle (red) line is the mean and the top and bottom (blue) lines are the mean* $\pm$ 2 *sd. The horizontal lines show the values where the mean and the variances are constant for different tolerances.*

ing the tolerance ϵ, the approximate samples are, roughly speaking more and more sensible. But we are aware that care is required with any sort of conclusion, given that we can not compare the approximate samples with the posterior density that is unknown. It is difficult to establish an appropriate tolerance value in order to obtain a representative approximate sample from the posterior density. Nonetheless, we observe that the sample standard deviation decreases and the sample mean changes progressively with the decreasing of the tolerances. In particular, for the four cases we have obtained respectively the values: $16.93(5.58)$, $18.53(4.77)$, $19.95(4.26)$ and $20.73(3.47)$ illustrated in Figure (5.4). Thus a criteria to select the tolerance for instance, can be obtained plotting the sample mean and standard deviation for the different tolerances (Bortot, Coles and Sisson, 2006). An example is provided in Figure (5.5). From the left panel we can see that the sample mean and standard deviation change with the decreasing of the tolerance ϵ, until they remain stabilized for a certain interval. This effect is better illustrated in the right panel focusing on the subset $(0, 2)$ of the considered tolerance range $(0, 10)$. We can select ϵ for example into the tolerance set $(0.5, 1.5)$. Because for the values of ϵ inside that interval it corresponds to sample mean and standard deviation values that are reasonably constant, in particular these are respectively 20 and 4. Instead for different values of ϵ the correspondent sample mean and standard deviation values change consistently. Moreover, the approximate samples have, for values of ϵ greater than the upper bound 1.5, large variability due to the high number of acceptances. Instead, for the values of ϵ smaller than about 0.5, the corresponding approximate samples are less reliable for the small number of accepted particles.

5.5 Application to the Florida rainfall data

The data consists of the daily precipitation totals for 5873 stations in the continental USA. The data are complete up to the end of 1999 and were obtained from Professor Richard Smith from the University of North Carolina who got them from Dr. Pavel Groisman of the National Climatic Data Center, `http://www.ncdc.noaa.gov`. The aim of the analysis is to asses the dependence of extreme rainfall levels between the locations where the data are collected. Moreover, also to estimate the rates at which high rainfall levels occur for the pairs of sites analyzed. In particular our analysis focuses on the states of North and South Carolina taking into account 35 stations sparse over a surface of about 500 km^2. For each site, data are blocked into sequences of observations corresponding to a time period of length, one year. The block maxima are the annual maxima of the rainfall levels from 1908 to 1999, so that the selected period of study consists of 91 observations per location. The selected region where the sites have been observed is illustrated in figure 5.6.

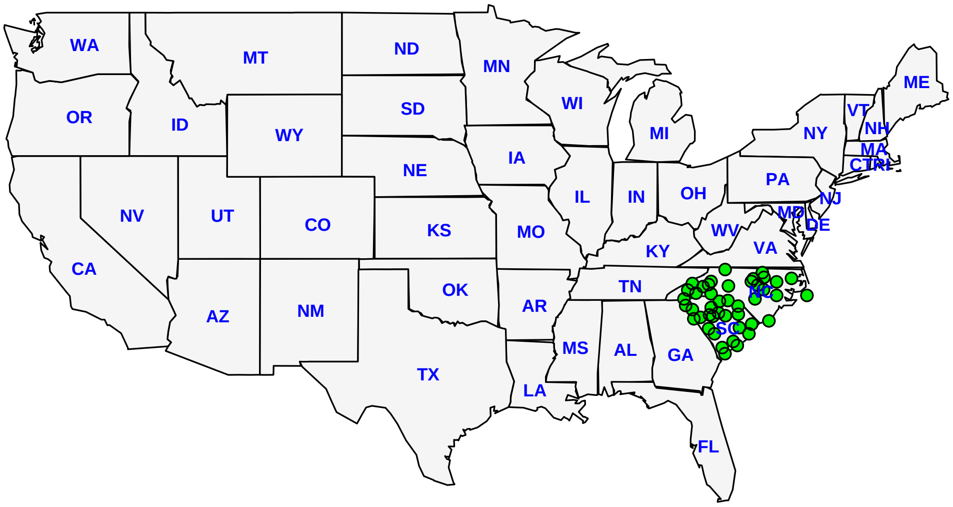

Figure 5.6: *USA map: the (green) circles represent the 35 stations spread over north and south Carolina. For those locations daily rainfall measurements are recorded.*

In order to conduct the analysis we assume that the data are correctly modeled by the Gaussian extreme value model introduced in Section (5.2), and more precisely the extended version explained in Section (5.3.2). In particular, we assumed the model (5.8) for the annual maxima, specifying for μ, ψ and ξ the polynomial surfaces:

$$\begin{aligned}
\mu(t_{1,k}, t_{2,k}) &= \sum_{i+j \leq p_1} \beta^{\mu}_{i,j}\, t^i_{1,k} t^j_{2,k} \\
\psi(t_{1,k}, t_{2,k}) &= \sum_{i+j \leq p_2} \beta^{\psi}_{i,j}\, t^i_{1,k} t^j_{2,k} \\
\xi(t_{1,k}, t_{2,k}) &= \sum_{i+j \leq p_3} \beta^{\xi}_{i,j}\, t^i_{1,k} t^j_{2,k},
\end{aligned}$$

for $k = 1, \ldots, K$ (the station's indexes) and for some p_1, p_2 and p_3 index values. The highest polynomial order considered here is 2 for all three spatial regressions (indicated by p_1, p_2 and p_3). We have not yet considered higher orders. However, note that higher order parametric regressions often do not provide substantial improvements and perhaps may be worth considering in future studies of spline regression alternatives. After computing the annual maxima with the proposal to asses the spatial dependencies, we need to estimate the model's parameters $(\zeta, \gamma, \rho, \boldsymbol{\beta}, \xi)$, where $\boldsymbol{\beta} = (\beta^{\mu}_{00} \, \ldots \, \beta^{\mu}_{02}\, \beta^{\psi}_{00} \, \ldots \, \beta^{\psi}_{02}\, \beta^{\xi}_{00} \, \ldots \, \beta^{\xi}_{02})$. Some regression models taken into account, and their regression coefficient estimates are summarized in Table 5.7. In parenthesis the deviation standards are reported.

β^{μ}_{00}	β^{μ}_{01}	β^{μ}_{10}	β^{μ}_{11}	β^{μ}_{20}	β^{μ}_{02}
β^{ψ}_{00}	β^{ψ}_{01}	β^{ψ}_{10}	β^{ψ}_{11}	β^{ψ}_{20}	β^{ψ}_{02}
β^{ξ}_{00}	β^{ξ}_{01}	β^{ξ}_{10}	β^{ξ}_{11}	β^{ξ}_{20}	β^{ξ}_{02}
661.56(7.96)	11.40(5.89)	-34.50(5.75)	-17.30(6.26)	1.41(4.79)	6.90(5.08)
5.25(0.03)	-0.041(0.02)	-0.042(0.02)	-0.063(0.03)	0.055(0.02)	0.054(0.02)
0.164(0.030)	0.022(0.021)	-0.054(0.017)	-0.000(0.030)	0.001(0.018)	-0.036(0.016)
661.59(8.31)	11.37(5.50)	-34.57(5.67)	-17.24(6.56)	1.41(5.46)	6.93(5.12)
5.23(0.03)	-0.041(0.02)	-0.041(0.02)	-0.064(0.02)	0.062(0.02)	0.054(0.01)
0.123(0.016)	-	-	-	-	-
665.90(4.32)	16.67(4.46)	-31.85(4.11)	-15.98(6.48)	-	-
5.25(0.03)	-	-	-0.08(0.02)	0.04(0.02)	0.05(0.02)
0.129(0.015)	-	-	-	-	-
668.99(7.53)	6.34(4.97)	-41.19(5.08)	-3.42(4.95)	-0.68(4.87)	1.07(4.84)
5.35(0.02)	-0.05(0.01)	-0.05(0.02)	-	-	-
0.121(0.014)	-	-	-	-	-
667.68(4.92)	5.85(4.98)	-41.34(5.35)	-	-	-
5.35(0.02)	-0.05(0.02)	-0.05(0.02)	-	-	-
0.124(0.015)	-	-	-	-	-
667.98(4.64)	13.01(3.64)	-34.45(3.80)	-	-	-
5.35(0.02)	-	-	-	-	-
0.129(0.016)	-	-	-	-	-

Table 5.7: *Regression coefficient estimates: each table's entry reports the estimates of the regression coefficients respectively for the spatial models of the location (first line), scale (second line) and shape (third line) parameters. Between parenthesis the standard deviations are reported.*

