UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXVII

Bayesian mixture models for extremes

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Ilaria Prodsdocimi

Co-supervisori: Prof. Isadora Antoniano-Villalobos,
Prof. Miguel de Carvalho

Revisori: Prof. Raffaele Argiento, Prof. Daniela Castro-Camilo

Dottoranda: Viviana Carcaiso

2 Dicembre 2024

Abstract

This thesis proposes novel methods to analyse and model extremely high values stemming from multiple data-generating processes. Extreme value theory deals with mathematically modelling the behaviour of the tails of a probability distribution. In many real-world scenarios, observed maximum values deviate from a single parametric distribution, challenging the conventional assumption of being realisations of independent and identically distributed random variables. Examples arise in various domains, such as hydrology (e.g., floods resulting from rainfall and snow-melt, precipitation maxima connected to different weather regimes) and finance (e.g., stock prices bursts linked to the impact of different economic cycles, maximum losses observed in bull and bear market conditions). Incorrectly assuming a single component for the right tail when extremes are actually organised into multiple groups can lead to model misspecification and inaccurate risk estimation for rare events based on return levels.

This dissertation focuses on the context of block maxima, in which extreme observations are maxima of non-overlapping blocks. Novel methodologies based on mixture models of distributions for extremes are exploited in a Bayesian framework to capture heterogeneity in the right tail. A first contribution is the development of a two-component mixture model of Gumbel distributions, in which the allocation is based on a set of relevant variables, not just the information on the underlying process, which is typically unknown or not useful in separating the right tail. A second contribution is the use of an infinite mixture model of generalised extreme value (GEV) distributions, which captures the grouped structure in the right tail by assigning a Dirichlet process prior to the mixing measure. This approach offers a highly flexible method capable of characterising a wide range of scenarios, without imposing restrictive assumptions on the number of mixture components in the data. The proposed methods are illustrated using both simulated and real-world data, with applications to precipitation.

Sommario

Questa tesi propone nuovi metodi per analizzare e modellare valori estremamente elevati che derivano da molteplici processi di generazione dei dati. La teoria dei valori estremi si occupa di modellare matematicamente il comportamento delle code di una distribuzione di probabilità. In molti scenari reali, i valori massimi osservati si discostano da una singola distribuzione parametrica, sfidando l'assunzione convenzionale che siano realizzazioni di variabili casuali indipendenti e identicamente distribuite. Esempi si presentano in vari settori, come l'idrologia (ad esempio, le inondazioni causate da pioggia e scioglimento della neve, i massimi di precipitazioni collegati a diversi regimi meteorologici) e la finanza (ad esempio, i picchi di prezzi delle azioni legati all'impatto di diversi cicli economici, le perdite massime osservate in condizioni di mercato rialzista e ribassista). Assumere erroneamente una singola componente per la coda destra quando i valori estremi sono in realtà organizzati in più gruppi può portare a una specificazione errata del modello e a una stima inaccurata del rischio di eventi rari basata sui livelli di ritorno.

Questa tesi si concentra sul contesto dei massimi di blocchi, in cui le osservazioni estreme sono i massimi di blocchi non sovrapposti. Vengono sviluppate nuove metodologie basate su modelli di mistura di distribuzioni per valori estremi in un contesto bayesiano per catturare l'eterogeneità nella coda destra. Un primo contributo è lo sviluppo di un modello mistura di distribuzioni di Gumbel con due componenti, in cui l'allocazione si basa su un insieme di variabili rilevanti e non solo sull'informazione sul processo sottostante, che è tipicamente sconosciuta o non utile per separare la coda destra. Un secondo contributo è l'utilizzo di un modello di mistura infinita di distribuzioni GEV (generalised extreme value), che cattura la struttura a gruppi nella coda destra assegnando un processo di Dirichlet alla misura alla base della mistura. Questo approccio offre un metodo estremamente flessibile, capace di caratterizzare una vasta

gamma di scenari senza imporre assunzioni restrittive sul numero di componenti della mistura nei dati. I metodi proposti sono illustrati utilizzando sia dati simulati che dati reali, con applicazioni alle precipitazioni.

Acknowledgements

I would like to deeply thank my supervisors, Ilaria and Isadora, for their invaluable guidance and support throughout my PhD. Your expertise and feedback have been essential to deal with the many challenges that come with research. Thank you for offering not only academic insight, but also personal encouragement. A heartfelt thank you also goes to Professor Miguel de Carvalho, whose generous invitation to the University of Edinburgh made for one of the most enriching experiences of my Ph.D. The time I spent there, engaging with you and others working on extreme value theory, was really appreciated. Thank you Miguel for always finding the time to help me and for teaching me so much about research. I would like to extend my thanks to the reviewers for their valuable insights and suggestions, which have definitely improved this work. I am also grateful to the wider extreme value community for their constructive feedback at conferences and discussions; I am happy to be part of such a collaborative field.

I would also like to acknowledge the contributions of many others who supported me during my Ph.D. Special thanks to my office mates (past and present), particularly Claudia, for her assistance with coding issues, and Daniele, for his support with GitHub. My time in Padova and Edinburgh was made memorable thanks to all the other young researchers I had the pleasure of meeting. Lastly, I would like to thank my family for their constant belief in me and their support throughout this journey.

Contents

List of Figures xi

List of Tables xiv

Introduction 1
- Overview . . . 1
- Main contributions of the thesis . . . 3
- Structure and organisation . . . 5

I Framework 7

1 Background on extreme value theory 9
- 1.1 Fundamentals on extreme value theory . . . 9
  - 1.1.1 Analysis of block maxima . . . 10
  - 1.1.2 Analysis of threshold exceedances . . . 15
  - 1.1.3 Extremes of non-stationary sequences . . . 17
- 1.2 Goodness of fit diagnostics for extreme value models . . . 19
  - 1.2.1 Graphical diagnostics . . . 19
  - 1.2.2 Metrics for model comparison . . . 21
    - 1.2.2.1 Traditional error-based measures . . . 22
    - 1.2.2.2 Information criteria . . . 22
    - 1.2.2.3 Scoring rules . . . 23

2 Background on mixture models 29
- 2.1 Finite mixture models . . . 29
  - 2.1.1 Main characteristics . . . 30
    - Dirichlet distribution. . . . 31
  - 2.1.2 Clustering properties . . . 32
  - 2.1.3 Inference for finite mixture models . . . 33
  - 2.1.4 Choice of the number of components . . . 34
- 2.2 Infinite mixture models . . . 35
  - 2.2.1 Dirichlet process . . . 35
  - 2.2.2 Dirichlet process mixtures . . . 38
  - 2.2.3 Dependent Dirichlet process . . . 40

3 A brief review on mixture models for extremes **43**
3.1 Finite mixture models for extremes . . . 43
3.2 Infinite mixture models for heavy-tails . . . 46
3.3 Further approaches . . . 48

II Main contributions **49**

4 Finite mixture models for extremes **51**
4.1 Introduction . . . 51
4.2 Exploring the generation of heterogeneous extremes . . . 52
4.2.1 Finite mixture models for fixed categories . . . 52
4.2.2 Finite mixture models for uncertain categories . . . 55
4.3 Simulation study on finite mixture models . . . 56
4.4 Proper scoring rules for distinguishing extreme-value distributions . . . 65
4.5 Application to ERA5 rainfall data in Venice . . . 69
4.6 Concluding remarks . . . 75

5 Infinite mixture models for heterogeneous extremes **77**
5.1 Introduction . . . 77
5.2 Infinite mixture models for extremes . . . 78
5.2.1 Heterogeneous extremes . . . 78
5.2.2 Bayesian modelling of grouped block maxima . . . 80
5.3 Gibbs sampler for infinite mixture models . . . 81
5.4 Simulation study . . . 84
5.4.1 Simulation setup and first experiments . . . 84
5.4.2 Monte Carlo simulation . . . 87
5.5 Application . . . 90
5.5.1 Data description . . . 90
5.5.2 Modelling extreme precipitation . . . 91
5.6 Concluding remarks . . . 94

Discussion **97**
Final remarks . . . 97
Future directions of research . . . 101

Appendix A **103**
A.1 Additional numerical results . . . 103
A.2 Additional algorithms . . . 105

Appendix B **107**
B.1 How to fit finite mixture models for extremes in R . . . 107
B.2 How to fit an infinite mixture models of GEV distributions in R . . . 112

Bibliography 117

List of Figures

1.1 Annual maxima (top, red) and values over a threshold (bottom, blue) for daily precipitation data. 11
1.2 Return level plot of the GEV distribution with shape parameters $\xi = 0.3$ (Frèchet), $\xi = 0$ (Gumbel), and $\xi = -0.3$ (Weibull). 21

2.1 Histograms of waiting time to next eruption and eruption time. 30
2.2 1000 realisations from a DP(α, H_0) prior with H_0 corresponding to the standard Normal distribution for different values of α. 37

4.1 Histograms of samples of size 1000 from the two-component mixture of Gumbel distributions with different ratios of the location and the scale parameters. The weight π for process 1 is always equal to 0.2. For each bin the area is coloured according to the mixture component: process 0 (light) and process 1 (dark). 54
4.2 One-sample experiments for Scenarios A.1 and A.2 with different sample sizes. Left: median posterior return levels, based on the empirical mean of π, with posterior credible interval (shaded) for the logistic weight model (red) and the constant weight model (blue), with true return levels (black). Right: box-plots of the posterior distribution of a sample of π_i. . 59
4.3 Comparison between true allocations (different shapes) and allocations estimated using the logistic weight mixture (different colours), based on one-sample experiments. Left: $n = 1000$; right: $n = 50$. 60
4.4 One-sample experiments for the scenarios B and C. Top: median posterior return level with posterior credible interval (shaded) for the label-based model (red) and the covariate-based model (green), compared to the true return levels (black, dashed) for different values of the covariates. Box-plots of the posterior distribution of a sample of π_i are also displayed. . . 61
4.5 Classification results for Scenarios B and C using the covariate-based model, with groups indicated by different colours, compared against the classifications determined by the labels (different shapes), based on one-sample experiment outcomes. Left: $n = 1000$; right: $n = 50$. 62
4.6 Monte Carlo simulation for Scenarios B and C. Top: median posterior return levels for the label-based model (red) and the covariate-based model (green), compared to the true return levels (black, dashed) for different values of the covariates. Box-plots of the posterior median of a sample of π_i are also shown. 64

4.7 Scenario 1. Absolute difference in expected proper scoring rules between two Gumbel distributions A and B, as the location and scale parameters of B vary. 67
4.8 Absolute difference in expected proper scoring rules between a Gumbel distribution A and mixture C of A and another Gumbel distribution B, as the weight of the mixture varies. In Scenario 2.1 the distributions A and B are well separated (pink), while in Scenario 2.2 the parameters are more similar (green). 68
4.9 Total precipitation (annual maxima) in Venice. Left: histogram by precipitation type; right: return level plot for a fitted GEV distribution with 95% credible interval. 70
4.10 Boxplots of the posterior distribution of π_i from the fitted mixture models with different choices of covariates, for $i = 1, \ldots, 83$. 72
4.11 Allocation based on the labels (different shapes) and model-based allocation (different colours). Values that change group when including the label in the model are highlighted. 74

5.1 One-sample experiments. Left: posterior median density with credible interval (shaded) for single GEV model (blue) and infinite mixture of GEV distributions (red), compared to the true density (black). Right: posterior median return level curve with credible interval (shaded) for the two fitted models (same colours as before), with true return levels (black), and empirical quantiles as grey points. 86
5.2 Monte Carlo results on $n = 1000$ data. Median posterior densities (left) and return levels (right) for $M = 100$ datasets based on the single GEV model (blue) and infinite mixture of GEV distributions (red), compared to the data-generating density (black). 88
5.3 Seasonal maximum precipitations in Lisbon 1863–2018. Top: time series; bottom: histogram. 91
5.4 Left: posterior distribution of the number of occupied components. Right: posterior distribution of the mixing weight of the occupied mixture components. 93
5.5 Left: median posterior densities for the infinite mixture model (red) and single GEV model (blue) with 95% credible interval, overlapping the histogram of the seasonal maxima of precipitation. Right: return level plot with posterior return levels and credible intervals for the two models, with empirical quantiles as grey points. 94

A.1 Distribution of 1000 simulations from the two-component mixture of Gumbel distributions with different ratios of the location parameters and the scale parameters. The mixing parameter π is always equal to 0.45. For each bin the area is coloured according to the proportion that is due to process 0 (light) and process 1 (dark). 103

A.2 One-shot experiments for the additional scenario. Top: median posterior return level with posterior credible interval (shaded) for the label-based model (red) and the covariate-based model (green), compared to the true return levels (black, dashed) for different values of the covariates. Boxplots of the posterior distribution of a sample of π_i are also displayed. . . 104

List of Tables

4.1 Average proportion of event from process 1 in the 10 most extreme ones obtained from 500 samples of size 1000 from the TCEV model with different ratios of location and scale parameters. The average proportion of events from process 1 in the whole sample is displayed in parenthesis. For reference, $\lambda_1 = 2$ always and $\theta_0 = 1$ everywhere except in the first row, where it is 2. 53

4.2 Average proportion of events from process 1 among the 10 most extreme ones obtained from 500 samples of size 1000 from the two-component mixture of Gumbel distributions with different ratios of location and scale parameters, and mixing weight for process 1 set to $\pi = 0.2$. The average proportion of events from process 1 in the whole sample is in parenthesis. 54

4.3 Simulation scenarios for experiments based on a dependent mixture model of two Gumbel distributions. Here x_{i1} denotes a binary variable and x_{i2} is a continuous covariate, $i = 1, \ldots, n$. 57

4.4 Results from Monte Carlo simulation study: number of groups identified by the model, missclassification rate with respect to the true allocations, and missclassification rate with respect to the groups created by the label. Averages across the M simulated datasets. 63

4.5 Scenarios for the analysis of proper scoring rules. If a value of a parameter is not specified, it means that the parameter is allowed to take multiple values within its support. 66

4.6 Posterior median with 95% credible interval in parenthesis for the parameters of the single GEV model (Model 0) and the two-component mixture models with different choices of the covariates. 71

4.7 Posterior median with 95% credible interval in brackets for the regression coefficients of the two-component mixture models with different choices of the covariates. 73

4.8 Posterior measures for model comparison, with standard error in parenthesis: average expected logarithmic score, average expected CRPS, and LOO information criterion. 74

5.1 Simulation scenarios. Here g, N and t respectively correspond to the probability density functions of the GEV, Normal and Student's t distribution. 85

5.2 Model comparison measures from the Monte Carlo simulations: MISE (mean integrated absolute error), average expected LogS (logarithmic score), average expected CRPS (continuous ranked probability score). Median and standard error across the $M = 100$ datasets are shown. . . . 89
5.3 Posterior median with 95% credible interval in parenthesis for the single GEV model and the first four components of the infinite mixture model fitted to the seasonal maxima of precipitation in Lisbon. 92

A.1 Average proportion of events from process 1 among the 10 most extreme ones obtained from 500 samples of size 1000 from the two-component mixture of Gumbel distributions with different ratios of location and scale parameters and mixing weight for process 1 set to $\pi = 0.45$. The average proportion of events from process 1 in the whole sample is in parenthesis. 104

Introduction

This chapter serves as an introduction to the framework and motivation behind this thesis, outlining the research problems and providing a brief description of the strategies employed to address these issues.

Overview

This thesis develops Bayesian mixture models of distributions for extreme values to describe the behaviour of maximum events that originate from more than one process. As such, the thesis combines methods and approaches from extreme value statistics and mixture models.

The typical assumption of any statistical analysis that the data are representative of a unique population of interest is often too restrictive in real-world problems. This limitation is also evident in the context of extreme value theory (e.g., Coles, 2001), where the focus is not, as in most statistical analysis, on the behaviour of the central part of the distribution of the process of interest, but on its tail behaviour. Extreme value theory aims at quantifying the stochastic behaviour of a process at extremely large (or small) values on the basis of asymptotic results and provides approaches that are specifically designed for the analysis of this kind of data.

Examples of extreme events where the data are actually generated by at least two distinct processes are rainfall generated by typhoon and non-typhoon weather systems, floods originating from a mixture of rainfall and snow melt (e.g., Tarasova *et al.*, 2019) or, in a non-environmental context, stock prices bursts linked to the impact of different economic cycles. An effective way to capture the behaviour of data stemming from multiple processes is the use of mixture models, which describe the data as drawn from a density modelled as a convex combination of components, each defined by a specified parametric form. A comprehensive review of these models can be found in Frühwirth-Schnatter (2006) and Fruhwirth-Schnatter *et al.* (2019). While mixture models are widely applied

in traditional data analysis, their application in the extreme value framework remains relatively unexplored. This thesis aims at bridging the gap in the literature between Bayesian mixture models and extreme value theory.

Bayesian mixture models, illustrated for instance in Gelman *et al.* (1995, Chapters 22–23), allow dealing with heterogeneity in the data and facilitate the borrowing of information between the mixture components when estimating model parameters. The Bayesian paradigm is particularly appealing since it allows the specification of prior distributions to encode any possible knowledge of the problem and to directly assess the quantification of the uncertainty in the model components and in the high quantiles.

The research presented here is mainly driven by applications to hydrology, where the estimation of the frequency of extreme events is crucial to prevent severe damages and to implement appropriate risk management strategies. In the literature, there are numerous studies on the application of mixture models for analysing hydrological extremes. Among others, Kjeldsen *et al.* (2018) proposed a mixture of two Gumbel distributions to model extreme events from two different phenomena. However, this model is not ideal in many situations, as it assumes a priori knowledge of the process originating each event, which is usually uncommon in practical applications. This approach involves dividing the series of annual maxima into sub-samples based on the originating process, leading to separate estimations and precluding the sharing of information between mixture components. Moreover, the use of information on the generating phenomenon for grouping units may not always be effective in separating the distribution of the right tail. To address this, rather than having predefined groups using this information, this thesis aims to create an allocation scheme based on a set of external variables.

Other examples of finite mixture models applied to extreme value analysis are Grego and Yates (2010) and Otiniano *et al.* (2017), who both deal with mixtures of GEV distributions, Bottolo *et al.* (2003), who use a non-fixed number of components, and, in the context of multivariate extremes, Tendijck *et al.* (2023). However, there is no need to restrict to a specific number of components in the right tail of the distribution. Tressou (2008) and Palacios Ramirez *et al.* (2024, to appear) apply infinite mixture models in the context of heavy tails. Infinite mixture models offer a compelling alternative to their finite counterparts by allowing an unbounded number of components. This is particularly suitable for scenarios where the complexity of the data structure is challenging to characterise with a fixed number of components. These nonparametric Bayesian models often rely on the Dirichlet process (DP) (Ferguson, 1973), a stochastic process used as a nonparametric prior over probability distributions, creating the foundation for infinite

mixture models. Subsequent developments, such as the Chinese restaurant process (Aldous, 1985) and the stick-breaking construction (Sethuraman, 1994), further expanded the understanding of infinite mixture models, providing insights into the distribution of weights assigned to each component in the mixture. Nowadays, the landscape of infinite mixture models is characterised by continuous research and advancements in Bayesian nonparametric methods. For an extended review of the main methodology of Bayesian nonparametric inference we refer to Ghosal and Van der Vaart (2017). Dirichlet process mixtures and their extensions are of particular interest in this thesis, as they are very flexible models and allow the bypass of common restrictive assumptions about the number of mixture components. Despite being a well-established Bayesian nonparametric framework, DP mixtures have not yet been fully exploited in the extreme value literature. One of the contribution of this thesis is to implement these well known approaches for the field of extreme values. Although recent advancements in Bayesian nonparametrics explore more refined methods, this dissertation focuses on providing a straightforward and approachable framework for practitioners, leveraging the Dirichlet Process due to its widely recognised properties. Additionally, while there is limited understanding of the tail behaviour of random probability measures, some details are known about the tails of DP and Normalised Generalised Gamma (NGG) mixtures (e.g., Palacios Ramirez *et al.*, 2024, to appear).

Taking into account the aforementioned issues, this thesis aims at using both finite and infinite dependent mixture models to address the complexity of extreme events arising from multiple processes. In doing so, this research aims to contribute to more robust and appropriate models for real-world situations. The relevance of this problem is underscored by the critical implications of understanding and accurately modelling extreme events in risk assessment, decision-making, and policy formulation. Furthermore, the fact that different physical phenomena may not be different in the tail represents an important problem from a practical perspective, since practitioners may expect the physical phenomenon to drive the distribution.

While the main emphasis of this dissertation is on the described case, there is room for further developments along the lines of our novel approach. This could involve exploring other related innovative possibilities, such as the occurrence of multivariate extreme values resulting from multiple processes, or spatial extremes manifesting in specific locations due to different processes. Therefore, this thesis aims not only to deepen our comprehension of these occurrences but also to unveil new insights and solutions across various domains.

Main contributions

The primary objective of this thesis is to provide novel methodologies for modelling block maxima, specifically addressing scenarios where assuming that they originate from a single stationary process is unrealistic. While modelling extreme values using mixtures has been investigated in the literature, it remains a relatively unexplored area, especially when dealing with block maxima. This is likely due to the complexities of the GEV distribution, which is the conventional choice for modelling block maxima and is notoriously challenging due to parameter-dependent support, as discussed, for instance, by Martins and Stedinger (2000). Our analysis predominantly focuses on exploring applications within the environmental domain, recognising the inherent presence of heterogeneous extremes in this area. This thesis encompasses two main contributions, which represent innovative uses of Bayesian methods within the framework of extreme value analysis, accounting for the difficulties inherent to this field. We detail below these two contributions.

Contribution 1: Dependent mixtures for block maxima

A typical scenario with extreme events is that there are two processes underlying the data: one occurring more frequently and another one, which is rarer but more intense. However, identifying subgroups in the right tail purely on the basis of the process that generates each maximum may be ineffective. This thesis introduces a novel approach that exploits two-component mixture models of distributions for extreme values, specifically Gumbel distributions, to model block maxima arising from multiple processes. Unlike previous methods that rely solely on the binary division based on physical processes, our approach assigns observations to the mixture components using relevant variables that inform the posterior distribution of the mixing weights. Thus, allocation is not solely based on the information about the underlying physical process, which may be unknown or uninformative. Moreover, a key advantage of the proposed model is the hierarchical structure, which facilitates the borrowing of information between groups when estimating parameters, improving the robustness of the results.

An important consideration is whether a mixture model is truly needed or if a single GEV distribution could sufficiently capture the data. This raises the challenge of selecting appropriate tools for comparing and ranking different models. Model comparison in the extreme value framework is notoriously challenging; here, we use proper scoring

rules (Gneiting and Raftery, 2007) to evaluate and compare models. We study the effectiveness of these scoring rules in distinguishing between extreme value distributions across various scenarios.

Contribution 2: Infinite mixture models for heterogeneous extremes

The interest of this part is again on describing the grouping structure that may appear in the extremes, giving rise to grouped block maxima. Unlike the previous approach, it is assumed that each entire block belongs to a single group, and this is reflected in the distribution of block maxima, resulting in heterogeneity in extremes. To handle the uncertainty regarding the number of underlying components, a novel approach based on infinite mixture models of GEV distributions is employed. The use of an infinite number of components enables the characterisation of every possible block behaviour, while at the same time defining similarities between observations based on their extreme behaviour. Unlike the previous contribution, where the support spans the entire real line for both components, the support of an infinite mixture of GEV distributions requires careful analysis to ensure the mixture model is well-defined.

By employing a Dirichlet process prior on the mixing measure, we can capture the complex structure of the data without the need to pre-specify the number of mixture components. The innovative combination of Bayesian nonparametric techniques and models from block maxima analysis allows to detect intricate patterns in the data, which would not be accounted for using the traditional single GEV model. Posterior inference is performed using a blocked Gibbs sampler with fixed truncation (Ishwaran and James, 2001), effectively addressing the challenges of fitting this more flexible model.

Structure and organisation

The rest of the thesis is organised as follows. Part I, which consists of the first three chapters, provides the necessary preparations and background to better understand the contributions presented in later chapters. Chapters 1 and 2 present some basic notions on extreme value theory and mixture models, respectively, which will be particularly useful in the following chapters. Chapter 3 offers a review of the main approaches for the analysis of extreme values based on mixture models. Part II focuses on the main contributions of the thesis, with Chapters 4 and 5 detailing the proposed solutions to the problems addressed in Contributions 1 and 2, respectively. Each chapter is self-contained in terms of notation, definitions, and results, with the aim of providing some continuity between the two. In more detail, Chapter 4 develops a Bayesian dependent

finite mixture model of Gumbel distributions to capture extremes generated by multiple physical processes, with a simulation study to assess its performance. This model is also applied to rainfall data from Venice. Methods for model comparison in the context of extremes are also discussed. Chapter 5 explores instead an infinite mixture model of GEV distributions, supported by a simulation study and applied to precipitation data from Lisbon.

Finally, a concluding chapter summarises the main findings of the thesis, discusses the key ideas, and points out possible directions for future research. Further details on additional simulation results and algorithms can be found in Appendix A, while Appendix B explains how to implement the discussed methodology in `R`.

Part I

Framework

Chapter 1

Background on extreme value theory

This chapter offers some preparations on extreme value theory, providing concepts and methods for the statistical analysis of extreme events.

1.1 Fundamentals on extreme value theory

Extreme value theory is a field of Statistics concerned with the study of the stochastic behaviour of a process at unusually large or small levels. Foundational works like Coles (2001), Beirlant *et al.* (2006), Haan and Ferreira (2006), and Resnick (2007) provide comprehensive insights into this field. In other words, extreme value theory represents a mathematical framework for analysing and modelling the tail behaviour of probability distributions, with the goal of characterising, predicting, and assessing the risk associated with extreme events. Unusual and extreme conditions tend to have much more substantial impacts despite occurring in a much smaller proportion of the time; therefore, it is important to be able to quantify rare and extreme behaviour. Statistical analysis of extremes finds applications in various fields where the occurrence of extreme events is of interest. Some of the key fields of application include:

- Hydrology, focusing on modelling and predicting floods, droughts and extreme rainfall events (e.g., Katz *et al.*, 2002; Kjeldsen *et al.*, 2018; de Carvalho *et al.*, 2022; Jóhannesson *et al.*, 2022).

- Environmental sciences, dealing with extreme temperature events, storms, sea level rise, and hurricanes (e.g., Clarkson *et al.*, 2023; Tendijck *et al.*, 2023).

- Finance and Insurance, with applications to extreme market movements, evaluation of the tail risk of financial assets, assessment of the risk of an insurance company (e.g., Bottolo *et al.*, 2003; Hambuckers and Kneib, 2023).
- Engineering and infrastructure, for designing structures against extreme loading and producing a reliability analysis of engineering systems (e.g., Castillo, 2012).
- Healthcare (e.g., Tressou, 2008; Vettori *et al.*, 2019; Castro-Camilo *et al.*, 2022).

There are two classical approaches to model extreme events: one that focuses on studying the distribution of maximum (or minimum) values of blocks of variables and one that analyses values that exceed a large threshold. The difference between these approaches is highlighted in Figure 1.1.

1.1.1 Analysis of block maxima

In the first approach the interest is on the distribution of the sample (or block) maximum

$$M_m = \max\{W_1, \ldots, W_m\},$$

where $W_1, \ldots, W_m$ is a random sample of size m with common distribution function F. Given that $W_1, \ldots, W_m$ are independent, it is straightforward to obtain the distribution of M_m from the distribution of $W_1, \ldots, W_m$, since

$$\begin{aligned} F_{M_m}(y) &= \mathbb{P}(M_m \leq y) \\ &= \mathbb{P}(W_1 \leq y, \ldots, W_m \leq y) \\ &= \prod_{i=1}^{m} \mathbb{P}(W_i \leq y) \\ &= \{F(y)\}^m. \end{aligned}$$

However, F is unknown in practice. It is not recommended to estimate F using conventional statistical methods and then to use that to estimate F_{M_m}, as even minor inaccuracies in the estimate of F can be significantly amplified when raised to the power m. It is, nonetheless, possible to approximate the behaviour of M_m for large m based on limit results (as $m \to \infty$). Note that for any $y < y^*$, where y^* is the upper end-point of F, $F_{M_m}(y)$ converges to 0 as $m \to \infty$. To obtain a non-degenerate limit distribution it is possible to exploit a linear standardisation of the maximum, in an extreme value analogous of the central limit theorem. Asymptotic results concerning the distribution

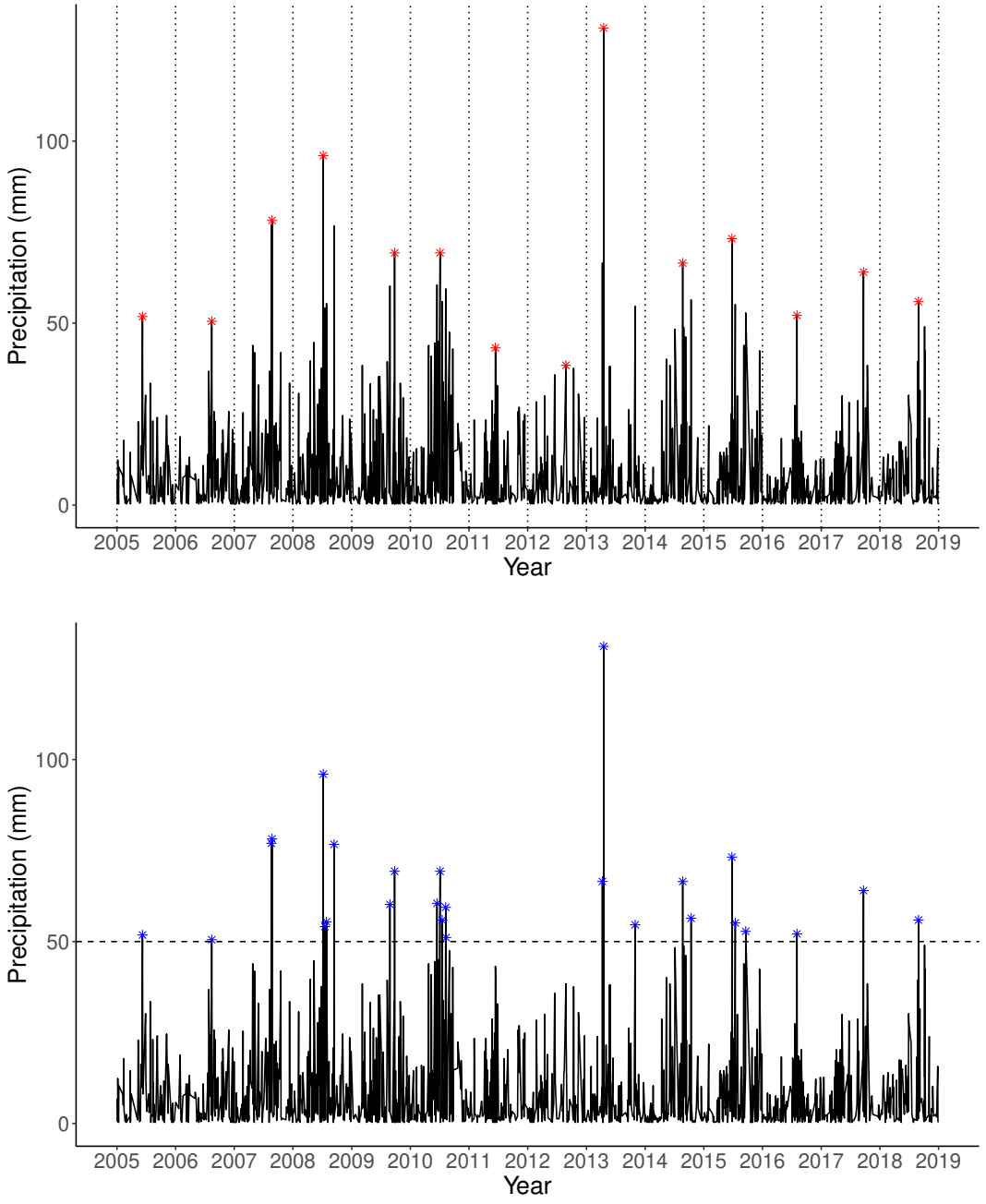


FIGURE 1.1: Annual maxima (top, red) and values over a threshold (bottom, blue) for daily precipitation data.

of the standardised maximum are given by the extremal types theorem, formulated by Fisher and Tippett (1928) and extended by Gnedenko (1948).

Theorem 1.1 (Extremal types)**.** *If there exist sequences of constants* $\{a_m > 0\}$ *and* $\{b_m\}$ *such that*

$$\mathbb{P}\left(\frac{M_m - b_m}{a_m} \leq z\right) \to G(y) \qquad \text{as } m \to \infty,$$

for a non-degenerate distribution function G, then G belongs to the family of generalised extreme value (GEV) distributions:

$$G(y) = \exp\left[-\left\{1+\xi\left(\frac{y-\mu}{\sigma}\right)\right\}^{-1/\xi}\right], \tag{1.1}$$

defined on $\{y : 1+\xi(y-\mu)/\sigma > 0\}$, where $\mu \in \mathbb{R}, \sigma > 0$ and $\xi \in \mathbb{R}$ are respectively the location, scale, and shape parameters.

The GEV distribution combines three distinct types of limit distributions, each reflecting different tail behaviours of the distribution function F for $W_1, \ldots, W_m$. The parameter ξ within the GEV distribution governs this tail behaviour, resulting in heavy tails ($\xi > 0$), light tails ($\xi = 0$), or short tails ($\xi < 0$), known respectively as the Fréchet, Gumbel, and Weibull distributions. The parameter ξ is commonly referred to as the extreme value index. The Gumbel distribution is obtain in the limiting case $\xi \to 0$, and it is defined as

$$G(y) = \exp\left\{-\exp\left(-\frac{y-\mu}{\sigma}\right)\right\}, \tag{1.2}$$

for $y \in \mathbb{R}$. The set of distributions whose normalised maximum converges to a specific extreme value distribution is called the domain of attraction (Pickands III, 1986). For example, Student-t distributions are in the Fréchet domain, Normal and Exponential distributions are in the Gumbel domain, and Beta distributions are in the Weibull domain.

The GEV family of distributions is used to model block maxima, such as weekly maximum losses in finance and annual peak flow maxima in hydrology. In the block maxima approach, data are divided into non-overlapping blocks or time intervals, and maximum values within each block are extracted. Selecting the block size poses a challenge as it entails a trade-off between bias and variance of the estimates of model parameters. Choosing very small blocks leads to poor model approximation, since Theorem 1.1 assumes that the block size approaches infinity; this results in bias during estimation and extrapolation. Conversely, opting for very large blocks produces small samples of block maxima, resulting in a high variance in the estimation of the parameters of the GEV distribution.

Moments of the GEV distribution and their extensions are used to understand and summarise the behaviour of extreme events, particularly in hydrology (Hosking and Wallis, 1997; Martins and Stedinger, 2000). The first two moments (mean and variance) help to describe the central tendency and spread of extreme values. Higher-order moments, such as skewness and kurtosis, provide information on the asymmetry and tail

behaviour of the distribution, which is critical in the context of extreme value theory. The method of probability-weighted moments (Greenwood *et al.*, 1979), a more robust generalisation of traditional moments, can be used to estimate parameters and quantiles of the GEV distribution, as discussed in Hosking *et al.* (1985). Another related and popular robust method for parameter estimation is represented by the use of L-moments (Hosking, 1990, 1992). L-moments are linear combinations of order statistics, making them more resistant to outliers than conventional moments. Another advantage is that they only require the distribution to have a finite mean. The first four L-moments correspond to L-location (analogous to the mean), L-scale, L-skewness, and L-kurtosis. L-moment diagrams (Vogel and Fennessey, 1993) can be used to assess the goodness of fit of the probability distributions, by visually comparing sample L-moments with theoretical values from candidate distributions, helping in the selection of the best-fitting model.

As another option, inference for the analysis of block maxima can be performed using likelihood-based approaches. The GEV model lacks the regularity conditions required for deriving the maximum likelihood estimator due to parameter-dependent support. This issue is studied in depth by Smith (1985), who proves that the usual asymptotic results hold for the estimators obtained by maximising the likelihood function in the case $\xi > -0.5$, which in practice is the usual scenario. A very short bounded upper tail (i.e., $\xi \leq -0.5$) is indeed considered rare in applications of extreme value analysis (Coles, 2001, p. 55). Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ be a random sample from the GEV distribution. The log-likelihood for the model with $\xi \neq 0$ is defined as

$$l(\mu, \sigma, \xi) = -n \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{n} \log \left\{1 + \xi \left(\frac{y_i - \mu}{\sigma}\right)\right\} - \sum_{i=1}^{n} \left\{1 + \xi \left(\frac{y_i - \mu}{\sigma}\right)\right\}^{-1/\xi}, \tag{1.3}$$

given that $1 + \xi(y_i - \mu)/\sigma > 0$, for $i = 1, \ldots, n$. In the case of the Gumbel distribution (i.e., $\xi = 0$),

$$l(\mu, \sigma) = -n \log \sigma - \sum_{i=1}^{n} \left(\frac{y_i - \mu}{\sigma}\right) - \sum_{i=1}^{n} \exp \left\{-\left(\frac{y_i - \mu}{\sigma}\right)\right\}. \tag{1.4}$$

There is no analytical solution for the maximum likelihood estimator, which needs to be derived using standard numerical optimisation algorithms.

The traditional approaches to extreme value theory, as discussed so far, primarily adopt a frequentist inference perspective. However, of greater relevance to this dissertation are Bayesian methods (e.g., Gelman *et al.*, 1995; Reich and Ghosh, 2019), which provide an alternative framework for fitting GEV models. Some classical examples of

Bayesian techniques applied to extreme value theory are Coles and Tawn (1996), Coles and Powell (1996), and Stephenson and Tawn (2004). The Bayesian framework allows for the inclusion of expert knowledge in the prior distribution, which is particularly useful in extreme value theory, where the amount of data available is often limited. It follows from Bayes theorem that the posterior distribution of $\boldsymbol{\theta} = (\mu, \sigma, \xi)$ is given by

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int f(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \tag{1.5}$$

Here, $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ is the likelihood corresponding to the log-likelihood in equations 1.3 or 1.4, and $p(\boldsymbol{\theta})$ is an appropriately chosen prior distribution. The choice of the prior distribution is indeed an important point of Bayesian methods for extreme value theory. Eliciting prior information using experts' knowledge on the application of interest is particularly useful for the prior on the shape parameter ξ, as information about it in the data is often limited. It is also common to restrict the prior support of ξ; for instance, a prior for the shape parameter with restricted support based on hydrological experience is employed by Martins and Stedinger (2000) in the context of hydrological data. Priors for the parameters of the GEV distribution constructed using formal rules and not based on subjective information are instead studied in Northrop and Attalides (2016). However, the Jeffreys prior and the maximal data information prior are shown to not yield proper posterior distributions, unlike independent uniform priors (when the sample size exceeds 4). Alternatively, Coles (2001, Chapter 9), uses independent Normal priors with very large variance on μ, $\log \sigma$ and ξ.

The posterior distribution in (1.5), for instance, can be optimised to obtain the maximum a posteriori (MAP) estimator of the parameters. However, the MAP provides only a limited summary of the posterior distribution. To summarise the full posterior distribution rather than just computing the mode, it is possible to exploit Monte Carlo sampling methods, such as the Gibbs sampler (Geman and Geman, 1984) and Metropolis–Hastings algorithms (Metropolis *et al.*, 1953; Hastings, 1970), to draw samples from the posterior distribution of $\boldsymbol{\theta}$. We refer to Robert *et al.* (1999) for a detailed review of Monte Carlo samplers. The Metropolis–Hastings algorithm is better suited to fit the GEV distribution, as the Gibbs sampler relies on the availability of full conditional distributions, which are not of any standard parametric form for the shape parameter. Additionally, since the GEV parameters are sometimes assumed to be correlated, it could be more efficient to sample them collectively using blocked Metropolis sampling rather than updating them sequentially at each iteration. An alternative option is to use the integrated nested Laplace approximation (INLA) (e.g., Opitz *et al.*, 2018). Once the

posterior distribution of the parameters is obtained, key quantities such as the posterior mean, median, and credible intervals can be computed.