In Figure 5.7 the plots show the location, scale and shape parameter estimates, computed assuming spatial regressions versus those estimates computed using individual sites (without assuming any regression models). In particular the regression models illustrated, in order from the top to the bottom panels, are those represented by the coefficient estimates of the rows 6, 3 and 1 in Table 5.7. An evident pattern is present in all three cases (location, scale and shape parameters) of Figure 5.7. In principle, it seems that the assumption of regression models from all three GEV parameters are appropriate. This consideration also appears to be supported by Figure 5.8. In fact, these plots show that all three GEV parameters change considerably with the spatial location. This considering a second degree polynomial surface. Although, with the same model we found that the spatial regression shows some deficiencies. In fact, from Figure 5.9 (left panel) we can see that with the location parameter case only 24 of 35 estimated values fall inside the confidence intervals computed using the dataset of each individual site. This could outline the insufficiency of the regression model to explain the spatial dependence of the location parameters. We also note that two site estimates (using individual sites) are particularly unusual respect the other estimates. So for them the regression estimates are especially inadequate. It is important to investigate the reason for such extreme rainfall levels in those two sites. May be some other additional effects should be taken into account which could have an impact on the extremes, such as the altitude. Middle and right panels of Figure 5.9 illustrate

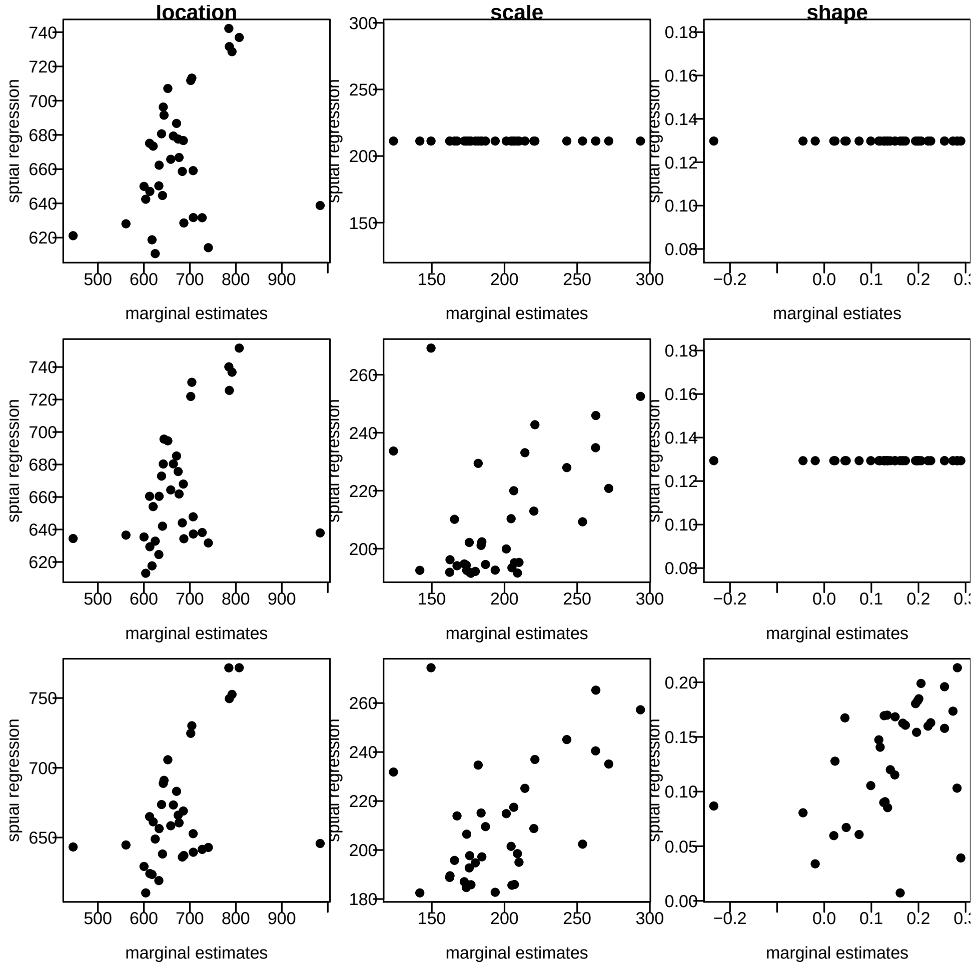

Figure 5.7: *Regression coefficient estimates vs. estimates of individual sites: the regression coefficient estimates of the models represented by the entries 6 (top row), 3 (middle row) and 1 (bottom row) of Table 5.7 are plotted versus the estimates obtained using the individual sites. The regression models involve the location (first column), scale (second column) and shape (third column) parameters.*

the scale and shape estimates. In these cases, respectively only 5 and 2 of 35 estimates fall outside the confidence intervals computed using the individual sites.

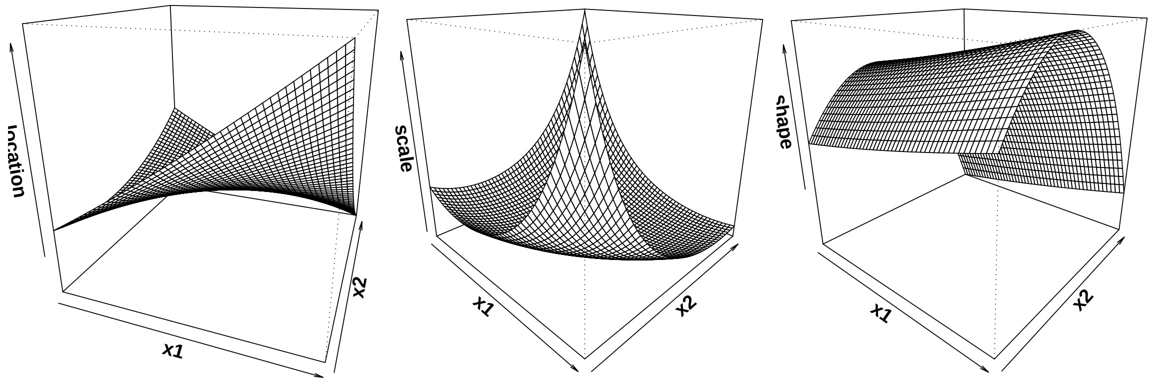

Figure 5.8: *Polynomial surfaces: the panels 1–3 show the second degree polynomial surfaces for the location (left), scale (middle) and shape (right) parameters.*

However, a deep analysis is required in order to establish the real adequacy of the model assumptions. We need some suitable procedures to test the compatibility of the model with the data and to determine the appropriate spatial regression. One example is the model selection technique.

The variability of the estimates are provided by the diagonal elements of the inverse of the "sandwich" information matrix introduced in Section (5.3.1). More precisely, the H part of the asymptotic covariance approximation is consistently estimated by the hessian matrix of composite log likelihood. An estimate can be provided by the numerical maximization routines that are used in order to obtain the parameter estimates. For example, we used the `optim` routine of the `R` statistical environment, Ihaka and Gentleman (1996). Instead, the J part is estimated by using a Monte Carlo estimate where the composite log likelihood gradient is obtained numerically by numeric differentiation routines as for example `fdHess` of `R`. Numerical methods are necessary because we can not easily derive the first-order derivative of the composite log likelihood associated to model (5.8). For a detailed discussion about H and J matrices to see Section (5.3.1).

We need some criteria in order to select a specific model among those available (some examples are illustrated in Table 5.7). In standard problems where the full likelihood is available the likelihood ratio test (e.g. Davison, 2003, p. 126) provides a useful procedure in order to test the adequacy of a particular model against a model alternative. The composite likelihood analogue of the likelihood ratio statistic can be used in order to test the hypothesis, determining whether a model $\mathcal{M}_0$ is a plausible reduction of a model $\mathcal{M}_1$ (where $\mathcal{M}_0$ is a subset of $\mathcal{M}_1$). The asymptotic distribution of the composite likelihood ratio statistic has non standard form and can be derived as a special case of the (profile) likelihood ratio test for a misspecified likelihood, Kent (1982). However, that distribution

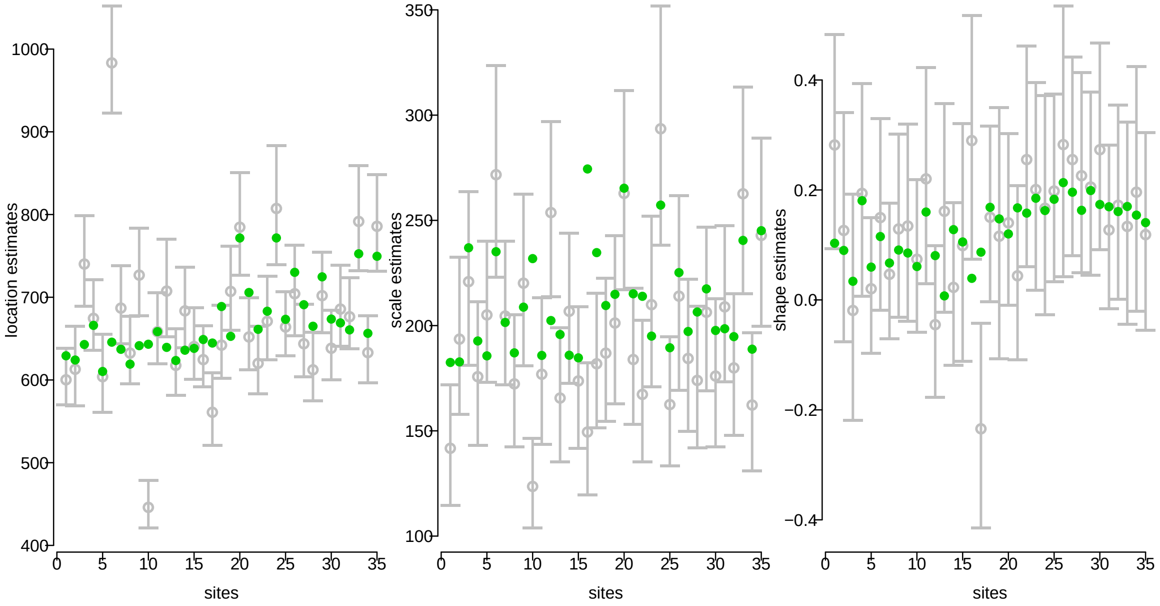

Figure 5.9: *GEV parameter estimates: the three panels illustrate the location, scale and shape estimates (green dots) of the 35 sites (Carolina) using second degree surface models. The (gray) dots show the estimates using the individual sites and the vertical lines their confidence intervals.*

in our case could not be valid given that we estimate the matrix J by a Monte Carlo estimate. Then the distribution of the composite likelihood ratio statistic can perhaps be more precisely evaluated by a parametric bootstrap, an example is illustrated by Bellio and Varin (2005). Another criterion for the model selection, based on the composite likelihood has been introduced by Varin and Vidoni (2005). This essentially consists of Akaike information criterion by using the composite likelihood.

We have no conclusive results that support any test for the model selection. Both methods described above are still under investigation.

However, given a specific model we can still test whether a single regression coefficient is zero by performing a hypothesis test. More specifically, the evidence of the null hypothesis, that says a regression coefficient is null against the alternative that the regressor is not zero, can be supported by testing the hypotheses:

$$H_0 : \beta = 0 \qquad \text{versus} \qquad H_1 : \beta \neq 0,$$

where the problem of hypothesis testing can be addressed through the result of the form

$$w \equiv \frac{\widehat{\beta} - \beta}{\widehat{\text{st.dev}}(\widehat{\beta})} \sim N(0, 1).$$

So computing the approximate p-value that is given by the tail area: p-value $\simeq 2\{1 - \Phi(|w|)\}$, the null hypothesis for the coefficients β is rejected if p-value $< \alpha$, where α is the significance level of the test. Alternatively, one can check if the confidence interval built for the coefficient contains zero. The $(1 - \alpha)\%$ approximate confidence interval for β is given by $CI = \widehat{\beta} \pm z_{\alpha/2}\, \widehat{\text{st.dev}}(\widehat{\beta})$, where $z_{\alpha/2}$ is the quantile of level $\alpha/2$ of the standard normal distribution and $\widehat{\text{st.dev}}(\widehat{\beta})$ is the standard deviation of the coefficient estimates $\widehat{\beta}$. The standard deviation is computed as the square root of the estimate variances. Then the null hypothesis for the coefficients β is rejected if zero falls into the approximate confidence interval. In Table (5.7) the regression coefficient estimates of some models that we have considered are illustrated. As we have mentioned before we can easily asses, for the models listed, the relevance of each regressor but, more complex is the model selection procedure. In our case, the latter is still under investigation.

Suppose that a second degree polynomial surface is the correct model for the location, scale and shape parameters. The regression coefficient estimates are reported in the first entry of Table 5.7. Instead, the estimates of the covariance matrix Σ_2 result in : $\hat{\gamma} = 23\,(1.8)$ km, $\hat{\zeta} = 11\,(1.1)$ km and correlation $\hat{\rho} = 0.19(0.09)$. In parenthesis the deviation standards are reported. We recall that the model parameter estimates are obtained by maximization of the composite log likelihood. Furthermore, in order to compute the deviation standard of the transformed estimated parameters we used, from the delta method, the approximate variance formula. In particular, by denoting with ψ the model parameter we have: $V\{f(\psi)\} \approx \{f'(\widehat{\psi})\}^2 V(\widehat{\psi})$, where $V(\widehat{\psi})$ is the variability of the estimated parameter.

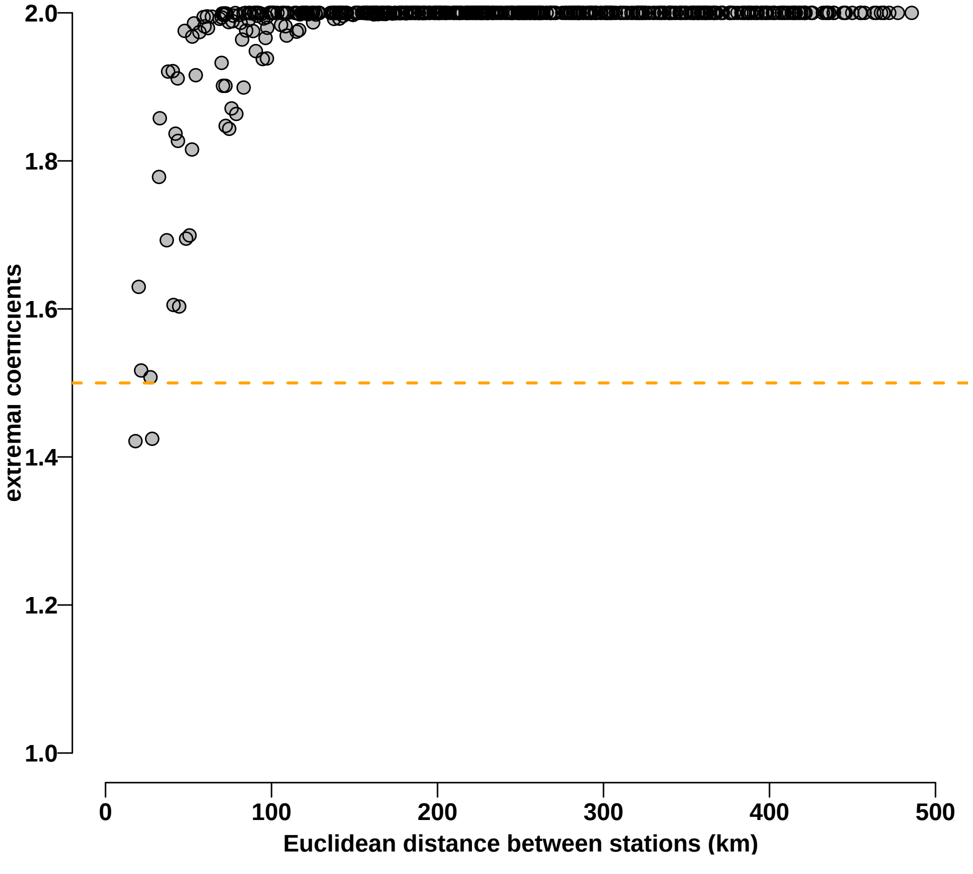

Figure 5.10: *Extremal coefficients: the extremal coefficient estimates against the Euclidean distances are plotted for each of the pairs involved, from the 35 stations of North and South Carolina.*

From the covariance matrix estimate (the model dependence structure) we can compute the extremal coefficient estimates for all the pairs of sites, that is $\widehat{\boldsymbol{\nu}}(\mathbf{h}) = 2\Phi(\widehat{\boldsymbol{\theta}}(\mathbf{h})/2)$, where $\widehat{\boldsymbol{\theta}}(\mathbf{h}) = (\mathbf{h}^T\widehat{\Sigma}^{-1}\mathbf{h})^{-1/2}$ and $\mathbf{h} = \mathbf{t}_1 - \mathbf{t}_2$ with $\mathbf{t}_1$ and $\mathbf{t}_2$ the coordinates of the j's and i's locations. The extremal coefficient estimates are illustrated by the cloud of points in the graph in Figure 5.10. The plot shows the extremal coefficients versus the Euclidean distances for each pair of sites. From Figure 5.10 we can see that only four pairs of locations can be classified as mildly or strongly dependent. Instead the other pairs can be classified as weakly dependent.

In order to illustrate the proposal of assessing the rates at which the extreme rainfall levels occur, from the complete set of pairs of sites, we have selected four of them. One is the most strongly dependent and the other three have been chosen randomly. In the univariate case the return level estimate associated with the return period $1/p$ consists of estimating the extreme quantile of the annual maximum distribution. After estimating the parameters of the GEV distribution, it follows that the quantile z_p of level p can be obtained inverting the equation $F(z_p) = 1 - p$ (where $F(z)$ is the GEV distribution). The related interpretation of all this is that the return level in any particular year with probability p is exceeded by the annual maximum, Coles (Chapter 3.1, 2001).