In the case where interest is in extremely small observations, the previously described methodologies can be readily adapted for block minima, with Theorem 1.1 yielding a version of the GEV distribution for minima (Coles, 2001, Theorem 3.3). Another possibility is to fit the GEV distribution for maxima after changing the sign of the data, applying the aforementioned inferential techniques.

1.1.2 Analysis of threshold exceedances

By only considering the maximum within each block, substantial information contained in the rest of the data is ignored. This can lead to less efficient use of the available data and a potential loss of valuable insights. Other aspects of extreme events, such as the analysis of peaks over thresholds (POT) (Davison and Smith, 1990), might provide more flexibility and better use of the data. However, this approach has some drawbacks; for instance, it assumes that exceedances are independent, which is not the case in practice. Accounting for this dependence can be challenging, and will be briefly discussed in Section 1.1.3. The analysis of threshold exceedances will not be the focus of this dissertation, but is discussed here for completeness.

The basic idea of threshold exceedance models is to model the distribution of data points that exceed a predefined high threshold u. These exceedances are often of great interest in risk assessment because they represent the most extreme observations that could have significant impacts. In fact, exceedances are defined as $Y = W - u$, where W denote the original data that satisfy $W > u$. The stochastic behaviour of extremes is described as

$$\mathbb{P}(W > u + y \mid W > u) = \frac{1 - F(u + y)}{1 - F(u)}, \quad y > 0,$$

where F is the distribution function of W. Again, if the distribution F was known, the distribution of threshold exceedances would also be known, but this is not the case in practice. Simply estimating F and then substituting it to compute this ratio is not advisable, since this is likely to introduce a large bias. Thus, analogously to the use of the GEV as an approximation for the distribution of sample maxima, we need to resort to asymptotic approximations for the distribution of the exceedances. The use of the generalised Pareto (GP) distribution is motivated by the following theorem (Pickands III, 1975; Balkema and De Haan, 1974).

Theorem 1.2 (Pickands–Balkema–de Haan)**.** *Let $W_1, W_2, \ldots$ be a sequence of independent and identically distributed random variables with distribution function F satisfying Theorem 1.1, meaning that for large m the distribution of $M_m = \max\{W_1, \ldots, W_m\}$ can be approximated as a GEV distribution with parameters $\mu, \sigma > 0$ and ξ. Then, for a sufficiently large threshold u, the distribution function of $Y = W - u$, conditional on $W > u$, is approximated by*

$$H(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi}, & \text{for } \xi \neq 0, \\ 1 - \exp\left(-\frac{y}{\tilde{\sigma}}\right), & \text{for } \xi = 0, \end{cases} \tag{1.6}$$

defined on $\{y : y > 0$ and $1 + \xi y/\tilde{\sigma} > 0\}$, with $\tilde{\sigma} = \sigma + \xi(u - \mu)$.

The distribution in equation (1.6) is known as generalised Pareto (GP) with scale parameter $\tilde{\sigma}$ and shape parameter ξ. Theorem 1.2 implies that, if block maxima approximately follow a GEV distribution, then the distribution of threshold exceedances is approximated by a generalised Pareto distribution. It also follows that there is a direct correspondence between the GP and GEV distribution parameters. As in equation (1.1), the shape parameter ξ, also called the extreme value index, is responsible for the tail behaviour. If $\xi < 0$, the distribution of exceedances has an upper bound $u - \tilde{\sigma}/\xi$; while, if $\xi \geq 0$, no upper limit exists.

Choosing an appropriate threshold is a delicate point, as it influences the accuracy and stability of the model. Like for the choice of the block size, there is a bias-variance trade-off which is linked to the size of the sample of extremes. A threshold that is too low implies a large sample of exceedances but may result in bias, since it is likely to violate the asymptotic assumptions of Theorem 1.2. Conversely, a threshold that is too high will produce a small sample size and not enough exceedances for reliable model estimation, yielding high variance in the parameter estimates. Basic techniques for threshold selection are based on an exploratory analysis before model fitting or on stability of the estimates after fitting models across a range of thresholds, but numerous other methods have been proposed.

The parameters of the GP distribution can be estimated using the method of moments or L-moments, maximum likelihood methods, or Bayesian inferential techniques. For a sample[1] $\boldsymbol{y} = (y_1, \ldots y_n)'$ of exceedances of a threshold u, the log-likelihood for

[1]In this dissertation $y_1, \ldots y_n$ denotes a sample of size n from a specified model. Although samples from GEV and GP models can differ in both size and values, we use this notation consistently throughout the thesis to maintain clarity.

$\xi \neq 0$ is given by

$$l(\tilde{\sigma}, \xi) = -n \log \tilde{\sigma} - (1 + 1/\xi) \sum_{i=1}^{n} \log(1 + \xi y_i/\tilde{\sigma}), \tag{1.7}$$

for $1 + \xi y_i/\tilde{\sigma} > 0$, $i = 1, \ldots, n$. The log-likelihood when $\xi = 0$ is instead

$$l(\tilde{\sigma}) = -n \log \tilde{\sigma} - \sum_{i=1}^{n} y_i/\tilde{\sigma}. \tag{1.8}$$

As for the GEV model, numerical methods are needed to obtain maximum likelihood estimates, since they cannot be derived analytically. Bayesian methods can be applied to fit a GP model, as in the GEV case, by employing the log-likelihood in (1.7) or (1.8) and a chosen prior distribution on the parameters (e.g., Castellanos and Cabras, 2007).

1.1.3 Extremes of non-stationary sequences

The previously discussed methods assume identically distributed observations. This section focuses on versions of these methods adapted for non-stationary settings. Non-stationarity implies that the statistical properties of the process, such as the mean, variance, or higher moments, change over time. This characteristic poses significant challenges for traditional extreme value analysis. For instance, in the context of finance, stock market returns and financial indices often exhibit non-stationarity due to changes in market conditions, economic policies, and investor behaviour. Moreover, non-stationarity is observed in phenomena like extreme rainfall, temperature, and sea level events. For example, increasing global temperatures lead to shifts in the distribution of temperature extremes, resulting in more frequent and severe heatwaves. Similarly, a change in precipitation patterns can cause more intense and frequent flooding events.

The study of extreme value theory for non-stationary random sequences has been explored in the literature (Coles, 2001; Beirlant *et al.*, 2006; Haan and Ferreira, 2006), but a comprehensive solution has not been developed. These approaches indeed rely on the concept of addressing non-stationarity retrospectively, i.e., they are applied to data that were originally assumed to be from a stationary process. Furthermore, data are often not treated as generated by a stochastic process, so changes in distribution are addressed separately from the lack of independence between consecutive observations.

A common strategy to handle non-stationarity is to model the time-varying behaviour explicitly. This can be achieved through the use of external variables that help explain

the changes in the distribution of the extremes. For example, in the context of environmental data, covariates such as temperature, seasonality, or other climate indicators can be incorporated into the model. Davison and Smith (1990) introduced generalised linear models for extremes, allowing the parameters of the extreme value distributions to vary with covariates. For instance, a non-stationary GEV model can be used to describe the limiting distribution of the normalised maximum M^*, as

$$M^* \sim \text{GEV}(\mu(\boldsymbol{x}), \sigma(\boldsymbol{x}), \xi(\boldsymbol{x})), \tag{1.9}$$

where $\boldsymbol{x}$ is a vector of covariates and $\mu(\boldsymbol{x})$, $\sigma(\boldsymbol{x})$, and $\xi(\boldsymbol{x})$ are covariate-depending location, scale and shape parameters, respectively. Each of them can be modelled using a generalised linear model (Coles, 2001, Chapter 6):

$$\mu(\boldsymbol{x}) = \psi_\mu(\boldsymbol{x}'\boldsymbol{\beta}_\mu), \quad \sigma(\boldsymbol{x}) = \psi_\sigma(\boldsymbol{x}'\boldsymbol{\beta}_\sigma), \quad \xi(\boldsymbol{x}) = \psi_\xi(\boldsymbol{x}'\boldsymbol{\beta}_\xi), \tag{1.10}$$

where $\psi_\mu, \psi_\sigma, \psi_\xi$ are inverse-link functions and $\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma, \boldsymbol{\beta}_\xi$ vectors of regression coefficients. The model in (1.10) might be adapted as necessary by allowing only some of the parameters to depend on one or more, possibly different, predictors. In many applications, the shape parameter ξ is typically assumed to remain constant and independent of covariates, with a few exceptions (e.g., Wang and Tsai, 2009).

Let $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ be the data, where y_i is the maximum observation and $\boldsymbol{x}_i$ the corresponding vector of covariates. Similarly to expression (1.3), the log-likelihood for model (1.9) in the case $\xi(\boldsymbol{x}_i) \neq 0, i = 1, \ldots, n$, is given by

$$\begin{aligned} l(\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma, \boldsymbol{\beta}_\xi) = &- \sum_{i=1}^n \log \sigma(\boldsymbol{x}_i) - \sum_{i=1}^n \left[\left(1 + \frac{1}{\xi(\boldsymbol{x}_i)} \right) \log \left\{ 1 + \xi(\boldsymbol{x}_i) \left(\frac{y_i - \mu(\boldsymbol{x}_i)}{\sigma(\boldsymbol{x}_i)} \right) \right\} \right] \\ &- \sum_{i=1}^n \left\{ 1 + \xi(\boldsymbol{x}_i) \left(\frac{y_i - \mu(\boldsymbol{x}_i)}{\sigma(\boldsymbol{x}_i)} \right) \right\}^{-1/\xi}, \end{aligned}$$

provided that $1 + \xi(\boldsymbol{x}_i)(y_i - \mu(\boldsymbol{x}_i))/\sigma(\boldsymbol{x}_i) > 0$, for $i = 1, \ldots, n$. Then $\mu(\boldsymbol{x})$, $\sigma(\boldsymbol{x})$ and $\xi(\boldsymbol{x})$ need to be replaced by their expressions in terms of $\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma, \boldsymbol{\beta}_\xi$ as in (1.10). If $\xi(\boldsymbol{x}_i) = 0$ for any i, a form analogous to the Gumbel log-likelihood (1.4) needs to be used. Again, numerical techniques are necessary to obtain maximum likelihood estimators of $\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma, \boldsymbol{\beta}_\xi$. Bayesian methods can also be applied to estimate model (1.9), using Monte Carlo Markov Chain (MCMC) algorithms (Gelman *et al.*, 1995; Robert *et al.*, 1999) to obtain samples from the posterior distributions of $\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma, \boldsymbol{\beta}_\xi$. Coles (2001, Chapter 6), considers model (1.9) to capture changes in time. Alternative approaches have been

proposed for this purpose, e.g., the use of dynamic linear models (Huerta and Sansó, 2007).

Similar methods have been proposed to model non-stationary threshold exceedances. In the case of non-stationarity, the threshold can be adjusted over time to reflect the changing nature of the underlying process. Davison and Smith (1990) consider a framework for peaks over threshold where the threshold and the parameters of the generalised Pareto distribution (1.6) depend on covariates. This involves defining a threshold $u(\boldsymbol{x}_i)$ and modelling the exceedances using a GPD$(\tilde{\sigma}(\boldsymbol{x}), \xi(\boldsymbol{x}))$, where, similarly to equations (1.10), the parameters can be formulated through generalised linear models. Chavez-Demoulin and Davison (2005) extends the idea of modelling the parameters of the GP distribution using additive models, providing a flexible approach to include covariates linearly.

Chapters 4 and 5 will present alternative approaches for non-stationary extremes, focusing on using mixture models of GEV distributions to describe block maxima generated by multiple processes. In contrast to model (1.9), in Chapter 4 covariates will be incorporated into the weights of the mixture model rather than directly influencing the parameters of the GEV distribution.

1.2 Goodness of fit diagnostics for extreme value models

Model comparison in extreme value theory is a critical step to ensure that the chosen model accurately captures the behaviour of extreme events. This process involves a combination of graphical and quantitative techniques to evaluate and compare the performance of different models.

1.2.1 Graphical diagnostics

Graphical methods for model comparison complement numerical tools (see Section 1.2.2) by providing visual insights that might be overlooked with numerical analysis alone. Graphical diagnostics are indeed crucial for assessing the fit of extreme value models. Plots allow practitioners to visually compare the observed data with the theoretical distribution, helping to identify any deviations from the model in a straightforward way. Common visual diagnostics, as discussed for example in Coles (2001), are:

- Q-Q plots (Quantile-Quantile plots), which compare the quantiles of the observed data with the quantiles of the theoretical distribution. If the model fits well, the points should lie approximately on the 45-degree line.
- P-P plots (Probability-Probability plots), which, similarly to Q-Q plots, compare the cumulative distribution function of the observed data with the theoretical one.
- Density (or distribution) plots, which overlay the empirical density (or distribution) functions with those predicted by different models to provide a visual indication of model fit.
- Return level plots, which show the return levels (quantiles corresponding to specified return periods) against return periods. Comparing empirical return levels with those predicted by different models helps to assess which model better captures the tail behaviour.

Among these types of plots, we will present density plots and return level plots in Chapters 4 and 5. Detailed explanation of the latter can be found below.

A critical concept in extreme value theory is return levels, representing the value expected to be exceeded once in a specified return period, such as 100 years, under the assumption of stationarity. These levels are quantiles of the distribution fitted to the extremes, providing valuable information for risk assessment and decision-making in fields like environmental science and finance. To visualise and interpret these values, return level plots are used. These plots display return levels against their corresponding return periods, often on a logarithmic scale, facilitating the detection of deviations from model assumptions and assessing model fit. By plotting the return levels, it is possible to understand the frequency and magnitude of extreme events, allowing for more informed decisions regarding safety measures and risk management strategies.

For both GEV and GP distribution expressions of return levels can be obtained analytically, since under the assumption of i.i.d. data return levels correspond to quantiles. In particular, quantiles of the block maximum distribution can be estimated by inverting the distribution function (1.1) as

$$q_p = \begin{cases} \mu + \frac{\sigma}{\xi}\left[\{-\log(1-p)\}^{-\xi} - 1\right] & \text{if } \xi \neq 0, \\ \mu - \sigma \log\{-\log(1-p)\} & \text{if } \xi = 0, \end{cases} \tag{1.11}$$

where $G(q_p) = 1-p$. The return level q_p corresponds to the return period $1/p$, meaning that q_p is expected to be exceeded once every $1/p$ blocks, assuming that the process is stationary. Plotting q_p against $\log\{-\log(1-p)\}$ is common practice, since it results in

a linear plot in the case of the Gumbel distribution. This can be observed in Figure 1.2, which shows return level curves for the different families belonging to the GEV distribution.

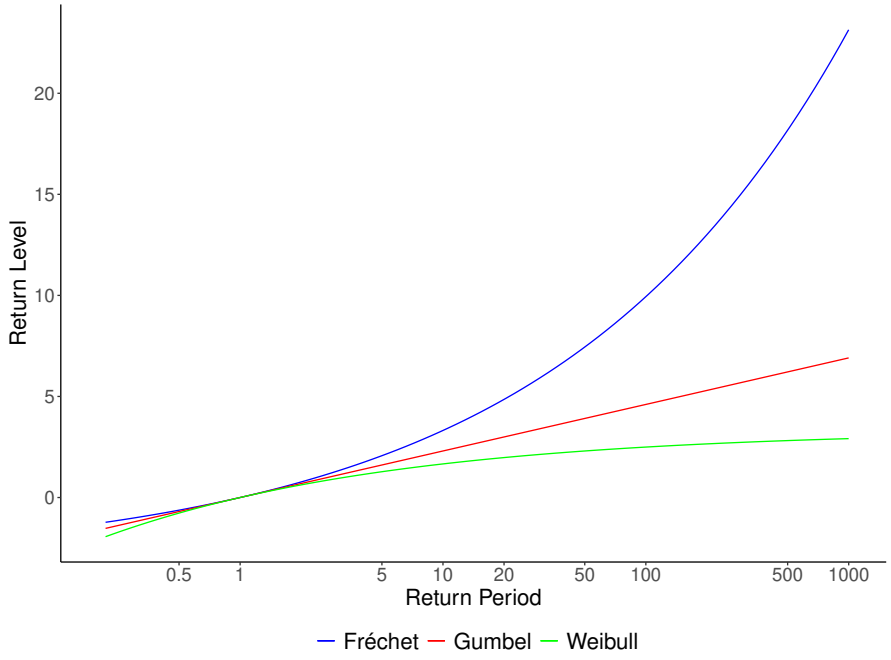


FIGURE 1.2: Return level plot of the GEV distribution with shape parameters $\xi = 0.3$ (Frèchet), $\xi = 0$ (Gumbel), and $\xi = -0.3$ (Weibull).

Inference for return levels can be performed by substituting maximum likelihood estimates of model parameters into equation (1.11) and approximating the corresponding variance using the delta method (Coles, 2001, p. 56). In a Bayesian framework, the posterior distribution of a return level can be obtained and key quantities of interest (e.g., posterior mean, median, credible interval) can be computed. Estimates of return level q_p can be plotted against $\log\{-\log(1-p)\}$ and compared to the empirical quantiles for a graphical visualisation of model performance.

Similarly, for the peaks over threshold approach, it is possible to define the s-observation return level. Again, the return level is usually plotted against s on a logarithmic scale, and the tail index is responsible for the linearity ($\xi = 0$), concavity ($\xi > 0$) or convexity ($\xi < 0$) of the curve.

1.2.2 Metrics for model comparison

1.2.2.1 Traditional error-based measures

Statistical measures provide a quantitative basis for model comparison. A classical approach is based on assessing the quality of estimates or predictions. Common examples, applicable across various statistical analyses and not limited to extreme value theory, are mean absolute error (MAE), mean squared error (MSE), and mean integrated squared error (MISE). In particular, MISE (Wand and Jones, 1994, p. 15) is used in the context of density estimation and it is defined by

$$\text{MISE} = \mathbb{E}\left[\int_{\mathcal{Y}} \{\hat{f}(y) - f(y)\}^2 \mathrm{d}y\right], \tag{1.12}$$

where $\hat{f}$ denotes the estimated density function and f is the true one, and the integral is computed over the sample space.

These measures do not take into account the complexity of the model. Below, we introduce criteria for goodness of fit that incorporate a penalty for the number of parameters.

1.2.2.2 Information criteria

A further approach for evaluating and comparing models involves estimating the out-of-sample predictive performance of a model without waiting for future data. Commonly used criteria include likelihood-based measures such as the Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz, 1978). They are defined as

$$\text{AIC} = 2k - 2l(\hat{\theta}_{MLE}), \quad \text{BIC} = k\log(n) - 2l(\hat{\theta}_{MLE}), \tag{1.13}$$

where k is the number of estimated parameters in the model, n is the number of data points and $l(\hat{\theta}_{MLE})$ is the value of the log-likelihood function for the model computed at the maximum likelihood estimate. These criteria penalise model complexity to prevent overfitting. Lower values indicate a better model considering both fit and complexity.

A fully Bayesian analogue of likelihood-based techniques is the widely applicable information criterion (WAIC) (Watanabe, 2009, 2010). It considers the log-posterior predictive density corrected for the effective number of parameters to account for overfitting. There are two approaches to define the WAIC, based on two different bias corrections:

$$\text{WAIC}_1 = -2\sum_{i=1}^{n} \log \mathbb{E}\{p(y_i \mid \theta) \mid \boldsymbol{y}\},$$

and

$$\text{WAIC}_2 = 2\sum_{i=1}^{n} \text{Var}\{\log p(y_i \mid \theta) \mid \boldsymbol{y}\} - 2\sum_{i=1}^{n} \log \mathbb{E}\{p(y_i \mid \theta) \mid \boldsymbol{y}\},$$

where $p(y_i \mid \theta)$ denotes the likelihood of the data point y_i with parameter value θ, and $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is the data vector.

When the model has highly influential data points, the information criterion based on leave-one-out cross-validation (LOO) can provide more accurate estimates, especially when approximated using Pareto-smoothed importance sampling (PSIS-LOO) (Vehtari *et al.*, 2017). LOO is a method for estimating the out-of-sample predictive performance of a model by iteratively leaving out each data point from the dataset, fitting the model to the remaining data and then predicting the left-out data point. The process is repeated for each data point in the data set and the predictive performance is assessed by averaging the prediction errors. The Bayesian information criterion based on LOO is

$$\text{LOO-IC} = -2\sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{y}_{-i}), \tag{1.14}$$

where y_i is the ith data point and $p(y_i \mid \boldsymbol{y}_{-i})$ is the posterior predictive density. Both WAIC and LOO-IC are known to provide nearly unbiased estimates of the predictive capacity of a model (Watanabe, 2010). Moreover, they have been shown to be asymptotically equal (Watanabe, 2010). Many other Bayesian predictive criteria have been proposed, such as the deviance information criterion (DIC) introduced by Spiegelhalter *et al.* (2002). Like WAIC, DIC estimates the effective number of parameters to account for overfitting. More details on information criteria for model comparison can be found in Gelman *et al.* (2014) and Piironen and Vehtari (2017b).

1.2.2.3 Scoring rules

Common evaluation methods such as MAE and MSE may not be reliable when focusing on extreme observations (Lerch *et al.*, 2017, Table 2). A valid alternative to assess predictive performance is provided by proper scoring rules (Gneiting and Raftery, 2007). A scoring rule is a measure of forecast quality that allows the ranking of competing procedures by assigning a numerical score based on the assumed distribution and a realising observation. It is proper if the expected score with respect to a distribution is minimised when the forecast and that distribution coincide, and strictly proper if the minimum is unique. The idea of propriety is that if a scoring rule is proper, a forecaster will achieve the highest expected score by reporting the probability distribution that matches the

true distribution. In practice, scores for the different observations are combined and the average score is used to compare and rank competing forecast procedures.

More specifically, a scoring rule is a function $\mathrm{S} : \mathcal{F} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ that returns a numerical value on the basis of a forecast $F \in \mathcal{F}$ and an observation $y \in \mathcal{Y}$. Like for the previous metrics, a smaller score implies a better performance. A scoring rule S is proper with respect to a class $\mathcal{F}$ if it satisfies

$$\mathbb{E}_F\{\mathrm{S}(F, Y)\} \geq \mathbb{E}_F\{\mathrm{S}(\tilde{F}, Y)\},$$

for all possible forecasted distributions $F, \tilde{F} \in \mathcal{F}$. The most popular choices are the logarithmic score (LogS) (Good, 1952)

$$\mathrm{LogS}(F, y) = -\log f(y), \tag{1.15}$$

where f is the density corresponding to F, and the continuous ranked probability score (CRPS) (Matheson and Winkler, 1976)

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} \{F(z) - I(y \leq z)\}^2 \, \mathrm{d}z, \tag{1.16}$$

where $I(\cdot)$ is the indicator function. CRPS is directly related to another proper scoring rule, the Brier score, (Brier, 1950), which is actually the mean squared error of the forecast. Indeed, CRPS can be understood as the integral of the Brier score calculated for a range of thresholds over all possible values of the forecasted variable (Matheson and Winkler, 1976).

Weighted scoring rules can be adopted to emphasise a region of interest, such as the right tail of the distribution. Let $w(\cdot)$ be a weight function such that $0 \leq w(z) \leq 1$ and $\int w(z) f(z) \mathrm{d}z > 0$ for every choice of density f. Examples of weighted scoring rules are the conditional likelihood (CL) score (Diks *et al.*, 2011)

$$\mathrm{CS}(F, y) = -w(y) \log \left(\frac{f(y)}{\int_{-\infty}^{+\infty} w(z) f(z) \mathrm{d}z} \right), \tag{1.17}$$

the censored likelihood (CSL) score

$$\mathrm{CSL}(F, y) = -w(y) \log f(y) - \{1 - w(y)\} \log \left(1 - \int_{-\infty}^{+\infty} w(z) f(z) \mathrm{d}z \right), \tag{1.18}$$

and the weighted continuous ranked probability score (wCRPS) proposed by Gneiting and Ranjan (2011) as

$$\text{wCRPS}(F, y) = \int_{-\infty}^{\infty} w(z) \left\{F(z) - I(y \leq z)\right\}^2 \, \mathrm{d}z. \tag{1.19}$$

When the weight function is equal to 1, the CL and CSL scores become the logarithmic score, while the wCRPS clearly reduces to CRPS. In extreme value theory, a practical choice of the weight functions to target the right tail of the distribution is $w(z) = I(z > t)$, where t represents a high threshold. Another recommended option is to use weight functions derived from the Normal distribution function (Gneiting and Ranjan, 2011), specifically $w(z) = \Phi(z \mid t, s^2)$. This choice of w ensures that the CL and CSL scores are always well defined. Additionally, in Chapter 4 we will explore a weight function based on the Gumbel distribution, expressed as $w(z) = G(z \mid t, s^2, 0)$, where G denotes the cumulative distribution function of a GEV distribution as defined in equation (1.1).

Numerous formal predictive performance tests have been developed (e.g. Diebold and Mariano, 2002). Assume that we want to test equal predictive performance between two forecasts A and B based on a scoring rule S and a set of observations $\boldsymbol{y} = (y_1, \ldots, y_n)'$. Let the respective average scores be defined as

$$\bar{\mathrm{S}}^A = \frac{1}{n} \sum_{i=1}^{n} \mathrm{S}(A, y_i), \quad \bar{\mathrm{S}}^B = \frac{1}{n} \sum_{i=1}^{n} \mathrm{S}(B, y_i).$$

The significance of the difference between forecasts can be evaluated using the test statistic proposed by Diebold and Mariano (2002):

$$t_{DM} = \frac{\bar{\mathrm{S}}^A - \bar{\mathrm{S}}^B}{\sqrt{n}\hat{\sigma}_{DM}}, \tag{1.20}$$

which follows an asymptotic Normal distribution under the null hypothesis of equal performance of A and B, given standard regularity conditions. Here, $\hat{\sigma}^2_{DM}$ is an estimator of the asymptotic variance of the score difference, which can simply be the sample variance of the differences $\mathrm{S}(A, y_1) - \mathrm{S}(B, y_1), \ldots, \mathrm{S}(A, y_n) - \mathrm{S}(B, y_n)$. For alternative estimators and further details in a time series context, see Gneiting and Ranjan (2011) and Diks *et al.* (2011). A significant result from the two-sided test based on statistic (1.20) indicates which forecast performs better; a negative t_{DM} implies that forecast A outperforms B, while a positive t_{DM} suggests the opposite. Diebold (2015) notes that this test was proposed to compare forecast, and not models.

Proper scoring rules are frequently used to compare forecasts of future extreme events

(e.g., Friederichs and Thorarinsdottir, 2012). However, the suitability of scoring rules in the context of extreme value theory has been questioned. Indeed, Brehmer and Strokorb (2019) showed that score expectations are not appropriate for distinguishing tail properties, highlighting the need for caution when using them to compare models. Specifically, Brehmer and Strokorb (2019) proved that proper scoring rules cannot effectively discriminate between models with different max-functional values, such as different tail indexes. They mainly focus on CRPS and wCRPS, but show that this limitation applies to all proper scoring rules, which may fail to detect differences between two models with quite different tail behaviours. Furthermore, Taillardat *et al.* (2023) continued to underline the drawbacks of using expected scores, particularly CRPS and its weighted counterpart, to compare models for extremes. Therefore, they propose an alternative index to assess tail equivalence based on a Cramér-von Mises criterion. Assume that the forecasts are in the domain of attraction of some distribution $C(\cdot \mid \tilde{\sigma}, \xi)$. Taillardat *et al.* (2023) proposed the tail-equivalent forecast performance index

$$\Omega(F, \boldsymbol{y}, u) = \frac{1}{12 m_u} + \sum_{i=1}^{m_u} \left(\frac{2i-1}{2m_u} - C(s_i^F \mid \tilde{\sigma}, \xi) \right)^2, \tag{1.21}$$

where m_u is the number of observations in $\boldsymbol{y} = (y_1, \ldots, y_n)$ exceeding a high threshold u such that the approximation to the GP distribution holds, $\tilde{\sigma}$ and ξ are the GP parameters based on threshold u, and $s_1^F, \ldots, s_{m_u}^F$ are the CRPS values based on F and $y \geq u$ in increasing order. A higher value of this index indicates a tail behaviour closer to the true one. To compare two forecasts A and B, they suggest to compute

$$1 - \frac{\Omega(B, \boldsymbol{y}, u)}{\Omega(A, \boldsymbol{y}, u)}.$$

Despite the noted criticisms, it is generally believed that there remains value in using proper scoring rules for extremes, if the focus is not solely on predicting max-functionals like the extreme value index (Olafsdottir *et al.*, 2024). To address some limitations, recent advancements have been introduced within the context of extreme value analysis, such as locally scale invariant scoring rules (Bolin and Wallin, 2023). Many scoring rules, like CRPS, are scale dependent, meaning that they give more importance to forecasts with higher uncertainty when observations vary in predictability. This results in different observations having unequal importance when computing average scores. Bolin and Wallin (2023) highlighted that this can lead to biased conclusions and introduced locally scale invariant scoring rules, ensuring that forecasts are weighted equally regardless of prediction uncertainty. While scale invariance is beneficial for average scores, it can

be too rigid when dealing with extremes. To address this, Olafsdottir *et al.* (2024) developed local weight-scale invariant scoring rules, which satisfy scale invariance within specific regions of interest. As a special case, local tail-scale invariance focuses on large events, and corresponds to defining the weight function as the indicator weight function $w(z) = I(z > t)$, with t large threshold. Building on the scaled CRPS scoring rule, which Bolin and Wallin (2023) demonstrated to be locally scale invariant, Olafsdottir *et al.* (2024) introduced the scaled weighted CRPS (swCRPS), defined as

$$\text{swCRPS}(F, y) = \frac{\int_{-\infty}^{+\infty} g_w(z, y) f(z) \mathrm{d}z}{\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g_w(z, y) f(z) \mathrm{d}z f(y) \mathrm{d}y} - \frac{1}{2} \log \left\{ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g_w(z, y) f(z) \mathrm{d}z f(y) \mathrm{d}y \right\}, \tag{1.22}$$

where $g_w(z, z') = |\int_z^{z'} w(t) \mathrm{d}t|$, with $w(\cdot)$ weight function. The swCRPS defined with the threshold weight function exhibits local tail-scale invariance, although it is not fully locally scale invariant. However, full local scale invariance may not be crucial if the property holds in the region of interest. The swCRPS is proved to be an effective option for evaluating extreme value models in scenarios with varying scales of extreme events. Details on its derivation for the GEV distribution are provided in Olafsdottir *et al.* (2024), whereas a more formal discussions on local scale invariance can be found in Bolin and Wallin (2023). As an additional note, scale dependence might occasionally be preferred in specific contexts, but it is rarely desirable in typical applications involving extremes.

In Chapter 4, we will investigate the use of proper scoring rules in the context of extremes. Our goal is to understand which scoring rules are most suitable for different scenarios and to evaluate their effectiveness in distinguishing between models for extremes to the extent that is relevant to our analysis. A simulation study will be conducted to investigate these aspects.

Chapter 2

Background on mixture models

This chapter offers preparations on mixture models, reviewing key concepts and methods related to finite and infinite mixture models, with a focus on nonparametric Bayesian inference.

2.1 Finite mixture models

Mixture models (e.g., Frühwirth-Schnatter, 2006; Fruhwirth-Schnatter *et al.*, 2019) are a powerful statistical tool for modelling complex distributions by representing them as a combination of simpler component distributions. One of the earliest mentions of these models is in Pearson (1894), but a pivotal contribution is the work of Dempster *et al.* (1977). Formally, a mixture model assumes that the observed data are generated from a mixture of different distributions, each associated with its own parameters. This approach is particularly useful when the data exhibit heterogeneity that cannot be adequately captured by a single parametric distribution. A traditional example of data arising from different sub-populations is one of the waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA, discussed by Azzalini and Bowman (1990). The data obtained from the `faithful` data set in `R` are visually summarised in Figure 2.1. It is clear that both waiting time and eruption time exhibit a bimodal behaviour. This is linked to the fact that shorter eruptions are typically followed by shorter waiting times, and longer eruptions by longer waiting times. If we were to model the distribution of either variables, we would need to to take into account this inherent heterogeneity, which can be effectively capture with a two-component mixture model.

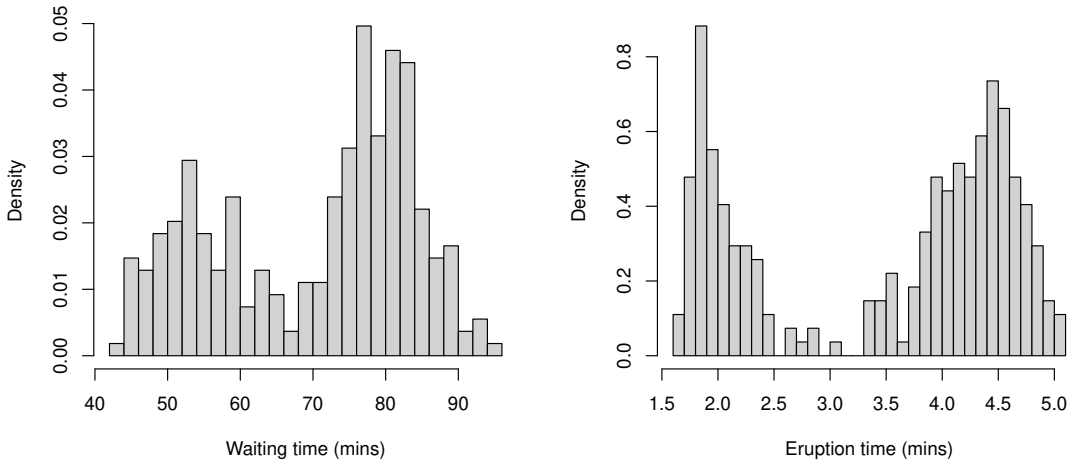


FIGURE 2.1: Histograms of waiting time to next eruption and eruption time.

2.1.1 Main characteristics

Consider a setting with K components, a vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$ of mixing proportions that sum to one, and component densities $\mathcal{K}(\cdot \mid \theta_h)$ with parameters θ_h such that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)'$. A finite mixture model can be defined as

$$f(y \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{h=1}^{K} \pi_h \mathcal{K}(y \mid \theta_h), \tag{2.1}$$

This is the standard finite mixture model considered by Everitt and Hand (1981), Titterington (1985) and Peel and McLachlan (2000). Equivalently, a finite mixture model can be written as

$$f(y \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \int \mathcal{K}(y \mid \boldsymbol{\theta}) H(\mathrm{d}\boldsymbol{\theta}), \tag{2.2}$$

with H a discrete mixing measure such that $H = \sum_{h=1}^{K} \pi_h \delta_{\theta_h}$, where δ_θ denotes the Dirac delta, a degenerate distribution with all its mass at θ. More generally, the functional form of the parametric family of each components could depend on the mixture component, as

$$f(y \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{h=1}^{K} \pi_h \mathcal{K}_h(y \mid \theta_h). \tag{2.3}$$

Bayesian nonparametric models (Ferguson, 1973) such as Dirichlet process mixture models (Lo, 1984), extend mixture models to potentially infinite components, allowing the data to determine the number of occupied components, as will be discussed later in Section 2.2.

A common approach to parameterise a mixture model involves introducing a latent categorical variable, known as allocation variable, that indicates the specific mixture component that generated each observed outcome. For each realisation y_i from mixture model (2.3) there exists a latent indicator $z_i \in \{1, \ldots, K\}$ such that

$$f(z_i = h \mid \boldsymbol{\pi}) := \mathbb{P}(Z_i = h \mid \boldsymbol{\pi}) = \pi_h, \quad h = 1, \ldots, K, \tag{2.4}$$

that is, z_i is drawn from a categorical distribution parameterised by $\boldsymbol{\pi}$. Knowing the allocations $z_1, \ldots, z_n$ allows the classification of the observations into the K mixture components.

In the Bayesian analysis of mixture models (e.g., Bernardo and Girón, 1988; Marin *et al.*, 2005; Gelman *et al.*, 1995, Chapter 22), it is necessary to add to the mixture model (2.3) the prior distributions for the parameters. The mixture model can then be written using the hierarchical structure

$$\begin{aligned} y_i \mid z_i = h &\sim \mathcal{K}_h(y_i \mid \theta_h), \\ z_i \mid \boldsymbol{\pi} &\sim f(z_i \mid \boldsymbol{\pi}), \\ \boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}), \quad \theta_h &\sim p(\theta_0). \end{aligned}$$

Here, $p(\theta_0)$ is an appropriate prior for the parameters of the component densities, and $\mathrm{Dir}(\boldsymbol{\alpha})$ denotes the Dirichlet distribution (Ferguson, 1974) with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)'$, which is the usual distribution of mixing weights due to some desirable properties that will be detailed below. An additional level of the model could be added by including hyperpriors for $\boldsymbol{\alpha}$ and θ_0. In the case of a two-component mixture, this prior choice reduces to $\pi_1 \sim \mathrm{Beta}(\alpha_0, \beta_0)$, and $\pi_2 = 1 - \pi_1$.

Dirichlet distribution. The Dirichlet distribution, denoted by $\mathrm{Dir}(\boldsymbol{\alpha})$, has probability density function

$$f(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{h=1}^{K} \alpha_h)}{\prod_{h=1}^{K} \Gamma(\alpha_h)} \prod_{h=1}^{K} \pi_h^{\alpha_h - 1}, \tag{2.5}$$

defined over the $(K-1)$-dimensional simplex

$$\Delta^{K-1} = \left\{ (\pi_1, \ldots, \pi_K) \in \mathbb{R}^K : \pi_h \geq 0 \text{ for all } h \text{ and } \sum_{h=1}^{K} \pi_h = 1 \right\}.$$

In expression (2.5) $\boldsymbol{\alpha}$ is a vector of positive concentration parameters, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)'$ is a vector in Δ^{K-1}, and Γ represents the Gamma function. The Dirichlet distribution is derived by sampling from a Gamma distribution and using the sum of the Gamma-distributed variables to normalise these draws. It serves as a conjugate prior for the Multinomial distribution. Specifically, if $\boldsymbol{\pi}$ is the parameter vector of the Multinomial distribution and $\boldsymbol{z}$ is the observed data vector, the posterior distribution is given by

$$p(\boldsymbol{\pi} \mid \boldsymbol{z}, \boldsymbol{\alpha}) \propto p(\boldsymbol{z} \mid \boldsymbol{\pi}) \cdot p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\alpha} + \boldsymbol{z}).$$

This conjugacy simplifies the computational aspects of Bayesian inference, allowing for an analytical update of the prior distribution as new data is observed. Hence, if $\boldsymbol{\pi} \sim \text{Dir}(\boldsymbol{\alpha})$, $z \mid \boldsymbol{\pi} \sim \text{Multinomial}(\boldsymbol{\pi})$ and $\mathbb{P}(z = h \mid \boldsymbol{\pi}) = \pi_h$, then $\mathbb{P}(\boldsymbol{\pi} \mid z = h, \boldsymbol{\alpha}) = \text{Dir}(\hat{\boldsymbol{\alpha}})$, where $\hat{\alpha}_h = \alpha_h + 1$ and $\hat{\alpha}_j = \alpha_j$ for each $j \neq h$. Furthermore, the marginals of the Dirichlet distribution follow Beta distributions, specifically $\pi_j \sim \text{Beta}(\alpha_j, \sum_{h \neq j} \alpha_h)$.

2.1.2 Clustering properties

Mixture models represent a common tool for model-based clustering, i.e., to find homogeneous groups within a dataset using probabilistic models; for a review on model-based clustering we refer to Bock (1996) and Fraley and Raftery (2002). Examples of finite mixture models applied for this purpose include the studies by Banfield and Raftery (1993) and Celeux and Govaert (1995). Bayesian methods for model-based clustering are explored in Malsiner-Walli *et al.* (2016).

Each occupied component of the mixture corresponds to a cluster defined by its own parameters, where an occupied component is defined as one that has at least one observation allocated to it. Therefore, we do not treat mixture components and clusters as interchangeable terms, following for instance Miller and Harrison (2018). Unlike traditional clustering methods, mixture models provide probabilistic assignments of data points to clusters. Indeed, it is possible to compute the posterior probability of observation y_i belonging to component h using the latent indicators as the probability that Z_i is equal to h conditional on the data:

$$\mathbb{P}(Z_i = h \mid y_i, \boldsymbol{\pi}, \boldsymbol{\theta}) = \frac{\pi_h \mathcal{K}_h(y_i \mid \theta_h)}{\sum_{j=1}^{K} \pi_j \mathcal{K}_j(y_i \mid \theta_j)}. \tag{2.6}$$

The soft clustering nature of mixture models, where each data point can belong to multiple clusters with varying membership probabilities, distinguishes them from hard

clustering methods such as k-means (Hartigan *et al.*, 1979).