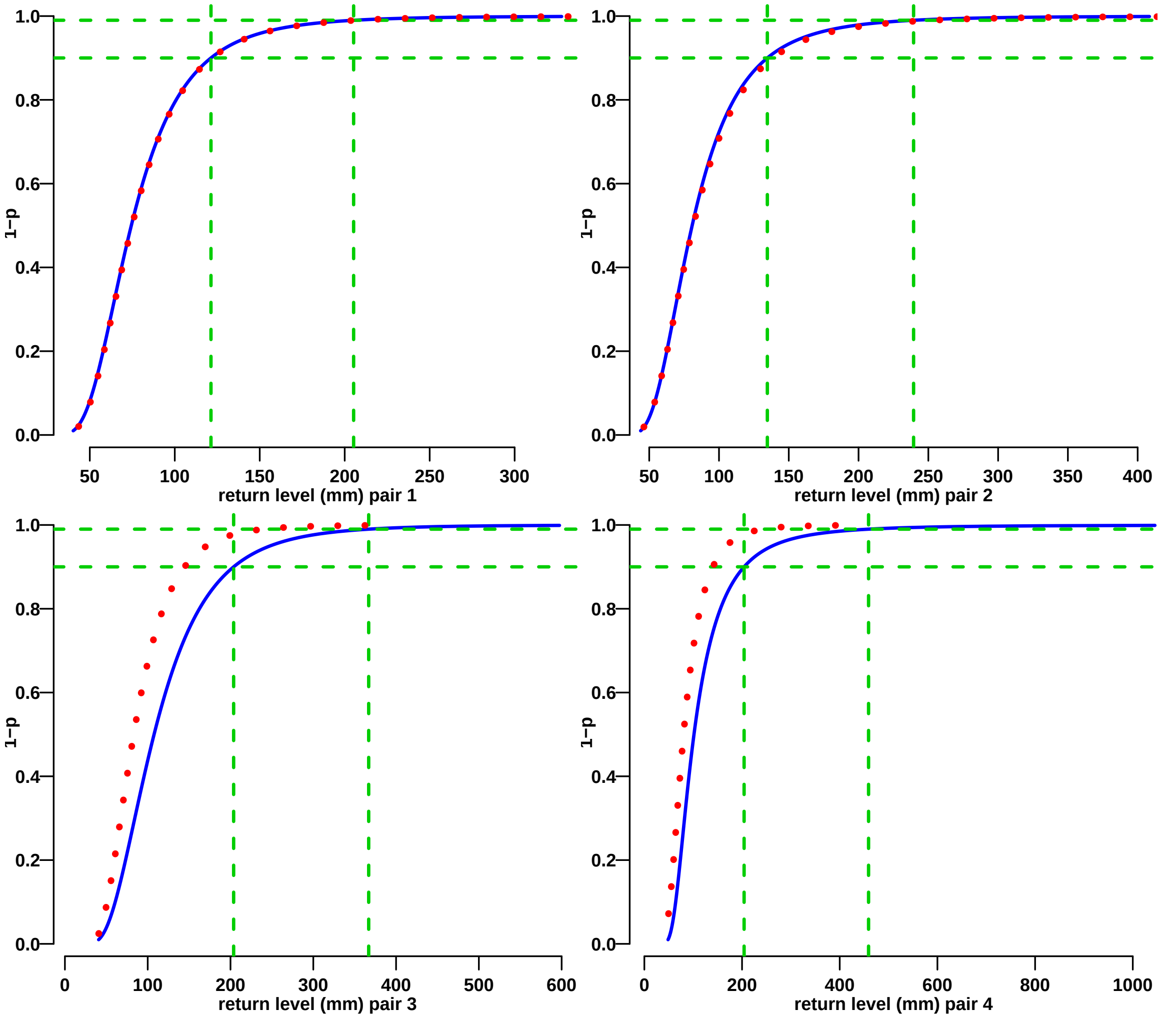

Figure 5.11: *Return level plots: the panels show the return levels for the selected four pairs. The continuous and dotted lines represent the return values of the two sites in each pair. The vertical axis represents the quantities 1-p and the horizontal, the extreme quantiles. The broken lines represent, respectively the quantile levels 0.1 and 0.01 .*

Our analysis of spatial extremes is driven by the composite likelihood approach with the pairwise setting. The bivariate marginal densities of all the pairs of observations take the form (5.7). For the pairs of sites we then use the expression of the marginal distribution $F(s_i, s_j)$, where s_i and s_j are the levels associated with the sites i and j, and the composite likelihood estimates determine the extremes quantiles.

Note that in the bivariate case many possible definitions and related interpretations of the return levels exist. For our problem we assume an analogue definition of the univariate case. Given the composite likelihood estimates, an estimate of the return levels, with return period $1/p$, can be derived inverting the equation: $F(s_i, s_j) = 1 - p$.

In this way we say that the return levels s_i and s_j in any particular year with probability p are exceeded by the annual maximum in at least one site, i or j. Unfortunately, the previous equation can not be easily solved analytically respect with s_i and s_j. We need another equation in order to have a system of two equations and two unknown quantities. A possibility is given by the system

$$\begin{cases} F(s_i, s_j) = 1 - p \\ F_i(s_i) = F_j(s_j), \end{cases} \tag{5.10}$$

where from the second equation we can derive that

$$s_i = \psi_i \left[\left\{ 1 + \xi_j \left(\frac{s_j - \mu_j}{\psi_j} \right) \right\}_{+}^{\frac{\xi_i}{\xi_j}} - 1 \right] \Bigg/ \xi_i, +\mu_i \quad \text{for} \quad \xi_i \neq 0,\, \xi_j \neq 0 \tag{5.11}$$

and then substituting this result into the first equation we get $F(s_j, s_j) = 1 - p$. However, the latter is still not easily solved analytically respect with s_j. Nonetheless a solution can be provided by numerical routines that search for the root of a function as for example `uniroot` of `R`. Thus, once we have obtained the estimate $\widehat{s_j}$ (with numerical methods), then by using the equation (5.11) we can determine the estimate $\widehat{s_i}$.

For the four pairs of sites we estimated the return levels corresponding to the values $p = 0.01$ and $p = 0.1$. The estimates results are reported in Table 5.8. We can see that the pairs with strong or mild dependence (as shown by the extremal coefficients) are related with similar return values for both sites (the first two rows). This occurs because the dependent sites have similar composite likelihood estimates due to their close proximity. The remaining two pairs show less similar return values, which is reasonable considering their larger distances.

Figure 5.11 illustrates the return level plots. Each panel shows the return levels of the pairs. With the first two horizontal plots the similarity between the return levels for the dependent sites is clearly evident. This could be motivated by the inadequacy of the polynomial surface order assumed.

The standard errors are obtained from the approximate sampling distribution of the $1/p$, return level as suggest by Coles (2001, p. 139). Broadly, denoting with $\boldsymbol{\psi}$ the model parameters, under regular conditions we know that the composite maximum likelihood estimator $\widehat{\boldsymbol{\psi}}_{CML}$, has approximate distribution $N(\boldsymbol{\psi}, H(\boldsymbol{\psi})^{-1}\, J(\boldsymbol{\psi})\, H(\boldsymbol{\psi})^{-1})$, where $H(\boldsymbol{\psi})^{-1}\, J(\boldsymbol{\psi})\, H(\boldsymbol{\psi})^{-1}$ is an approximation of the asymptotic covariance matrix, see Section (5.3.1).

Plugging the composite likelihood estimates and their approximate asymptotic covariance matrix estimates into the normal distribution, we can generate samples from the approximate sampling distribution of the composite maximum

sites	distance	extr. coef.	return levels (0.1)	return levels (0.01)
30, 31	20	1.44	121.3(35.2), 121.8(37.3)	205.2(25.2), 206.6(29.3)
26, 29	72	1.89	134.6(37.6), 138.5(89.1)	239.4(30.0), 252.7(84.7)
2, 20	435	2.00	203.8(30.3), 144.5(47.8)	366.9(26.4), 239.9(39.5)
5, 33	405	2.00	204.2(42.9), 140.4(26.8)	459.0(21.4), 243.0(12.7)

Table 5.8: *Return level estimates: the sites, distances (km), extremal coefficients, return levels (mm) for p=0.1 and return levels for p=0.01 are reported for the four selected pairs. In parenthesis the standard deviations are reported.*

likelihood estimator. For the set of simulated values $\boldsymbol{\psi}_1^{sim}, \ldots, \boldsymbol{\psi}_M^{sim}$ we compute the associated return levels by solving the system of equations (5.10). Therefore, we obtain a set of return levels that form an approximate sampling distribution of the return level. This can be used in order to calculate approximate variances and standard deviations.