2.1.3 Inference for finite mixture models

If the allocations are known, inference both from the frequentist and Bayesian points of view is based on the complete-data likelihood function

$$f(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{h=1}^{K} \left(\prod_{i:z_i=h} \mathcal{K}(y_i \mid \theta_h) \right) \left(\prod_{h=1}^{K} \pi_h^{n_h} \right), \tag{2.7}$$

where n_h is the number of observations in component h. It is relatively straightforward to obtain maximum likelihood estimates of the weights, while the estimates of the θ_h depend on the parametric family and may need to be obtained numerically. Complete-data Bayesian estimation, instead, involves treating each component parameter θ_h and the posterior $p(\boldsymbol{\pi} \mid \boldsymbol{\theta})$ independently . However, allocations are typically unknown and inference for finite mixture models commonly requires the mixture likelihood function, which takes the form

$$f(\boldsymbol{y} \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{h=1}^{K} \pi_h \mathcal{K}_h(y_i \mid \theta_h). \tag{2.8}$$

The maximum likelihood approach to inference is based on maximising (2.8) and has to rely to numerical methods, since dealing with a product of sums is very complicated. A common numerical technique for maximum likelihood estimation is the EM algorithm (Dempster *et al.*, 1977). This works by treating the z_i as missing data and alternating between two main steps: expectation (E) and maximisation (M). In the E-step, the algorithm calculates the conditional expectation of the complete-data log-likelihood (the logarithm of (2.7)), given the observed data and the current estimates of the parameters. In the M-step, the algorithm updates the parameter estimates by maximising the expected complete-data log-likelihood obtained from the E-step.

While we include fundamentals on frequentist inference for completeness, in this thesis we focus on the Bayesian approach, which makes use of algorithms for posterior sampling to estimate the parameters and weights of finite mixtures. The Gibbs sampler (Lavine and West, 1992; Diebolt and Robert, 1994; Escobar and West, 1995) is typically used to sample $\boldsymbol{z}$ and update $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ based on their full conditional distributions given the data and the sampled value of $\boldsymbol{z}$. A standard alternative is the Metropolis–Hastings algorithm (Marin *et al.*, 2005, p. 190). Other methods include sequential Monte Carlo (Del Moral *et al.*, 2006) and nested sampling (Skilling, 2004).

2.1.4 Choice of the number of components

Until now, we implicitly assumed that the number of mixture components K was known. In frequentist inference, a common approach to choose K is to fit many models with different values of K and to use a model selection criterion (illustrated in section 1.2.2), such as the AIC (e.g., Bozdogan and Sclove, 1984) and BIC (e.g., Fraley and Raftery, 1998), to select the best one. Other likelihood-based methods have been explored, such as using the likelihood ratio statistic (Wolfe, 1970; Hartigan and Hartigan, 1985) to test the number of components of a mixture model. The Bayesian approach offers the advantage of directly dealing with the uncertainty around K by treating it as a random variable and setting a prior distribution $p(K)$ for it. Then, inference about the number of components can be carried out alongside estimating the mixture weights and component-specific parameters. There are two categories of MCMC algorithms used in this setting. Within-model sampling involves choosing a range of values for K from 1 to a chosen maximum K^* and performing separate MCMC runs for every value. Then, results can be combined to compute the posterior distribution of K:

$$p(K \mid \boldsymbol{y}) = \frac{p(K) f(\boldsymbol{y} \mid K)}{\sum_{J=1}^{K^*} p(J) f(\boldsymbol{y} \mid J)}.$$

However, this practice could be computationally expensive. In across-model sampling, a single MCMC algorithm is instead implemented to sample from the joint posterior of $(K, \{\pi_h, \theta_h\}_{h=1}^{K})$, using, for instance, a reversible jump MCMC (Green, 1995; Richardson and Green, 1997). However, designing effective proposals for moves between models of different dimensions is complex and often results in inefficient sampling. This complexity, along with the need to assess models with different numbers of components, can lead to a high computational burden. Recently, methods for dealing with mixtures of an unknown number of components without resorting to the reversible jump MCMC algorithm have been proposed. For instance, Miller and Harrison (2018) suggested using methods designed for Dirichlet process mixture model (which will be introduced in Section 2.2.2), and in particular split-merge samplers (Dahl, 2003, 2005; Jain and Neal, 2004, 2007). Another contribution which connects finite mixture models and Bayesian nonparametric techniques is the one of Argiento and De Iorio (2022), who introduced a new prior process for the mixing measure H of model (2.2), and developed an auxiliary variable MCMC algorithm to sample from the proposed model. Frühwirth-Schnatter *et al.* (2021) considered instead a generalised class of mixture models with random number of components where the Dirichlet prior on the weights depends on the number of components, and proposed telescoping sampler for posterior inference.

Other Bayesian techniques have been proposed to select an appropriate value of K in mixture models, e.g., posterior predictive model checking (Gelfand *et al.*, 1992; Gelman *et al.*, 1996) and the use of non-local priors (Fúquene *et al.*, 2019). However, determining the number of mixture components is a delicate task, as results can be heavily influenced by this choice. Additionally, if the model is not well specified, the prior on the component-specific parameters θ_h can have a significant asymptotic impact on the inference, and hence it has to be chosen carefully (Fruhwirth-Schnatter *et al.*, 2019, Chapter 4). These challenges often motivate the use of alternative approaches, such as infinite mixture models, which infer the number of components in a more data-driven manner without explicitly sampling across models with different dimensions and avoid the need to choose a priori a specific number of components.

2.2 Infinite mixture models

As explained at the beginning of this chapter, a mixture model can be written as $f(y \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \int \mathcal{K}(y \mid \boldsymbol{\theta}) H(\mathrm{d}\boldsymbol{\theta})$, with H a mixing measure. When H is discrete and involves a finite number of parameter values, we get a finite mixture model (Section 2.1). Instead, Bayesian nonparametric models (e.g., Hjort *et al.*, 2010; Müller *et al.*, 2015; Ghosal, 2010) incorporate an infinite number of parameters in order to more flexibly represent uncertainty in H. More precisely, in a Bayesian setting, it is necessary to complete the general mixture model (2.2) with a prior model on the unknown mixing measure H. An infinite mixture model results from placing a prior probability model on H, which is a random probability measure defined on an infinite-dimensional space; see Kallenberg (1983) for an introduction to random probability measures. Complications arise in this framework as it is relatively easy to define probability measures (priors) on Euclidean spaces (i.e., in the parametric case) but it is more difficult to do it in infinite-dimensional spaces. The traditional prior in Bayesian nonparametric inference, proposed for this problem by Bush and MacEachern (1996), is the Dirichlet process prior (Ferguson, 1973).

2.2.1 Dirichlet process

A draw from a Dirichlet process is a probability measure on an Euclidean space $\mathbb{X}$, and it has the property that, for any finite partition of $\mathbb{X}$, the distribution of the vector of probabilities assigned to the elements of the partition follows a Dirichlet distribution. A Dirichlet process is denoted by $\mathrm{DP}(\alpha, H_0)$, where α is called the precision parameter and the measure H_0 is called baseline or centring measure. In particular, the realisations

of H are centred around the measure H_0, while α has the role of controlling the spread around H_0. Indeed, Ferguson (1973) showed that

$$\mathbb{E}\{H(A)\} = H_0(A), \quad \text{Var}\{H(A)\} = \frac{H_0(A)\{1 - H_0(A)\}}{1 + \alpha},$$

where A is a measurable subset of $\mathbb{X}$.

A remarkable property of the Dirichlet process is the simplicity of obtaining a posterior update. Suppose $X_1, \ldots, X_n \mid H \overset{\text{iid}}{\sim} H$ with $H \sim \text{DP}(\alpha, H_0)$, and we are interested in characterising the posterior distribution of H given $X_1, \ldots, X_n$. The posterior distribution is again a Dirichlet process (Ghosal, 2010, Section 2.2.3), given by

$$H \mid X_1, \ldots, X_n \sim \text{DP}\left(\alpha + n, \frac{\alpha}{\alpha + n} H_0 + \frac{n}{\alpha + n} F_n\right),$$

where $F_n(\cdot)$ is the empirical distribution function of $X_1, \ldots, X_n$. Moreover, a Dirichlet process restricted on a subset of its support is still distributed as a Dirichlet process, and it is independent of what happens on the rest of the support.

A key characteristic of the Dirichlet process is that its realisations are almost surely discrete random probability measures. Any $H \sim \text{DP}(\alpha, H_0)$ can be written as

$$H = \sum_{h=0}^{\infty} \pi_h \delta_{\theta_h}, \quad \theta_h \mid H_0 \overset{\text{iid}}{\sim} H_0, \tag{2.9}$$

where $\pi_h = V_h \prod_{j<h}(1 - V_j)$, with $V_h \sim \text{Beta}(1, \alpha)$ for every non-negative integer h. Here, $\{\theta_h\}_{h=1}^{\infty}$ are the atoms generated independently from the base distribution H_0, with probability mass at atom θ_h given by π_h. H_0 needs to be non-atomic (essentially continuous), so that the probability of ties is zero. This is known as the stick-breaking representation (Sethuraman, 1994; Ishwaran and James, 2001), since it uses the metaphor of sequentially breaking a stick into smaller pieces to construct a probability distribution. Imagine having a stick of unit length, which represents the total mass 1 that we want to distribute. The stick-breaking process involves the following steps:

- First, the stick is broken at a random point V_1, generating two pieces: one of length V_1 and the other of length $1 - V_1$.
- The piece of length V_1 is allocated to the first part of the probability distribution, i.e., to the first atom θ_1.

- The remaining piece of length $1 - V_1$ is then subjected to another random break. Let V_2 be the proportion of the remaining piece (of length $1 - V_1$) that will be allocated to θ_2. The actual length of the second piece is $(1 - V_1)V_2$.
- The process continues recursively, with each subsequent break being applied to the remaining piece.

Because $\mathbb{E}[V_h] = 1/(1+\alpha)$, for small values of α, such as $\alpha = 1$, most of the probability is allocated to the first few atoms, while for large α, each of the atoms is assigned an extremely small weight, so that H resembles H_0. Realisations of the Dirichlet process with base measure equal to the standard Normal distribution and different values of α are displayed in Figure 2.2. Each of these plots shows the values of atoms generated according to a standard Normal distribution H_0, that are drawn with probability defined by the stick-breaking process with the respective value of α.

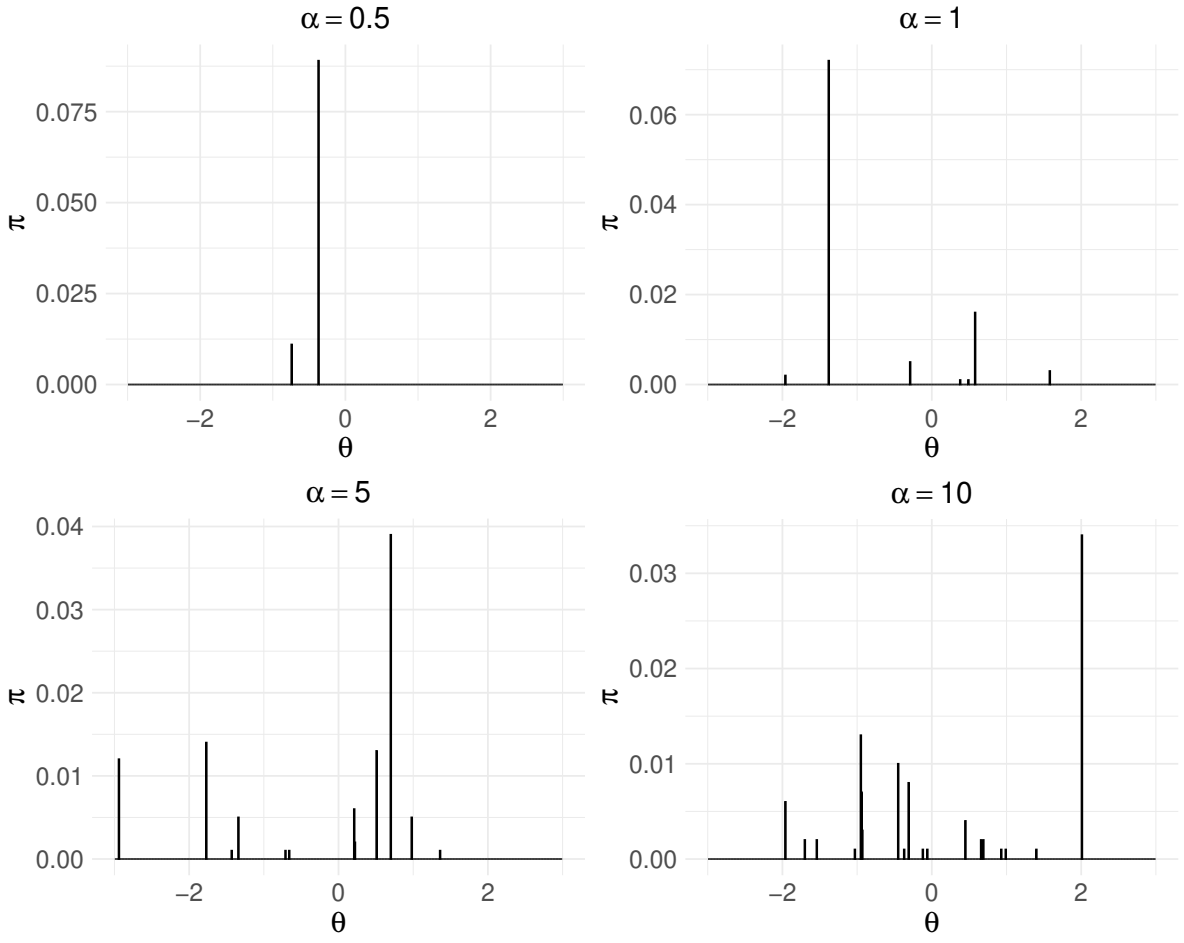


FIGURE 2.2: 1000 realisations from a DP(α, H_0) prior with H_0 corresponding to the standard Normal distribution for different values of α.

An important feature of the Dirichlet process is the clustering property (Dunson, 2010). Consider parameters θ_{z_i}, where $z_i \in \{1, \ldots, k\}$, is a latent indicator, with $i =$

$1, \ldots, n$ and $k \leq n$. Due to the discrete nature of H, as implied by (2.9), ties are created among $\theta_{z_1}, \ldots, \theta_{z_n}$ with positive probability. Let $S_j = \{i : \theta_{z_i} = \theta_j^*\}$, then $\rho_n = \{S_1, \ldots, S_k\}$ is a random partition of $\{1, \ldots, n\}$ inducing clustering. Here, $\theta_1^*, \ldots, \theta_k^*$ denote the unique values of $\theta_{z_1}, \ldots, \theta_{z_n}$, i.e. θ_j^* is the value of the parameter in cluster j. The generative scheme that illustrates how samples from the Dirichlet process lead to a partition of the data into clusters is known as the Polya urn scheme (Blackwell and MacQueen, 1973). For the first observation, a value θ_1^* is drawn from H_0. For each subsequent observation $n+1$, the value $\theta_{z_{n+1}}$ is determined by either selecting an existing value θ_j^* (i.e., joining an existing cluster), with probability proportional to the number of times θ_j^* has already been drawn, or by drawing a new value from the base distribution H_0, with probability proportional to α. Specifically, the probability that the $(n+1)$th draw is a previously drawn θ_j^* is:

$$\mathbb{P}(\theta_{z_{n+1}} = \theta_j^* \mid \theta_{z_1}, \ldots, \theta_{z_n}) = \frac{n_j}{n + \alpha},$$

where n_j is the number of times θ_j^* has been observed among the first n draws. The probability of drawing a new value from H_0 is:

$$\mathbb{P}(\theta_{z_{n+1}} \text{ is new} \mid \theta_{z_1}, \ldots, \theta_{z_n}) = \frac{\alpha}{n + \alpha}.$$

This process results in a distribution over partitions of the observations, where clusters are formed dynamically. The distribution of values generated from this scheme conforms to a Dirichlet process, where the probabilities of the clusters follow the stick-breaking construction. The probability of forming a new cluster is proportional to α, while the probability of joining an existing cluster is proportional to its current size. This results in a "rich-get-richer" phenomenon where larger clusters grow faster, but new clusters can still form.

2.2.2 Dirichlet process mixtures

The Dirichlet process is not well suited for direct density estimation of continuous data due to its discrete nature and clustering tendency. Instead of being used as a direct prior on the distribution of the data, the Dirichlet process is widely used as a priori for mixing distributions (Ferguson, 1973; Lo, 1984; Escobar and West, 1995).

Let Θ be a finite-dimensional parameter space. A general kernel mixture model can be expressed as in equation (2.2). In the specific case where H is considered discrete, with masses concentrated at a finite number of K atoms, the result is a finite mixture

model (see Section 2.1). In an infinite mixture model a prior is assigned to the random probability measure H; specifically a DP prior can be chosen (Bush and MacEachern, 1996), leading to a DP mixture model:

$$\begin{aligned} f(y \mid \boldsymbol{\pi}, \boldsymbol{\theta}) &= \sum_{h=1}^{\infty} \pi_h \mathcal{K}(y \mid \theta_h), \\ \pi_h &= V_h \prod_{j<h} (1 - V_j), \end{aligned} \tag{2.10}$$

where $V_h \sim \text{Beta}(1, \alpha)$, meaning that the weights of the mixture are sampled from a stick-breaking process with parameter α and $\theta_h \overset{\text{iid}}{\sim} H_0$, for $h = 1, \ldots, \infty$. Assuming an infinite number of mixture components in the overall population does not mean that infinitely many components are occupied by subjects in the sample. Instead, an unknown number is observed in a finite sample of size n. As mentioned in Section 2.2.1, when an additional unit is included in the sample, a new mixture component can be formed with probability $\alpha/(\alpha + n)$ (Blackwell and MacQueen, 1973). Therefore, model (2.10) avoids assuming that observations can be allocated to a predetermined number of components, unlike the finite mixture model (2.1).

As for Bayesian finite mixture models, MCMC algorithms (Robert *et al.*, 1999) are the standard approach for posterior computation of DP mixture models. Computation is, however, much more complicated due to the theoretical infinite-dimensional nature of the process. Many variations of MCMC methods have been proposed, such as the collapsed Gibbs sampler (MacEachern, 1994), the blocked Gibbs sampler with fixed truncation (Ishwaran and James, 2001), the slice sampler (Walker, 2007; Neal, 2003; Kalli *et al.*, 2011), retrospective sampling (Papaspiliopoulos and Roberts, 2008) and reversible jump algorithms (e.g., Jain and Neal, 2004). In Chapter 5, we will employ a blocked Gibbs sampler with fixed truncation for posterior sampling of the proposed model, with detailed steps explained in that section. For now, we provide a brief overview. In the blocked Gibbs sampler of Ishwaran and James (2001) the infinite sum in the stick breaking representation (2.9) is truncated to a large enough value K by letting $V_K = 1$, which is reasonable considering that π_h tends to decrease as h increases. The algorithm then iterates these steps: (i) sample the allocation indicators z_i based on the full conditional probabilities $\mathbb{P}(Z_i = h \mid \text{else})$, for $i = 1, \ldots, n$ and for $h = 1, \ldots, K$; (ii) update the stick-breaking weights from conditionally conjugate Beta posterior distributions; (iii) separately for each h, update the parameters of the component specific densities sampling from their full conditional. For this last step, a Metropolis–Hastings update may be needed in case of non-conjugacy.

2.2.3 Dependent Dirichlet process

Most Bayesian nonparametric priors designed to incorporate the dependence of probability distributions on predictors are extensions and generalisations of the Dirichlet process (Ferguson, 1973) and Dirichlet process mixture models (Lo, 1984), as defined in Sections 2.2.1 and 2.2.2.

The dependent Dirichlet process (DDP) (MacEachern, 1999) is an extension of the Dirichlet process designed to handle situations where the underlying distributions are influenced by covariates or predictors. While the standard Dirichlet process provides a flexible prior for random probability measures, it assumes that observations are exchangeable. The DDP, on the other hand, introduces dependence on covariates, resulting in partially exchangeable observations, with a distribution that varies with changes in these covariates.

In particular, assume that we have a set of observations y_i associated with covariates $\boldsymbol{x}_i$, for $i = 1, \ldots, n$. The conditional distribution of y_i given $\boldsymbol{x}_i$ is modelled as

$$y_i \mid F_{\boldsymbol{x}_i} \overset{\text{ind}}{\sim} F_{\boldsymbol{x}_i},$$

where each $F_{\boldsymbol{x}_i}$ is a random probability measure that varies with $\boldsymbol{x}_i$. The collection of these random measures, $\mathcal{F} = \{F_{\boldsymbol{x}} : \boldsymbol{x} \in \mathcal{X}\}$, is such that for each fixed $\boldsymbol{x}$, $F_{\boldsymbol{x}}$ follows a Dirichlet process. The key idea behind the dependent Dirichlet process is to construct these predictor-dependent random measures so that they retain the marginal properties of a Dirichlet process while introducing dependence on the covariates. A model that introduces covariate-dependence in a Dirichlet process mixture is given by

$$F(y \mid \boldsymbol{x}) = \int_{\Theta} G(y \mid \mu, \sigma, \xi) H_{\boldsymbol{x}}(\mathrm{d}\mu, \mathrm{d}\sigma, \mathrm{d}\xi). \tag{2.11}$$

The most common approach, according to Quintana *et al.* (2022), is to consider the single-weights DDP, where the mixing measure $H_{\boldsymbol{x}}$ is defined as

$$H_{\boldsymbol{x}} = \sum_{h=0}^{\infty} \pi_h \delta_{\boldsymbol{\theta}_h(\boldsymbol{x})} = \sum_{h=0}^{\infty} V_h \prod_{j<h} (1 - V_j) \delta_{\boldsymbol{\theta}_h(\boldsymbol{x})}. \tag{2.12}$$

Here, the variables V_h are common across all the levels of the covariates, and $\boldsymbol{\theta}_h(\boldsymbol{x})$ are independent stochastic processes with marginal distribution $\boldsymbol{\theta}_h(\boldsymbol{x}) \sim H_{0,\boldsymbol{x}}$. Posterior simulation can be carried out using sampling algorithms similar to those used for the standard Dirichlet process, making this variant of DPP widely used. A possible option is to incorporate dependence on factorial covariates by applying a simple linear regression

to the atoms, as implemented in the ANOVA-DDP model proposed by De Iorio *et al.* (2004) and applied, for instance, by De la Cruz-Mesía *et al.* (2007) and Gutiérrez *et al.* (2019). This model is extended to linear combinations of any generic set of covariates in the linear DPP (Jara *et al.*, 2010). Another proposed single-weight DDP is that of Gelfand *et al.* (2005), who define a spatial DDP with atoms that depend on spatial locations. Furthermore, Caron *et al.* (2007) propose a dynamic DDP to model a random distribution that changes over time. In the same context, Rodriguez and Ter Horst (2008) consider a model with atoms that vary with time.

A different approach to DDPs is to consider the atoms fixed across all values of $\boldsymbol{x}$, i.e., only the weights of the stick-breaking representation $\pi(\boldsymbol{x})$ depend on the covariates. The single-atoms DDP is indeed characterised by the mixing measure

$$H_x = \sum_{h=0}^{\infty} \pi(\boldsymbol{x})\delta_{\theta_h} = \sum_{h=0}^{\infty} V_h(\boldsymbol{x}) \prod_{j<h} \{1 - V_j(\boldsymbol{x})\}\delta_{\theta_h}, \tag{2.13}$$

with $V_h(\boldsymbol{x}) \sim \text{Beta}\,(1, \alpha_h(\boldsymbol{x}))$. In this case $V_h(\boldsymbol{x})$ are independent stochastic processes, and atoms $\boldsymbol{\theta}_h$ are simply distributed as $\boldsymbol{\theta}_h \overset{\text{iid}}{\sim} H_0$. Thus, all the covariate dependence is in the weights of the stick-breaking representation. In this context Griffin and Steel (2006) propose an order-based DDP, in which the ordering in the stick-breaking weights depends on covariates. Another example of single-atoms DPP is the one of Duan *et al.* (2007) for spatial applications, which allows the weights to vary spatially. A single-atoms DDP to model time-dependence is instead proposed by Gutiérrez *et al.* (2016).

The single-atoms DDP may be preferred over the single-weights DDP in cases where it could more effectively capture the underlying structure of the data. It is particularly useful when components are considered to be consistent across covariate values but their importance varies with the covariates. In contrast, the single-weights DDP might be too restrictive for some applications (Dunson, 2010).

Many other models extend the traditional framework of the dependent Dirichlet process of MacEachern (1999). Some examples are the weighted mixture of Dirichlet processes (Dunson *et al.*, 2007), the kernel stick-breaking process (Dunson and Park, 2008), the probit stick-breaking process (Chung and Dunson, 2009), the logit stick-breaking process (Ren *et al.*, 2011; Rigon and Durante, 2021), the hierarchical DP of Teh *et al.* (2006) and the nested DP of Rodriguez *et al.* (2008). For a comprehensive overview on DDP and its extensions, refer to the recent review of Quintana *et al.* (2022).

Chapter 3

A brief review on mixture models for extremes

This chapter reviews the principal current literature on the use of mixture models for extreme value theory.

In the framework of extreme value theory described in Chapter 1, mixture models offer a robust methodology for addressing heterogeneity in the tails, which can arise from extremes generated by different physical processes (e.g., summer and winter rainfall). Several examples of mixture models for extremes can be found in the literature; here, we review the most significant ones for this dissertation.

3.1 Finite mixture models for extremes

This thesis is driven by the application to hydrology, where the estimation of the frequency of the extreme events is crucial to prevent severe damages and to implement appropriate risks management strategies. We start by considering data coming from two different processes. The first contribution in this direction is that of Rossi *et al.* (1984), who defined the two-component extreme value (TCEV) distribution, which assumes that individual floods originate from a mixture of two components, representing two different types of storms, one responsible for ordinary events that are more common and less extreme in magnitude, and another one that generates more rare but severe floods. In particular, they assume the presence of two independent sequences of events in a year, each defining a compound Poisson process with parameters λ_0 and λ_1, with $\lambda_0 > \lambda_1 > 0$. Moreover, the events in each process follow an Exponential distribution with expectation, respectively, $\theta_0 > 0$ and $\theta_1 > 0$. The TCEV distribution for the

annual maximum has distribution function

$$F(y) = \exp\left\{-\exp\left(-\frac{y-\theta_0\log\lambda_0}{\theta_0}\right)\right\}\exp\left\{-\exp\left(-\frac{y-\theta_1\log\lambda_1}{\theta_1}\right)\right\}, \tag{3.1}$$

which is exactly the product of two Gumbel distribution functions, as defined in equation (1.2), one with location parameter $\theta_0\log\lambda_0$ and scale parameter θ_0, and the other with location parameter $\theta_1\log\lambda_1$ and scale parameter θ_1. We studied this distribution and also implemented an algorithm for maximum likelihood estimation. However, the practical implementation of the estimation process proved to be challenging. The behaviour of the TCEV model still requires further in-depth analysis, which is one of the reasons why this distribution will not be a primary focus of this dissertation.

The TCEV distribution (3.1) is a multiplicative mixture model of two Gumbel distributions which describes maxima originating from two processes. Kjeldsen *et al.* (2018) proposed instead a two-component mixture of Gumbel distributions to model precipitation maxima, distinguishing between events generated by typhoon and non-typhoon rainfall. Again, they focus on the case where one event occurs every year (process 0) and the other takes place infrequently (process 1). The distribution function of the annual maximum is defined as

$$F(y) = (1-\pi)\exp\left\{-\exp\left(-\frac{y-\mu_0}{\sigma_0}\right)\right\} + \pi\exp\left\{-\exp\left(-\frac{y-\mu_1}{\sigma_1}\right)\right\}, \tag{3.2}$$

with $0 < \pi < 0.5$, since as in Rossi *et al.* (1984) events from process 0 are expected to occur more frequently. The parameters $\mu_0, \sigma_0, \mu_1, \sigma_1$ are the location and scale parameters of the two Gumbel distributions of events from process 0 and 1, respectively. This model, unlike the TCEV, allows scenarios in which events from one of the phenomena do not occur every year. It also exploits more information than the TCEV, taking into account the label that identifies the process associated with the annual maximum; however, there is an additional parameter to estimate. The main drawback of this model is that the information on which process generated each annual maximum is assumed to be known when estimating model parameters, which is often not the case in practice. Even when the data are labelled with the type of process that generates each observation, it is not always the case that these labels are actually useful in identifying the subgroups in the tail population. In fact, simulation studies and applications to real data show that these labels may fail to describe the distribution of the right tail, as will be discussed in Chapter 4. The fact that different physical phenomena may not be different in the tail represents a very important problem from a practical perspective, since practitioners may expect the division based on the originating phenomenon, often informed by the

typical behaviour of events of different types, to be relevant also for the behaviour of the tail. To overcome the mentioned issues, in Chapter 4 we will propose a mixture model that exploits a set of external variables as covariates to inform the weights of the mixture.

Grego and Yates (2010) also studied the use of finite mixture models for flood frequency analysis, developing an accelerated version of the EM algorithm to estimate model parameters, observed information matrix, the 0.99th quantile of the mixture distribution, and its standard error. In particular, they focused of finite mixtures of K Gumbel distributions, defined by

$$F(y) = \sum_{h=1}^{K} \pi_h \exp\left\{-\exp\left(-\frac{y-\mu_h}{\sigma_h}\right)\right\},$$

and, unlike Kjeldsen *et al.* (2018), they did not estimate the parameters separately for each component, but used a EM algorithm, as described in Section 2.1, which is the standard for mixtures in frequentist inference. Moreover, convergence of the EM algorithm is accelerated by computing an observed information matrix. The application of this model is again based on the idea that there is a component responsible for the extremely large flooding events and another one for the more typical events, but a third component for unusually small peaks is also included. Similarly, Otiniano *et al.* (2017) study finite mixtures of multiple GEV distributions and estimate the parameters in the two-component case using the EM algorithm. They also prove the identifiability property of mixture models of Gumbel distributions and mixtures of Fréchet distributions.

From a different perspective, Liu *et al.* (2024) defined a family based on a mixture of a Gumbel distribution for the maximum and a Gumbel distribution for the minimum. The proposed distribution, called the flexible Gumbel (FG) distribution, is defined by

$$F(y) = (1-\pi)\exp\left\{-\exp\left(-\frac{y-\mu_0}{\sigma_0}\right)\right\} + \pi\exp\left\{-\exp\left(\frac{y-\mu_1}{\sigma_1}\right)\right\},$$

where π is the mixing proportion. Inference can be carried out using either a proposed EM algorithm or Bayesian methods (Metropolis-within-Gibbs sampler). Covariates can also be added into the model in the same way that they are included in the GEV distribution to model non-stationary extremes (see Section 1.1.3). This model is designed for situations where extremes are present in both tails of the distribution, which is different from the research problem of this thesis, but it is worth mentioning.

In a Bayesian framework, Bottolo *et al.* (2003) employed finite mixture models with number of mixture components determined by the data to model exceedances over a

given threshold (see Section 1.1.2) in the presence of heterogeneity. The authors highlighted the importance of Bayesian methods for modelling extreme values, particularly for handling extreme data generated by multiple sources. More specifically, data are organised into C pre-specified categories. In every category, exceedances over a high threshold are modelled by a Poisson process with category-specific parameters. In particular, the intensity of the Poisson process for exceedances in category j is given by

$$\lambda_j(y) = \left(1 + \xi \frac{y - \mu_j}{\sigma_j}\right)_+^{-1/\xi},$$

where $a_+ = \max(0, a)$ and $\mu_j \in \mathbb{R}$, $\sigma_j > 0$ and $\xi_j \in \mathbb{R}$ are respectively the location, scale and shape parameters of the Poisson process specific to category j. A distinct mixture prior distribution is specified for each parameter, with different number of components, weights, means, and variances. For instance, the model for the location μ_j is:

$$\begin{aligned}
\mu_j \mid K^\mu, \pi^\mu, \theta^\mu, \psi^\mu &\sim \sum_{h=1}^{K^\mu} \pi_h^\mu \mathrm{N}(\theta_h^\mu, (\psi_h^\mu)^2), \\
K^\mu &\sim \mathbb{P}(K^\mu = k), \quad k = 1, \ldots, C, \\
\pi^\mu \mid K^\mu &\sim \mathrm{Dir}(\alpha_1^\mu, \ldots, \alpha_{K_n}^\mu),
\end{aligned}$$

with an additional Normal hyperprior for θ_h^μ and inverse Gamma for $(\psi_h^\mu)^2$. The models for σ_j and ξ_j are analogous. This hierarchical structure enables a borrowing-of-information approach and provides considerable flexibility in taking into account heterogeneity. Inference is performed using reversible jump MCMC (Green, 1995), which is designed for Bayesian mixture models with unknown number of components.

3.2 Infinite mixture models for heavy-tails

Bayesian nonparametric approaches (Section 2.2) further extend the idea of a non-specified number of mixture components. Tressou (2008) used infinite mixtures of Pareto distributions to approximate heavy-tailed distributions, with a focus on estimating the tail index. Their aim was to identify clusters with different risk levels, defining similarity based on extreme behaviour, without knowing a priori the number of clusters. The focus was on heavy-tailed distributions, which belong to the Fréchet domain of attraction (see Section 1.1.1). Mixtures of Pareto distributions are used to represent heavy-tailed distributions. Two methods are exploited: Bayesian model-based clustering (Fraley and Raftery, 2002) (see Section 2.1.2) and a nonparametric extension with a Dirichlet process

prior (Ferguson, 1973, 1974); for more details on the Dirichlet process, see Section 2.2.1. The infinite mixture of Pareto distributions is given by

$$f(y \mid H) = \int \int g(y \mid \xi, \tau) H(\mathrm{d}\xi, \mathrm{d}\tau),$$

where $g(y \mid \xi, \tau)$ is the density of a Pareto distribution with tail index ξ and precision parameter τ, and H is a mixing measure which is assigned a Dirichlet process prior. The quantity of interest is the tail probability

$$\mathbb{P}(Y > y) = \int \int \mathbb{P}(Y > y \mid \xi, \tau) H(\mathrm{d}\xi, \mathrm{d}\tau),$$

for which an estimation algorithm is provided. Therefore, while the idea aligns with our research, the perspective and the interest are quite different.

Alternatively, Palacios Ramirez *et al.* (2024, to appear) developed heavy-tailed mixture models based on the class of normalised generalised Gamma (NGG) processes (Lijoi *et al.*, 2007). Examples of NNG processes are the Dirichlet process, the stable process, and the normalised inverse Gaussian process. The right tail of an NGG process is heavy-tailed if the centring distribution also exhibits heavy tails. However, the Dirichlet process is the only member of the NGG class that does not follow this property, as its tail is exponentially lighter than that of the centring distribution, as noted by Ghosal and Van der Vaart (2017). Palacios Ramirez *et al.* (2024, to appear) showed that NGG scale mixtures may be preferred over Dirichlet process mixtures (or NGG mixtures) of heavy-tailed kernels, like the one proposed by Tressou (2008). Two classes of heavy-tailed mixture models are developed: heavy-tailed NGG scale mixtures and heavy-tailed NGG shape mixtures. To keep it concise, we focus on the first class, which is defined as

$$f(y) = \int_0^\infty \mathcal{K}_\sigma(y \mid \eta_\sigma) \mathrm{d}H(\sigma),$$

where H is a NGG process with centring distribution H_0, which is assumed heavy tailed, and $\mathcal{K}_\sigma(\cdot) = \mathcal{K}(\cdot/\sigma \mid \eta_\sigma)/\sigma$ with $\mathcal{K}$ kernel, $\sigma > 0$ scale parameter and η_σ additional parameters. A noteworthy aspect is the extension to a dependent version, which allows analysing the effect of covariates on heavy-tailed observations. Similarly to the context of the dependent Dirichlet process (see Section 2.2.3), a single-atoms dependent stable process is constructed. All these models are fitted using a slice sampler (Neal, 2003; Kalli *et al.*, 2011). A multivariate extension of the two classes of mixture models is also proposed.

3.3 Further approaches

Shifting focus to the point process approach for extreme value analysis (Pickands III, 1971), Kottas and Sansó (2007) proposed a Bayesian nonparametric method within this framework. Intensities of a spatial nonhomogeneous Poisson process are modelled using an infinite mixture with a bivariate Beta distribution for the kernel and a Dirichlet process prior for the mixing distribution.

It is also worth mentioning the work of Tendijck *et al.* (2023), who concentrated on mixture models for multivariate extremes, expanding the application of mixture models to more complex, multidimensional settings.

Mixture models are also used to simultaneously capture both the bulk and the tail of a distribution. This approach has been applied in the univariate setting (e.g., Frigessi *et al.*, 2002; Behrens *et al.*, 2004) and has been extended to multivariate contexts (e.g., Hu *et al.*, 2024). In a similar vein, the so-called extreme value mixture models (Scarrott and MacDonald, 2012) have been developed to formally incorporate the entire dataset. For instance, do Nascimento *et al.* (2012) proposed a finite mixture of Gamma distributions for the bulk paired with a generalised Pareto distribution for the tail. It is important to note that this class of extreme value mixture models differs from the mixture models for extreme value theory explored in this thesis.

The discussed studies offer some valuable approaches on the use of mixture models to quantify the behaviour of extreme events. Nonetheless, to our knowledge Bayesian mixture models for block maxima analysis have not been extensively explored in the literature. In this dissertation, we will study mixtures of GEV distributions, focusing on the specific case of Gumbel distributions in Chapter 4. Finite and infinite mixture models will be investigated in Chapters 4 and 5, respectively.

Part II

Main contributions

Chapter 4

Finite mixture models for extremes

This chapter proposes modelling heterogeneous extremes via a Bayesian dependent mixture model. It covers the initial framework, the proposed model, a simulation study, and an application to annual rainfall maxima in Venice. The use of proper scoring rules for assessing goodness of fit is also explored.

4.1 Introduction

In numerous real-world applications, the conventional statistical assumption that data are realisations of independent and identically distributed variables is often untenable. This constitutes an issue that is also present in the extreme value framework, illustrated in Chapter 1. To model scenarios where extreme events are believed to stem from multiple processes, we extend traditional methods for modelling the behaviour of block maxima, discussed in Section 1.1.1. Chapter 3 highlighted some contributions on the use of finite mixtures for this purpose (e.g., Grego and Yates, 2010; Otiniano *et al.*, 2017; Kjeldsen *et al.*, 2018). We aim to build upon and expand these studies as there is still limited research in this area. With a focus on hydrology, we start from the model of Kjeldsen *et al.* (2018), which assumes that the physical process originating each element of the sample is known and that therefore two groups of maximum values can be defined. However, this information is typically unknown or difficult to obtain. Moreover, even when it is possible to separate the data into predefined groups, such groups may not be effective in separating extreme observations. In other words, the factors that characterise different data-generating processes for the majority of the data may not correspond to different distributional behaviours for the tails. To address this, we want to exploit other relevant variables to inform the mixing weights, thus allowing for a data-driven allocation to mixture components.

A related issue is that different data-generating processes may share the same tail behaviour. In such a scenario, a mixture model may not be necessary for extremes and a single GEV distribution, as defined in (1.1), may be sufficient to describe the data. This raises the problem of using appropriate measures to compare and rank different models for extremes. We consider proper scoring rules, illustrated in Section 1.2.2.3, to compare different models. In particular, they are employed here to determine when a mixture model with component allocation based on physical considerations is preferable to simply using a single distribution for extreme values. These scoring rules will also be used to evaluate mixture models based on different choices of variables that influence mixing weights. As mentioned in Section 1.2.2.3, the use of proper scoring rules, and particularly of CRPS, to compare different models for extremes is criticised by Brehmer and Strokorb (2019) and Taillardat *et al.* (2023). However, to the best of our knowledge, there has been limited research on the performance of different scoring rules in various scenarios. In this chapter, we aim to explore this further.

4.2 Exploring the generation of heterogeneous extremes

In this section, we explore how heterogeneous extremes can emerge on the basis of some of the finite mixture model characterisations discussed in Chapter 3. We then introduce a mixture model where the mixing weights depend on a set of covariates.

This section frequently references part of the literature and methodologies introduced in Section 3.1. Therefore, readers may find it helpful to revisit Section 3.1 for additional context and theoretical background as needed.

4.2.1 Finite mixture models for fixed categories

A typical scenario which might lead to samples of non-identically distributed extreme events occurs when there are two data-generating processes: one that occurs with a much higher frequency and another which takes place more rarely but can lead to larger magnitudes. For example, in the tropics, extreme rainfall can stem from slow and steady rain or from a typhoon (or hurricane). The distribution of rainfall measurements would be a mixture of the single distributions corresponding to the two physical processes generating them. Consequently, the block maxima extracted from the continuous records may preserve some of the distinctive features of these diverse phenomena, leading to

	$\lambda_0/\lambda_1 = 1.5$	$\lambda_0/\lambda_1 = 2$	$\lambda_0/\lambda_1 = 5$
$\theta_1/\theta_0 = 0.5$	0.0018 (0.2213)	0.0012 (0.1565)	0.0000 (0.0357)
$\theta_1/\theta_0 = 1$	0.4042 (0.3969)	0.3244 (0.3321)	0.1732 (0.1664)
$\theta_1/\theta_0 = 1.5$	0.9090 (0.5135)	0.8794 (0.4591)	0.7624 (0.2999)
$\theta_1/\theta_0 = 2$	0.9924 (0.5902)	0.9892 (0.5450)	0.9722 (0.4055)

TABLE 4.1: Average proportion of event from process 1 in the 10 most extreme ones obtained from 500 samples of size 1000 from the TCEV model with different ratios of location and scale parameters. The average proportion of events from process 1 in the whole sample is displayed in parenthesis. For reference, $\lambda_1 = 2$ always and $\theta_0 = 1$ everywhere except in the first row, where it is 2.

some grouping structure in the distribution of extremes. We begin by considering mixture models (see Chapter 2) designed to take into account this scenario. We will refer to the two physical processes that generate the data as process 0 and process 1.

The TCEV distribution, as defined in Equation (3.1), results from modelling the annual number of events as being generated by two independent Poisson processes, while the magnitudes of events from each process are assumed to be exponentially distributed. Designed to model events generated by two processes, the TCEV distribution corresponds to a multiplicative mixture of two Gumbel distributions. Table 4.1 shows the average proportion of observations from process 1 among the 10 most extreme values in 500 samples of size 1000 from a TCEV model with different choices of ratios of parameters, satisfying $\lambda_0 > \lambda_1$ (i.e., events of type 0 are on average more frequent than events of type 1, by definition of the TCEV). It is not straightforward to understand how the combination of parameters controls which of the underlying processes produces larger observations, and it is difficult also to acknowledge the role of the parameters in determining the total proportion of events of each type (displayed in parenthesis in the table). Indeed, for some choices of ratios, observations from process 1 are not the rarest ones, but are still the strongest in terms of magnitude.

An alternative approach for the same scenario is the mixture model of two Gumbel distributions proposed by Kjeldsen *et al.* (2018) and defined in (3.2). Figure 4.1 shows how the distribution of 1000 simulated data from this model (with $\pi = 0.2$) changes with different ratios between the location and the scale parameters. The location parameter μ_0 is equal to 10, while the scale parameter σ_0 is set equal to 3, except in the case $\mu_1/\mu_0 = 1.5$ and $\sigma_1/\sigma_0 = 0.5$, when it is 4, and in the case $\mu_1/\mu_0 = 2$ and $\sigma_1/\sigma_0 = 0.5$, with $\sigma_0 = 5$. It is possible to recognise that the most present process in the tail of the mixture can change depending on the parameter combinations: with some choices of the parameters, the rare observations (process 1) are not the most common in the right tail, and frequent ones (process 0) can correspond to the most extreme values. Moreover,

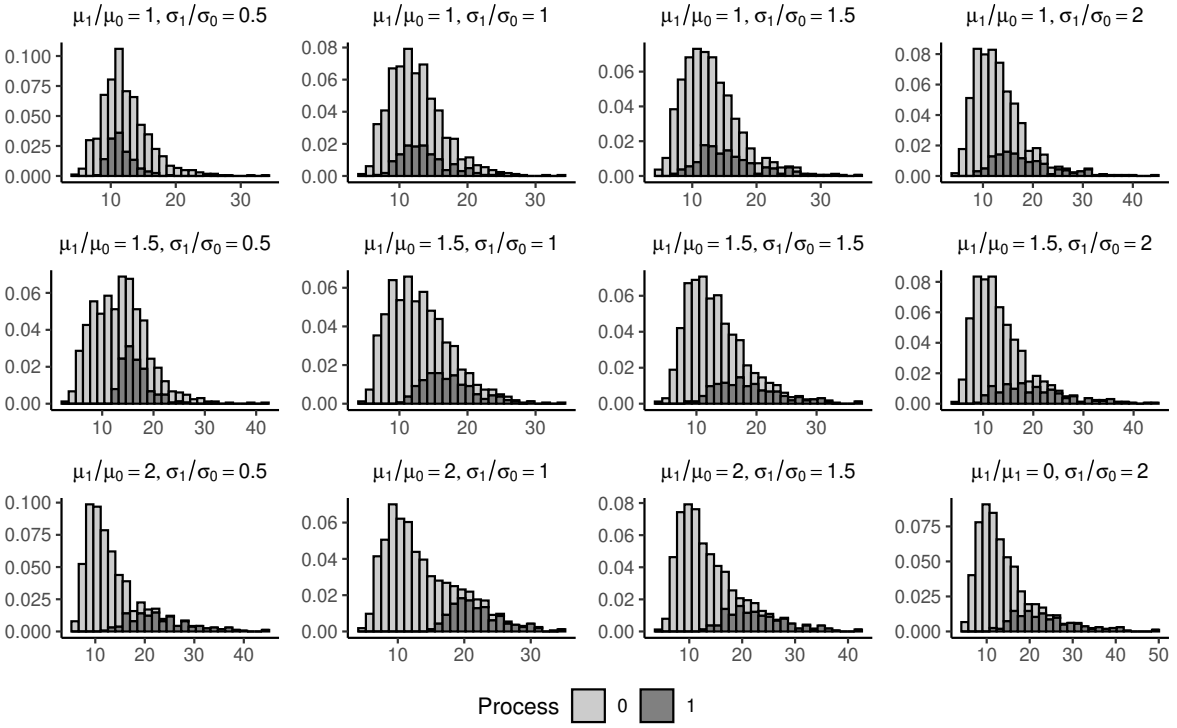


FIGURE 4.1: Histograms of samples of size 1000 from the two-component mixture of Gumbel distributions with different ratios of the location and the scale parameters. The weight π for process 1 is always equal to 0.2. For each bin the area is coloured according to the mixture component: process 0 (light) and process 1 (dark).

the largest observations come all or mostly from one of the processes. This is more noticeable in Table 4.2, which is analogous to Table 4.1 for model (3.2) and shows the estimated probability of one of the 10 largest order statistics to originate from process 1, compared to the overall probability of any observation being from process 1. This aims to highlight how the frequency at which the most extremes events are coming from the rare process is related to the different properties of the distribution of events from process 0 and process 1. As expected, events from process 1 are the majority when

	$\mu_1/\mu_0 = 1$	$\mu_1/\mu_0 = 1.5$	$\mu_1/\mu_0 = 2$
$\sigma_1/\sigma_0 = 0.5$	0.0062 (0.2002)	0.0350 (0.2003)	0.9902 (0.2002)
$\sigma_1/\sigma_0 = 1$	0.2800 (0.1994)	0.5408 (0.2005)	0.8482 (0.1998)
$\sigma_1/\sigma_0 = 1.5$	0.7254 (0.1997)	0.9276 (0.2004)	0.9632 (0.1999)
$\sigma_1/\sigma_0 = 2$	0.9296 (0.2012)	0.9634 (0.2005)	0.9902 (0.2005)

TABLE 4.2: Average proportion of events from process 1 among the 10 most extreme ones obtained from 500 samples of size 1000 from the two-component mixture of Gumbel distributions with different ratios of location and scale parameters, and mixing weight for process 1 set to $\pi = 0.2$. The average proportion of events from process 1 in the whole sample is in parenthesis.

μ_1/μ_0 or σ_1/σ_0 are large, i.e., when the two data-generating processes are indeed quite different from each other. The same plots and table for a different value of π can be found in Appendix A.1.

The model proposed by Kjeldsen *et al.* (2018) is more flexible than the TCEV distribution, as the occurrence of events of each type in each block is not required. With appropriate choices of the parameters, it can represent multiple diverse scenarios, which is appealing when representing the behaviour of real-life data. Furthermore, it allows the characterisation of the distribution of the process for each of the two groups, whereas in the TCEV a single fitted distribution is derived. However, this is also a drawback, since we need to have predefined groups and finding data on extreme events that include information about the generating processes is often difficult. Even when such information is available, it does not necessarily result in distinct groups within the extremes, as seen in many of the plots of Figure 4.1. Moreover, when groups do exist, this information may not be sufficient to effectively characterise them. For this reason, we aim to develop a model in which the allocation of data points to groups is not solely based on distinguishing a priori between process 0 and process 1.

4.2.2 Finite mixture models for uncertain categories

As discussed, although the division of the observations used in the previous models may make sense from a physical point of view, it is not necessarily appropriate for describing the behaviour of extremes. Hence, rather than having pre-fixed groups using the division based on the originating process, we want to allow the data to be informative and let the allocation to be data-driven. We assume that the data come from two populations, as in many practical scenarios there are two physical phenomena underlying the data.

Consider the series of maximum observations $\boldsymbol{y} = (y_1, \ldots, y_n)'$ and the latent allocation variables $\boldsymbol{z} = (z_1, \ldots, z_n)'$, where z_i identifies the mixture component to which y_i belongs. Let $\boldsymbol{x}_i$ be a vector of observed covariates for the ith maximum, which may include, when available, a binary indicator of the process generating y_i, for $i = 1 \ldots, n$. We define the model

$$\begin{aligned} y_i \mid z_i = j &\sim \text{Gumbel}\left(y_i \mid \boldsymbol{\theta}_j\right), \quad \boldsymbol{\theta}_j = (\mu_j, \sigma_j), \\ z_i \mid \boldsymbol{x}_i &\sim \text{Bernoulli}\left(z_i \mid \pi_i\right), \\ \pi_i &= \frac{\exp(\alpha + \boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\alpha + \boldsymbol{x}_i'\boldsymbol{\beta})}, \\ p(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \alpha, \boldsymbol{\beta}) &= p_\theta(\boldsymbol{\theta}_0) p_\theta(\boldsymbol{\theta}_1) p_\alpha(\alpha) p_\beta(\boldsymbol{\beta}), \\ i = 1, \ldots, n; &\quad j = 0, 1. \end{aligned} \tag{4.1}$$

Thus, π_i is derived from a logistic regression as a function of a set of covariates that inform the allocation of the data in the tail. This hierarchical structure allows for a more flexible modelling of extreme events that do not originate from a single process, and since the allocation to mixture components is data-driven, it does not require the process originating the data to be known. Moreover, it facilitates the borrowing of information between groups when estimating the model parameters. The Bayesian paradigm allows the specification of prior distributions that can encode the possible understanding we have of the problem, and to directly quantify the uncertainty connected to the allocation.

Posterior computation for model (4.1) is carried out using `Stan` (Carpenter *et al.*, 2017), a platform for Bayesian computation that implements Hamiltonian Monte Carlo (HMC) methods, particularly the no u-turn sampler (NUTS) (Hoffman *et al.*, 2014). Specifically, we use `RStan` (Stan Development Team, 2023), the `R` interface for `Stan`. Alternatively, `Stan` can be used as an optimisation tool to maximise an objective function, such as a log-likelihood.

4.3 Simulation study on finite mixture models

We conduct a simulation study to assess the performance of model (4.1) under various conditions. The study considers several scenarios, each corresponding to a distinct data-generating process, as outlined in Table 4.3. For each scenario we consider the sample sizes $n = 1000$ and $n = 50$, with data $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$. In Scenarios A.1 and A.2 observations y_i are independently simulated from a mixture of two Gumbel distributions:

$$F(y) = (1-\pi)\exp\left\{-\exp\left[-\frac{y-\mu_0}{\sigma_0}\right]\right\} + \pi\exp\left\{-\exp\left[-\frac{y-\mu_1}{\sigma_1}\right]\right\},$$

and $\boldsymbol{x}_i = x_{i1}$ is a binary variable which will be called "label". In Scenario A.1 the label perfectly identifies which of the two Gumbel distributions originated each datapoint, i.e. it perfectly corresponds to the true allocations z_i. In Scenario A.2 the label produces instead 10% of errors in replicating z_i. In these scenarios we aim to compare two approaches: fitting a mixture of two Gumbel distributions by separately estimating the parameters for each component, based on the assumption that the label correctly identifies the component, and fitting model (4.1) using the label as the only covariate.

We extend this analysis to scenarios where the label, i.e., the binary identifier of the data-generating process, is not useful for allocating observations to components. In Scenarios B and C, observations y_i are simulated conditionally on $\boldsymbol{x}_i$ from a mixture of

TABLE 4.3: Simulation scenarios for experiments based on a dependent mixture model of two Gumbel distributions. Here x_{i1} denotes a binary variable and x_{i2} is a continuous covariate, $i = 1, \ldots, n$.

Scenario	μ	σ	π	$\boldsymbol{x}$
A.1	$\mu_0 = 7$ $\mu_1 = 20$	$\sigma_0 = 3$ $\sigma_1 = 4$	$\pi_i = 0.2 \ \forall i$	$x_{i1} = z_i$
A.2	$\mu_0 = 7$ $\mu_1 = 20$	$\sigma_0 = 3$ $\sigma_1 = 4$	$\pi_i = 0.2 \ \forall i$	$x_{i1} \neq z_i$
B	$\mu_0 = 10$ $\mu_1 = 20$	$\sigma_0 = 2$ $\sigma_1 = 3.5$	$\alpha = -5, \beta_1 = 0.2, \beta_2 = 10$ $\pi_i = \frac{\exp(\alpha+\beta_1 x_{i1}+\beta_2 x_{i2})}{1+\exp(\alpha+\beta_1 x_{i1}+\beta_2 x_{i2})}$	$x_{i1} \sim \text{Bern}(0.2)$ $x_{i2} \sim \text{U}(0,1)$
C	$\mu_0 = 10$ $\mu_1 = 16$	$\sigma_0 = 2$ $\sigma_1 = 3.5$	$\alpha = -5, \beta_1 = 0.2, \beta_2 = 10$ $\pi_i = \frac{\exp(\alpha+\beta_1 x_{i1}+\beta_2 x_{i2})}{1+\exp(\alpha+\beta_1 x_{i1}+\beta_2 x_{i2})}$	$x_{i1} \sim \text{Bern}(0.2)$ $x_{i2} \sim \text{U}(0,1)$

two Gumbel distributions:

$$F(y \mid \boldsymbol{x}) = (1 - \pi(\boldsymbol{x})) \exp\left\{-\exp\left[-\frac{y-\mu_0}{\sigma_0}\right]\right\} + \pi(\boldsymbol{x}) \exp\left\{-\exp\left[-\frac{y-\mu_1}{\sigma_1}\right]\right\},$$

where $\boldsymbol{x} = (x_1, x_2)$. Here, x_2 is a continuous covariate drawn from a $\text{U}(0,1)$ distribution, which provides useful information about the allocation of observations to the mixture components, while x_1 is a binary label which is largely uninformative. In the following, we will use π_i to denote $\pi(\boldsymbol{x}_i)$, as in equation (4.1), $i = 1, \ldots, n$. Scenario C is analogous to scenario B but with mixture components that are more closely spaced in terms of their locations. In these scenarios we are interested in comparing the results of fitting a mixture model based solely on a binary label, representing a physical process, versus a model where the mixture weights also depend on the additional covariate.

The prior distributions for the regression coefficients of model 4.1 are those recommended by Gelman *et al.* (2008), i.e., $\boldsymbol{\beta} \overset{\text{iid}}{\sim} \text{N}(0, 2.5^2)$, whereas the intercept is assigned the prior $\alpha \sim \text{N}(0, 5^2)$. The label is transformed in order to have 0 average and a difference of 1 between the two values it can assume, whereas the continuous covariate is simply scaled in order to have mean 0 and unit variance. In all the fitted models, the prior distribution for the location parameters is specified as a Normal distribution, centred on the sample mean with a large variance, which is equal to a quarter of the range of the data. Meanwhile, the scale parameters are assigned a wide Gamma prior distribution; specifically $\sigma_j \sim \text{Gamma}(1, 0.1)$ in Scenario A.1, $\sigma_j \sim \text{Gamma}(3, 1)$ in Scenario A.2, and $\sigma_j \sim \text{Gamma}(3, 0.1)$ in Scenarios B and C, $j = 0, 1$. All models are

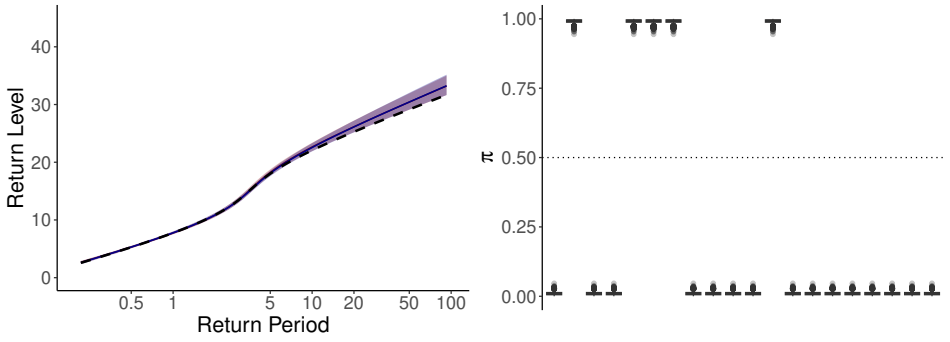
fitted employing `stan`, using 4 Markov chains with 2000 iterations each. The first 1000 iterations of each chain are treated as warm-up and discarded, leaving a total of 4000 post-warm-up iterations available for computing posterior quantities of interest.

For Scenarios A.1 and A.2, we consider the following models:

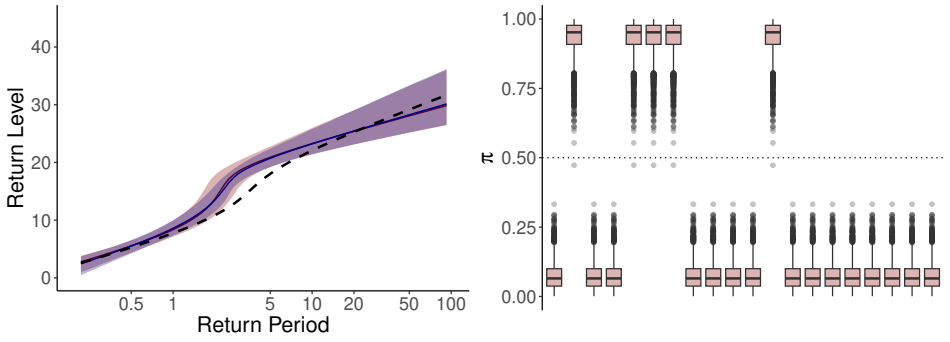
- A Bayesian version of model (3.2), fitted by grouping the observations with $x_{i1} = 0$ and $x_{i1} = 1$, and then estimating the model parameters separately for each component.
- Model (4.1) with covariate x_1.

Note that Scenario A.1 corresponds to perfectly separated data, and fitting model (4.1) with covariate x_1 to this scenario is likely to fail, as the logistic regression framework struggles to handle perfect separation (Lewis and Battey, 2024).

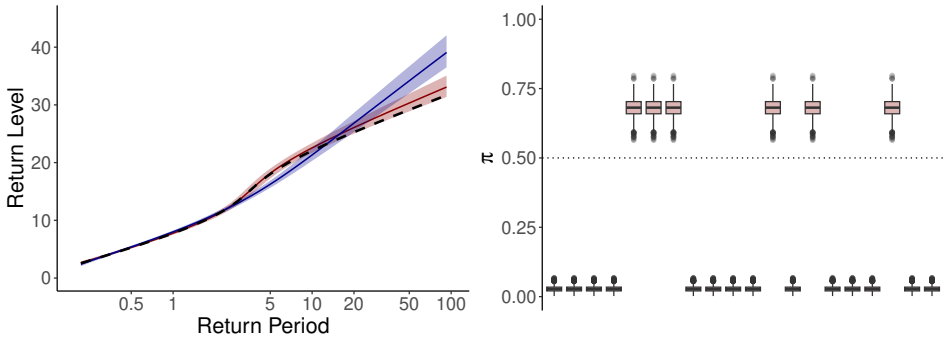
The performance of the two models with different sample sizes is displayed in Figure 4.2, which shows return level plots (see Section 1.2.1), together with the posterior distribution of a sample of 20 π_i from the second model (i.e. the logistic weight mixture). For the model with constant weights we estimate the weight of the second Gumbel as the proportion of $x_{i1} = 1$ in the sample, while for the model with label-dependent mixing weights we use the average of the estimated π_i. When the label actually identifies the originating distribution (i.e., $x_{i1} = z_i$ for $i = 1, \dots, n$), the two model performs almost identically and are able to capture the tail behaviour, especially with high sample size. When this information is wrong, even by a small percentage as in scenario A.2, the model in which different distributions are fitted separately for each value of x_1 fails to accurately capture the highest quantiles. We obtain more robust results when exploiting the logistic regression proposed in equation 4.1. The case with $n = 50$ yields similar results, but the uncertainty around the estimates is, not surprisingly, much bigger. In all cases the model with label-dependent mixing weights recognises the presence of two mixture components. As an additional information, Figure 4.3 shows the allocations estimated on the basis of this model, $\hat{z}_i$, $i = 1, \dots, n$, using a single run of the models. In particular, $\hat{z}_i^{(j)}$ is set equal to 0 if $\pi_i^{(j)} \leq 0.5$, and equal to 1 in the opposite case, where j is an iteration of the MCMC algorithm, $j = 1, \dots, 4000$. Then the estimate of z_i, which in Figure 4.3 is called "group", is the MAP of $\hat{z}_i^{(j)}$, $j = 1, \dots, 4000$, $i = 1, \dots, n$. In Scenario A.1 the two identified groups perfectly align with the true allocations z_i, while in Scenario A.2 there is some missclassification due to the label being different from z_i. This simulation study is not exhaustive, as numerous variations of Scenario A.2 could be explored to examine what occurs when x_1 differs even more from the true allocations. Furthermore, it is important to emphasise that the primary objective is to



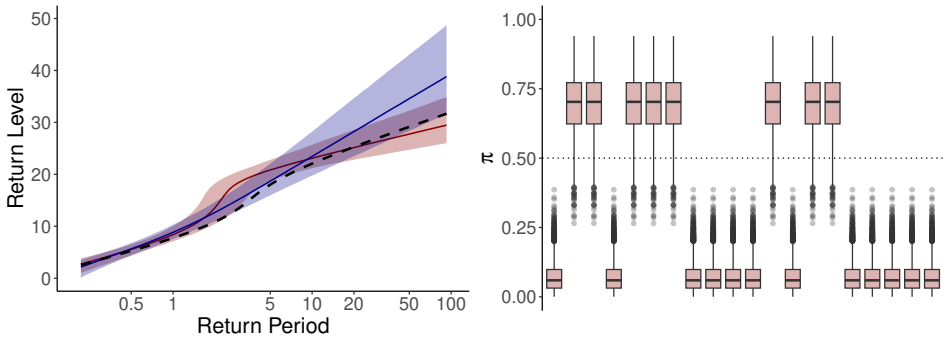
(a) Scenario A.1, $n = 1000$



(b) Scenario A.1, $n = 50$



(c) Scenario A.2, $n = 1000$



(d) Scenario A.2, $n = 50$

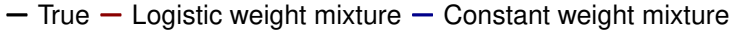
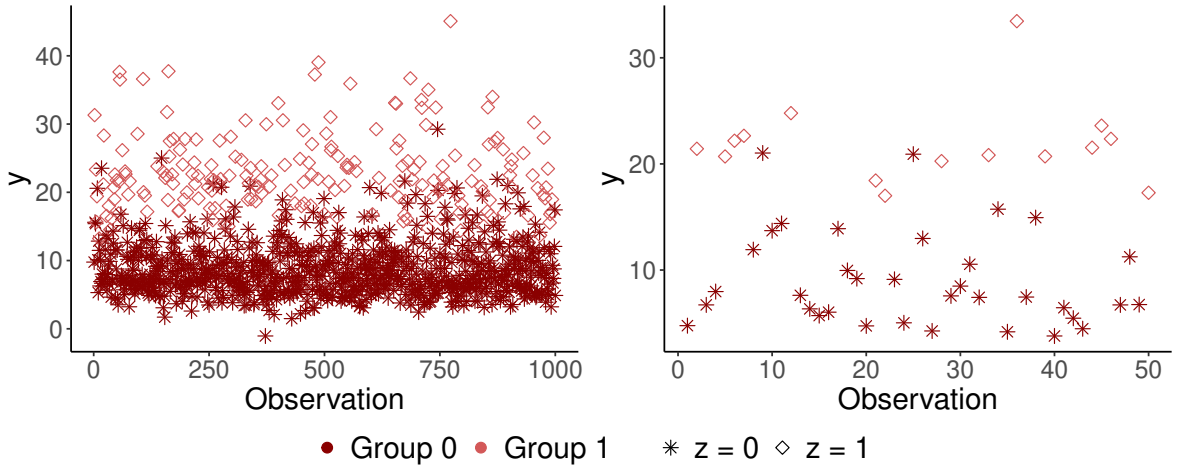
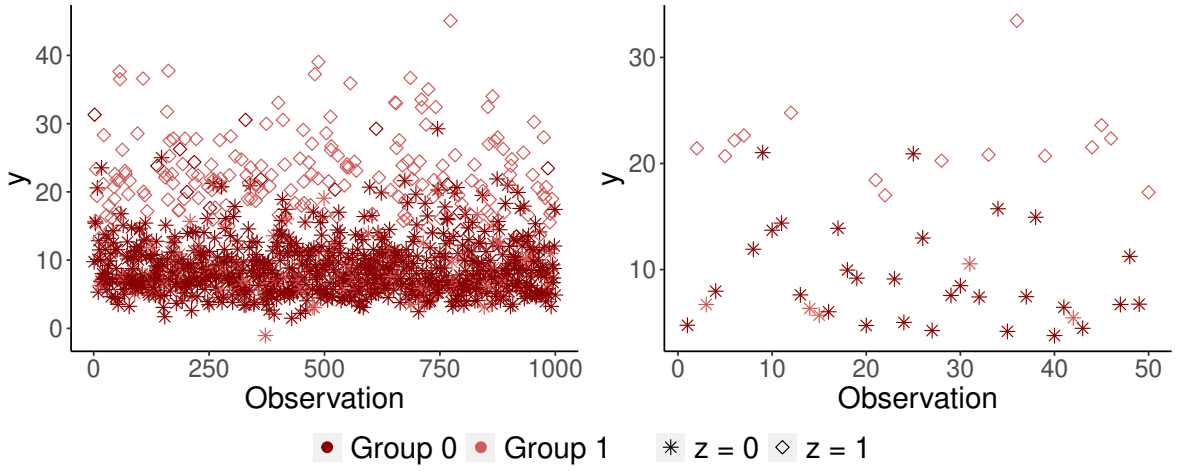


FIGURE 4.2: One-sample experiments for Scenarios A.1 and A.2 with different sample sizes. Left: median posterior return levels, based on the empirical mean of π, with posterior credible interval (shaded) for the logistic weight model (red) and the constant weight model (blue), with true return levels (black). Right: box-plots of the posterior distribution of a sample of π_i.



(a) Scenario A.1

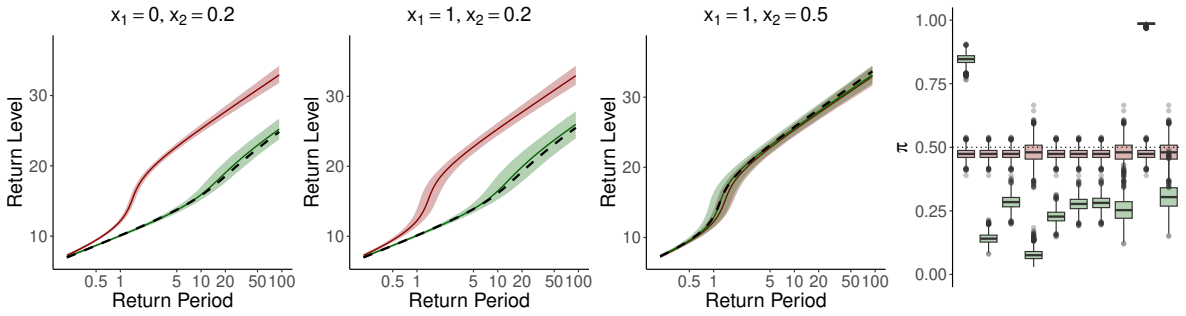


(b) Scenario A.2

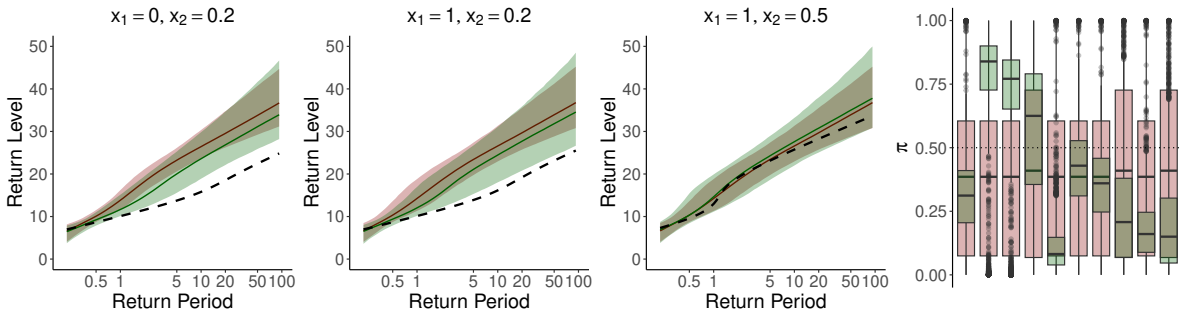
FIGURE 4.3: Comparison between true allocations (different shapes) and allocations estimated using the logistic weight mixture (different colours), based on one-sample experiments. Left: $n = 1000$; right: $n = 50$.

estimate a distribution believed to characterise heterogeneous maxima, rather than to cluster the observations. The focus is thus on return levels, while the capacity of the model to allocate observations to components is considered an additional information provided by the model.

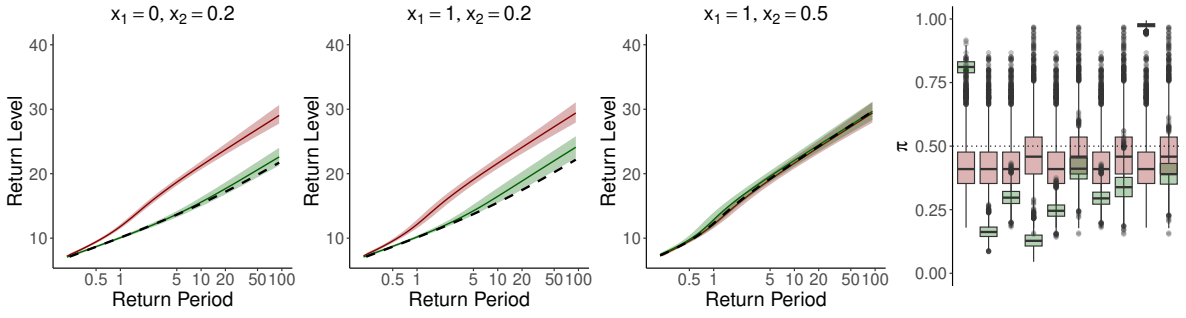
In Scenarios B and C the model (4.1) with x_1 as only covariate and the same model with x_1 and x_2 as covariates are fitted to the simulated data. Return level plots based on the predictive distributions of the models are shown in Figure 4.4. Results for an additional scenario can be found in Appendix A.1. Since x_1 has little influence on the distribution of the mixture weights, changing its value has minimal impact on the results. In contrast, varying x_2 leads to much bigger changes. The model that incorporates both x_1 and x_2 captures the return levels reasonably well when $n = 1000$, despite a general



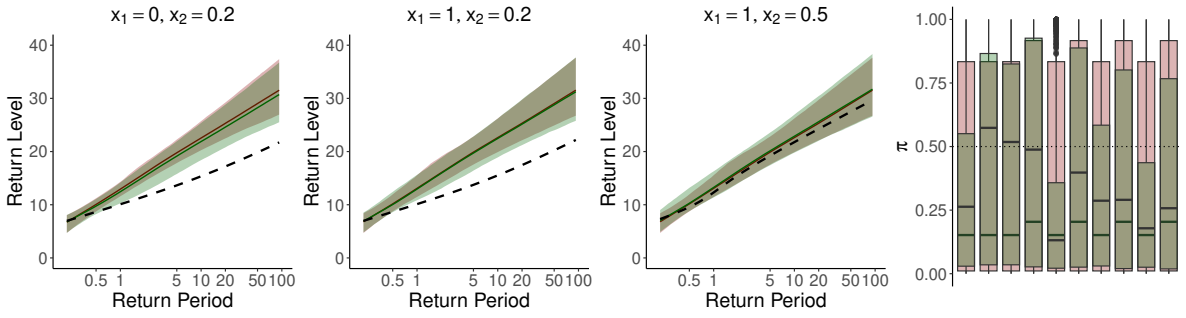
(a) Scenario B, $n = 1000$



(b) Scenario B, $n = 50$



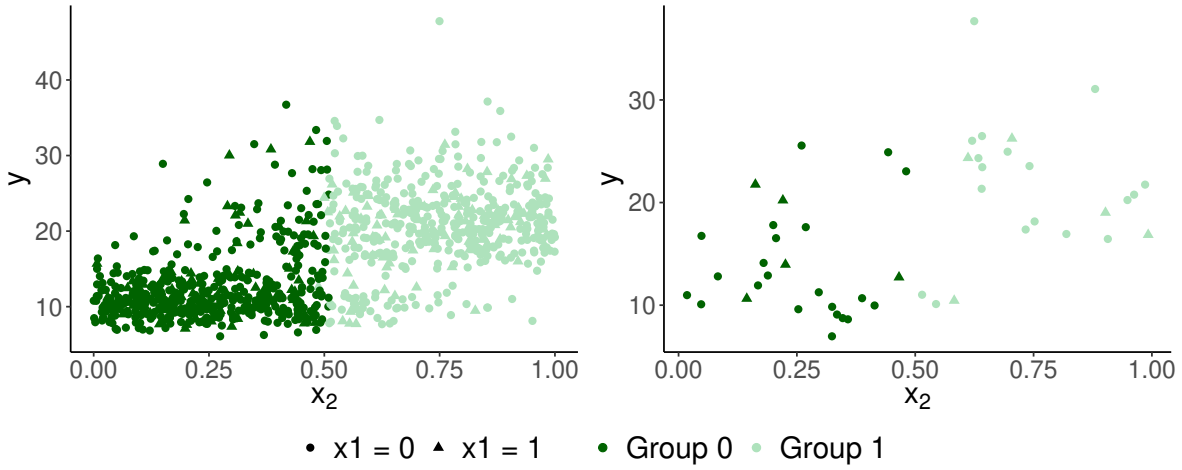
(c) Scenario C, $n = 1000$



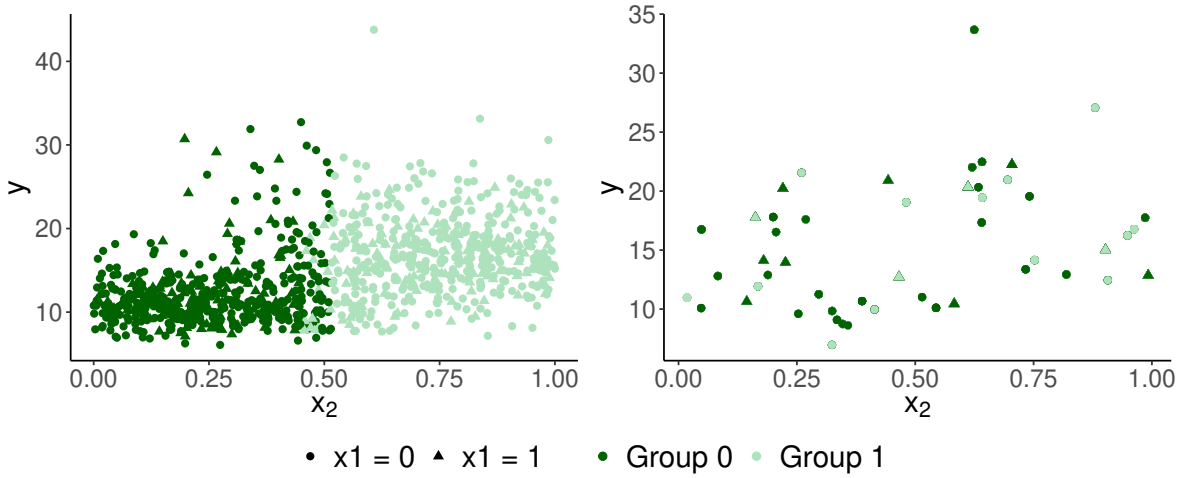
(d) Scenario C, $n = 50$

True Label–based mixture Covariate–based mixture

FIGURE 4.4: One-sample experiments for the scenarios B and C. Top: median posterior return level with posterior credible interval (shaded) for the label-based model (red) and the covariate-based model (green), compared to the true return levels (black, dashed) for different values of the covariates. Box-plots of the posterior distribution of a sample of π_i are also displayed.



(a) Scenario B



(b) Scenario C

FIGURE 4.5: Classification results for Scenarios B and C using the covariate-based model, with groups indicated by different colours, compared against the classifications determined by the labels (different shapes), based on one-sample experiment outcomes. Left: $n = 1000$; right: $n = 50$.

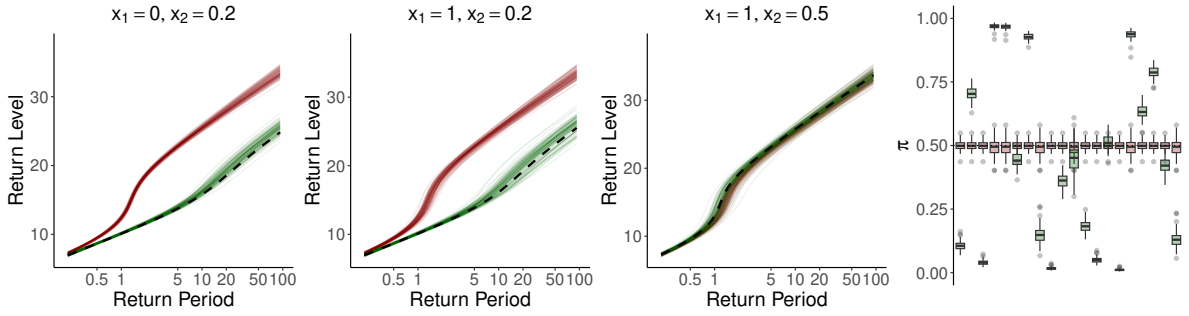
overestimation when x_2 is far from 0.5. However, its performance declines with smaller sample sizes. On the other hand, the model that only accounts for the label always struggles to describe the tail behaviour, except in the average case with $x_2 = 0.5$, where both models perform well. Both models entail considerably more uncertainty when the sample size is small, especially the misspecified one. Similarly to Scenarios A.1 and A.2, we evaluate how each model assigns observations to mixture components by examining the posterior distribution of a sample of estimates of π_i. We also compare the groups based on the logistic weights with those identified by the label, as shown in Figure 4.5. As expected, the groups produced by the model incorporating x_2 do not align well with the label-based ones. By construction different values of x_2 lead to significant differences

in the weights, as it is possible to see from the green box-plots in Figure 4.4. This results in the identification of two distinct groups, even when the location parameters of these groups are not very different, if the sample size is high enough. There is no actual allocation in two groups when relying solely on x_1, since the posterior distribution of the weights from this model is centred around 0.5. When the sample size is low, the posterior distribution of the weights from both models is much more spread, especially in the scenario with more similar location parameters, where the allocation to mixture components is uncertain even for the model that includes x_2.