Conclusion

The spline mixed model approach for extremes that we propose can be considered a suitable alternative to the current nonparametric methods (e.g. Chavez-Demoulin and Davison, 2005). Our goal is not to supplant the already proposed methods but rather to outline the potential advantages offered by the spline mixed model paradigm. The novelty of this thesis is their the extension to the generalized extreme value distribution. The study focuses on the analysis of series of block maxima, for instance the annual maxima, to which the GEV distribution can be fitted. Non stationary sequences of extremes can be modelled incorporating spline smoother into the GEV model. The particular attractiveness of this approach is the flexibility of the model and because model fitting and inference use extensions of standard likelihood methods. The benefit of the mixed model formulation is the inclusion of the smoothing parameter into the model framework enabling its estimation to occur concurrently with the other parameters. Initially by adopting the simplest spline mixed model setup (location covarite-dependent) we illustrate with examples that complex patterns can be easily accommodated. Fitting and parameter estimates are based on penalized maximum likelihood estimation. Our estimation procedure is simple and relatively little computational effort is required. Simulation study confirms that for an opportune range of scale and shape values the three-type estimation results have only little bias and variance. Analysis of the annual maximum temperatures using real data suggests that our model fits well, properly taking into account the trend in the extremes.

Some drawbacks and points for future study. Non-negligible bias can occur for a particular range of the shape parameter that correspond to the heavy tail distributions, especially in moderately small observations. Analysis of environmental extreme processes (our primary goal) could require more realistic models than the location covariate-dependent case. A less crude approximation could be provided by including the scale and shape covariate-dependent cases. We consider here the scale covariate-dependent case. We found that the fitting based on penalized maximum likelihood estimation in the present fashion is not easy to adapt to further modifications. We explore for the scale covariate-dependent case the alternative Bayesian approach and use MCMC for estimation and inference. We found that this proposal facilities the fitting, and that uncertainty in variance components is more easily taken into account. But it is interesting to explore the potential solutions offered by the likelihood approach. Mixed model approach facilities the incorporation of space-time extensions and missing data complications, and this could be of interest for future studies.

The analysis of spatial extremes is also a primary interest of this thesis. Most of the models of spatial extremes based on max-stable processes have difficulty handling estimation methods of the parameters. The study focuses on the analysis of series of block maxima, for instance the annual maxima, spread over a region to which the models that arise from the max-stable processes can be fitted. We discuss a method of inference based on the composite likelihood approach for the estimation of the spatial model parameters. Simulation study shows the sound estimation results of the tail dependencies. The spatial extreme value framework can incorporate spatial dependent regression models with the GEV margins, allowing the assessment of the rate at which the extreme events occur at the sites under the composite likelihood umbrella.

We analyze the annual maximum rainfall levels at different sites in North and South Carolina. We are interested in the extreme dependence between the weather-stations. It is also interesting to explore the use of the composite likelihood with an emphasis on the flexible inference framework it provides in order to treat the GEV margins. The purpose is to predict the rate of the extreme rainfall levels that occur at the sites. The estimation results of the dependencies between the extremes at the sites are reasonable. Although the regression models of the GEV parameter provide a flexible spatial framework without being too computationally demanding, their adequacy with data need to be opportunely examined. Explorative analysis for adequacy of the regression models may be misleading in situations with spatial data. This is a potential drawback of our analysis and points for further development. In fact a deep study of the concurrent estimation method of dependence and regression coefficients is necessary. Furthermore, more flexible nonparametric alternatives could be an initial point of reference.

The intractability of the likelihood function derived from the max-stable process formulation of spatial models can be attacked also by Bayesian analysis. We explore the class of computational intensive methods known as Approximate Bayesian Computation (ABC). Simulation exercises expose the reasonable approximation of the posterior densities that can be achieved with this approach. Simple cases do not require excessive computational effort. Highly structured problems such as spatial extremes analysis are more difficult to manage with these techniques. Approximation accuracy favors large sampling numbers, computational considerations favor small numbers. It is of interest for future study to determine pragmatic methods in order to establish the good compromise between computational intensity and the approximation accuracy.

Appendix

In this part of the appendix we present the concept of Laplace approximation with an example related to the extreme value distribution.

A.1 Laplace approximation

Let $f(y)$ be a positive function. Then one of the simplest methods for approximation of f is given by the first few terms of its Taylor series expansion. For simplicity $f(y)$ is a univariate function. The approximation can be applied straightforwardly to f or to a transformation of itself like $h(y) \equiv \log f(y)$ without changing anything. For instance, considering $f(y) = \exp\{h(y)\}$ and choosing y_0 as the point to expand around, then an approximation of f is given by

$$f(y) \approx \exp\left\{ h(y_0) + (y - y_0)\, h'(y_0) + \frac{(y-y_0)^2}{2}\, h''(y_0) \right\}, \tag{5.12}$$

where $h'(y) = \partial h(y)/\partial y$ is the first-order differentiation and $h''(y) = \partial^2 h(y)/\partial^2 y$ is the second-order differentiation. For chosen $y_0 = \hat{y}$ (the value that maximizes the function), it follows that $h'(\hat{y}) = 0$ and thus the expression 5.12 can be written equivalently to

$$f(y) \approx \exp\left\{ h(\hat{y}) + \frac{(y-y_0)^2}{2} h''(\hat{y}) \right\}. \tag{5.13}$$

Note that the approximation (5.13) is no longer an approximation, but an exact expression in the special case that the function $h(y)$ is quadratic. Therefore when $h(y)$ is not quadratic, for any values y far away from $\hat{y}$, the approximation may not to be close to $h(y)$, so the omitted terms of order $(x - \hat{x})^3$ and higher will be important to guarantee a good approximation also for those points. Care should be taken in any approximation procedure, however many authors like Qind and Pierce (1993) argue that such an approximation is easy and accurate in inferential terms. The same approximation method can be applied to compute integrals of positive functions in the real line, such as $\int f(y)\,\mathrm{d}y$. As in expression (5.13) the integral can be written as follows

$$\int f(y)\,\mathrm{d}y \approx \int \exp\left\{ h(\hat{y}) + \frac{(y-y_0)^2}{2}\, h''(\hat{y}) \right\} \mathrm{d}y. \tag{5.14}$$

If the quantity $\hat{y}$ is the maximum, it follows that $h''(\hat{y})$ is negative and the right side of 5.14 can be explicitly computed by recognizing that the kernel of the integral is the same as the kernel of a normal density with mean $\hat{y}$ and variance $-1/h''(\hat{y})$. With opportune adjustments the above expression becomes

$$\int f(y)\,\mathrm{d}y \approx \exp\{h(\hat{y})\} \left[-\frac{2\pi}{h''(\hat{y})} \right]^{-1/2}, \tag{5.15}$$

which is a method known as *Laplace approximation* of the first order.

Example: Generalized extreme value distribution.

Of interest is to look at a practical example that could be the approximation of the GEV distribution. Let the generalized extreme value distribution with expression (1.1). For $\xi \neq 0$, when $\xi = 0$ the the distribution is defined by continuity and it has to be treated separably, the maximum is obtained at the value $\hat{y} = \psi\{(1+\xi)^{-\xi} - 1\}/\xi + \mu$ for $\xi > -1$. Instead of $\xi < -1$ the solution can not be exist. The second derivative of the natural logarithm of (1.1) evaluated at $\hat{y}$ is equal to $(1+\xi)^{2\xi}(\xi(1+\xi)^{\xi} - (1+\xi)^2)/\psi^2$. Therefore the Taylor series expansion (5.13) of (1.1) yields

$$f(y) \approx \exp\left\{2(\log 2 - 1) - \log\psi - \frac{4}{\psi^2}\left(y - \mu + \frac{\psi}{2}\right)^2\right\}. \tag{5.16}$$

In this part of the appendix we present explicit expressions for the derivatives required for the likelihood-based fitting scheme given in Section 3.2–3.2.1.

A.2 Vector notation

Let f be a real-valued function in the $d \times 1$ vector $\mathbf{x} = (x_1, \ldots, x_d)$. Then the derivative vector $\mathsf{D}_{\mathbf{x}} f(\mathbf{x})$, is the $1 \times d$ with ith entry $\partial f(\mathbf{x})/\partial x_i$. The corresponding Hessian matrix is given by $\mathsf{H}_{\mathbf{x}} f(\mathbf{x}) \mathsf{D}_{\mathbf{x}}\{\mathsf{D}_{\mathbf{x}} f(\mathbf{x})^T\}$.

If $\boldsymbol{a} = (a_1, \ldots, a_d)$ and $\mathbf{b} = (b_1, \ldots, b_d)$ are two $d \times 1$ vectors then element-wise multiplication is denoted by $\boldsymbol{a} \odot \mathbf{b} = (a_1 b_1, \ldots, a_d b_d)$ The expression $\boldsymbol{a}/\mathbf{b}$ denotes element-wise division $(a_1/b_1, \ldots, a_d/b_d)$. Scalar functions applied to vectors are also evaluated element-wise. For example, $\boldsymbol{a}^{-1/\xi} = (a_1^{-1/\xi}, \ldots, a_d^{-1/\xi})$.