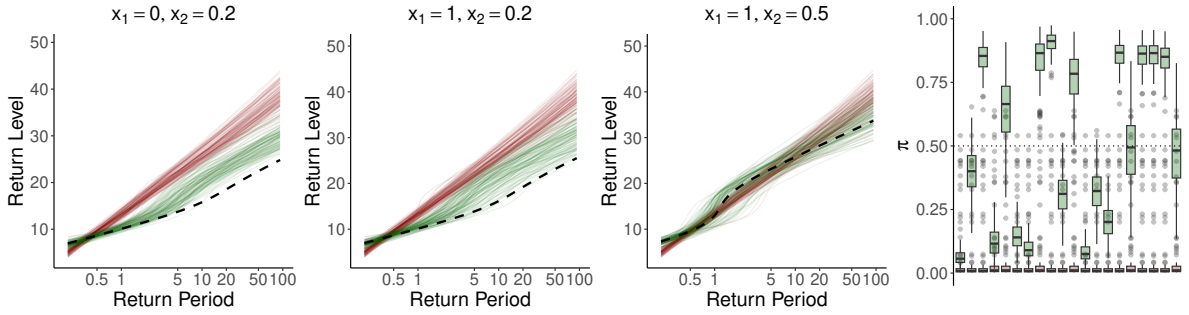
The discussed results are derived from a single simulation experiment. Additionally, we conduct a Monte Carlo simulation study by repeating the experiment for $M = 100$ different datasets. Table 4.4 presents the misclassification rate between the actual allocations and those obtained from various fitted models, averaged across the M datasets. It also illustrates the discrepancies between model-based groupings and those created by the label. Here we employ the missclassification rate as a summary of the goodness of fit of the model based on the logistic regression, even if the model is not aimed at classification. Model 1 corresponds to (4.1) with x_1 as the sole covariate, while Model 2 includes x_2 as an additional continuous covariate. In Scenarios A.1 and A.2, the model based on x_1 is mostly able to recognise the presence of two groups and to estimate correctly the true allocations. When mixing weights are generated mainly based on x_2 (Scenarios B and C), the model that uses only x_1 fails to distinguish between two groups, while the model that incorporates x_2 can identify two groups, even with a small sample size. Although the model with x_2 does not perfectly match the true allocations, it correctly classifies between 70% and 90% of the data. When the sample size is low, misclassification is slightly higher in Scenario C, where the location parameters of the

TABLE 4.4: Results from Monte Carlo simulation study: number of groups identified by the model, missclassification rate with respect to the true allocations, and missclassification rate with respect to the groups created by the label. Averages across the M simulated datasets.

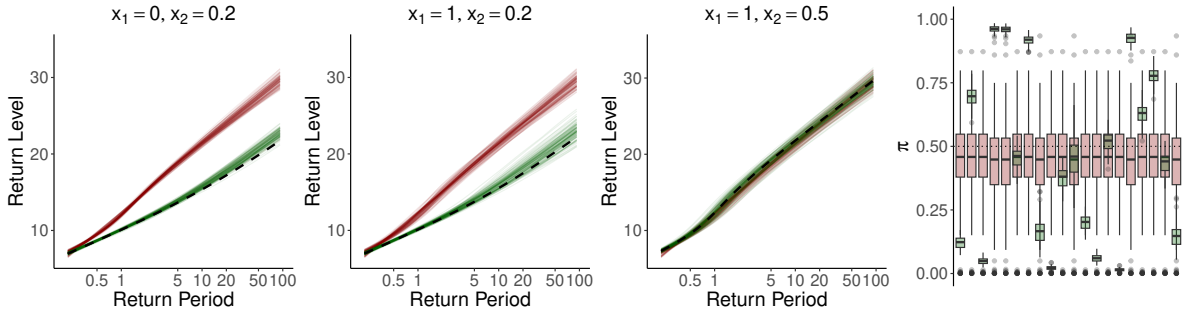
		$n = 1000$				$n = 50$			
Scenario		**A.1**	**A.2**	**B**	**C**	**A.1**	**A.2**	**B**	**C**
Number of groups	Model 1	1.72	1.66	1.35	1.10	2	1.96	1.06	1
	Model 2			2	2			2	1.57
Misscl. rate	Model 1	0.196	0.168	0.498	0.499	0.020	0.101	0.474	0.477
	Model 2			0.144	0.145			0.122	0.289
Misscl. rate (w.r.t. label)	Model 1	0.196	0.129	0.319	0.248	0.101	0.009	0.217	0.220
	Model 2			0.505	0.508			0.431	0.334



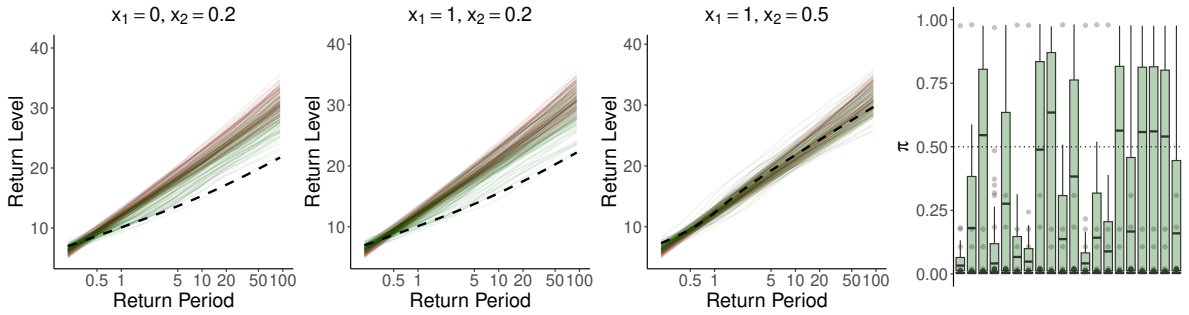
(a) Scenario B, $n = 1000$



(b) Scenario B, $n = 50$



(c) Scenario C, $n = 1000$



(d) Scenario C, $n = 50$

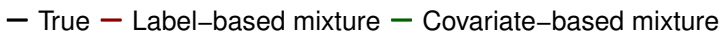


FIGURE 4.6: Monte Carlo simulation for Scenarios B and C. Top: median posterior return levels for the label-based model (red) and the covariate-based model (green), compared to the true return levels (black, dashed) for different values of the covariates. Box-plots of the posterior median of a sample of π_i are also shown.

two Gumbel distributions are closer. As anticipated, the groups based on the model with both x_1 and x_2 differ significantly from those imposed by the label. The model based solely on x_1 also does not reproduce this division.

For a visual inspection, Figure 4.6 presents the median return levels across 100 datasets, along with box-plots of the posterior median of a sample of mixture weights. These results are shown for Scenarios B and C, where the models exhibit greater numerical stability compared to Scenarios A.1 and A.2, which face challenges arising from issues related to data separation. The findings for Scenarios B and C align closely with those observed in the single-experiment analysis, reinforcing the earlier conclusions. Additionally, the Monte Carlo distribution of the median mixture weights supports the insights derived from Table 4.4, particularly regarding the number of components identified by the models.

4.4 Proper scoring rules for distinguishing extreme-value distributions

In any real application, once more than one model is fitted to a sample of data, the question arises of finding a way to compare the goodness of fit of the different models. Among the several metrics that have been proposed, we focus here on investing how differences in proper scoring rules can be used for comparison for extreme value models.

Specifically, since we wish to not depend on a specific sample, we use the expected scoring rules:

$$S(F) = \int \mathrm{S}(F, y) f(y) \mathrm{d}y,$$

where $\mathrm{S}(F, y)$ is a proper scoring rule based on a forecast (or model) F and on an observation y, as defined in Section 1.2.2.3, and f is the density function corresponding to F. The scoring rules we consider are logarithmic score (LogS), continuous ranked probability score (CRPS), conditional likelihood (CL) score, censored likelihood (CSL) score and weighted continuous ranked probability score (wCRPS). The latter three scores depend on a weight function w. We focus on three options: (i) $w(z) = I(z \geq t)$, that is, truncation below a certain threshold t; (ii) $w(z) = \Phi(z \mid t, s^2)$, where $\Phi(\cdot \mid a, b^2)$ denotes the distribution function of a Normal distribution with mean a and variance b^2; (iii) $w(z) = G(z \mid t, s^2, 0)$, where G is the distribution function of a GEV distribution as defined in (1.1), which in this case is a Gumbel distribution. The first choice is the most practical, but restricting to the interval $[t, +\infty)$ can be misleading. The second choice has been proposed in Amisano and Giacomini (2007) and Gneiting and Ranjan

TABLE 4.5: Scenarios for the analysis of proper scoring rules. If a value of a parameter is not specified, it means that the parameter is allowed to take multiple values within its support.

Scenario	**Distribution** A	**Distribution** B	**Distribution** C
1	Gumbel$(\mu_A = 1, \sigma_A = 1)$	Gumbel(μ_B, σ_B)	
2.1	Gumbel$(\mu_A = 20, \sigma_A = 2)$	Gumbel$(\sigma_B = 30, \sigma_B = 3.5)$	$(1-\pi)A + \pi B$
2.2	Gumbel$(\mu_A = 20, \sigma_A = 2)$	Gumbel$(\mu_B = 22, \sigma_B = 3)$	$(1-\pi)A + \pi B$

(2011), and ensures that all scores are well defined. We explore the third choice as we are working on Gumbel distributions.

We examine some scenarios of interest where we compare two models using proper scoring rules, aiming to understand how these scoring rules perform in each case. First, we consider two Gumbel distributions with different location and scale parameters to determine whether expected scoring rules can effectively distinguish between them. Then, we compare a Gumbel distribution to a mixture of itself and another Gumbel distribution, using proper scoring rules to quantify the difference between the mixture and the individual Gumbel distribution as the mixing weight changes. We assess this for both a scenario where the two Gumbel distributions are very different and one where they are more similar. These scenarios are summarised in Table 4.5. In all of them, we set t in the weighted scoring rules to the 75th percentile of the first Gumbel distribution (distribution A) and choose large values for s, ensuring the weight function spans most of this distribution.

Figure 4.7 shows how the difference in expected proper scoring rules between two Gumbel distributions A and B changes with different choices of the parameters of B (Scenario 1). Interestingly, the unweighted scoring rules yield the same values for different choices of the location parameter μ_B, thus capturing only differences in scale between the two distributions. Indeed, the absolute difference in LogS and CRPS is zero not only when A and B are identical (i.e., $\mu_A = \mu_B = 1$ and $\sigma_A = \sigma_B = 1$), but also when they share the same scale, regardless of differences in location. When looking at weighted scoring rules, expected CL and CSL scores better quantify the differences between the two models compared to weighted CRPS. Again, while the scores are zero when A and B are identical, they can also be zero, or nearly zero, when the distributions differ. For example, these scores fail to distinguish between a Gumbel distribution with location and scale both equal to 1 and another with a lower location parameter if the scale is large enough, or a Gumbel distribution with a higher location and a smaller scale.

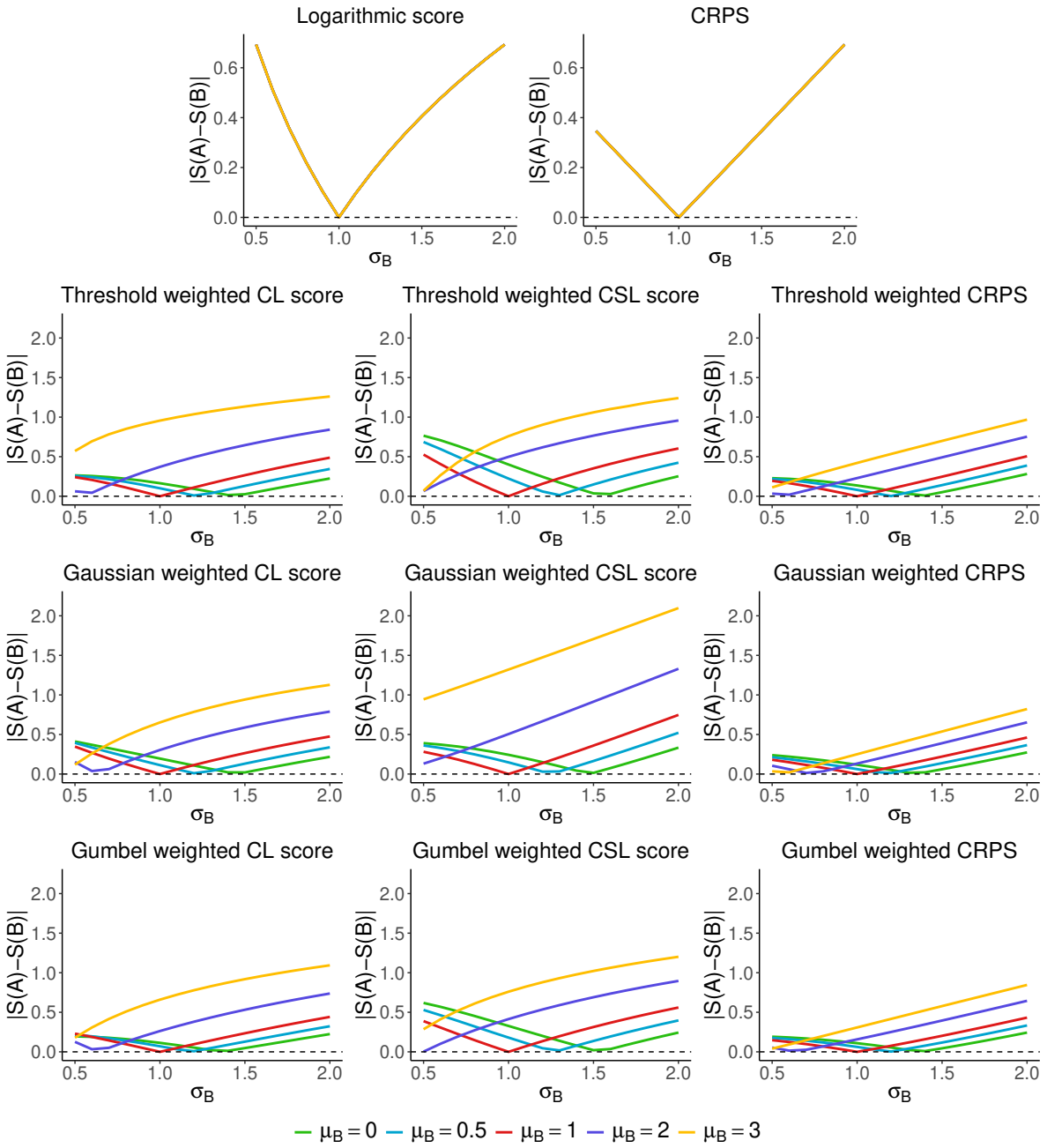


FIGURE 4.7: Scenario 1. Absolute difference in expected proper scoring rules between two Gumbel distributions A and B, as the location and scale parameters of B vary.

Figure 4.8 displays the absolute difference in expected proper scoring rules between a mixture distribution C of two Gumbel distributions A and B and one of the individual distribution, specifically A (Scenarios 2.1 and 2.2). The difference is much higher in the scenario when the parameters of A and B are closer, i.e., Scenario 2.2, which is reassuring. However, according to LogS and CRPS, the difference between A and B –which is the difference between A and C when $\pi = 1$– remains nearly the same across both scenarios. This aligns with the earlier observation that these two scores primarily

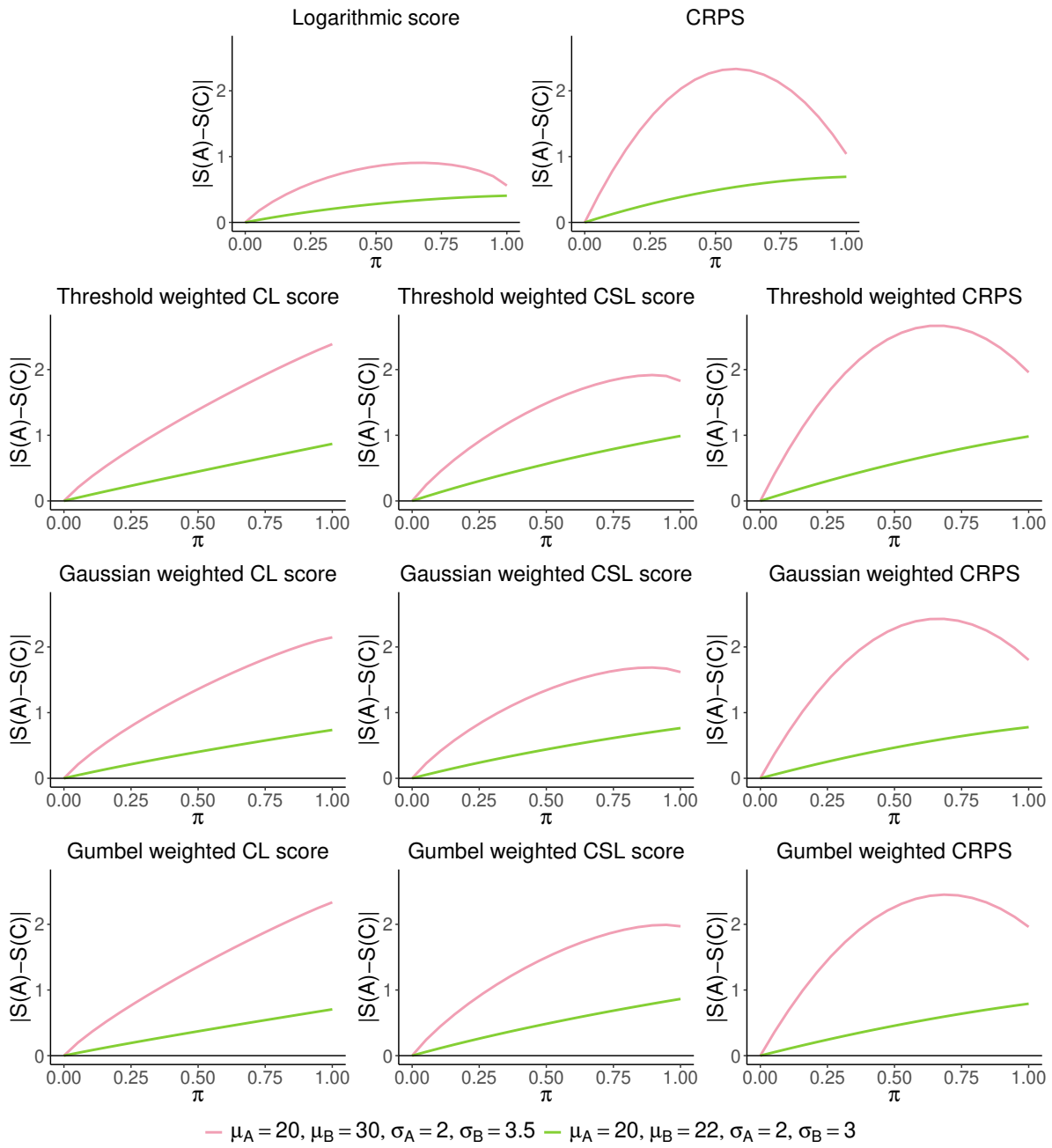


FIGURE 4.8: Absolute difference in expected proper scoring rules between a Gumbel distribution A and mixture C of A and another Gumbel distribution B, as the weight of the mixture varies. In Scenario 2.1 the distributions A and B are well separated (pink), while in Scenario 2.2 the parameters are more similar (green).

reflect differences in the scale parameters, rather than in the location. When A and B have similar parameter values, the difference between the mixture and the individual distribution increases steadily with π across all scores. However, in the scenario where the parameters of A and B differ significantly, this trend is only observed with the CL score. For all other scores, there is a value of π between 0.5 and 1 (not included) where the difference is greatest. This suggests that a Gumbel distribution can be measured to

be more distinct from a mixture of itself and another Gumbel distribution than from the other Gumbel distribution alone if its mixing proportion is low.

With a better understanding of how to interpret proper scoring rules for model comparison, we move on to an application to rainfall data, where we will apply these scoring rules to evaluate and compare the fitted models.

4.5 Application to ERA5 rainfall data in Venice

We are interested on the case study based on ERA5 reanalysis data about climate and weather in the area of Venice from Climate Data Store (Copernicus Climate Change Service, 2023). Reanalysis data, which are often used in environmental applications, are based on advanced assimilation techniques (e.g., Dee *et al.*, 2011) which combine observations with model forecasts to generate a more accurate and complete picture of atmospheric conditions. This helps to fill gaps in observational data and improve the temporal and spatial resolution of weather and climate data. The data of interest consist of a series of $n = 83$ annual maximum values of total hourly precipitation. These maxima can be labelled according to their precipitation generation scheme, which can be convective or large-scale. This information is available as they are reanalysis data rather than pure real-world observations. In most cases, however, it is challenging to obtain this classification from actual data. The two types of precipitation are quite different: convective rainfalls have usually very big magnitudes in a short temporal window, while large-scale precipitations are less intense but can last longer. From the data store it is possible to retrieve a set of additional variables about weather conditions; we select wind characteristics, dew point temperature, cloud coverage, pressure measure and convective available potential energy, as they could affect the precipitation scheme. The dataset consists of multiple simulations (specifically, 10 separate runs or "ensemble members") of the same model, performed at every three-hour interval. This 10-member ensemble, used in the data assimilation procedure, offers an estimate of the uncertainty in the outputs of the climate model. For simplicity, the data we use is derived taking only one ensemble member (namely the first). Figure 4.9 shows the distribution of the data, i.e. the precipitation maxima and, since this can be directly derived from the ERA5 output, the information on the type of precipitation that originated each observation: convective precipitation (46% of the maximum events) or large-scale precipitation (54%).
We fit a number of different extreme value models to the sample. All models are fitted using `Stan`, with 4 Markov chains of 2000 iterations each, half of which are considered as

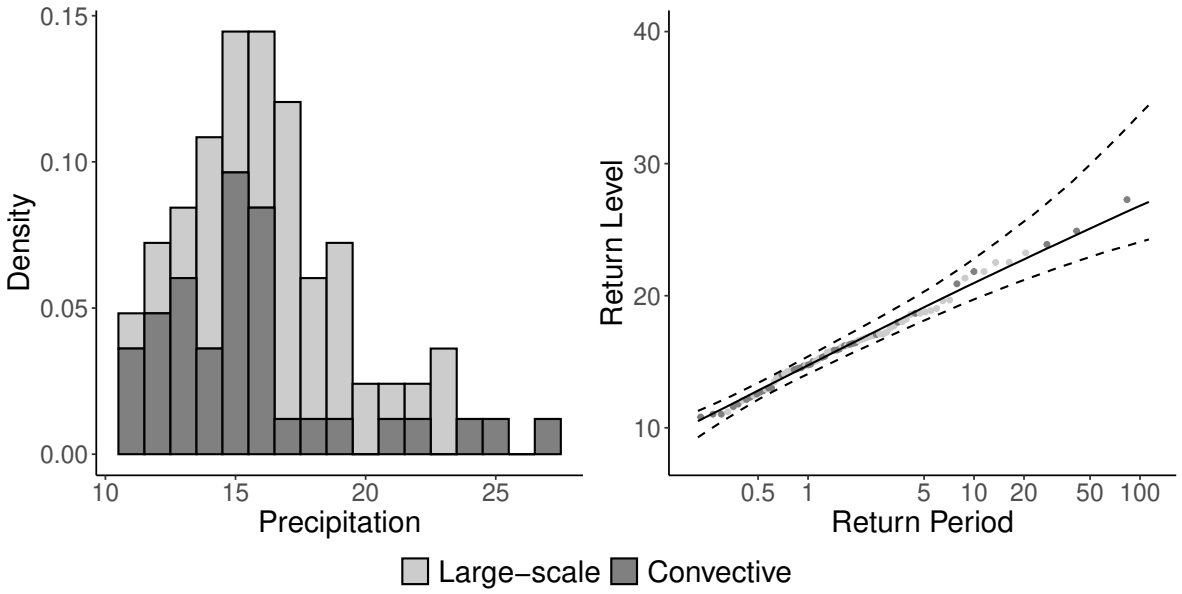


FIGURE 4.9: Total precipitation (annual maxima) in Venice. Left: histogram by precipitation type; right: return level plot for a fitted GEV distribution with 95% credible interval.

warm up and discarded. The posterior distribution of the fitted models is summarised in Table 4.6. We begin by fitting a single GEV distribution to the data (Model 0), with priors

$$\mu \sim \mathrm{N}\left(\bar{y}, [\{\max(y) - \min(y)\}/2]^2\right), \quad \log(\sigma) \sim \mathrm{N}(0, 25), \quad \xi \sim \mathrm{N}(0, 25),$$

where $\bar{y}$ denotes the sample mean and $\boldsymbol{y}$ is the sample of annual maxima. Thus, we use a prior based on the data for the location, and general wide priors for the scale and shape parameters. From Table 4.6, it is possible to notice that the fitted GEV distribution is approximately Gumbel, as the posterior distribution of ξ is concentrated around 0. It is important to note that testing the hypothesis $\xi = 0$ is not a straightforward task. This is due to the fact that the support of the GEV distribution is parameter-dependent, making such tests particularly complex. Consequently, this test is not performed here.

From the return level plot based on the GEV model shown in Figure 4.9, it is possible to recognise that a single GEV distribution that not takes into account the specific physical process originating the data still provides an adequate fit. This raises the question of whether a more complex mixture model, which can be more challenging to estimate, is really necessary.

We proceed by fitting the dependent mixture model (4.1) with different choices of the covariates. In particular, we consider:

TABLE 4.6: Posterior median with 95% credible interval in parenthesis for the parameters of the single GEV model (Model 0) and the two-component mixture models with different choices of the covariates.

Model	μ	σ	ξ
0	14.72 (14.03; 15.43)	2.75 (2.31; 3.37)	-0.02 (-0.16; 0.18)
1	14.14 (11.37; 15.13)	2.60 (1.16; 4.67)	
	16.03 (14.47; 22.72)	2.49 (1.04; 5.77)	
2	13.93 (13.12; 14.76)	2.44 (1.78; 3.11)	
	17.09 (16.06; 18.05)	1.84 (1.16; 2.89)	
3	13.92 (13.08; 14.75)	2.43 (1.79; 3.17)	
	17.11 (15.80; 18.03)	1.86 (1.19; 3.01)	

1. only the binary variable on the precipitation scheme, which we will refer to as "label";

2. some physical and meteorological variables that are supposed to affect the precipitation scheme;

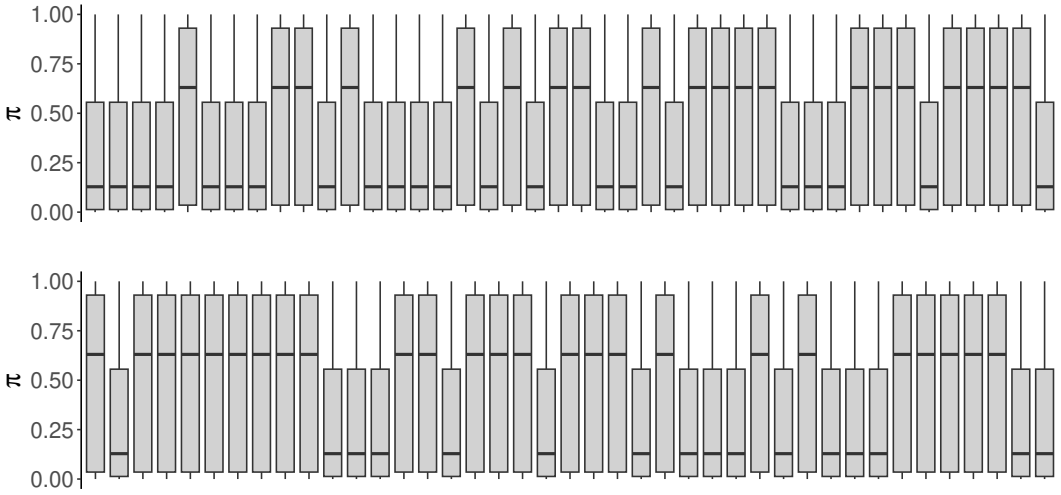
3. the label and these additional variables.

All the continuous covariates are scaled in order to have mean 0 and unit variance, whereas, as suggested by Gelman *et al.* (2008), binary inputs are adjusted so that their average becomes 0 and they are set to have a unit difference between their lower and upper states. The prior distributions of the location and scale parameters are determined using information from the sample (i.e., sample mean and standard deviation), whereas the regression coefficients $\boldsymbol{\beta}$ and intercept α follow a default prior specification for logistic regression (Gelman *et al.*, 1995; Ghosh *et al.*, 2018), which is

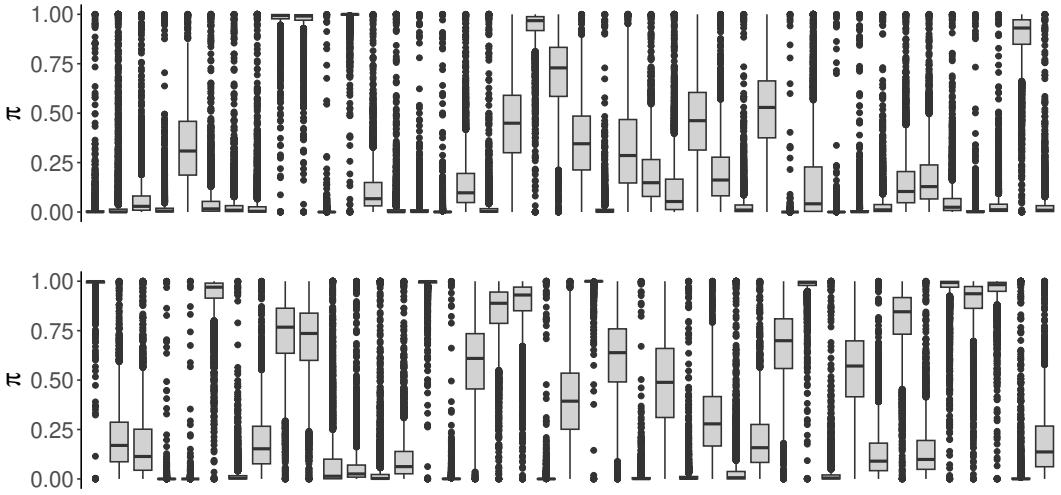
$$\alpha \sim \mathrm{N}(0, 5^2), \quad \boldsymbol{\beta} \overset{\text{iid}}{\sim} \mathrm{N}(0, 2.5^2).$$

We also experimented with Student-t and Cauchy prior distributions, which yielded very similar results.

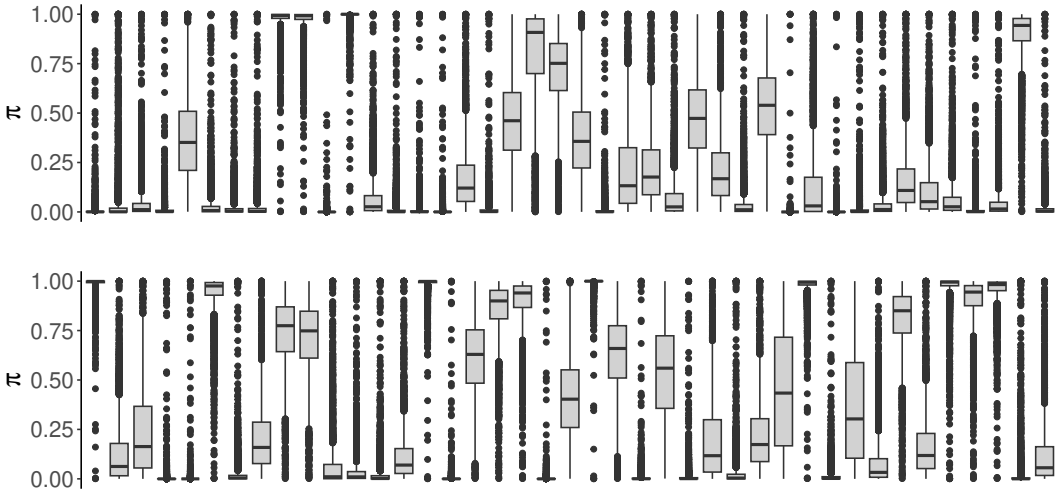
The mixture model which uses just the binary variable of precipitation scheme to inform the allocation (Model 1) does not result in two actual mixture components. Indeed, as displayed in the top panel Figure 4.10, the posterior distribution of π_i, is widely spread across the domain, indicating that this model struggles to accurately assign observations to two distinct components. The inadequacy of this model justifies the inclusion of other variables that can bring useful information in separating the data.



(a) Model with only the indicator of precipitation scheme as a covariate.



(b) Model with temperature anomaly and convective available potential energy as covariates.



(c) Model with indicator of the precipitation scheme, temperature anomaly and convective available potential energy as covariates.

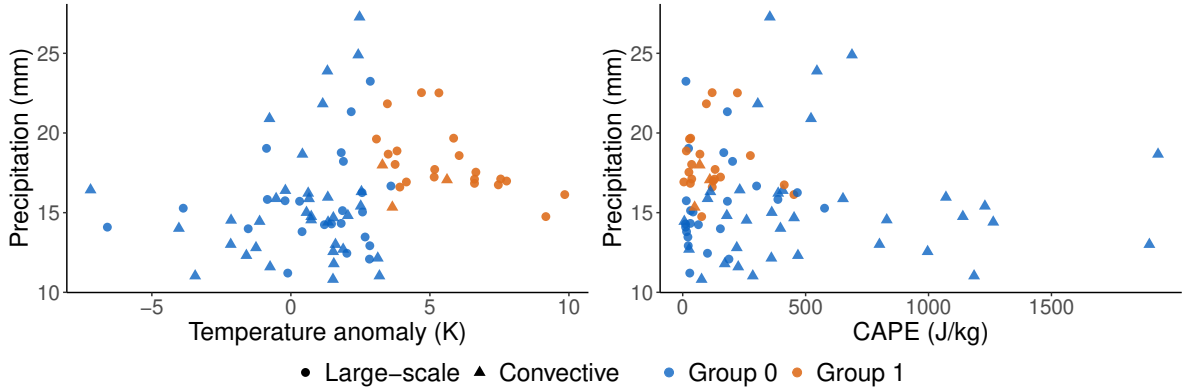
FIGURE 4.10: Boxplots of the posterior distribution of π_i from the fitted mixture models with different choices of covariates, for $i = 1, \ldots, 83$.

TABLE 4.7: Posterior median with 95% credible interval in brackets for the regression coefficients of the two-component mixture models with different choices of the covariates.

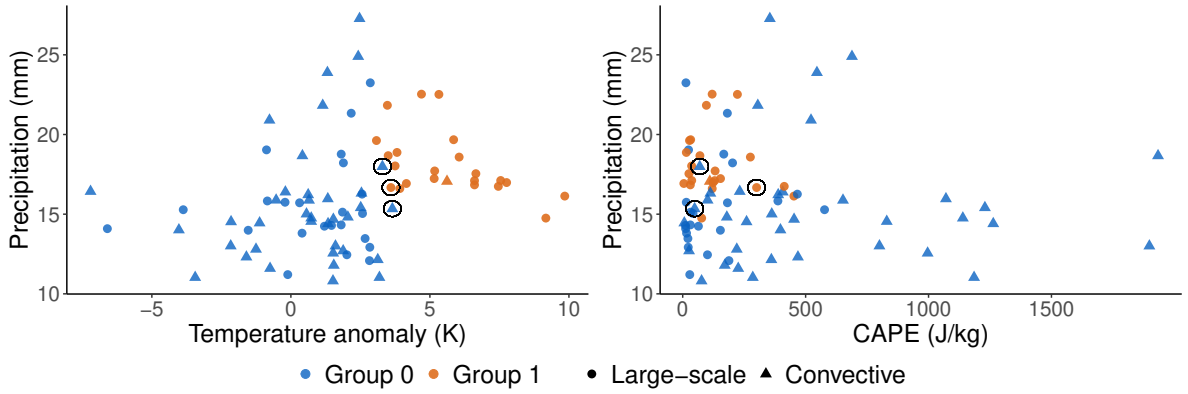
Model	α	β_1 (label)	β_2 (temp. anomaly)	β_3 (CAPE)
1	-0.91 (-9.90; 7.53)	-1.86 (-5.92; 3.12)		
2	-2.52 (-5.62; -0.09)		4.49 (1.78; 7.87)	-1.17 (-4.34; 0.48)
3	-2.65 (-5.86; 0.55)	-1.17 (-5.19; 1.84)	4.37 (1.59; 7.83)	-0.79 (-3.90; 1.38)

We fit a model using two covariates relevant for distinguishing between the two precipitation regimes (Model 2): anomaly from average dew point temperature and convective available potential energy (CAPE). The motivation is that the binary label may impose an overly rigid classification, whereas these variables could help define groups that remain physically understandable. The posterior distribution of the weights π_i is displayed in the central panel Figure 4.10. Unlike previously, this time the data suggest the presence of two distinct mixture components. From Table 4.6 it emerges that the location parameters μ_0 and μ_1 are very different and their posterior distributions do not overlap, which confirms the presence of two different mixture components. The scale parameters are instead quite similar and a specification of model (3.2) with a common σ may be a sensible choice. Interestingly, the groups formed based on whether π_i is more often below or equal to 0.5 (group 0) or greater than 0.5 (group 1) do not align with the label-based division into large-scale and convective precipitation. This discrepancy is clearly illustrated in Figure 4.11, which compares the allocations imposed by the label (different shapes) with the ones suggested by the model (different colours). This suggests that the binary classification into large-scale and convective precipitation does not explain the presence of two distinct mixture components in the tail. Instead, other factors play a role in creating a more meaningful allocation scheme. As shown in Figure 4.11, and supported by the estimates of the regression coefficients in Table 4.7, the temperature anomaly appears to be the primary driver behind the groups.

We combine the two previous models with a specification that incorporates the two meteorological features alongside the label of the physical precipitation process (Model 3). The posterior distribution of π_i, depicted in the bottom panel of Figure 4.10, is practically unchanged from the previous model. This confirms that the label does not really play a role, and it is also evident in the credible interval of the coefficient of the label (β_1) containing 0 (see Table 4.7). Thus the model is over-parametrised. Moreover, when the label is introduced CAPE becomes much less influential due to the strong association between the two variables. The two groups based on this model are the



(a) Model with temperature anomaly and CAPE.



(b) Model with label, temperature anomaly and CAPE.

FIGURE 4.11: Allocation based on the labels (different shapes) and model-based allocation (different colours). Values that change group when including the label in the model are highlighted.

same of the ones created without the label, with the exception of three observations, as shown in Figure 4.11.

We aim to compare the fitted models, acknowledging the challenges of model comparison in extreme value theory. Among the set of proper scoring rules explored in Section 4.4, we employ the logarithmic score and CRPS. Additionally, we use the leave-one-out

TABLE 4.8: Posterior measures for model comparison, with standard error in parenthesis: average expected logarithmic score, average expected CRPS, and LOO information criterion.

Model	LogS	CRPS	LOO-IC
0	2.576 (0.016)	1.852 (0.021)	430.3 (13.8)
1	2.534 (0.0.22)	1.822 (0.032)	427.8 (13.9)
2	**2.449** (0.027)	**1.693** (0.043)	**412.6** (16.0)
3	2.452 (0.029)	1.694 (0.045)	413.7 (16.0)

information criterion (LOO-IC) defined in (1.14), computed with the R package `loo` (Vehtari *et al.*, 2023). Table 4.8 presents these metrics for model comparison based on the posterior estimates of the fitted models. All diagnostics consistently indicate that the mixture model based on temperature anomaly and CAPE is the preferable option. The small sample size and high data variability complicate model selection, but the consistency across all measures suggests that Model 2 works well. While this appears to be the best modelling structure, there remains potential for improvement, possibly by incorporating other variables and using a variable selection procedure, as temperature is the only truly informative factor.

4.6 Concluding remarks

This chapter is motivated by practical considerations, specifically the insight from practitioners that multiple components may influence individual events and, consequently, annual maxima. Starting from the literature based on this idea, taking into account that the binary information on the type of phenomenon may not drive the two components in the tail, we developed two-component mixture models with covariate-informed weights. The proposed model was applied to precipitation maxima in Venice, a city highly vulnerable to flooding, even if not only from rainfall. While coastal flooding, related to high tides, is the most notorious flooding to which the city of Venice is exposed to, understanding precipitation patterns is important for improving risk assessment in the city and the surrounding area. Rather than relying on a binary variable to categorise rainfall events, the model employed a set of explanatory variables to better capture the mixture behaviour. In particular, the convection available potential energy (CAPE) has been used by climate experts for the analysis of rainfall extremes (e.g., Barbero *et al.*, 2019). This approach resulted in a well-rounded representation, revealing two distinct components. Our proposed method provides practitioners with a flexible tool that eliminates the need to define rigid and time-consuming binary categories, allowing the inclusion of variables deemed relevant for the analysis. However, determining which covariates are most influential for the mixture model is challenging. Enhancing the proposed model with Bayesian techniques for variable selection (e.g., Tadesse and Vannucci, 2021) would improve its utility and represents an important direction for future research. It is however challenging in the context of extreme value theory (e.g., de Carvalho *et al.*, 2022).

In this chapter, we assumed that the distribution of block maxima could be explained by two components, a choice motivated by insights from real-world scenarios

and by existing literature. However, recognising that two components may not always be sufficient, we will relax this assumption in the next chapter (Chapter 5), where a more flexible approach will be introduced, allowing for an unspecified number of components. A more detailed summary and critical discussion of the key points from this chapter will be provided in the concluding chapter.

Chapter 5

Infinite mixture models for heterogeneous extremes

This chapter presents our research on infinite mixture models for extreme values, detailing the proposed model, posterior computation methods, simulation studies, and an application to precipitation data.

5.1 Introduction

In the block maxima approach for extreme value theory, introduced in Section 1.1.1, extreme values are assumed to be approximately independent and identically distributed, and are modelled using extreme value distributions, such as the generalised extreme value (GEV) distribution. As already discussed in Chapter 4, in practical situations observed maximum values may not follow a single parametric (e.g., GEV) distribution; instead, they present a grouped structure. In other words, similar to many statistical problems, it is believed that there is some degree of heterogeneity in the data, which can be assessed using mixture models, presented in Chapter 2. Extending the research presented in Chapter 4, we assume that there is an infinite number of mixture components determining the distribution the overall population. As explained in Section 2.2, infinite mixture models offer a compelling alternative to mixtures with fixed number of components by allowing an unbounded number of components, making them particularly suitable for scenarios where the complexity of the data structure is challenging to characterise with a fixed number of components. In the scenario of interest the use of infinite mixture model is also motivated by the assumption that each block of units,

characterised by a maximum value, belongs to a certain component, and using an infinite number of components allows to account for all the diversity across blocks. In particular, we rely on Dirichlet process mixtures (see Section 2.2.2).

The assumption of this chapter is different from Chapter 4, as it is not assumed that the original sequence of variables, from which block maxima are extracted, follows a mixture distribution. Assuming a mixture distribution for the entire data does not necessarily imply that the tails are described by a mixture distribution, and frequently the characteristics responsible for grouping in the entire dataset are not useful for fitting a mixture in the right tail, as discussed in Chapter 4. The aim of this part of the dissertation is instead to model the behaviour of heterogeneous maximum values by using mixtures of an infinite number of GEV distributions. As illustrated in Chapter 3, modelling extreme values using infinite mixtures has been explored in the literature (Kottas and Sansó, 2007; Tressou, 2008; Palacios Ramirez *et al.*, 2024, to appear), but it remains a relatively unexplored area, especially when dealing with block maxima data. Alternatively, Bottolo *et al.* (2003) proposed a finite mixture model with a non-fixed number of components, focusing however on modelling exceedances over thresholds.

5.2 Infinite mixture models for extremes

5.2.1 Heterogeneous extremes

Let $M_m = \max\{W_1, \dots, W_m\}$ be the sample maximum of a random sample $W_1, \dots, W_m$. The extremal types theorem 1.1 states that as $m \to \infty$ the distribution of the normalised maximum $(M_m - b_m)/a_m$—where $a_m > 0$ and b_m are appropriate normalising sequences—converges to a generalised extreme value (GEV) distribution, as defined in (1.1). Our starting point for modelling is the following domain-extended GEV distribution function that covers all real numbers, while still being coherent with the definition in (1.1):

$$G(y) = \begin{cases} 0, & y \leq \inf \mathbb{S}, \\ \mathbb{G}(y), & y \in \mathbb{S}, \\ 1, & y \geq \sup \mathbb{S}, \end{cases} \tag{5.1}$$

where $\mathbb{G}$ is the GEV distribution function defined in (1.1) and $\mathbb{S} = \{y : 1+\xi(y-\mu)/\sigma > 0\}$ is its support. Equation (5.1) aligns with most numerical implementations of the GEV distribution function in mainstream extreme value packages (e.g., Stephenson,

2002), which return 0 for $y \leq \inf \mathbb{S}$ and 1 for $y \geq \sup \mathbb{S}$. For our purposes, the extended-domain GEV distribution function is convenient for the formal definition of the proposed mixture model.

Assume for every sample a latent indicator Z in $\{0, \ldots, K\}$ identifying the group to which the sample belongs. Hence, a sample can be characterised by the value of Z: $W_{h,1}, \ldots, W_{h,m}$ is a random sample of size m belonging to group h, since $Z = h$. It follows that $M_{h,m} = \max\{W_{h,1}, \ldots, W_{h,m}\}$ is the maximum of a sample belonging to group h. Any maximum can thus be characterised by the value of Z corresponding to its originating sample:

$$M_m = \sum_{h=1}^{K} I(Z = h) M_{h,m},$$

where $I(\cdot)$ is the indicator function. To allow for different domains of attractions in different groups, we define the normalising sequences of random variables $A_m > 0$ and B_m to be group-dependent as follows:

$$A_m = \sum_{h=1}^{K} I(Z = h) a_{h,m}, \quad B_m = \sum_{h=1}^{K} I(Z = h) b_{h,m},$$

where $a_{h,m} > 0$ and $b_{h,m}$ are the normalising sequences for $M_{h,m}$. Then, by the law of total probability and the extremal types theorem

$$\begin{aligned} \mathbb{P}\left(\frac{M_m - B_m}{A_m} \leq y\right) &= \sum_{h=1}^{K} \mathbb{P}(Z = h) \mathbb{P}\left(\frac{M_m - B_m}{A_m} \leq y \mid Z = h\right) \\ &= \sum_{h=1}^{K} \pi_h \mathbb{P}\left(\frac{M_{h,m} - b_{h,m}}{a_{h,m}} \leq y\right) \xrightarrow[m \to \infty]{} \sum_{h=1}^{K} \pi_h G(y \mid \mu_h, \sigma_h, \xi_h). \end{aligned} \tag{5.2}$$

In simpler terms, when the samples, from which the maxima are derived, are organised into groups identified by a latent indicator taking infinite integers values, the distribution of the maximum of any sample of observations converges to a mixture of GEV distributions. There is a trade-off between number of mixture components and size of the samples from which maxima are computed: allowing for many groups permits to find similar behaviours across samples, but at the same time m needs to be large enough for the approximation based on (5.2) to be valid. When K is random, the limit in (5.2) has to be interpreted in the sense of almost sure convergence.

We now introduce a complementary formulation that interprets the previous one as describing heterogeneous extremes. While mathematically equivalent to the original,

this perspective may result more intuitive in the context of block maxima. Suppose we have a series of data, which we divide into n blocks of fixed size m, and we assume that within each block the distribution of the data is the same, but it is allowed to change between blocks. We also assume that the blocks could be grouped based on their distribution. Consider now for each block a latent variable Z_i, that takes values in $\{1, \dots, K\}$ and identifies the group to which the block belongs, for $i = 1, \dots, n$. Let $\boldsymbol{W}_i = (W_{i,1}, \dots, W_{i,m})'$ be the ith block, which is a random sample of size m, for $i = 1, \dots, n$. The distribution of the ith block given the latent variable is

$$\boldsymbol{W}_i = (W_{i,1}, \dots, W_{i,m})' \mid Z_i = h \sim (F_h)^m,$$

where F_h is the distribution of a single $W_{i,j}$ given that $Z_i = h$, for $j \in \{1, \dots, m\}$. It follows that the distribution of a block $\boldsymbol{W}_i$ is

$$\boldsymbol{W}_i \overset{\text{iid}}{\sim} \sum_{h=1}^{K} \pi_h (F_h)^m, \quad i = 1, \dots, n,$$

where π_h is the probability that $Z_i = h$. This mixture model results in a mixture model for the maxima extracted from each block, which we are interest on modelling. It extends the previous formulation by allowing different components to belong to different domains of attraction. Therefore, there is no need to define a single domain of attraction for the overall mixture distribution. By focusing on the maximum of any of these blocks, it is possible to derive the result in (5.2).

5.2.2 Bayesian modelling of grouped block maxima

Following the results of the previous sections, we propose the following approach for modelling extremes from a series of observations originating from distinct random samples (blocks) organised into groups. The data are indeed divided into blocks of size m, where m is sufficiently large for asymptotic approximations to hold. In practical applications, these blocks often correspond to specific time periods (e.g., years, months, weeks). We model the block maxima using mixtures. Indeed, Section 5.2.1 implies that a mixture model based on GEV-distributed components is a sensible choice to approximate the distribution of grouped block maxima.

To model the GEV mixture model in (5.2) in a Bayesian nonparametric fashion (Section 2.2), we first rewrite it as

$$F(y) = \int_{\Theta} G(y \mid \mu, \sigma, \xi) H(\mathrm{d}\mu, \mathrm{d}\sigma, \mathrm{d}\xi), \tag{5.3}$$

where $(\mu, \sigma, \xi) \in \Theta$ and H is an almost surely discrete random probability measure. A Bayesian treatment of heterogeneous extremes entails setting a prior for the mixing measure H. As discussed in Section 2.1, the choice of the number of mixture components is non-trivial and it has been highly investigated in the literature. We decide to use an infinite mixture model, which allows for an infinite number of components[1]. Many options are available for the prior specification on H, including the stick-breaking representation (2.9). It follows that H follows a Dirichlet process (Ferguson, 1973, 1974) with precision parameter $\alpha > 0$ and baseline measure H_0. Thus, our model is a Dirichlet process mixture (Escobar and West, 1995) with a GEV kernel. For an overview on Dirichlet process and Dirichlet process mixtures refer to Sections 2.2.1 and 2.2.2. The proposed model is

$$\begin{gathered} f(y \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{h=1}^{\infty} \pi_h g(y \mid \boldsymbol{\theta}_h), \\ \boldsymbol{\theta}_h \mid H_0 \overset{\text{iid}}{\sim} H_0, \quad \pi_h = V_h \prod_{j<h} (1 - V_j), \\ V_h \mid \alpha \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha), \end{gathered} \tag{5.4}$$

where g is the density of the extended-domain GEV distribution (5.1), $\boldsymbol{\theta}_h = (\mu_h, \sigma_h, \xi_h)'$, with possible additional prior $\alpha \sim \text{Gamma}(a_\alpha, b_\alpha)$.

5.3 Gibbs sampler for infinite mixture models

The parameters of model (5.4) can be fitted using a blocked Gibbs sampler with fixed truncation (Ishwaran and James, 2001), mentioned in Section 2.2. This algorithm truncates the number of components in (2.9) to a sufficiently high value K, so that the infinite mixture model in (5.2) is approximated as

$$\hat{F}(y) = \sum_{h=1}^{K} \pi_h G(y \mid \mu_h, \sigma_h, \xi_h). \tag{5.5}$$

Typical choices of this upper bound are $K = 50$ or $K = 100$, but other values can be adopted to satisfy a certain criterion. It is worth clarifying that truncating is not equivalent to fitting a standard finite mixture model, as K does not represent the actual number of components occupied by the observations, but rather serves as an upper bound on the possible number of components. Nonetheless, the model fitted in this way is not truly

[1]The result in (5.2) is valid when $K = \infty$, relying on Tannery's lemma (Tannery, 1910). This is supported by the bounded nature of $\pi_h G(y \mid \mu_h, \sigma_h, \xi_h) \leq \pi_h$ and the convergence of the infinite sum of the weights in the stick-breaking process. For brevity and to keep the discussion clear and focused on the main interest of this chapter, we do not delve into the detailed technical derivations here.

infinite-dimensional but represents a sparse, finite approximation (Frühwirth-Schnatter *et al.*, 2021). Other algorithms, which are mentioned in Sections 2.1.4 and 2.2.2, can be used for posterior computation of Dirichlet process mixture models without resorting to such approximation. However, the algorithm of Ishwaran and James (2001) is used here for its relative simplicity.

The main challenge of fitting model (5.4) is that it is not possible to simulate directly from the full conditional distributions of μ_h, σ_h and ξ_h, since no conjugate forms are available for the parameters of a GEV distribution. Therefore, we combine Gibbs sampling and Metropolis—Hastings schemes, using a single step of Metropolis algorithm to jointly simulate the current value of $\boldsymbol{\theta}_h$, for $h = 1, \dots, K$. In particular, at each step of the Gibbs sampler we make a random walk proposal, which is accepted according to the Metropolis acceptance rate. The adaptive method for the automatic scaling of the proposal distribution of Garthwaite *et al.* (2016) is exploited.

Let the data be the series of maxima $\boldsymbol{y} = (y_1, \dots, y_n)'$. The algorithm starts from some initial values for π_h, $\boldsymbol{\theta}_h$, and for the proposal covariance matrix $\boldsymbol{S}_h$, for every h. Repeat the following steps for $t = 1, \dots, T$, from which B burn-in iterations are then discarded.

a) New allocation indicators $\boldsymbol{z} = (z_1, \dots, z_n)'$ are sampled based on the posterior probabilities conditional on the data:

$$\mathbb{P}(Z_i = j \mid y_i) = \frac{\pi_j g\left(y_i \mid \boldsymbol{\theta}_j\right)}{\sum_{h=1}^{K} \pi_h g\left(y_i \mid \boldsymbol{\theta}_h\right)},$$

where $g(\cdot \mid \boldsymbol{\theta})$ is the probability density function of a GEV distribution with parameters $\boldsymbol{\theta} = (\mu, \sigma, \xi)$.

b) Once obtained the $\boldsymbol{z}$, the counts $n_1, \dots, n_K$ of units in every components can be computed and the stick-breaking variables are updated as

$$V_h \sim \text{Beta}\left(1 + n_h, \alpha + \sum_{j>h}^{K} n_j\right), \quad h = 1, \dots, K-1.$$

Then $\pi_h = V_h \prod_{j<h} (1 - V_j)$.

c) Define a component $\boldsymbol{y}_h$ by all the observations y_i such that $z_i = h$, for every h. For each $\boldsymbol{y}_h$ update the GEV parameter vector $\boldsymbol{\theta}_h$ via a step of adaptive Metropolis-Hasting scheme:

1. Generate from random-walk proposal distribution:

$$(\mu_h^*, \phi_h^*, \xi_h^*)' \mid (\mu_h, \phi_h, \xi_h)' \sim \mathrm{N}\left((\mu_h, \phi_h, \xi_h)', \boldsymbol{S}_h\right),$$

   where $\phi_h = \log(\sigma_h)$.

2. Compute the acceptance probability

$$p_{\mathrm{ACC},h} = \min\left\{1, \frac{p(\mu_h^*, \phi_h^*, \xi_h^* \mid \boldsymbol{y})}{p(\mu_h, \phi_h, \xi_h \mid \boldsymbol{y})}\right\},$$

   where $p(\cdot \mid \boldsymbol{y})$ is the posterior density.

3. Update $(\mu_h, \phi_h, \xi_h) = (\mu_h^*, \phi_h^*, \xi_h^*)$ if $p_{\mathrm{ACC},h} > u$, with $u \sim \mathrm{Unif}(0, 1)$.

4. Update the covariance matrix $\boldsymbol{S}_h$ following the adaptive strategy of Garthwaite *et al.* (2016):

$$\begin{aligned}
\boldsymbol{\Sigma}_h &= \begin{cases} \mathbf{I}_3\{1 + (\tau_h)^2/t\}, & t \leq 100 \\ \mathbf{I}_3(\tau_h)^2/t + \frac{1}{t-1}\sum_{j=1}^{t}(\boldsymbol{\theta}_h^{(j)} - \bar{\boldsymbol{\theta}}_h)(\boldsymbol{\theta}_h^{(j)} - \bar{\boldsymbol{\theta}}_h)' & t > 100 \end{cases} \\
\tau_h &= \exp\left\{\log(\tau_h) + c\left(p_{\mathrm{ACC},h} - p_{\mathrm{ACC}}^*\right)/t\right\}, \\
\boldsymbol{S}_h &= (\tau_h)^2\,\boldsymbol{\Sigma}_h.
\end{aligned}$$

   Here, t is the current iteration, $\bar{\boldsymbol{\theta}}$ is the average of $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(t)}$, and

$$c = \frac{2}{3}\frac{(2\pi)^{1/2}\exp(\zeta^2/2)}{2\zeta} + \frac{1}{3p_{\mathrm{ACC}}^*(1 - p_{\mathrm{ACC}}^*)}$$

   is a step-length constant, where $\zeta = -\Phi^{-1}(p_{\mathrm{ACC}}^*/2)$ and $p_{\mathrm{ACC}}^* = 0.234$ is the target overall acceptance probability (Gelman *et al.*, 1997), and $\Phi(\cdot)$ the probability function of a standard Normal distribution. The update of τ_h starts at iteration n_0, corresponding to the closest integer to $5/\{p_{\mathrm{ACC}}^*(1 - p_{\mathrm{ACC}}^*)\}$.

In Step b α is considered fixed, but to add more flexibility to the model it is possible to let the data inform the choice of α, by assigning an hyperprior to it. Choosing $\alpha \sim \mathrm{Gamma}(a_\alpha, b_\alpha)$, Step b is now given by:

b1) Update the stick-breaking variables as

$$V_h \sim \mathrm{Beta}\left(1 + n_h, \alpha + \sum_{j>h}^{K} n_j\right), \quad h = 1, \ldots, K-1,$$

and compute each weight as $\pi_h = V_h \prod_{j<h} (1 - V_j)$.

b2) Sample α from its full conditional distribution as follows:

$$\alpha \sim \text{Gamma}\left(a_\alpha + K - 1, b_\alpha - \sum_{h=1}^{K-1} \log(1 - V_h)\right).$$

5.4 Simulation study

We evaluate the performance of the proposed model (5.4) through a simulation study. First, we outline the data-generating processes and present findings from single-sample experiments. Then, we move to illustrating results from a Monte Carlo simulation study.

5.4.1 Simulation setup and first experiments

We examine the performance of the proposed infinite mixture model of GEV distributions across four simulation scenarios. First, we assess the model when data are generated from a mixture of three GEV distributions (Scenario A). Next, we check its performance when the data originate from a single GEV distribution, which can also be viewed as a mixture with one component (Scenario B). Then, we examine the performance when data are derived from a mixture of Normal and Student's t distributions (Scenario C), which are not typically used for modelling extremes. Lastly, we assess whether the model can identify the presence of two GEV mixture components which are similar in terms of location and scale parameters, but with significantly different shape parameters (Scenario D). Table 5.1 provides an overview of these four data-generating processes, including the specific parameter values for each scenario.

For each scenario a random sample of size $n = 1000$ is simulated from the corresponding data-generating process. The blocked Gibbs sampler from Section 5.3 is applied for 10 000 iterations, among which the first half is considered as burn-in. The number of mixture components is truncated at $K = 50$, which is checked to be high enough compared to the resulting number of occupied components. The prior distributions assigned to the parameters of the GEV distributions are very wide distributions:

$$\mu_h \sim \text{N}(0, 10^8), \quad \phi_h \sim \text{N}(0, 10^8), \quad \xi_h \sim \text{N}(0, 10^4),$$

for every h. The precision parameter α is not kept fixed, but its distribution is influenced by the data and its value at every iteration is updated as described in step b2 of Section 5.3. A Gamma prior distribution is assigned to this parameter. The choice

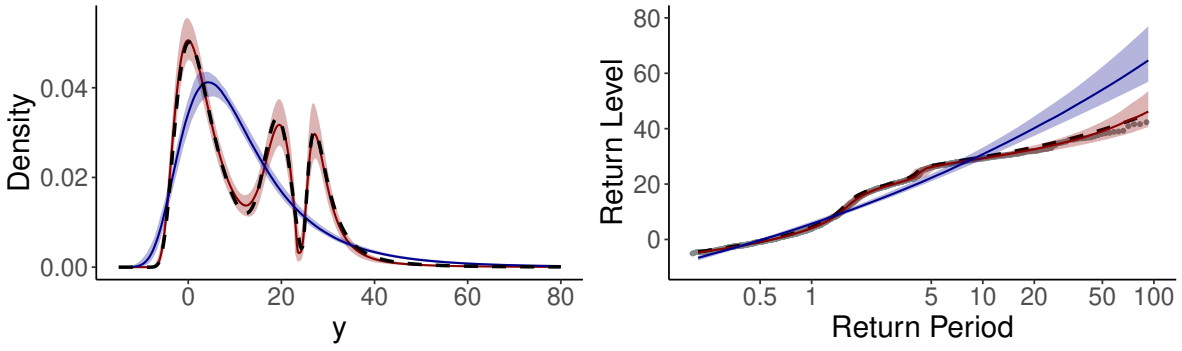
of the hyperparameters a_α (shape) and b_α (rate) is found to impact the results of the algorithm and, in particular, the estimated number of occupied components. Here, for illustration we use $a_\alpha = 1$ and $b_\alpha = 1$. We also experimented with smaller values of the hyperparameters, such as $a_\alpha = 0.01$ and $b_\alpha = 0.01$, which result in higher probability on small values of α and thus fewer groups (see Figure 2.2).

We compare the fit of the infinite mixture model to that of a single GEV distribution, as defined in (1.1). This simpler model is fitted using a Metropolis–Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970), with the same very wide priors of the mixture model for the three parameters of the GEV distribution. The employed algorithm can be found in Appendix A.2.

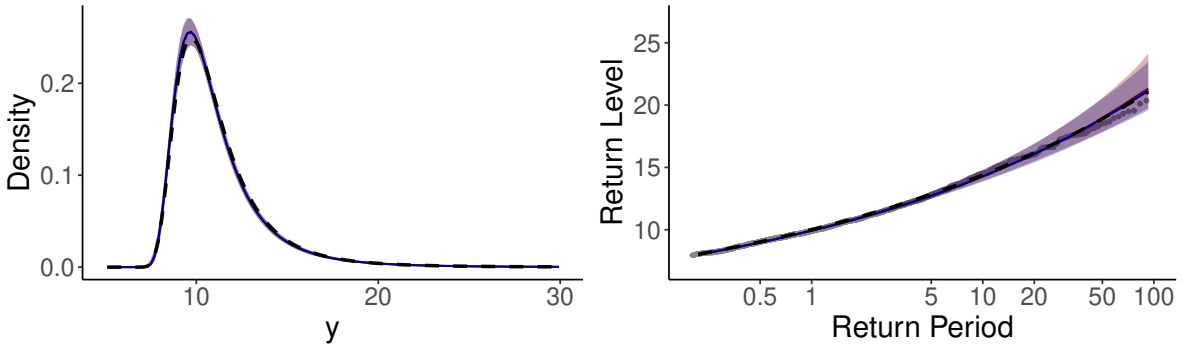
Figure 5.1 shows results of one-sample experiments for all the scenarios, with a comparison of the proposed model (5.4) to a single GEV distribution. Each graph on the left represents the median posterior density for both models compared to the true one, together with the corresponding posterior credible interval. The plots on the right of Figure 5.1 display return level plots. For both models posterior median return levels and posterior credible interval are computed based on a sample of 100 parameter-values, and are compared to the true ones and to the empirical quantiles. To produce return level plots quantile estimates for the infinite mixture of GEV distributions are derived by inverting (5.5). Since the inverse function $\hat{F}^{-1}$ has no closed form, quantiles need to

TABLE 5.1: Simulation scenarios. Here g, N and t respectively correspond to the probability density functions of the GEV, Normal and Student's t distribution.

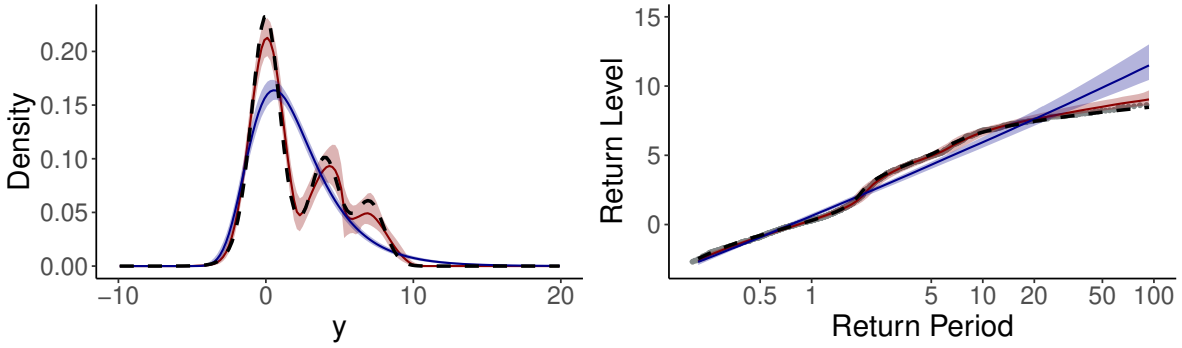
Scenario	**Kernel**	**Density function**	**Parameters**
A	GEV	$\sum_{h=1}^{3} \pi_h g(y; \mu_h, \sigma_h, \xi_h)$	$\pi_1 = 0.6, \pi_2 = 0.2, \pi_3 = 0.2$ $\mu_1 = 1, \phi_1 = 0.5, \xi_1 = 0.2$ $\mu_2 = 18, \phi_2 = 1, \xi_2 = -0.4$ $\mu_3 = 28, \phi_3 = 1, \xi_3 = 0.4$
B	GEV	$g(y; \mu, \sigma, \xi)$	$\mu = 10, \sigma = 1.5, \xi = 0.2$
C	Normal, Student's t	$\sum_{j=1}^{2} \pi_j N(y; \mu_j, \sigma_j) + \pi_3 t(y; \nu)$	$\pi_1 = 0.25, \pi_2 = 0.15, \pi_3 = 0.6$ $\mu_1 = 4, \sigma_1 = 1$ $\mu_2 = 7, \sigma_2 = 1$ $\nu = 10$
D	GEV	$(1 - \pi_2) g(y; \mu_1, \sigma_1, \xi_1) + \pi_2 g(y; \mu_2, \sigma_2, \xi_2)$	$\pi_2 = 0.6$ $\mu_1 = 0.5, \sigma_1 = 0.4, \xi_1 = -0.2$ $\mu_2 = 1.5, \phi_2 = 0.5, \xi_2 = 0.2$



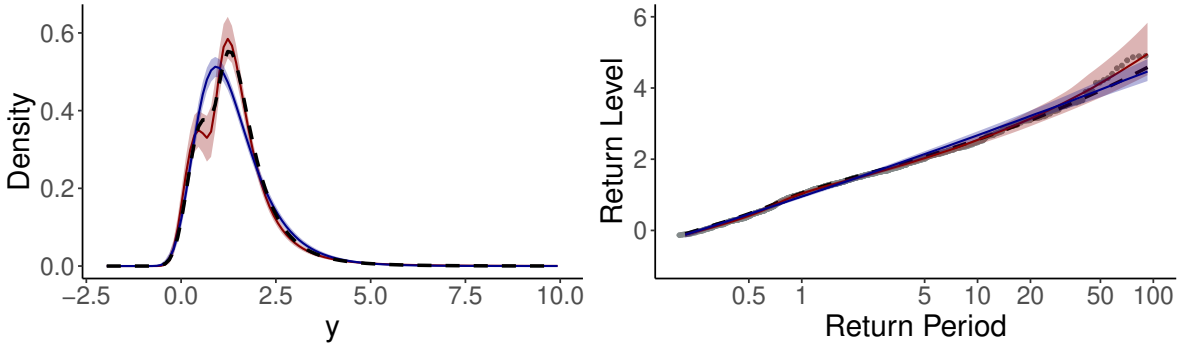
(a) Scenario A



(b) Scenario B



(c) Scenario C



(d) Scenario D

True Single GEV Mixture of GEV

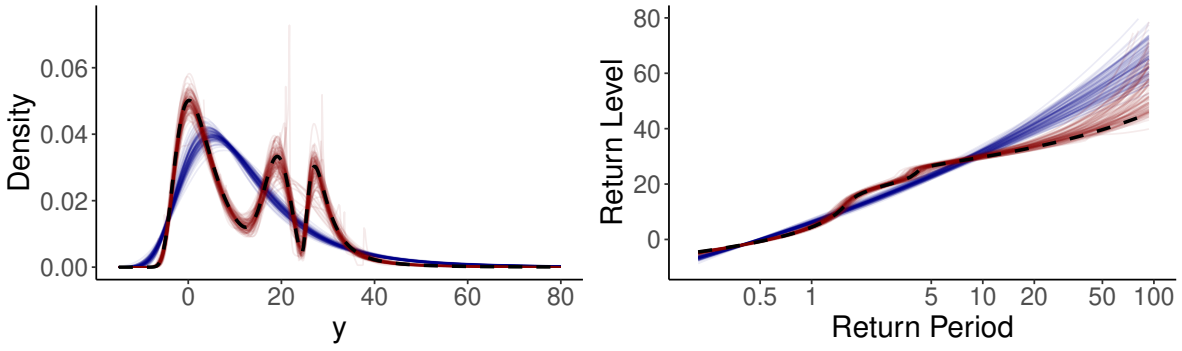
FIGURE 5.1: One-sample experiments. Left: posterior median density with credible interval (shaded) for single GEV model (blue) and infinite mixture of GEV distributions (red), compared to the true density (black). Right: posterior median return level curve with credible interval (shaded) for the two fitted models (same colours as before), with true return levels (black), and empirical quantiles as grey points.

be numerically computed by finding the values y that solve $\hat{F}(y) = p$ for various choices of p.

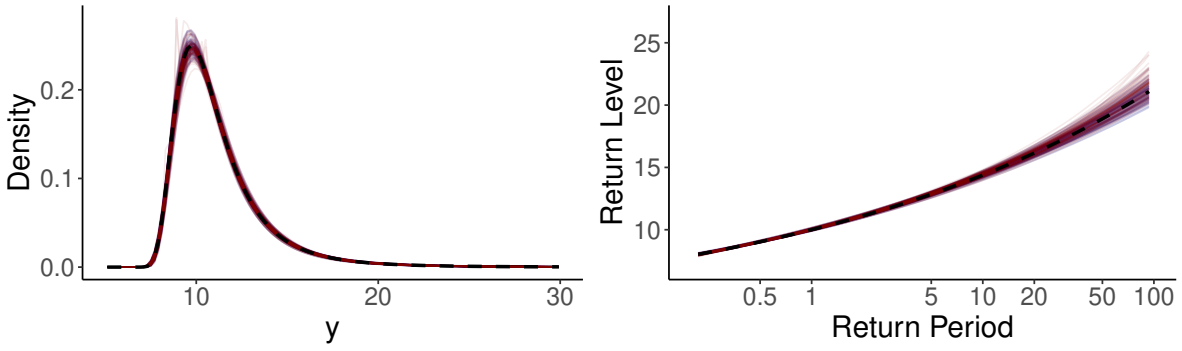
As illustrated in Figure 5.1, the mixture model performs exceptionally well in Scenario A, fitting both the density and quantile levels accurately, as expected. In Scenario B, the mixture model adapts effectively to the case of a single component, achieving a fit that closely resembles that of the single GEV model. Despite the higher flexibility of the infinite mixture model, it maintains a comparable level of uncertainty to the simpler single GEV model, without the need to assume a specific number of components a priori (one in this case). Additionally, in Scenario C the infinite mixture model of GEV distributions captures the behaviour of the mixture of Normal and Student's t distributions quite well, even though it is a misspecified model. The proposed infinite mixture model is able to identify the three modes and the mixing weights, but tends to overestimate the upper quantiles. Finally, in Scenario D, the mixture model seems to align well to the bimodal behaviour of the true density and to the true return levels, again with a slight overestimation of the highest quantiles.

5.4.2 Monte Carlo simulation

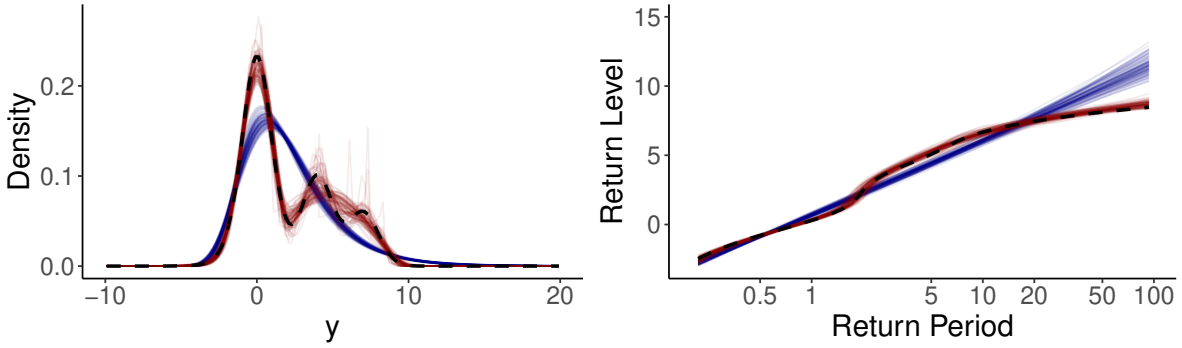
To validate the promising results from the one-sample experiments, we conduct a Monte Carlo simulation study. This involves replicating the one-sample experiments for $M = 100$ datasets simulated under each scenario. Figure 5.2 presents the Monte Carlo median densities and return levels for each scenario, compared with the true values. The findings from the previous one-sample analysis are supported by this Monte Carlo study. However, in Scenario A, the mixture model tends to overestimate the highest quantiles and occasionally struggles to distinguish between the two smallest components. In Scenario C, which is particularly challenging for the model to fit accurately, there is a noticeable difficulty in separating the components with the smallest data proportions, but the return plot shows a reasonably good fit, despite the overestimation at the highest quantiles. In Scenario D, which is also particularly challenging, on average the mixture model performs better than a single GEV distribution, although there are instances where it fails to detect the presence of two components, resulting in a fit comparable to that of a single GEV. Nonetheless, despite the challenges due to the components not being well separated, the model demonstrates high chances of success in this scenario, and it is particularly more accurate in estimating return levels. It is worth noting that defining what constitutes "well-separated" in the context of extremes remains challenging, as discussed in Chapter 4 when addressing proper scoring rules.



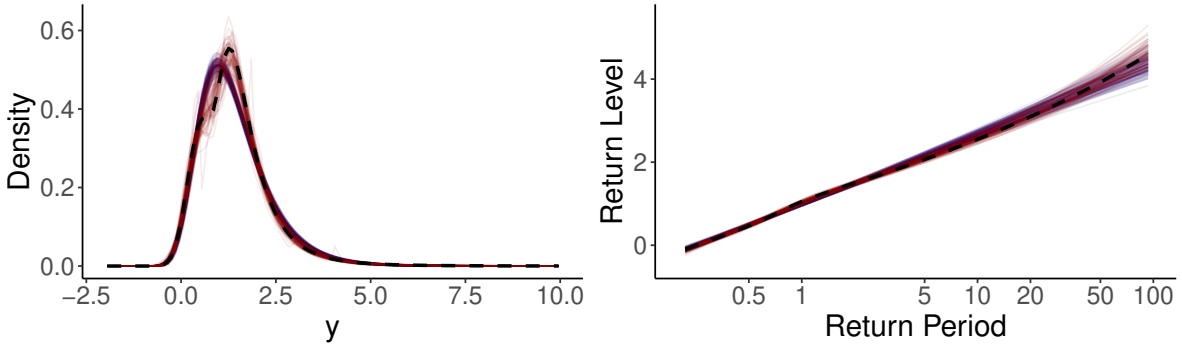
(a) Scenario A



(b) Scenario B



(c) Scenario C



(d) Scenario D

FIGURE 5.2: Monte Carlo results on $n = 1000$ data. Median posterior densities (left) and return levels (right) for $M = 100$ datasets based on the single GEV model (blue) and infinite mixture of GEV distributions (red), compared to the data-generating density (black).

TABLE 5.2: Model comparison measures from the Monte Carlo simulations: MISE (mean integrated absolute error), average expected LogS (logarithmic score), average expected CRPS (continuous ranked probability score). Median and standard error across the $M = 100$ datasets are shown.

Scenario		**A**	**B**	**C**	**D**
MISE	**Single**	0.0053 (0.0001)	9.09e-5 (3.57e-6)	6.58e-7 (1.12e-6)	2.7e-10 (1.3e-11)
	Mixture	8.71e-5 (0.0001)	9.56e-6 (3.46e-6)	9.89e-9 (1.32e-5)	5.48e-9 (2.34e-8)
LogS	**Single**	3.8903 (0.0241)	2.1011 (0.0291)	2.4195 (0.0217)	1.2445 (0.0262)
	Mixture	3.7331 (0.0305)	2.1030 (0.0293)	2.2988 (0.0210)	1.2373 (0.0263)
CRPS	**Single**	7.3105 (0.2095)	1.2888 (0.0498)	1.6443 (0.0305)	0.4953 (0.0152)
	Mixture	7.1133 (0.2041)	1.2914 (0.0501)	1.6241 (0.0311)	0.4950 (0.0151)

As a measure of model performance we use the mean integrated squared error (MISE), as defined in (1.12). It allows to compare the true density, as reported in Table 5.1, to the fitted one, which corresponds to the truncated version of model (5.4). In particular, the MISE is approximated as

$$\text{MISE} \approx \frac{1}{T-B} \sum_{t=B+1}^{T} \left[\int_{\mathcal{Y}} \{f^{(t)}(y) - f_0(y)\}^2 \mathrm{d}y \right],$$

where $f^{(t)}(y)$ is the posterior density corresponding to the tth iteration of the Gibbs sampler, with $t = B+1, \ldots, T$, where B is the number of iterations in the burn-in phase and T is the total number of length of the chain. The integral has to be computed using numerical methods, e.g., using the function `integrate` in `R`. As described in Section 1.2.2, further methods of model comparison for extremes involves proper scoring rules. Among them, logarithmic score (1.15) and CRPS (1.16) are used here as metrics to compare the fitted models. We compute the posterior average of a sample of 100 expected scores for each dataset. Comparing models only using the expected score has some drawbacks, as described in 1.2.2. However, proper scoring rules are not employed here for model selection, as the single GEV model is fitted solely for illustrative purposes, and a comparison based on the expectation is considered sufficient. Alternatively, the index in (1.21) could be considered.

For each scenario, fitted models are compared using posterior model comparison measures (Monte Carlo MISE, expected LogS, expected CRPS) summarised across the M datasets, as displayed in Table 5.2. All measures indicate a better performance of the mixture model in Scenarios A and C, as it also evident in Figure 5.5. In Scenario B, as expected, these metrics are nearly identical. In Scenario D, the scoring rules suggest

that the mixture model performs better, whereas the Monte Carlo MISE indicates the opposite. However, the differences between these results are minimal, making it difficult to draw a definitive preference. In this context, Figure 5.2 becomes particularly useful for further insight.