A.3 Expression for $\mathcal{L}(\boldsymbol{\beta}, \psi, \xi, \sigma^2)$

Let $f(\mathbf{y}|\mathbf{u})$ and $f(\mathbf{u})$ be respectively the GEV conditional and the normal densities distributions. The likelihood for $\mathcal{L}(\boldsymbol{\beta}, \psi, \xi, \sigma^2)$ is given from

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}, \psi, \xi, \sigma^2) &= f(\mathbf{y}; \boldsymbol{\beta}, \psi, \xi, \sigma^2) = \int_{\mathbb{R}^K} f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}, \psi, \xi) f(\mathbf{u}; \sigma^2)\, d\mathbf{u} \\
&= (2\pi)^{-K/2} |\mathbf{G}_{\boldsymbol{\sigma}^2}|^{-1/2} \int_{\mathbb{R}^K} \exp\{b(\mathbf{u}; \boldsymbol{\beta}, \psi, \xi, \sigma^2)\}\, d\mathbf{u}
\end{aligned}$$

where

$$\begin{aligned}
b(\mathbf{u}; \boldsymbol{\beta}, \psi, \xi, \sigma^2) &= -n\log(\psi) - \tfrac{1+\xi}{\xi} \mathbf{1}^T \log\{\mathbf{1} + \xi(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})/\psi\} \\
&- \mathbf{1}^T\{\mathbf{1} + \xi(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})/\psi\}^{-\frac{1}{\xi}} - \frac{1}{2}\mathbf{u}^T \mathbf{G}_{\boldsymbol{\sigma}^2}^{-1} \mathbf{u}.
\end{aligned}$$

The same arguments of appendix A.1 yields the integral approximation

$$\int_{\mathbb{R}^K} \exp\{b(\mathbf{u})\}\, d\mathbf{u} \simeq (2\pi)^{K/2} | - \mathsf{H}_{\mathbf{u}} b(\widehat{\mathbf{u}}; \boldsymbol{\beta}, \psi, \xi, \sigma^2)|^{-1/2} \exp\{b(\widehat{\mathbf{u}}; \boldsymbol{\beta}, \psi, \xi, \sigma^2)\}$$

where

$$\mathsf{H}_{\mathbf{u}} b(\mathbf{u}; \boldsymbol{\beta}, \psi, \xi, \sigma^2) = \mathbf{X}^T \mathrm{diag}\{h_{\mathbf{u}\mathbf{u}}(\mathbf{u}; \boldsymbol{\beta}, \psi, \xi, \sigma^2)\}\mathbf{X} - \mathrm{blockdiag}(\mathbf{G}_{\boldsymbol{\sigma}^2}^{-1})$$

and

$$h_{\mathbf{uu}}(\mathbf{u};\boldsymbol{\beta},\psi,\xi,\sigma^2)\equiv\frac{(1+\xi)[\xi\mathbf{1}-\{\mathbf{1}+\xi(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Zu})/\psi\}^{-1/\xi}]}{\psi^2\{\mathbf{1}+\xi(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Zu})/\psi\}^2}.$$

Now considering $\mathcal{I}_{\mathbf{uu}}(\mathbf{u};\boldsymbol{\beta},\psi,\xi,\sigma^2)=-\mathsf{H}_{\mathbf{u}}b(\mathbf{u};\boldsymbol{\beta},\psi,\xi,\sigma^2)$ the final approximate expression of the *integrated* log likelihood results in

$$\mathcal{L}(\boldsymbol{\beta},\psi,\xi,\sigma^2)\simeq|\mathbf{G}_{\boldsymbol{\sigma}^2}|^{-1/2}\exp\{b(\widehat{\mathbf{u}};\boldsymbol{\beta},\psi,\xi,\sigma^2)\}\mathcal{I}_{\mathbf{uu}}(\widehat{\mathbf{u}};\boldsymbol{\beta},\psi,\xi,\sigma^2)$$

A.4 Expression for $\mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\boldsymbol{\nu},\psi,\xi,\boldsymbol{\sigma}^2)$

For the additive model extension, the penalized log-likelihood may be written as

$$\ell_{\text{PL}}(\boldsymbol{\nu},\psi,\xi,\sigma^2)=h(\mathbf{y},\boldsymbol{\nu},\psi,\xi)-\tfrac{1}{2}\mathbf{u}^T\mathbf{G}_{\boldsymbol{\sigma}^2}^{-1}\mathbf{u}$$

where

$$h(\mathbf{y},\boldsymbol{\nu},\psi,\xi)\equiv-n\log(\psi)-\tfrac{1+\xi}{\xi}\mathbf{1}^T\log\{\mathbf{1}+\xi(\mathbf{y}-\mathbf{C}\boldsymbol{\nu})/\psi\}-\mathbf{1}^T\{\mathbf{1}+\xi(\mathbf{y}-\mathbf{C}\boldsymbol{\nu})/\psi\}^{-\frac{1}{\xi}}.$$

Vector differential calculus methods (e.g. Wand, 2002) lead to

$$\mathsf{D}_{\boldsymbol{\nu}}\ell_{\text{PL}}(\boldsymbol{\nu},\psi,\xi,\sigma^2)=h_{\boldsymbol{\nu}}(\mathbf{y},\boldsymbol{\nu},\psi,\xi)^T\mathbf{C}-\begin{bmatrix}\mathbf{0}\\ \mathbf{u}^T\mathbf{G}_{\boldsymbol{\sigma}^2}^{-1}\end{bmatrix}$$

and

$$\mathsf{H}_{\boldsymbol{\nu}}\ell_{\text{PL}}(\boldsymbol{\nu},\psi,\xi,\sigma^2)=\mathbf{C}^T\text{diag}\{h_{\boldsymbol{\nu}\boldsymbol{\nu}}(\mathbf{y},\boldsymbol{\nu},\psi,\xi)\}\mathbf{C}-\text{blockdiag}(\mathbf{0},\mathbf{G}_{\boldsymbol{\sigma}^2}^{-1})$$

where

$$h_{\boldsymbol{\nu}}(\mathbf{y},\boldsymbol{\nu},\psi,\xi)\equiv\frac{(1+\xi)\mathbf{1}-\{\mathbf{1}+\xi(\mathbf{y}-\mathbf{C}\boldsymbol{\nu})/\psi\}^{-1/\xi}}{\psi\{\mathbf{1}+\xi(\mathbf{y}-\mathbf{C}\boldsymbol{\nu})/\psi\}}$$

and

$$h_{\boldsymbol{\nu}\boldsymbol{\nu}}(\mathbf{y},\boldsymbol{\nu},\psi,\xi)\equiv\frac{(1+\xi)[\xi\mathbf{1}-\{\mathbf{1}+\xi(\mathbf{y}-\mathbf{C}\boldsymbol{\nu})/\psi\}^{-1/\xi}]}{\psi^2\{\mathbf{1}+\xi(\mathbf{y}-\mathbf{C}\boldsymbol{\nu})/\psi\}^2}.$$

The required observed information matrix expression is then

$$\mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\boldsymbol{\nu},\psi,\xi,\boldsymbol{\sigma}^2)=\mathbf{C}^T\text{diag}\{-h_{\boldsymbol{\nu}\boldsymbol{\nu}}(\mathbf{y},\boldsymbol{\nu},\psi,\xi)\}\mathbf{C}+\text{blockdiag}(\mathbf{0},\mathbf{G}_{\boldsymbol{\sigma}^2}^{-1}).$$

The (penalised likelihood-based) information matrix is

$$E\{\mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\boldsymbol{\nu},\psi,\xi,\boldsymbol{\sigma}^2)\}=\{(1-\xi)/\psi\}^2\Gamma(1+2\psi)\mathbf{C}^T\mathbf{C}+\text{blockdiag}(\mathbf{0},\mathbf{G}_{\boldsymbol{\sigma}^2}^{-1})$$

which is consistent with results in Prescott and Walden (1980) and Tawn (1988).