5.5 Application

We now present an application of the proposed method to precipitation data in Lisbon, Portugal. Extreme value theory is often applied to climate studies because accurately assessing the risk associated with natural hazards is essential for public safety and infrastructure planning. There are significant motivations to study precipitation in Portugal, given the increasing frequency and severity of recent flood events. For instance, on March 27, 2024, Lisbon experienced flooding caused by heavy rainfall, highlighting the need to understand and predict extreme weather events. Extreme precipitation events in Lisbon have been extensively studied. For example, Ferreira *et al.* (2024) analyse a record-breaking precipitation event in Lisbon in December 2022. Additionally, several studies have addressed flood risks in Portugal, including works by Liberato (2014) and Trigo *et al.* (2016). Our analysis aims at increasing the understanding of precipitation patterns in Lisbon by employing a Bayesian nonparametric model to determine whether rainfall maxima can be attributed to multiple underlying components.

5.5.1 Data description

The data consist of a series of daily precipitation measurements from December 1863 to March 2018 in Lisbon; see Valente *et al.* (2008) and Gallego *et al.* (2011) for details on the data. Data spanning from 1863 to 1940 have been digitised from the archives of IDL (Infante D. Luiz) meteorological observatory in Lisbon. Data from 1941 to 2006 were provided by IPMA (Instituto Português do Mar e da Atmosfera), while data from April 2006 onwards were again obtained from IDL, based on digitised maps of the classical station. Our goal is to study the behaviour of seasonal maxima of precipitation and to use the proposed approach to assess whether or not there is evidence in favour of a patter of heterogeneous extremes. We use the northern hemisphere meteorological seasons division (winter: December-February; spring: March-May; summer: June-August; autumn: September-November). The choice of blocks of 3 months is guided by three considerations. Using a longer block (e.g., an entire year) would be too wide, as the model assumes that all observations within the block belong to the same group, which may be an overly strong assumption. At the same time, the block must contain enough

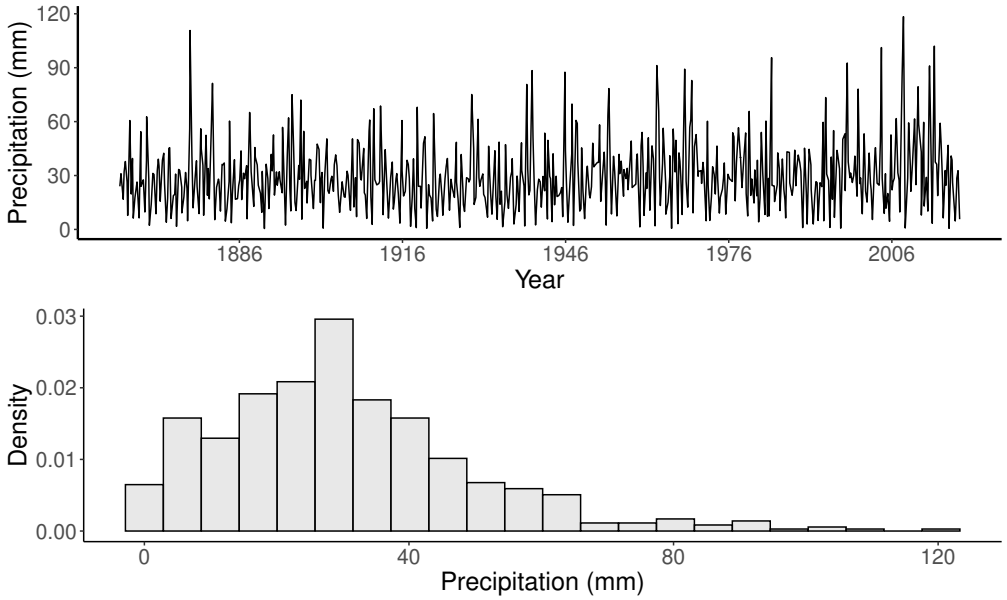


FIGURE 5.3: Seasonal maximum precipitations in Lisbon 1863–2018. Top: time series; bottom: histogram.

observations for the approximation of the extremal types theorem (Theorem 1.1) to be valid. Finally, this choice is natural from an applied viewpoint as it coincides with the length of seasons.

The time series of the 619 seasonal maxima is displayed in Figure 5.3, where it is possible to observe an increasing trend in the magnitude of peaks. The histogram of this series of maxima is also shown in Figure 5.3. To overcome the traditional assumption that a single distribution would be appropriate to model a rich dataset, we explore fitting a mixture model to uncover potential insights on the data.

5.5.2 Modelling extreme precipitation

The first step of this Bayesian analysis is fitting a single GEV distribution to the seasonal data using a Metropolis–Hastings algorithm with random-walk proposal, which can be found in Appendix A.2, for 10 000 iterations, half of which are considered warm-up. The initial values for the three parameters are set to their maximum likelihood estimates $\hat{\mu} = 21.61, \hat{\sigma} = 15.12, \hat{\xi} = -0.0017$, while the tuning parameter for the proposal distribution is set to the inverse of the Fisher information matrix, evaluated at those estimates. The prior distributions are chosen to be very wide priors: $\mu \sim \mathrm{N}(0, 10^8), \log(\sigma) \sim \mathrm{N}(0, 10^8)$ and $\xi \sim \mathrm{N}(0, 10^4)$. Table 5.3 shows the estimated posterior median of the parameters, along 95% credible intervals. The estimate of the shape parameter is very close to 0, which indicates an approximate fit to the Gumbel distribution.

TABLE 5.3: Posterior median with 95% credible interval in parenthesis for the single GEV model and the first four components of the infinite mixture model fitted to the seasonal maxima of precipitation in Lisbon.

Model	μ	σ	ξ	π
GEV	21.54 (20.37; 22.66)	15.14 (14.29; 16.16)	0.0004 (-0.051; 0.086)	
Mixture 1	24.74 (22.30; 27.05)	13.24 (11.10; 14.60)	0.049 (-0.015; 0.162)	0.878 (0.769; 0.943)
Mixture 2	4.304 (2.409; 6.663)	2.926 (0.972; 5.940)	0.157 (-0.487; 0.849)	0.105 (0.002; 0.213)
Mixture 3	1.016 (0.605; 4.033)	0.022 (0.001; 6.602)	2.208 (-1.933; 3.012)	0.007 (0.000; 0.078)
Mixture 4	0.609 (-2.816; 1.700)	0.022 (0.001; 16.07)	-2.269 (-2.309; 4.162)	0.010 (0.000; 0.026)

We proceed by fitting the infinite mixture model of GEV distribution (5.4) using the blocked Gibbs sampler described in 5.3 with number of components truncated at $K = 50$. This threshold is checked to be high enough, since the weights of the last components are extremely close to 0. As in the simulation study, the algorithm runs for 10 000 iterations, with the first half considered as burn-in. The algorithm is sensitive to the choice of the prior distributions, particularly the one for α, which strongly informs the number of occupied components. We assign the prior distributions:

$$\mu_h \sim \mathrm{N}(0, 100), \quad \log(\sigma_h) \sim \mathrm{N}(0, 100), \quad \xi_h \sim \mathrm{N}(0, 100), \quad \alpha \sim \mathrm{Gamma}(0.01, 0.01),$$

for $h = 1, \ldots, K$. We use prior predictive checks to assess whether data simulated from a mixture of K GEV distributions with parameters sampled from the specified priors aligns with the observed data, similarly to Gabry *et al.* (2019). We check the consistency of the chosen priors by examining simulations from the prior marginal distribution of the data under varying values of α. The priors we employ are intentionally vague, allowing to cover a wide range of scenarios, including cases aligned with the real data.

An important characteristic of infinite mixture models is the ability of making the data inform about the number of mixture components, avoiding the need of specifying it a priori. Figure 5.4 shows on the left the posterior distribution of the number of occupied components, defined as the components which are assigned at least one observation. To address the issue of label switching (Redner and Walker, 1984) in MCMC estimation, where the labels of mixture components can change throughout the sampling process, we adopt the strategy of relabelling the draws post-processing based on descending order of their location parameters. Thus, for each iteration, component 1 is the occupied component with the highest value of μ_h, component 2 is the one with the

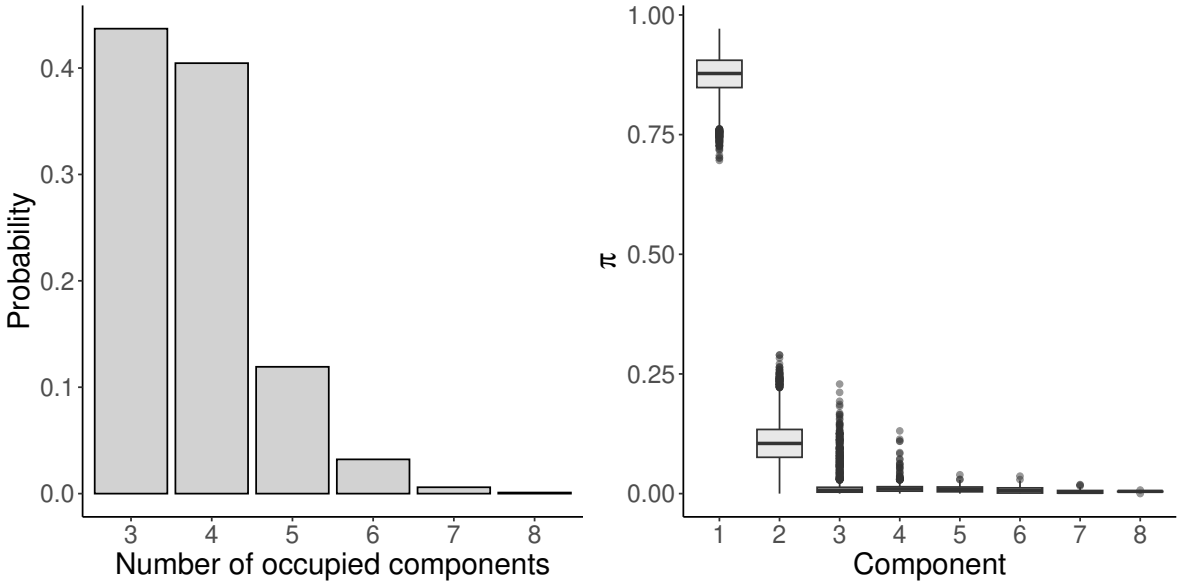


FIGURE 5.4: Left: posterior distribution of the number of occupied components. Right: posterior distribution of the mixing weight of the occupied mixture components.

second highest μ_h, and so forth. Figure 5.4 also shows the posterior distribution of π_h for these components in order of decreasing location parameter. These plots indicate that the model recognises the presence of 3 or 4 groups in the data, thus providing evidence in favour of heterogeneous extremes. The estimated posterior median for the parameters of the first 4 components in order of decreasing location parameter can be found in Table 5.3 with 95% credible intervals. The biggest component is the one in the centre of the distribution and it is the most similar to fitting a GEV to all the data, but it has a much bigger shape parameter. The second component seems to correspond to a lower mode that can be seen in the histogram in Figure 5.3, and the other very small components are located near 0 and capture a spike of the data in correspondence of very small values.

A straightforward way to assess model performance and compare the two fits is represented once again by graphical tools (see Section 1.2.1). Figure 5.5 shows the median posterior density with corresponding credible interval for the two fitted models. To better understand the tail behaviour, Figure 5.5 also includes the return plot with median posterior return levels and interval along with the empirical quantiles. The two models provide a similar overall fit; however, the infinite mixture model is significantly more flexible, especially in the left tail of the distribution, where it is able to capture the presence of at least two distinct groups, in addition to one for the rest of the data. Figure 5.4 indeed suggests that there are 3 or 4 actual groups in the data. Allowing the model to incorporate multiple components leads to a much closer alignment with

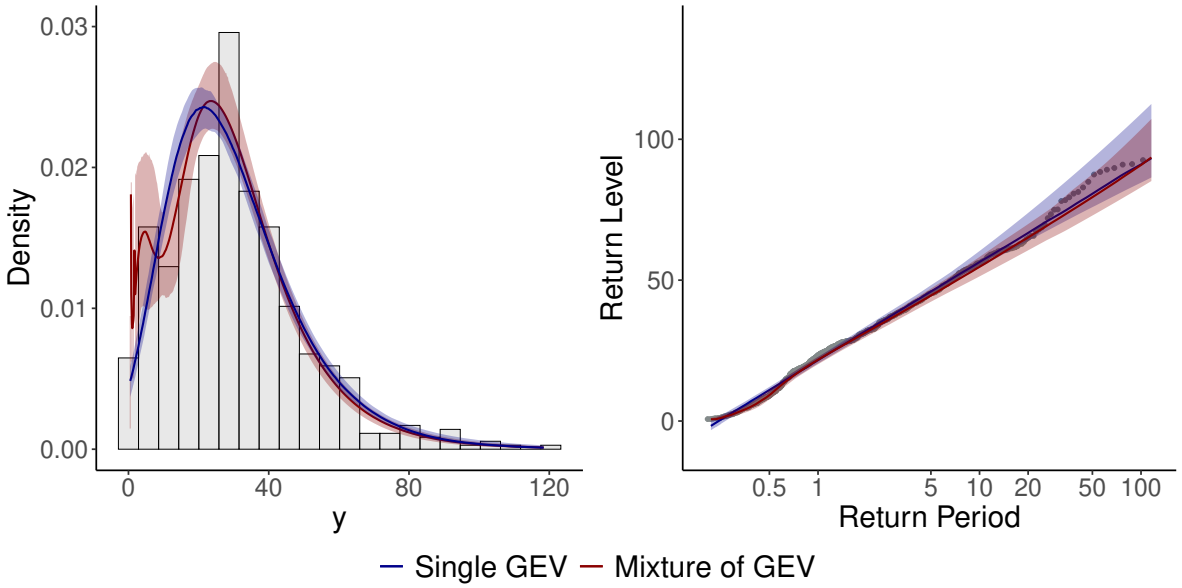


FIGURE 5.5: Left: median posterior densities for the infinite mixture model (red) and single GEV model (blue) with 95% credible interval, overlapping the histogram of the seasonal maxima of precipitation. Right: return level plot with posterior return levels and credible intervals for the two models, with empirical quantiles as grey points.

the empirical quantiles compared to the single GEV model, which, being approximately Gumbel, results in a return level plot that is nearly linear (see Figure 1.2). The mixture model, however, struggles to accurately fit the highest quantiles, probably due to the presence of an additional group that cannot be clearly identified given the limited data available. The model still provides a reasonably good fit and, despite being highly flexible, results in a level of uncertainty very similar to the one of the much simpler single GEV model. Therefore, the infinite mixture represents a valid choice that avoids assumptions that may be unrealistic, i.e., the assumption that the data follow a single GEV distribution. Nonetheless, it is significantly more computationally expensive than the simple single GEV model.

5.6 Concluding remarks

This chapter has demonstrated how Bayesian nonparametric methods can be effectively applied to the analysis of extreme events, and specifically of a series of maxima. A Dirichlet process mixture of GEV distributions was developed, using the Dirichlet process as it is the most established approach within the framework of infinite mixture models. Despite the irregularity of the GEV distribution, the proposed model successfully fitted a range of scenarios, even in cases where the mixture components were not

"well separated". However, the large sample size of the simulations may be unrealistic in practical scenarios.

Motivated by the understanding within the climate science community that multiple phenomena may influence precipitation, and thus precipitation maxima, we applied this infinite mixture model to real precipitation data, aiming to capture the presence of multiple components in rainfall maxima. The results indeed suggested the presence of such components, which could reflect distinct characteristics of the data worthy of further exploration by climate experts. However, it is important to emphasize that the objective here is not to perform clustering of the observations, but to model precipitation maxima more flexibly, relaxing traditional assumptions and improving estimates of return levels.

A more detailed discussion, including a deeper analysis of critical points, will be provided in the following chapter.

Discussion

This chapter summarises and critically discusses the main contributions, and outlines some directions for future research.

Final remarks

This thesis focused on addressing some challenges in the context of extreme value theory, that are very important from an applied perspective. Indeed, one of the main goals of this analysis was to improve estimates of the risk associated with rare events in real world situations. Indeed, the common assumption that observed maxima are drawn from an i.i.d. sequence of random variables following a single parametric distribution can lead to inaccurate risk estimates when the data instead arise from a mixture of multiple components. Using ideas, concepts and methods from mixture models, Bayesian nonparametrics and extreme value theory, this thesis aimed at presenting novel contributions to the problem.

While assuming more than one component in the extremes could look like a simple application of mixture models to extreme value analysis, there are several challenges that we noted in this dissertation. A first important consideration involves the challenges inherent to block maxima analysis. The GEV distribution is particularly challenging because its support depends on its parameters, making estimation complex. Additionally, this distribution represents an asymptotic approximation of the distribution of maxima as the block size tends to infinity. However, in practical applications, the block sizes are often quite small, leading to model misspecification when the GEV is applied to finite block data, as noted by Dombry and Ferreira (2019). These complexities also help to explain the limited literature on mixtures of GEV distributions, as previously discussed, since extending the GEV framework to mixture models introduces additional layers of difficulty. Moreover, the subtleties involved in the construction of heavy-tailed mixtures are not trivial, and recent research has been dedicated to them (e.g., Tressou, 2008; Li *et al.*, 2019; Palacios Ramirez *et al.*, 2024, to appear). This thesis kept all

such challenges and developments in mind and strongly contributed to the development of statistical modelling of heterogeneous extremes. In particular, we focused on the analysis of heterogeneous block maxima, a relatively unexplored area.

In Chapter 4, a dependent finite mixture model of two Gumbel distributions was proposed to describe real-world situations where assuming a single GEV distribution for the data is too restrictive. An important point that was discovered in a preliminary analysis is that information on which physical process originated the data (e.g., which weather regime generated each maximum precipitation event) is often not relevant to characterise the heterogeneous structure of the extremes. To address this issue, the proposed model includes covariate information in the mixing weight. This feature allows to advance the current methodology on finite mixtures for block maxima analysis, reviewed in Section 3.1. Furthermore, the incorporation of Bayesian methods introduces a novelty, offering several advantageous properties that have been previously discussed. Data-driven priors are used in both the simulations and reanalysis data application. This choice aligns with the literature on empirical Bayes methods (e.g., Carlin and Louis, 1997; Petrone *et al.*, 2014), which provide a practical framework using data-informed hyperparameters. However, we acknowledge that such an approach may not be ideal in scenarios with limited sample size. To address this concern, alternative options for specifying prior distributions for the parameters of the Gumbel distributions could be considered, such as the ones described in Section 1.1.1. For instance, priors based on domain knowledge or expert elicitation could be employed to incorporate external information. This would complement the proposed approach, which offers practitioners a method to deal with extremes that originate from multiple processes, without solely relying on knowing the physical process originating the observations and without needing to create a-priori categories with time-consuming and data intensive pre-processing. We do not extend the proposed model to a two-component mixture of GEV distributions because the Gumbel kernel adequately meets our objectives and applications. This decision also simplifies the estimation process, as incorporating a non-null shape parameter in the GEV distribution is likely to introduce challenges for estimation, as previously highlighted.

Measures of goodness of fit in extremes are also an important aspect of Chapter 4. As already pointed out, one challenge is selecting the appropriate measure from many options, each with distinct properties. In this chapter, we used a combination of scoring rules and an information criterion to compare different models, which can be an effective approach. However, different measures may yield conflicting model preferences. We opted for the traditional CRPS and logarithmic score, as weighted versions do not offer

substantial advantages (Lerch *et al.*, 2017). While we acknowledge their limitations, these rules also have benefits; for instance, the logarithmic score has the property of being locally scale invariant (Bolin and Wallin, 2023). We recognise that other scoring rules could have been applied, and we will discuss some alternative approaches in the next section.

Chapter 5 introduced a novel infinite mixture model using GEV kernels to handle heterogeneity in block maxima, offering a possibly more flexible alternative to the method in Chapter 4. Each block is assumed to belong to a specific mixture component, influencing the distribution of block maxima and leading to heterogeneity in extremes. We employed a Dirichlet process mixture, merging Bayesian nonparametrics with extreme value theory to characterise unknown complex structures in the right tail. A key innovation is the unbounded number of components, which eliminates the need to pre-specify how many components exist in the data. Indeed, to the best of our knowledge, infinite mixture models with a GEV kernel have not been previously proposed. An important point of our proposed model is also the possibility of capturing different domains of attraction in the data. Although inference is computationally challenging and is complicated by the irregularity of the GEV distribution, the model is successfully fitted using a blocked Gibbs sampler with random walk Metropolis steps and adaptive scaling. We would like to point out the sensitivity of the results to the specification of the baseline distribution in the Dirichlet Process, which could significantly affect model performance. More details on the choice of the baseline distribution can be found, for instance, in Hanson *et al.* (2005). A detailed exploration of this issue, while valuable, falls outside the scope of this chapter. The focus here was indeed on applying Bayesian nonparametric methods to the domain of extreme value analysis, taking into account its challenges. The proposed infinite mixture model is able to flexibly capture complex scenarios where extremes are organised into multiple mixture components, while also accommodating the traditional single GEV case without a noticeable increase in the variability of estimates. Additionally, the model adapts to new data without having to specify a different number of components, further enhancing its practical utility. However, in some cases, such as the application presented in Chapter 5, there is no clear superiority of the infinite mixture model over the simpler single GEV model. In fact, the single GEV model performs reasonably well, particularly given that it is much less computationally intensive than an infinite mixture model. Furthermore, the proposed infinite mixture model runs the risk of over-fitting, as it emerged from the real data analysis. However, this approach corresponds to fitting a complex model in order to explore the full range of possibilities. The idea behind using an infinite mixture model is

to relax many of the standard modelling assumptions, such as treating maxima as i.i.d. from a single distribution or having a fixed number of mixture components. Once this flexibility is introduced, we can then determine if a simpler model might be sufficient, but starting directly with the simple model may risk overconfidence in its adequacy. By relaxing modelling assumptions, the infinite mixture model helps mitigate the risk of model misspecification – or at least part of it, as the GEV model is known to be misspecified for a finite block size (Dombry and Ferreira, 2019), since it is an asymptotic model. We would like to also note that it is possible that the mixture components found in the data have been artificially created due to the combination of different data sources. Pre-processing the data could be an effective solution to mitigate the impact of these differences. An example of pre-processing techniques to integrate multiple sources of data to improve the accuracy and reliability of analysis is data fusion (e.g., Kalnay, 2003), which often uses reanalysis data (e.g., Dee *et al.*, 2011). When successful, pre-processing can reduce, if not remove, heterogeneity in the data. Nevertheless, our approach provides an alternative to the need for pre-processing techniques, offering a direct way of addressing heterogeneity in the data.

While the novel developments in this dissertation pioneer statistical modelling of heterogeneous extremes and offer several novel insights, there are a variety of limitations that should be acknowledged. For instance, if the model proposed in Chapter 4 incorporates many covariates, this may lead to over-fitting, especially in a scenario with low sample size, which is very common in block maxima analysis. Moreover, model comparison in the extreme value setting is proved to be challenging, as also proper scoring rules have been discussed to not be reliable when comparing models for extreme values (Brehmer and Strokorb, 2019; Taillardat *et al.*, 2023), since the expected score may not be able to distinguish differences in the tails, a concern we have also confirmed. When scoring rules are still believed to be useful, choosing the most appropriate one can be challenging, as each scoring rule comes with its unique strengths and weaknesses, and different rules can yield different model rankings. Furthermore, data analysis results based on the model in Chapter 5 can be largely affected by the choice of hyper-parameters. The prior choice for the parameters of the GEV distribution is especially important, since assigning flat prior leads to creating very few groups. At the same time, being too informative is not recommended. The model is also very sensitive to the prior on the precision parameter of the Dirichlet distribution, which strongly affects the resulting number of groups. Moreover, the assumption that the whole block belongs to the same component may be restrictive in practice, especially when considering large blocks (e.g., annual maxima). At the same time, a large block size is needed for the assumptions of

the model to hold.

We conclude this chapter by discussing possible future research avenues linked to some of the drawbacks and issues identified in the methods proposed throughout the thesis.

Future directions of research

Finally, some comments on directions for future research are pointed out below.

Contribution 1: Dependent mixtures for block maxima

To address the issue of overfitting that arises when using too many covariates, variable selection within the extreme value framework can be explored. In this setting, de Carvalho *et al.* (2022) contributed to regularisation and shrinkage methods by introducing a Bayesian Lasso-type model tailored for the lower and upper values of a potentially heavy-tailed positive outcome variable. Additional approaches involve the application of shrinkage priors, such as the spike-and-slab (George and McCulloch, 1993), the horseshoe prior (Carvalho *et al.*, 2009), and its various extensions (e.g., Piironen and Vehtari, 2017a). Another open question is the development of effective measures of goodness of fit for model comparison, which is a relatively unexplored area in the extreme value framework. The index proposed by Taillardat *et al.* (2023) offers a novel alternative to comparisons based on the expectation of proper scoring rules. Another approach worth considering is the use of the scaled CRPS developed by Bolin and Wallin (2023) and the scaled weighted CRPS introduced by Olafsdottir *et al.* (2024). Finally, there is no clear reason why one should only assume that the number of components of the data generating process should be exactly two, and it could be interesting to experiment with model specifications in which another finite number of components is assumed. Indeed, for the real-data analysis the assumption of two types of underlying physical process was found to not be entirely useful in describing the tail behaviour. However, this different specification would be more complicated and, more importantly, would entail an even more complex process of model selection.

Contribution 2: Infinite mixture models for heterogeneous extremes

To simplify the complexity of estimating a mixture of GEV distributions compared to a single GEV, one approach which could be explored is to assume that certain parameters are the same across groups, such as keeping the coefficient of variation constant.

Additionally, to mitigate sensitivity to the choice of prior, the shape parameter of the GEV distribution can be restricted to a suitable interval, with a transformation applied across the entire real line (see Jóhannesson *et al.*, 2022).

The proposed model can be extended to include covariate-dependence, moving to a dependent Dirichlet process (DDP) framework (MacEachern, 1999). Two main options are available: modelling the parameters of the GEV distributions as functions of a set of relevant variables (single-weights DDP), or using covariates to inform the distribution of the mixing weights (single-atoms DDP). Including additional variables, as done in Chapter 4, could enhance the identification of further mixture components that are not captured by the limited information in the series of maxima. Nonetheless, this would reflect in an even more complex model. Another future option would be to move from a Dirichlet process to a Pitman–Yor process (Pitman and Yor, 1997), with the construction of the weights employing a different stick-breaking definition. The Pitman–Yor process allows for even more flexibility in modelling the heterogeneity in the data.

Finally, an intriguing avenue for future research involves investigating the rate of convergence of mixtures of GEV distributions. While this topic holds considerable theoretical interest, it is believed to be very challenging and lies beyond the scope of this dissertation.

As a final point, both proposed models could potentially be extended into the framework of multivariate extremes (Beranger and Padoan, 2016). Such extensions are not straightforward, as incorporating dependence structures introduces additional complexity to model specification and fitting. This challenge becomes particularly significant when attempting to add the flexibility of infinite mixture models. Mixtures of extreme value copulas (e.g., Gudendorf and Segers, 2010) might provide a feasible approach that could be investigated. The proposed extension refer to scenarios with a moderate number of dimensions, aligning with specific applications of interest. However, extending the models to high-dimensional contexts would be exceedingly complex. A possible extension to to spatial extremes (Davison *et al.*, 2012), which could be believed to arise in specific locations due to different processes, is subject to the same considerations.

Appendix A

A.1 Additional numerical results

In this section we present some results related to the numerical experiments conducted in Chapter 4. In particular, Figure A.1 and Table A.1 are equivalent to Figure 4.1 and Table 4.2, respectively, but with $\pi = 0.45$.

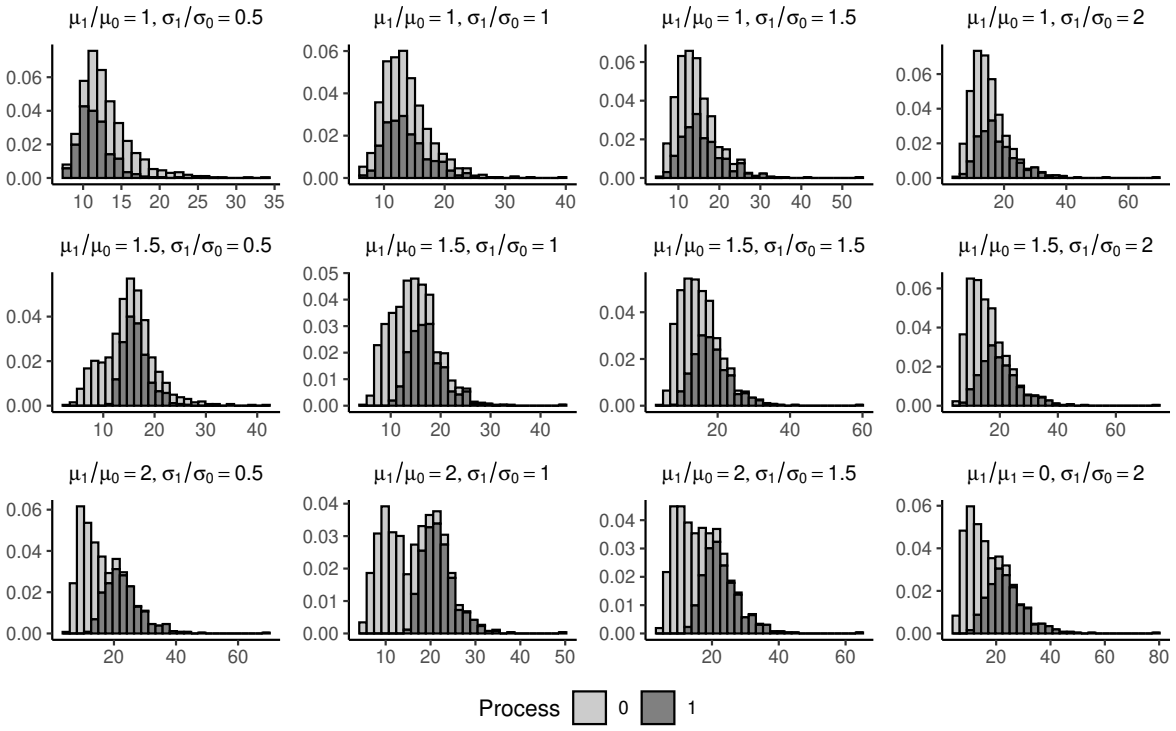
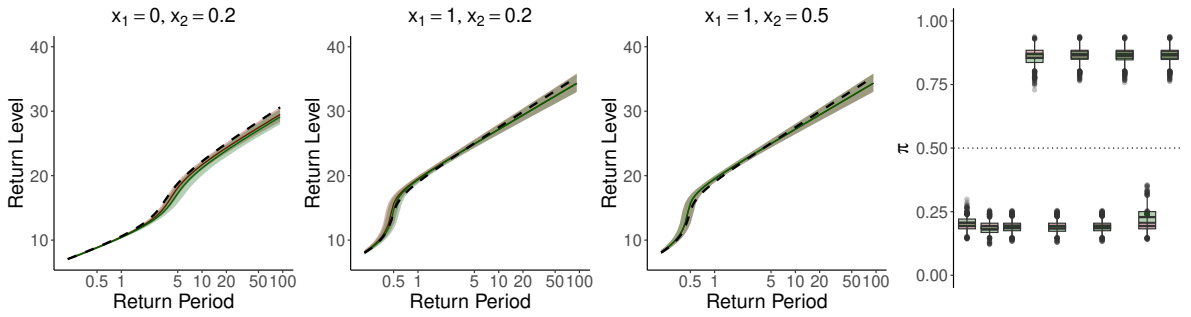


FIGURE A.1: Distribution of 1000 simulations from the two-component mixture of Gumbel distributions with different ratios of the location parameters and the scale parameters. The mixing parameter π is always equal to 0.45. For each bin the area is coloured according to the proportion that is due to process 0 (light) and process 1 (dark).

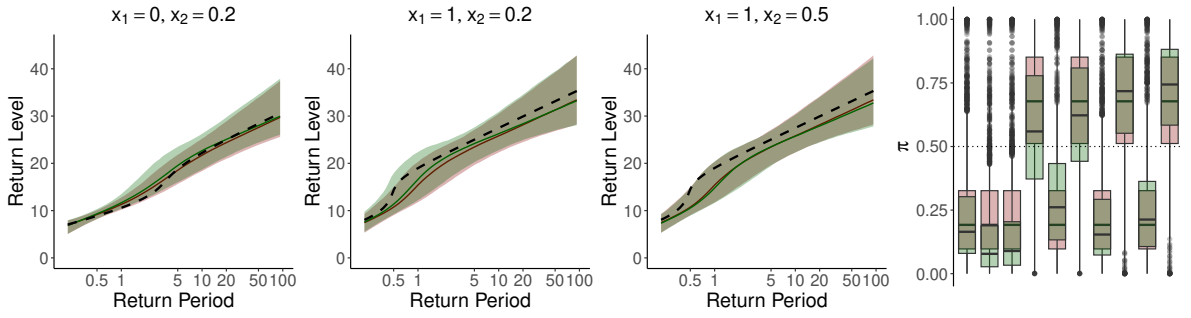
	$\mu_1/\mu_0 = 1$	$\mu_1/\mu_0 = 1.5$	$\mu_1/\mu_0 = 2$
$\sigma_1/\sigma_0 = 0.5$	0.0096 (0.4593)	0.0644 (0.4512)	0.9992 (0.4507)
$\sigma_1/\sigma_0 = 1$	0.4520 (0.4666)	0.7262 (0.4499)	0.9326 (0.4498)
$\sigma_1/\sigma_0 = 1.5$	0.8890 (0.4516)	0.9558 (0.4504)	0.9890 (0.4508)
$\sigma_1/\sigma_0 = 2$	0.9842 (0.4516)	0.9928 (0.4517)	0.9990 (0.4534)

TABLE A.1: Average proportion of events from process 1 among the 10 most extreme ones obtained from 500 samples of size 1000 from the two-component mixture of Gumbel distributions with different ratios of location and scale parameters and mixing weight for process 1 set to $\pi = 0.45$. The average proportion of events from process 1 in the whole sample is in parenthesis.

Figure A.2 refers instead to the simulation study in Section 4.3, presenting one-sample results of an additional simulation scenario which reflects the opposite case of Scenario B of Table 4.3, where the label x_1 is informative ($\beta_1 = 3$) and the continuous covariate x_2 is not ($\beta_2 = 0.1$).



(a) $n = 1000$



(b) $n = 50$

True Covariate-based mixture Label-based mixture

FIGURE A.2: One-shot experiments for the additional scenario. Top: median posterior return level with posterior credible interval (shaded) for the label-based model (red) and the covariate-based model (green), compared to the true return levels (black, dashed) for different values of the covariates. Box-plots of the posterior distribution of a sample of π_i are also displayed.

A.2 Additional algorithms

Metropolis–Hastings algorithm for GEV distribution

In the simulation study of Section 5.4 and in the application of Section 5.5, we use MCMC methods to estimate the parameters of the GEV distribution, defined in (1.1), from data $\boldsymbol{y} = (y_1, \ldots, y_n)'$. Setting $\phi = \log(\sigma)$, we specify a prior distribution

$$p(\mu, \phi, \xi) = N(\mathbf{0}_3, \boldsymbol{V}),$$

where $\boldsymbol{V}$ is a 3×3 covariance matrix. We assume a priori parameter independence, thus $\boldsymbol{V}$ is a diagonal matrix, with $\text{diag}(\boldsymbol{V}) = (v_\mu, v_\phi, v_\xi)$. We use a Metropolis–Hastings method with random walk proposals for posterior inference. We start by initialising the chain at some values of μ, ϕ and ξ, which we update for a total of T iterations, then discarding a burn-in period of B iterations. Given a covariance matrix $\boldsymbol{\nu}$ for the proposal distribution, the algorithm proceeds with the following steps:

1. Generate new values of the parameters from the proposal

$$(\mu^*, \phi^*, \xi^*)' \mid (\mu, \phi, \xi)' \sim \text{N}((\mu, \phi, \xi)', \boldsymbol{\nu}).$$

2. Compute the acceptance probability

$$p_{\text{ACC}} = \min\left\{1, \frac{p(\mu^*, \phi^*, \xi^* \mid \boldsymbol{y})}{p(\mu, \phi, \xi \mid \boldsymbol{y})}\right\},$$

   where $p(\cdot \mid \boldsymbol{y}) \propto p(\cdot) \prod_{i=1}^{n} g(y_i \mid \cdot)$ is the posterior density and g is the density function of the GEV distribution.

3. If $p_{\text{ACC}} > u$, with $u \sim \text{Unif}(0, 1)$, accept and update $(\mu, \phi, \xi) = (\mu^*, \phi^*, \xi^*)$.

Appendix B

B.1 How to fit finite mixture models for extremes in R

In this section, we show how to implement in `R` the methodology proposed in Chapter 4. As previously mentioned, posterior computation is performed using `RStan`. All the related `.stan` files can be found in https://github.com/VivianaCarcaiso/thesis/tree/main/stan.