A.5 Expression for $\mathcal{I}_{(\psi,\xi)(\psi,\xi)}(\psi,\xi,\boldsymbol{\sigma}^2)$

First note that

$$\mathcal{I}_{(\psi,\xi)(\psi,\xi)}(\psi,\xi,\boldsymbol{\sigma}^2)=-\begin{bmatrix}H_{\psi\psi}&H_{\psi\xi}\\ H_{\psi\xi}&H_{\xi\xi}\end{bmatrix}$$

where

$$\begin{aligned} H_{\psi\psi} &\equiv \frac{\partial^2}{\partial\psi^2}\left\{\ell_{\text{PL}}(\boldsymbol{\nu},\psi,\xi,\sigma^2) - \tfrac{1}{2}|\mathcal{I}_{\boldsymbol{\nu\nu}}(\boldsymbol{\nu},\psi,\xi,\sigma^2)|\right\} \\ &= \frac{\partial^2}{\partial\psi^2}\left\{h(\mathbf{y},\boldsymbol{\nu},\psi,\xi) - \tfrac{1}{2}\left|\mathbf{C}^T\text{diag}\{-h_{\boldsymbol{\nu\nu}}(\mathbf{y},\boldsymbol{\nu},\psi,\xi)\}\mathbf{C} + \text{blockdiag}(\mathbf{0},\mathbf{G}_{\boldsymbol{\sigma}^2}^{-1})\right|\right\} \end{aligned}$$

and $H_{\psi\xi}$ and $H_{\xi\xi}$ are defined analogously as the other second-order partial derivatives. Then vector calculus methods lead to

$$\begin{aligned} H_{\psi\psi} &= h_{\psi\psi}(\mathbf{y},\psi,\xi) + \tfrac{1}{2}\text{tr}\left[\mathcal{I}_{\boldsymbol{\nu\nu}}(\boldsymbol{\nu},\psi,\xi,\boldsymbol{\sigma}^2)^{-1}\mathbf{C}^T\text{diag}\{h_{\psi\psi\boldsymbol{\nu\nu}}(\mathbf{y},\psi,\xi)\}\mathbf{C}\right], \\ H_{\psi\xi} &= h_{\psi\xi}(\mathbf{y},\psi,\xi) + \tfrac{1}{2}\text{tr}\left[\mathcal{I}_{\boldsymbol{\nu\nu}}(\boldsymbol{\nu},\psi,\xi,\boldsymbol{\sigma}^2)^{-1}\mathbf{C}^T\text{diag}\{h_{\psi\xi\boldsymbol{\nu\nu}}(\mathbf{y},\psi,\xi)\}\mathbf{C}\right] \quad \text{and} \\ H_{\xi\xi} &= h_{\xi\xi}(\mathbf{y},\psi,\xi) + \tfrac{1}{2}\text{tr}\left[\mathcal{I}_{\boldsymbol{\nu\nu}}(\boldsymbol{\nu},\psi,\xi,\boldsymbol{\sigma}^2)^{-1}\mathbf{C}^T\text{diag}\{h_{\xi\xi\boldsymbol{\nu\nu}}(\mathbf{y},\psi,\xi)\}\mathbf{C}\right]. \end{aligned}$$

Here $\mathbf{r} \equiv (\mathbf{y} - \mathbf{C}\boldsymbol{\nu})/\psi$,

$$\begin{aligned} h_{\psi\psi}(\mathbf{y},\psi,\xi) &\equiv \frac{n}{\psi^2} \\ &+\mathbf{1}^T\left[\frac{\{(1+\xi)\mathbf{r}^2\}\odot\left\{\xi\mathbf{1} - (\mathbf{1}+\xi\mathbf{r})^{-\frac{1}{\xi}}\right\} + 2\mathbf{r}\odot(\mathbf{1}+\xi\mathbf{r})\odot\left\{(\mathbf{1}+\xi\mathbf{r})^{-\frac{1}{\xi}} - (1+\xi)\mathbf{1}\right\}}{\psi^2(\mathbf{1}+\xi\mathbf{r})^2}\right], \end{aligned}$$

$$\begin{aligned} h_{\psi\xi}(\mathbf{y},\psi,\xi) &\equiv \mathbf{1}^T\left[\frac{\mathbf{r}\odot(\mathbf{1}+\xi\mathbf{r})\odot\left[\mathbf{1} - (\mathbf{1}+\xi\mathbf{r})^{-\frac{1}{\xi}}\odot\left\{\frac{\log(\mathbf{1}+\xi\mathbf{r})}{\xi^2} - \frac{\mathbf{r}}{\xi(\mathbf{1}+\xi\mathbf{r})}\right\}\right]}{\psi(\mathbf{1}+\xi\mathbf{r})^2}\right] \\ &\quad +\mathbf{1}^T\left[\frac{\mathbf{r}^2\odot\left\{(\mathbf{1}+\xi\mathbf{r})^{-\frac{1}{\xi}} - (1+\xi)\mathbf{1}\right\}}{\psi(\mathbf{1}+\xi\mathbf{r})^2}\right], \end{aligned}$$

$$\begin{aligned} h_{\xi\xi}(\mathbf{y},\psi,\xi) &\equiv -\mathbf{1}^T\left[\frac{\log(\mathbf{1}+\xi\mathbf{r})\odot\{(\mathbf{1}+\xi\mathbf{r})\odot\log(\mathbf{1}+\xi\mathbf{r}) - 2\xi(\mathbf{r}+\mathbf{1})\} + 2\xi^2\mathbf{r}}{\xi^4(\mathbf{1}+\xi\mathbf{r})^{1+\frac{1}{\xi}}}\right] \\ &\quad +\mathbf{1}^T\left[\frac{\xi\mathbf{r}\odot\{\xi\mathbf{r}(\xi+3)+2\} - 2(\mathbf{1}+\xi\mathbf{r})\odot\log(\mathbf{1}+\xi\mathbf{r})}{\xi^3(\mathbf{1}+\xi\mathbf{r})^2}\right], \end{aligned}$$

$$\begin{aligned} h_{\psi\psi\boldsymbol{\nu\nu}}(\mathbf{y},\psi,\xi) &\equiv \frac{6(1+\xi)(\xi\mathbf{r})^2\odot\left\{\xi\mathbf{1} - (1+\xi\mathbf{r})^{-\frac{1}{\xi}}\right\} - (1+6\xi+5\xi^2)(\mathbf{r}^2)\odot(1+\xi\mathbf{r})^{-\frac{1}{\xi}}}{\psi^4(\mathbf{1}+\xi\mathbf{r})^4} \\ &\quad +\frac{6(1+\xi)\left\{\xi\mathbf{1} - (1+\xi\mathbf{r})^{-\frac{1}{\xi}}\right\}}{\psi^4(\mathbf{1}+\xi\mathbf{r})^2} \\ &\quad +\frac{(1+\xi)\mathbf{r}\odot\left[6(\mathbf{1}+\xi\mathbf{r})^{-\frac{1}{\xi}} - 12\xi\left\{\xi\mathbf{1} - (\mathbf{1}+\xi\mathbf{r})^{-\frac{1}{\xi}}\right\}\right]}{\psi^4(1+\xi\mathbf{r})^3}, \end{aligned}$$

$$
\begin{aligned}
h_{\psi\xi\boldsymbol{\nu}\boldsymbol{\nu}}(\mathbf{y},\psi,\xi) \quad \equiv \quad & \frac{3\xi(2+\xi)\mathbf{r}-(7+8\xi)\mathbf{r}\odot(\mathbf{1}+\xi\mathbf{r})^{-\frac{1}{\xi}}}{\psi^3(1+\xi\mathbf{r})^3} \\
& + \frac{2(1+\xi)\left[\mathbf{1}-(\mathbf{1}+\xi\mathbf{r})^{-1/\xi}\odot\left\{\frac{\log(\mathbf{1}+\xi\mathbf{r})}{\xi^2}-\frac{\mathbf{r}}{\xi(\mathbf{1}+\xi\mathbf{r})}\right\}\right]\odot(\xi\mathbf{r}-1)}{\psi^3(\mathbf{1}+\xi\mathbf{r})^3} \\
& - \frac{(1+\xi)\mathbf{r}\odot(\mathbf{1}+\xi\mathbf{r})^{-1/\xi}\odot\{\frac{\log(\mathbf{1}+\xi\mathbf{r})}{\xi^2}-\frac{\mathbf{r}}{\xi(\mathbf{1}+\xi\mathbf{r})}\}}{\psi^3(\mathbf{1}+\xi\mathbf{r})^3} \\
& + \frac{3(1+\xi)\mathbf{r}\odot(2\xi\mathbf{1}+\mathbf{r}\psi)\odot(\mathbf{1}+\xi\mathbf{r})^{-1/\xi}-6\xi^2(1+\xi)\mathbf{r}}{\psi^4(\mathbf{1}+\xi\mathbf{r})^4} \\
& - \frac{2\left\{\xi\mathbf{1}-(\mathbf{1}+\xi\mathbf{r})^{-\frac{1}{\xi}}\right\}}{\psi^3(1+\xi\mathbf{r})^2}
\end{aligned}
$$

and

$$
\begin{aligned}
h_{\xi\xi\boldsymbol{\nu}\boldsymbol{\nu}}(\mathbf{y},\psi,\xi) \quad \equiv \quad & \frac{-(1+\xi)(\mathbf{1}+\xi\mathbf{r})^{-1/\xi}\odot\left\{\frac{\log(\mathbf{1}+\xi\mathbf{r})}{\xi^2}-\frac{\mathbf{r}}{\xi(\mathbf{1}+\xi\mathbf{r})}\right\}^2}{\psi^2(\mathbf{1}+\xi\mathbf{r})^2} \\
& - \frac{(1+\xi)(\mathbf{1}+\xi\mathbf{r})\odot\left\{\frac{2\mathbf{r}+\mathbf{r}^2\xi}{\xi^2(\mathbf{1}+\xi\mathbf{r})}-\frac{2\log(\mathbf{1}+\xi\mathbf{r})}{\xi^3}\right\}}{\psi^2(\mathbf{1}+\xi\mathbf{r})^2} \\
& + \frac{\left[\mathbf{1}-(\mathbf{1}+\xi\mathbf{r})^{-1/\xi}\odot\left\{\frac{\log(\mathbf{1}+\xi\mathbf{r})}{\xi^2}-\frac{\mathbf{r}}{\xi(\mathbf{1}+\xi\mathbf{r})}\right\}\right]\odot(2\mathbf{1}-4\mathbf{r}-2\xi\mathbf{r})}{\psi^2(\mathbf{1}+\xi\mathbf{r})^3} \\
& + \frac{2\mathbf{r}\odot\left\{\xi\mathbf{1}-(\mathbf{1}+\xi\mathbf{r})^{-1/\xi}\right\}\odot(3\xi\mathbf{1}-2\xi\mathbf{r}+\mathbf{1})}{\psi^2(\mathbf{1}+\xi\mathbf{r})^4}.
\end{aligned}
$$

References

Barndorff-Nielson, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343–365.

Barndorff-Nielson, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*. London: Chapman & Hall.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Bllio, R., Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling*, **5**, 217–227.

Besag, J. and Green, P. J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, Series B*, **55**, 25–37.

Bortot, P., Coles, S. G., Sisson, S. A. (2007). Inference for stereological extremes. *Journal of the American Statistical Association*, **102**, 84–92.

Bownman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford: Clarendon.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Broyden, C. G. (1967). Quasi-Newton methods and their application to function minimization. *Mathematics of Computation*, **21**, 368–81.

Casella, G. and Edward I. G. (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**, 167–174.

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

Coles, S. G. (1993). Regional modelling of extreme storms via max-stable processes. *Journal of the Royal Statistical Society, Series B*, **55**, 797–816.

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

Coles, S. G. and Tawn, J. A.(1990). Statistics of coastal flood prevention. *Phil. Trans. R. Soc. Lond.*, **332**, 457–476.

Coles, S. G. and Tawn, J. A.(1991). Modelling multivariate extreme events. *Journal of the Royal Statistical Society, Series B*, **53**, No. 1.

Coles, S. G. and Tawn, J. A.(1996). Modelling extremes of the areal rainfall process. *Journal of the Royal Statistical Society, Series B*, **58**, 329–347.

Coles, S. G. and D. Walshaw (1994). Directional Modeling of Extreme Wind Speeds. *Journal of Applied Statistics*, **33**, 139–158.

Crainiceanu, C. and Ruppert, D. (2002). Asymptotic distribution of the likelihood ratio tests in linear mixed models. Unpublished manuscripts.

Crainiceanu, C., Ruppert, D. and Wand, M.P. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software*, **14**.

Davison, A. C. and Ramesh, N. I. (2000). Local likelihood smoothing of sample extremes. *Journal of the Royal Statistical Society, Series B*, **62**, 191–208.

Davison, A. C. (2003). *Statistical Models*. Cambridge: University Press.

De Haan, L. (1984). A spectral representation for max-stable processes. *The Annals of Probability*, **12**, 1194–1204.

De Haan, L. and Pereira, T. T. (2006). Spatial extremes: models for stationary case. *The Annals of Statistics*, **34**, 146–168.

De Haan, L. and Pickands, J. (1986). Stationary min-stable stochastic processes. *Probability Theory Related Fields*, **74**, 477–492.

De Haan, L. and Resnick, S. I. (1977). Limit Theory for Multivariate Sample Extremes. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, **40**, 317–337.

Denison, D. G. T., Holmes, C.C., Mallick, B.K. and Smith, A.F.M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Chichester, UK: Wiley.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, **11**, 89–121.

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

Gnedenko, B. V. (1943). Sur la distribution limite du terme maximum d'une série aléatorie. *Annals of Mathematics*, **44**, 423–453.

Gelman A., Carlin J. B., Stern H. S. and Rubin D. B. (2004). *Bayesian Data Analysis*. 2nd. Edn. Chapman & Hall

Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*. 2nd. Edn. Krieger, Melbourne, Fl.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–348.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, **55**, 245–259.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. London: Chapman & Hall.

Hall, P. and Tajvidi, N. (2000). Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. *Statistical Science*, **15**, 153–167.

Harville, D. A. and Mee, R. W. (1984). A mixed model procedure for analyzing ordered categorical data. *Biometrics*, **40**, 393–408.

Hastie, T., Tibshirani, R. (1970). Monte Carlo sampling methods using markov chains and their applications. *Biometrika*, **57**, 97–109.

Hastings, W. K. (1990). *Generalized additive models*. Chapman and Hall, London.

Hoerl A. E. and Kennard R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Ihaka, R. Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological events . *Quarterly Journal of the Royal Meteorological Society*, **81**, 158–272.

Kent, J. T. (1982). Information gain and a general measure of correlation . *Biometrika*, **70**, 163–173.

Lee Y. and Nelder, J. A. (1996). Hierarchical generalized linear models . *Journal of the Royal Statistical Society, Series B*, **58**, 619–678.

Lee Y., Nelder, J. A. and Pawitan, Y. (2006). *Generalized Linear Models With Random Effects: Unified Analysis Via H-Likelihood*. Chapman & Hall.

Lindsay, B. (1988). Composite likelihood methods. *Statistical Inference from Stochastic Processes. American Mathematical Society, Providence RI.*

Lindstrom, M. J. and Bates D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, **46**, 673–687.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *PNAS*, **100**, 15324–15328.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. Chapman and Hall, London.

Miller, J. J. (1977). Asymptotic proprieties of maximum likelihood estimates in the mixed model of the analysis of variance. *The Annal od Statistics*, **5**, 746–762.

Nott, D. J. and Rydén, T. (1999). Pairwise likelihood methods for inference in image models. *Biometrika*, **86**, 661–667.

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

Pickands, J. (1981). Multivariate extreme value distribution. *In Proceedings of the 43rd Session of the I.S.I.*, pages 859–878. The Hague. International Statistical Institute.

Prescott, P. and Walden, A. T. (2001). Maximum likelihood of the parameters of the generalized extreme-value distribution. *Biometrika*, **67**, 723–724.

Resnick, S. (1987). *Extreme Values, Point Processes and Regular Variation.*,Springer Verlag, New York.

Ripley, B. D. (1987). *Stochastic Simulation*, Wiley.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, **6**, 15–51.

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

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, New York.

Self, S. G. and Liang, K. Y. (1997). Asymptotic proprieties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.

Severini, T. A. (2005). *Elements of Distribution Theory*. Cambridge: University Press.

Schlather, M. (2002). Models for stationary max-stable random fields. *Extremes*, **5**, 33–44.

Schlather, M. and Tawn J. A. (2003). A dependence measure for multivariate and spatial extreme value: Properties and inference. *Biometrika*, **90**, 139–154.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, **72**, 67–90.

Smith, R. L. (1990). Max-Stable processes and spatial extremes. *Unpublished manuscript*.

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

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, **22**, 1701–1728.

Varin, C., Høst, G. and Skare, Ø. (2004). Pairwise likelihood inference in spatial generalized linear mixed models. *Computational Statistics and Data Analysis*, **49**, 1173–1191.

Varin, C., Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, **92**, 519–528.

Von Mises, R. (1954). La distribution de la plus grande de n valeurs. *In Selected Papers, Volume II*, pages 271–294. American Mathematical Society, Providence, RI.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.

Wand, M. P. (2002). Vector differential calculus in statistics. *The American Statistician*, **56**, 55–62.

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics.* **18**, 223–249.

Wand, M. P. and Ormerod, J. T. (2007). On semiparametric regression with O′ Sullivan penalised splines. *Australian and New Zealand Journal of Statistics*, to appear.

Welham, S. J., Cullis, B. R., Kenward, M. G. and Thompson, R. (2006). A comparison of mixed model splines. *Australian and New Zealand Journal of Statistics*, **49**, 1–23.

Wolfinger, R. and O′′Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, **48**, 233–243.

Yee, T. W. and Stephenson, A. G. (2007). Vector generalized linear and additive extreme value models. *Extremes.* **10**, 1–19.

Zhao, Y., Staudenmayer, J., Coull B. A. and Wand, M. P. (2006). General design Bayesian generalized linear mixed model. *Statistical Science.* **21**, 35–51.

Zhao, Y., Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics.* **33**, 335–356.