We start by defining a set of function to compute density, distribution function, quantile function and random generation for a mixture of two Gumbel distributions, as it is usually done for standard distributions. The corresponding functions for the Gumbel distribution can be retrieved from the package `VGAM`.

```
# Auxiliary functions
require(VGAM)

dmix2gumb <- function(x, mu0, sigma0,  mu1, sigma1, pi)
  (1-pi) * dgumbel(x, location = mu0, scale = sigma0) +
  pi * dgumbel(x, location = mu1, scale = sigma1)

pmix2gumb <- function(x, mu0, sigma0,  mu1, sigma1, pi)
  (1-pi) * pgumbel(x, location = mu0, scale = sigma0) +
  pi * pgumbel(x, location = mu1, scale = sigma1)

qmix2gumb <- function(p, mu0, sigma0,  mu1, sigma1, pi){
  out <- NULL
  # numerical procedure to invert the CDF
  for(i in 1:length(p)){
    out <- c(out, uniroot(function(x)
```

```
      pmix2gumb(x, mu0, sigma0,  mu1, sigma1, pi) - p[i],
                           lower=-1e7, upper=1e7)$root)
  }
  out
}

rmix2gumb <- function(n, mu0, sigma0, mu1, sigma1, pi){

  # generate the latent indicator z
  z <- rbinom(n, 1, prob = pi)
  n1 <- sum(z == 1)
  n0 <- n - n1

  # simulate from the two Gumbel distributions
  y <- rep(NA, n)
  y[z==0] <- rgumbel(n0, mu0, sigma0)
  y[z==1] <- rgumbel(n1, mu1, sigma1)


  return(list("y" = y, "z" = z))
}
```

We can now generate a sample from a mixture of two Gumbel distributions, specifying the location and scale parameters, and the mixing weight π. In particular, we reproduce Scenarios A.1 and A.2 of table 4.3.

```
# Generating a mixture of 2 Gumbel distrbutions
n <- 1000
mu0=7; mu1=20; sigma0=3; sigma1=4; pi=0.2 # values of the parameters
set.seed(123)

samp <- rmix2gumb(n, mu0=mu0, mu1=mu1, sigma0=sigma0, sigma1=sigma1,
                  pi=pi)

y <- samp$y # observations
z <- samp$z # latent indicator of the mixture component
```

We now assume that a variable x_1 perfectly identifies the true allocations, like in Scenario A.1.

```
# Defining x1
x1 <- z
pr1 <- sum(z==1)/n   # basic estimate of pi
pr1

## [1] 0.198
```

We start by fitting to the data a mixture of two Gumbel distributions: one for the observations with $x_1 = 0$ and one for $x_1 = 1$. To do so we can use the `stan` model in `mixsep.stan`. We first need to create a list of data for the model, specifying the number of mixture components `K`, the sample size `n`, the observations `y`, and the hyper-parameters for the prior distributions of the location and scale parameters. We recall that we assign a Normal prior to μ_0 and μ_1 and a Gamma prior to σ_0 and σ_1. For more details on the prior distributions we refer to Section 4.3. Moreover, we need to specify a vector `nvec` which contains the cumulative number of observations per component.

```
# Fitting the model with fixed mixing weights

require(rstan)
options(mc.cores = 8) # specify the number of cores
rstan_options(auto_write = TRUE)

# order y based on x1
y.ord <- y[order(x1)]
# number of maxima of each type
nvec <- c(sum(x1==0), sum(x1==1))
# cumulate sum with a 0 as first element
nvec <- c(0, cumsum(nvec))

# create the data for the stan model
mix_data0 <- list(K=2, n=n, y=y.ord, nvec=nvec, mu_mu0=mean(y),
                  mu_sigma0=(max(y)-min(y))/4,
                  sigma_alpha0=3, sigma_beta0=1)

# fit the stan model
mod0 <-  stan_model(file = "mixsep.stan")
fit0 <- sampling(mod0, data = mix_data0, seed = 28)
```

Then we can print and plot the results of the model with the following code:

```
# Checking the results

print(fit0)
pairs(fit0, pars= c("mu", "sigma"))
```

We can now fit the dependent finite mixture model in (4.1) with x_1 as covariate. The `stan` code to fit this model is in `mixcov.stan`. Again, we need to create a list of data for the model, containing the sample size `n`, the number of covariates `p`, the observations `y`, the matrix of covariates `x` (scaled), and the hyper-parameters for the prior distributions of the location and scale parameters, and of the regression coeffiecients and intercept.

```
# Fitting the model with weights based on a logistic regression

# create the data for the stan model
mix_data1 <- list(n=n, p=1, y=y, x=as.matrix(x1-pr1),
       mu_mu0=mean(y), mu_sigma0=(max(y)-min(y))/4,
       sigma_alpha0=3, sigma_beta0=1, beta0_mu0=0, beta0_sigma0=5,
       beta_mu0=0, beta_sigma0=2.5)

# fit the stan model
mod1 <- stan_model(file = "mixcov.stan")
fit1 <- sampling(mod1, data = mix_data1, seed = 28)
```

Again, we can obtain a first summary of the results:

```
# Checking the results

print(fit1, pars = c("mu", "sigma", "beta0", "beta", "lp__"))
pairs(fit1, pars = c("mu", "sigma", "beta0", "beta"))
```

Another model that we may want to fit for comparison is a single GEV distribution. This is not shown in the simulation study of Chapter 4, but it is done in the application (Section 4.5). The `stan` code for a GEV distribution is in `gev.stan`, and can be fitted similarly to the previous two models using the code below:

```
# Fitting a single GEV model

# create the data for the stan model
gev_data <- list(n = n, y = y, mu_mu0 = mean(y),
                 mu_sigma0 = (max(y)-min(y))/4,
```

```
                  lsigma_mu0 = 0, lsigma_sigma0 = 5,
                  xi_mu0 = 0, xi_sigma0 = 5)

# fit the stan model
modGEV <- stan_model(file = "gev.stan")
fitGEV <- sampling(modGEV, data = gev_data, seed = 28)

# summary of the posterior distributions
print(fitGEV, pars=c('mu', 'sigma', 'xi'))
pairs(fitGEV, pars=c('mu', 'sigma', 'xi'))
```

Another possibility is to generate data from a mixture of two Gumbel distributions with mixing weight that depends on some covariates, like in Scenarios B and C of Table 4.3.

```
# Generating data from a mixture of two Gumbel distributions

require(boot)

n <- 1000
mu0 <- 10; mu1 <- 20; sigma0 <- 2; sigma1 <- 3.5
set.seed(123)
x1 <- rbinom(n, 1, 0.2) # label
x2 <- runif(n) # continuous variable that is informative about z

# choose the betas
beta1 <- 0.2
beta2 <- 10
beta0 <- -5


samp <- rmix2gumb(n, mu0=mu0, mu1=mu1, sigma0=sigma0, sigma1=sigma1,
                          pi=inv.logit(beta0 + beta1*x1 + beta2*x2))
y <- samp$y
z <- samp$z
```

Then we can proceed to fit model (4.1) like above.

```
# Fitting the model with weights based on a logistic regression

# create the data for the stan model
x <- cbind(x1-mean(z==1), scale(x2))
mix_data2 <- list(n=n, p=2, y=y, x=x,
          mu_mu0=mean(y), mu_sigma0=(max(y)-min(y))/4,
          sigma_alpha0=3, sigma_beta0=1, beta0_mu0=0, beta0_sigma0=5,
          beta_mu0=0, beta_sigma0=2.5)

# fit the stan model
fit2 <- sampling(mod1, data = mix_data2, seed = 28)
```

B.2 How to fit an infinite mixture models of GEV distributions in R

In this section, we show how to implement in `R` the infinite mixture model (5.4) proposed in Chapter 5, and also how to fit a single GEV distribution without using `stan`, but instead with the Metropolis–Hastings algorithm described in Appendix A.2.

Let us first define functions to compute the density, distribution function, quantile function and random generation for a mixture of K GEV distribution with user-specified location, scale and shape parameters.

```
# Auxiliary functions
require(ismev)
require(mev)

dmixgev <- function(x, K, mu, sigma, xi, pi){
  out <- 0
  for (h in 1:K) {
    out <- out + pi[h] * dgev(x, mu[h], sigma[h], xi[h])
  }
  return(out)
}


pmixgev <- function(x, K, mu, sigma, xi, pi){
```

```
  out <- 0
  for (h in 1:K) {
    out <- out + pi[h] * pgev(x, mu[h], sigma[h], xi[h])
  }
  return(out)
}



qmixgev <- function(p, K, mu, sigma, xi, pi){
  out <- NULL
  # numerical procedure to invert the CDF
  for(i in 1:length(p)){
    out <- c(out, uniroot(function(x)
                pmixgev(x, K, mu, sigma, xi, pi) - p[i],
                          lower=-1e50, upper=1e50)$root)
  }
  out
}

rmixgev <- function(n, K, mu, sigma, xi, pi){

  # generate the allocations
  z <- sample(1:K, n, prob = pi, replace = TRUE)

  # simulate from gev distributions
  out <- rep(NA, n)
  for (h in 1:K) {
    nh <- sum(z==h)
    if(nh>0) out[z==h] <- rgev(nh, mu[h], sigma[h], xi[h])
  }

  return(list("y" = out, "z" = z))
}
```

Now we can generate data from a mixture of three GEV distributions, following Scenario A from Table 5.1. We recall that we use the transformation $\phi = \log(\sigma)$.

```
# Sampling from a mixture of 3 GEV distributions
n <- 1000

mu.true <- c(1, 18, 28)
phi.true <- c(1.5, 1, 1)
xi.true <- c(0.2, -0.4, 0.4)
pi.true <- c(0.6, 0.2, 0.2)

set.seed(123)

y <- rmixgev(n, 3, mu.true, exp(phi.true), xi.true, pi.true)$y
```

Then we can fit a single GEV distribution to compare to the mixture model. The Metropolis–Hastings algorithm for this purpose can be implemented using the function `gev.est`, which can be found in https://github.com/VivianaCarcaiso/thesis/blob/main/R/algorithms_ch5.R. We can fit this algorithm to our simulated data, using as starting values the maximum likelihood estimates obtained with the `ismev` package, and as tuning parameter the inverse of the Fisher information. We set the number of iterations to 10 000, with the first half discarded as burn-in.

```
# Fitting a single GEV distribution
niter <- 10000
mle.gev <- gev.fit(y, show=FALSE)$mle
gev <- gev.est(y, mle.gev[1], mle.gev[2], mle.gev[3],
                  v.mu=10000, v.phi=10000, v.xi=100,
                  nu = gev.fit(y, show=FALSE)$cov, niter = niter)
```

The main challenge is fitting the infinite mixture model (5.4). In Section 1.1.1 we described a blocked Gibbs sampler with fixed truncation, which we implemented in the function `mix.est`, which can be found in https://github.com/VivianaCarcaiso/thesis/blob/main/R/algorithms_ch5.R. We set the upper bound on the number of components to 50. Again, we run the algorithm for 10 000 iterations.

```
# Fitting the infinite mixture model
niter <- 10000
K.tr <- 50 # upper bound for the number of components
mix <- mix.est(y, K = K.tr, niter=niter, seed=123)
```

Posterior quantities of interest from the two models can be obtained from the values contained in `mix`. For example, posterior return levels can be computed using the following code:

```
# Computing return levels
a <- seq(log(-1/log(0.01)), log(-1/log(0.992)), length.out = 90)
p <- exp(-1/exp(a))

quant.gev <- quant.mix <- matrix(nrow = (niter/2), ncol = length(p))
for (j in 1:(niter/2)) {
    quant.gev[j,] <- qgev(p, gev$mu[j+niter/2], gev$sigma[j+niter/2],
                          gev$xi[j+niter/2])
    quant.mix[j,] <- qmixgev(p, 50, mix$mu[j+niter/2,],
                exp(mix$phi[j+niter/2,]),
                mix$xi[j+niter/2,], mix$pi[j+niter/2,])
}
```

Parallelising the cycle over the chain is recommended if possible, as the computation can result slow. Posterior summaries of return levels can then be obtained to produce return level plots like those in Figure 5.1:

```
# Return level plots
quant.gev.median <- apply(quant.gev, 2, median)
quant.gev.q025 <- apply(quant.gev, 2, quantile, probs=0.025)
quant.gev.q975 <- apply(quant.gev, 2, quantile, probs=0.975)

quant.mix.median <- apply(quant.mix, 2, median)
quant.mix.q025 <- apply(quant.mix, 2, quantile, probs=0.025)
quant.mix.q975 <- apply(quant.mix, 2, quantile, probs=0.975)

quant.true <- qmixgev(p, 3, mu.true, exp(phi.true), xi.true, pi.true)

require(ggplot2)

ggplot() +
  geom_point( aes(x=log(-1/log((1:length(y))/(length(y) + 1))),
              y=sort(y)), size=1, color='azure4') +
  geom_line(aes(x = a, y = quant.true, color='true'),
            linewidth=1, linetype='dashed') +
  geom_line(aes(x = a, y = quant.mix.median, color='mix'),
            linewidth=0.6) +
  geom_ribbon(aes(x = a, ymax=`quant.mix.q975`, ymin=`quant.mix.q025`),
              fill="darkred", alpha=0.3) +
```

```
geom_line(aes(x = a, y = quant.gev.median, color = 'gev'),
          linewidth=0.6) +
geom_ribbon(aes(x = a, ymax=`quant.gev.q975`, ymin=`quant.gev.q025`),
            fill="darkblue", alpha=0.3) +
xlab('Return Period') + ylab('Return Level') +
theme(panel.background = element_rect(fill=NA, color=NA),
      plot.background = element_rect(fill=NA, color=NA),
      axis.line = element_line(color="black"),
      legend.title =  element_blank(),
      legend.text = element_text(size=18),
      axis.title = element_text(size=18),
      axis.text=element_text(size=15),
       legend.position='bottom') +
scale_x_continuous(limits = log(c(0.2, 99)),
     breaks = log(c(0.1, 0.5, 1, 5, 10, 20, 50, 100, 500, 1000)),
     labels = c(0.1, 0.5, 1, 5, 10, 20, 50, 100, 500, 1000)) +
scale_y_continuous(limits = c(-10, 80)) +
scale_colour_manual("",
     breaks = c("true", "gev", "mix"),
     values = c("black", "darkblue", "darkred"),
     labels = c('True', 'Single GEV', 'Mixture of GEV'))
```

Bibliography

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Aldous, D. (1985) Exchangeability and related topics. *Ecole d'Eté de Probabilités de Saint-Flour XIII-1983* **1117**, 1–198.

Amisano, G. and Giacomini, R. (2007) Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics* **25**(2), 177–190.

Argiento, R. and De Iorio, M. (2022) Is infinity that far? a bayesian nonparametric perspective of finite mixture models. *The Annals of Statistics* **50**(5), 2641–2663.

Azzalini, A. and Bowman, A. W. (1990) A look at some data on the old faithful geyser. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **39**(3), 357–365.

Balkema, A. A. and De Haan, L. (1974) Residual life time at great age. *The Annals of probability* **2**(5), 792–804.

Banfield, J. D. and Raftery, A. E. (1993) Model-based gaussian and non-gaussian clustering. *Biometrics* pp. 803–821.

Barbero, R., Fowler, H. J., Blenkinsop, S., Westra, S., Moron, V., Lewis, E., Chan, S., Lenderink, G., Kendon, E., Guerreiro, S. *et al.* (2019) A synthesis of hourly and daily precipitation extremes in different climatic regions. *Weather and Climate Extremes* **26**, 100219.

Behrens, C. N., Lopes, H. F. and Gamerman, D. (2004) Bayesian analysis of extreme events with threshold estimation. *Statistical modelling* **4**(3), 227–244.

Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. L. (2006) *Statistics of extremes: theory and applications*. John Wiley & Sons.

Beranger, B. and Padoan, S. (2016) Extreme dependence models. In *Extreme Value Modeling and Risk Analysis*. Chapman and Hall/CRC.

Bernardo, J. and Girón, F. (1988) A bayesian analysis of simple mixture problems. *Bayesian statistics* **3**(3), 67–78.

Blackwell, D. and MacQueen, J. B. (1973) Ferguson distributions via pólya urn schemes. *The annals of statistics* **1**(2), 353–355.

Bock, H. H. (1996) Probabilistic models in cluster analysis. *Computational Statistics & Data Analysis* **23**(1), 5–28.

Bolin, D. and Wallin, J. (2023) Local scale invariance and robustness of proper scoring rules. *Statistical Science* **38**(1), 140–159.

Bottolo, L., Consonni, G., Dellaportas, P. and Lijoi, A. (2003) Bayesian analysis of extreme values by mixture modeling. *Extremes* **6**, 25–47.

Bozdogan, H. and Sclove, S. L. (1984) Multi-sample cluster analysis using akaike's information criterion. *Annals of the Institute of Statistical Mathematics* **36**(1), 163–180.

Brehmer, J. R. and Strokorb, K. (2019) Why scoring functions cannot assess tail properties. *Electronic Journal of Statistics* **13**(2), 4015 – 4034.

Brier, G. W. (1950) Verification of forecasts expressed in terms of probability. *Monthly weather review* **78**(1), 1–3.

Bush, C. A. and MacEachern, S. N. (1996) A semiparametric bayesian model for randomised block designs. *Biometrika* **83**(2), 275–285.

Carlin, B. P. and Louis, T. A. (1997) Bayes and empirical bayes methods for data analysis.

Caron, F., Davy, M., Doucet, A., Duflos, E. and Vanheeghe, P. (2007) Bayesian inference for linear dynamic models with dirichlet process mixtures. *IEEE Transactions on Signal Processing* **56**(1), 71–84.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017) Stan: A probabilistic programming language. *Journal of statistical software* **76**(1).

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2009) Handling sparsity via the horseshoe. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 73–80.

de Carvalho, M., Pereira, S., Pereira, P. and de Zea Bermudez, P. (2022) An extreme value bayesian lasso for the conditional left and right tails. *Journal of Agricultural, Biological and Environmental Statistics* pp. 1–18.

Castellanos, M. E. and Cabras, S. (2007) A default bayesian procedure for the generalized pareto distribution. *Journal of Statistical Planning and Inference* **137**(2), 473–483.

Castillo, E. (2012) *Extreme value theory in engineering*. Elsevier.

Castro-Camilo, D., Huser, R. and Rue, H. (2022) Practical strategies for generalized extreme value-based regression models for extremes. *Environmetrics* **33**(6), e2742.

Celeux, G. and Govaert, G. (1995) Gaussian parsimonious clustering models. *Pattern recognition* **28**(5), 781–793.

Chavez-Demoulin, V. and Davison, A. C. (2005) Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society Series C: Applied Statistics* **54**(1), 207–222.

Chung, Y. and Dunson, D. B. (2009) Nonparametric bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association* **104**(488), 1646–1660.

Clarkson, D., Eastoe, E. and Leeson, A. (2023) The importance of context in extreme value analysis with application to extreme temperatures in the us and greenland. *Journal of the Royal Statistical Society Series C: Applied Statistics* **72**(4), 829–843.

Coles, S. (2001) *An introduction to statistical modeling of extreme values*. Springer.

Coles, S. and Tawn, J. (1996) A bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **45**(4), 463–478.

Coles, S. G. and Powell, E. A. (1996) Bayesian methods in extreme value modelling: a review and new developments. *International Statistical Review/Revue Internationale de Statistique* pp. 119–136.

Copernicus Climate Change Service, C. D. S. (2023) Era5 hourly data on single levels from 1940 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)* .

De la Cruz-Mesía, R., Quintana, F. A. and Müller, P. (2007) Semiparametric bayesian classification with longitudinal markers. *Journal of the Royal Statistical Society Series C: Applied Statistics* **56**(2), 119–137.

Dahl, D. B. (2003) An improved merge-split sampler for conjugate dirichlet process mixture models. *Technical R eport* **1**, 086.

Dahl, D. B. (2005) Sequentially-allocated merge-split sampler for conjugate and non-conjugate dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **11**(6), 4–2.

Davison, A. C., Padoan, S. A. and Ribatet, M. (2012) Statistical modeling of spatial extremes .

Davison, A. C. and Smith, R. L. (1990) Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)* **52**(3), 393–442.

De Iorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004) An anova model for dependent random measures. *Journal of the American Statistical Association* **99**(465), 205–215.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d. P. *et al.* (2011) The era-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society* **137**(656), 553–597.

Del Moral, P., Doucet, A. and Jasra, A. (2006) Sequential monte carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **68**(3), 411–436.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1), 1–22.

Diebold, F. X. (2015) Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests. *Journal of Business & Economic Statistics* **33**(1), 1–1.

Diebold, F. X. and Mariano, R. S. (2002) Comparing predictive accuracy. *Journal of Business & economic statistics* **20**(1), 134–144.

Diebolt, J. and Robert, C. P. (1994) Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society: Series B (Methodological)* **56**(2), 363–375.

Diks, C., Panchenko, V. and van Dijk, D. (2011) Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics* **163**(2), 215–230.

Dombry, C. and Ferreira, A. (2019) Maximum likelihood estimators based on the block maxima method. *Bernoulli* **25**(3), 1690–1723.

Duan, J. A., Guindani, M. and Gelfand, A. E. (2007) Generalized spatial dirichlet process models. *Biometrika* **94**(4), 809–825.

Dunson, D. B. (2010) Nonparametric bayes applications to biostatistics. *Bayesian nonparametrics* **28**, 223–273.

Dunson, D. B. and Park, J.-H. (2008) Kernel stick-breaking processes. *Biometrika* **95**(2), 307–323.

Dunson, D. B., Pillai, N. and Park, J.-H. (2007) Bayesian density regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **69**(2), 163–183.

Escobar, M. D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**(430), 577–588.

Everitt, B. and Hand, D. (1981) Mixtures of normal distributions. In *Finite Mixture Distributions*, pp. 25–57. Springer.

Ferguson, T. S. (1973) Bayesian analysis of some nonparametric problems. *The Annals of Statistics* pp. 209–230.

Ferguson, T. S. (1974) Prior distributions on spaces of probability measures. *The annals of statistics* pp. 615–629.

Ferreira, T. M., Trigo, R. M., Gaspar, T. H., Pinto, J. G. and Ramos, A. M. (2024) The record-breaking precipitation event of december 2022 in portugal. *Natural Hazards and Earth System Sciences Discussions* **2024**, 1–21.

Fisher, R. A. and Tippett, L. H. C. (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical proceedings of the Cambridge philosophical society*, volume 24, pp. 180–190.

Fraley, C. and Raftery, A. E. (1998) How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal* **41**(8), 578–588.

Fraley, C. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association* **97**(458), 611–631.

Friederichs, P. and Thorarinsdottir, T. L. (2012) Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics* **23**(7), 579–594.

Frigessi, A., Haug, O. and Rue, H. (2002) A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes* **5**, 219–235.

Frühwirth-Schnatter, S. (2006) *Finite mixture and Markov switching models*. Springer.

Fruhwirth-Schnatter, S., Celeux, G. and Robert, C. P. (2019) *Handbook of mixture analysis*. CRC press.

Frühwirth-Schnatter, S., Malsiner-Walli, G. and Grün, B. (2021) Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Analysis* **16**(4), 1279–1307.

Fúquene, J., Steel, M. and Rossell, D. (2019) On choosing mixture components via non-local priors. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **81**(5), 809–837.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M. and Gelman, A. (2019) Visualization in bayesian workflow. *Journal of the Royal Statistical Society Series A: Statistics in Society* **182**(2), 389–402.

Gallego, M., Trigo, R., Vaquero, J., Brunet, M., García, J., Sigró, J. and Valente, M. (2011) Trends in frequency indices of daily precipitation over the Iberian Peninsula during the last century. *Journal of Geophysical Research: Atmospheres* **116**(D2).

Garthwaite, P. H., Fan, Y. and Sisson, S. A. (2016) Adaptive optimal scaling of metropolis–hastings algorithms using the Robbins–Monro process. *Communications in Statistics-Theory and Methods* **45**(17), 5098–5111.

Gelfand, A. E., Kottas, A. and MacEachern, S. N. (2005) Bayesian nonparametric spatial modeling with dirichlet process mixing. *Journal of the American Statistical Association* **100**(471), 1021–1035.

Gelfand, A. E., Smith, A. F. and Lee, T.-M. (1992) Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling. *Journal of the American Statistical Association* **87**(418), 523–532.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995) *Bayesian data analysis*. Chapman and Hall/CRC.

Gelman, A., Gilks, W. R. and Roberts, G. O. (1997) Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability* **7**(1), 110–120.

Gelman, A., Hwang, J. and Vehtari, A. (2014) Understanding predictive information criteria for bayesian models. *Statistics and computing* **24**, 997–1016.

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y.-S. (2008) A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* **2**(4), 1360 – 1383.

Gelman, A., Meng, X.-L. and Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica* pp. 733–760.

Geman, S. and Geman, D. (1984) Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence* (6), 721–741.

George, E. I. and McCulloch, R. E. (1993) Variable selection via gibbs sampling. *Journal of the American Statistical Association* **88**(423), 881–889.

Ghosal, S. (2010) The dirichlet process, related priors and posterior asymptotics. *Bayesian nonparametrics* **28**, 35.

Ghosal, S. and Van der Vaart, A. (2017) *Fundamentals of nonparametric Bayesian inference*. Volume 44. Cambridge University Press.

Ghosh, J., Li, Y. and Mitra, R. (2018) On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression. *Bayesian Analysis* **13**(2), 359 – 383.

Gnedenko, B. V. (1948) On a local limit theorem of the theory of probability. *Uspekhi Matematicheskikh Nauk* **3**(3), 187–194.

Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association* **102**(477), 359–378.

Gneiting, T. and Ranjan, R. (2011) Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics* **29**(3), 411–422.

Good, I. J. (1952) Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)* **14**(1), 107–114.

Green, P. J. (1995) Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika* **82**(4), 711–732.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C. and Wallis, J. R. (1979) Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. *Water resources research* **15**(5), 1049–1054.

Grego, J. M. and Yates, P. A. (2010) Point and standard error estimation for quantiles of mixed flood distributions. *Journal of hydrology* **391**(3-4), 289–301.

Griffin, J. E. and Steel, M. J. (2006) Order-based dependent dirichlet processes. *Journal of the American statistical Association* **101**(473), 179–194.

Gudendorf, G. and Segers, J. (2010) Extreme-value copulas. In *Copula Theory and Its Applications: Proceedings of the Workshop Held in Warsaw, 25-26 September 2009*, pp. 127–145.

Gutiérrez, L., Barrientos, A. F., González, J. and Taylor-Rodriguez, D. (2019) A bayesian nonparametric multiple testing procedure for comparing several treatments against a control. *Bayesian Analysis* **14**(2), 649–675.

Gutiérrez, L., Mena, R. H. and Ruggiero, M. (2016) A time dependent bayesian nonparametric model for air quality analysis. *Computational Statistics & Data Analysis* **95**, 161–175.

Haan, L. and Ferreira, A. (2006) *Extreme value theory: an introduction*. Volume 3. Springer.

Hambuckers, J. and Kneib, T. (2023) Smooth-transition regression models for non-stationary extremes. *Journal of Financial Econometrics* **21**(2), 445–484.

Hanson, T., Sethuraman, J. and Xu, L. (2005) On choosing the centering distribution in dirichlet process mixture models. *Statistics & probability letters* **72**(2), 153–162.

Hartigan, J. A. and Hartigan, P. M. (1985) The dip test of unimodality. *The annals of Statistics* pp. 70–84.

Hartigan, J. A., Wong, M. A. *et al.* (1979) A k-means clustering algorithm. *Applied statistics* **28**(1), 100–108.

Hastings, W. K. (1970) Monte carlo sampling methods using markov chains and their applications. *Biometrika* .

Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G. (2010) *Bayesian nonparametrics*. Volume 28. Cambridge University Press.

Hoffman, M. D., Gelman, A. *et al.* (2014) The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.* **15**(1), 1593–1623.

Hosking, J. R. (1990) L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **52**(1), 105–124.

Hosking, J. R. (1992) Moments or l moments? an example comparing two measures of distributional shape. *The American Statistician* **46**(3), 186–189.

Hosking, J. R. M. and Wallis, J. R. (1997) *Regional frequency analysis*. Cambridge University Press.

Hosking, J. R. M., Wallis, J. R. and Wood, E. F. (1985) Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics* **27**(3), 251–261.

Hu, C., Castro-Camilo, D. and Swallow, B. (2024) A bayesian multivariate extreme value mixture model. *arXiv preprint arXiv:2401.15703* .

Huerta, G. and Sansó, B. (2007) Time-varying models for extreme values. *Environmental and Ecological Statistics* **14**, 285–299.

Ishwaran, H. and James, L. F. (2001) Gibbs sampling methods for stick-breaking priors. *Journal of the American statistical Association* **96**(453), 161–173.

Jain, S. and Neal, R. M. (2004) A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of computational and Graphical Statistics* **13**(1), 158–182.

Jain, S. and Neal, R. M. (2007) Splitting and merging components of a nonconjugate dirichlet process mixture model. *Bayesian Analysis* pp. 445–472.

Jara, A., Lesaffre, E., De Iorio, M. and Quintana, F. (2010) Bayesian semiparametric inference for multivariate doubly-interval-censored data. *Annals of Applied Statistics* .

Jóhannesson, Á. V., Siegert, S., Huser, R., Bakka, H. and Hrafnkelsson, B. (2022) Approximate bayesian inference for analysis of spatiotemporal flood frequency data. *The Annals of Applied Statistics* **16**(2), 905–935.

Kallenberg, O. (1983) *Random measures*. De Gruyter.

Kalli, M., Griffin, J. E. and Walker, S. G. (2011) Slice sampling mixture models. *Statistics and computing* **21**, 93–105.

Kalnay, E. (2003) *Atmospheric Modeling, Data Assimilation and Predictability*. Volume 341. Cambridge University Press.

Katz, R. W., Parlange, M. B. and Naveau, P. (2002) Statistics of extremes in hydrology. *Advances in Water Resources* **25**(8-12), 1287–1304.

Kjeldsen, T. R., Ahn, H., Prosdocimi, I. and Heo, J.-H. (2018) Mixture gumbel models for extreme series including infrequent phenomena. *Hydrological Sciences Journal* **63**(13-14), 1927–1940.

Kottas, A. and Sansó, B. (2007) Bayesian mixture modeling for spatial poisson process intensities, with applications to extreme value analysis. *Journal of Statistical Planning and Inference* **137**(10), 3151–3163.

Lavine, M. and West, M. (1992) A bayesian method for classification and discrimination. *Canadian Journal of Statistics* **20**(4), 451–461.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F. and Gneiting, T. (2017) Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science* pp. 106–127.

Lewis, R. and Battey, H. (2024) On inference in high-dimensional logistic regression models with separated data. *Biometrika* **111**(3), 989–1011.

Li, C., Lin, L. and Dunson, D. B. (2019) On posterior consistency of tail index for Bayesian kernel mixture models. *Bernoulli* **25**(3), 1999 – 2028.

Liberato, M. L. (2014) The 19 January 2013 windstorm over the north atlantic: large-scale dynamics and impacts on Iberia. *Weather and Climate Extremes* **5**, 16–28.

Lijoi, A., Mena, R. H. and Prünster, I. (2007) Controlling the reinforcement in bayesian non-parametric mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **69**(4), 715–740.

Liu, Q., Huang, X. and Zhou, H. (2024) The flexible gumbel distribution: A new model for inference about the mode. *Stats* **7**(1), 317–332.

Lo, A. Y. (1984) On a class of bayesian nonparametric estimates: I. density estimates. *The annals of statistics* pp. 351–357.

MacEachern, S. N. (1994) Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics-Simulation and Computation* **23**(3), 727–741.

MacEachern, S. N. (1999) Dependent nonparametric processes. *Statistica Sinica* pp. 441–455.

Malsiner-Walli, G., Frühwirth-Schnatter, S. and Grün, B. (2016) Model-based clustering based on sparse finite gaussian mixtures. *Statistics and computing* **26**(1), 303–324.

Marin, J.-M., Mengersen, K. and Robert, C. P. (2005) Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics* **25**, 459–507.

Martins, E. S. and Stedinger, J. R. (2000) Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research* **36**(3), 737–744.

Matheson, J. E. and Winkler, R. L. (1976) Scoring rules for continuous probability distributions. *Management Science* **22**(10), 1087–1096.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *The journal of chemical physics* **21**(6), 1087–1092.

Miller, J. W. and Harrison, M. T. (2018) Mixture models with a prior on the number of components. *Journal of the American Statistical Association* **113**(521), 340–356.

Müller, P., Quintana, F. A., Jara, A. and Hanson, T. (2015) *Bayesian nonparametric data analysis*. Volume 1. Springer.

do Nascimento, F. F., Gamerman, D. and Lopes, H. F. (2012) A semiparametric bayesian approach to extreme value estimation. *Statistics and Computing* **22**, 661–675.

Neal, R. M. (2003) Slice sampling. *The annals of statistics* **31**(3), 705–767.

Northrop, P. J. and Attalides, N. (2016) Posterior propriety in bayesian extreme value analyses using reference priors. *Statistica Sinica* pp. 721–743.

Olafsdottir, H. K., Rootzén, H. and Bolin, D. (2024) Locally tail-scale invariant scoring rules for evaluation of extreme value forecasts. *International Journal of Forecasting* .

Opitz, T., Huser, R., Bakka, H. and Rue, H. (2018) Inla goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles. *Extremes* **21**(3), 441–462.

Otiniano, C., Gonçalves, C. and Dorea, C. (2017) Mixture of extreme-value distributions: identifiability and estimation. *Communications in Statistics-Theory and Methods* **46**(13), 6528–6542.

Palacios Ramirez, V., de Carvalho, M. and Gutierrez, L. (2024, to appear) Heavy-tailed NGG-mixture models. *Bayesian Analysis* .

Papaspiliopoulos, O. and Roberts, G. O. (2008) Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika* **95**(1), 169–186.

Pearson, K. (1894) Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A* **185**, 71–110.

Peel, D. and McLachlan, G. J. (2000) Robust mixture modelling using the t distribution. *Statistics and computing* **10**, 339–348.

Petrone, S., Rousseau, J. and Scricciolo, C. (2014) Bayes and empirical bayes: do they merge? *Biometrika* **101**(2), 285–302.

Pickands III, J. (1971) The two-dimensional poisson process and extremal processes. *Journal of applied Probability* **8**(4), 745–756.

Pickands III, J. (1975) Statistical inference using extreme order statistics. *the Annals of Statistics* pp. 119–131.

Pickands III, J. (1986) The continuous and differentiable domains of attraction of the extreme value distributions. *The Annals of Probability* pp. 996–1004.

Piironen, J. and Vehtari, A. (2017a) Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics* **11**(2), 5018 – 5051.

Piironen, J. and Vehtari, A. (2017b) Comparison of bayesian predictive methods for model selection. *Statistics and Computing* **27**, 711–735.

Pitman, J. and Yor, M. (1997) The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25**(2), 855–900.

Quintana, F. A., Müller, P., Jara, A. and MacEachern, S. N. (2022) The dependent Dirichlet process and related models. *Statistical Science* **37**(1), 24–41.

Redner, R. A. and Walker, H. F. (1984) Mixture densities, maximum likelihood and the em algorithm. *SIAM review* **26**(2), 195–239.

Reich, B. J. and Ghosh, S. K. (2019) *Bayesian statistical methods*. Chapman and Hall/CRC.

Ren, L., Du, L., Carin, L. and Dunson, D. B. (2011) Logistic stick-breaking process. *Journal of Machine Learning Research* **12**(1).

Resnick, S. I. (2007) *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media.

Richardson, S. and Green, P. J. (1997) Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**(4), 731–792.

Rigon, T. and Durante, D. (2021) Tractable bayesian density regression via logit stick-breaking priors. *Journal of Statistical Planning and Inference* **211**, 131–142.

Robert, C. P., Casella, G. and Casella, G. (1999) *Monte Carlo statistical methods*. Volume 2. Springer.

Rodriguez, A., Dunson, D. B. and Gelfand, A. E. (2008) The nested dirichlet process. *Journal of the American statistical Association* **103**(483), 1131–1154.

Rodriguez, A. and Ter Horst, E. (2008) Bayesian dynamic density estimation. *Bayesian Analysis* .

Rossi, F., Fiorentino, M. and Versace, P. (1984) Two-component extreme value distribution for flood frequency analysis. *Water Resources Research* **20**(7), 847–856.

Scarrott, C. and MacDonald, A. (2012) A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Statistical journal* **10**(1), 33–60.

Schwarz, G. E. (1978) Estimating the dimension of a model. *The Annals of Statistics* **6**(2), 461–464.

Sethuraman, J. (1994) A constructive definition of dirichlet priors. *Statistica Sinica* pp. 639–650.

Skilling, J. (2004) Nested sampling. *Bayesian inference and maximum entropy methods in science and engineering* **735**, 395–405.

Smith, R. L. (1985) Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72**(1), 67–90.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4), 583–639.

Stan Development Team (2023) RStan: the R interface to Stan. R package version 2.26.24.

Stephenson, A. and Tawn, J. (2004) Bayesian inference for extremes: accounting for the three extremal types. *Extremes* **7**, 291–307.

Stephenson, A. G. (2002) evd: Extreme value distributions. *R news* **2**(2), 31–32.

Tadesse, M. G. and Vannucci, M. (2021) *Handbook of Bayesian variable selection*. CRC Press.

Taillardat, M., Fougères, A.-L., Naveau, P. and De Fondeville, R. (2023) Evaluating probabilistic forecasts of extremes using continuous ranked probability score distributions. *International Journal of Forecasting* **39**(3), 1448–1459.

Tannery, J. (1910) *Introduction à la théorie des fonctions d'une variable*. Volume 1. A. Hermann.

Tarasova, L., Merz, R., Kiss, A., Basso, S., Blöschl, G., Merz, B., Viglione, A., Plötner, S., Guse, B., Schumann, A. *et al.* (2019) Causative classification of river flood events. *Wiley Interdisciplinary Reviews: Water* **6**(4), e1353.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006) A hierarchical bayesian language model based on pitman–yor processes. *Journal of Machine Learning Research* **7**, 399–421.

Tendijck, S., Eastoe, E., Tawn, J., Randell, D. and Jonathan, P. (2023) Modeling the extremes of bivariate mixture distributions with application to oceanographic data. *Journal of the American Statistical Association* **118**(542), 1373–1384.

Titterington, D. (1985) Common structure of smoothing techniques in statistics. *International Statistical Review/Revue Internationale de Statistique* pp. 141–170.

Tressou, J. (2008) Bayesian nonparametrics for heavy tailed distribution. application to food risk assessment. *Bayesian Analysis* .

Trigo, R. M., Ramos, A. M., Pereira, S. S., Ramos, P., Zêzere, J. L. and Liberato, M. L. (2016) The deadliest storm of the 20th century striking Portugal: Flood impacts and atmospheric circulation. *Journal of Hydrology* **541**, 597–610.

Valente, M. A., Trigo, R., Barros, M., Nunes, L. F., Alves, E., Pinhal, E., Coelho, F., Mendes, M. and Miranda, J. (2008) Early stages of the recovery of portuguese historical meteorological data. In *MEDARE-Proceedings of the International Workshop on Rescue and Digitization of Climate Records in the Mediterranean Basin*, number 67, pp. 95–102.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T. and Gelman, A. (2023) loo: Efficient leave-one-out cross-validation and waic for bayesian models. R package version 2.6.0.9000.

Vehtari, A., Gelman, A. and Gabry, J. (2017) Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing* **27**, 1413–1432.

Vettori, S., Huser, R. and Genton, M. G. (2019) Bayesian modeling of air pollution extremes using nested multivariate max-stable processes. *Biometrics* **75**(3), 831–841.

Vogel, R. M. and Fennessey, N. M. (1993) L moment diagrams should replace product moment diagrams. *Water resources research* **29**(6), 1745–1752.

Walker, S. G. (2007) Sampling the dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation* **36**(1), 45–54.

Wand, M. P. and Jones, M. C. (1994) *Kernel smoothing*. CRC press.

Wang, H. and Tsai, C.-L. (2009) Tail index regression. *Journal of the American Statistical Association* **104**(487), 1233–1240.

Watanabe, S. (2009) *Algebraic geometry and statistical learning theory*. Cambridge University Press.

Watanabe, S. (2010) Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* **11**, 3571–3594.

Wolfe, J. H. (1970) Pattern clustering by multivariate mixture analysis. *Multivariate behavioral research* **5**(3), 329–350.

Viviana Carcaiso

CURRICULUM VITAE

Personal Details

Date of Birth: January 28, 1998
Place of Birth: Florence, Italy
Nationality: Italian

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.
e-mail: viviana.carcaiso@phd.unipd.it

Current Position

Since November 2024; (expected completion: April 2025)
Research engineer, Unité BioSP – Biostatistique et Processus Spatiaux, INRAE, Centre Provence-Alpes Côte d'Azur.
Project title: Extrapolation of extreme covariates in predictive models with application to climate-change projections. Funded by H2020 FIRE-RES.
Supervisor: Thomas Opitz
Co-supervisors: Sebastian Engelke, Juliette Legrand.

Since October 2021; (expected completion: January/February 2025)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Bayesian mixture models for extremes
Supervisor: Ilaria Prosdocimi
Co-supervisors: Isadora Antoniano-Villalobos, Miguel de Carvalho.

Research interests

- Extreme value theory
- Environmental extremes
- Bayesian methods
- Bayesian Nonparametrics

Education

November 2019 – July 2021
Master degree in Statistics and Data Science.
University of Florence, School of Economics and Management
Title of dissertation: "The impact of remote teaching on university students' gained credits: an analysis based on parametric modeling of quantile regression coefficient functions"
Supervisor: Leonardo Grilli
Final mark: 110/110 cum laude

September 2016 – November 2019
Bachelor degree in Statistics.
University of Florence, School of Economics and Management
Title of dissertation: "Un Sistema Informativo a supporto del CRR per le epilessie dell'AOU Careggi" (*An information system to support the Regional Reference Center for Epilepsy at Careggi University Hospital*)
Supervisor: Bruno Bertaccini
Final mark: 110/110 cum laude

Visiting periods

January 2024 – June 2024
University of Edinburgh,
Edinburgh, UK.
Supervisor: Prof. Miguel de Carvalho

Awards and Scholarships

October 2021
Ph.D. scholarship (3 years), University of Padova.
Academic year 2020/2021
DSU TOSCANA scholarship (1 year).
Academic year 2019/2020
DSU TOSCANA scholarship (1 year).
Academic year 2018/2019
DSU TOSCANA scholarship (1 year).
Academic year 2016/2017
DSU TOSCANA scholarship (1 year).

Computer skills

- Programming languages: R, Python, Fortran
- Markup languages: LaTeX
- Statistical analysis: R, SAS, Stata
- Geographic Information System: QGIS
- LimeSurvey (Online Survey Tool)
- Microsoft Office environment

Language skills

Italian: native; English: fluent; French: basic.

Publications

Articles in journals

Richards, J., Lee, M. W., Carcaiso, V., de Carvalho, M. (2024). Contribution to the Discussion of 'Inference for extreme spatial temperature events in a changing climate with application to Ireland'

by Healy, D., Tawn, J., Thorne, P., and Parnell, A. *Journal of the Royal Statistical Society Series C: Applied Statistics* (accepted).

Carcaiso, V., Prosdocimi, I., Antoniano-Villalobos, I. (2023). Regression for mixture models for extremes. *Book of the Short Papers SIS 2023 - Statistical Learning, Sustainability and Impact Evaluation*, 629-634.

Carcaiso, V., Grilli, L. (2022). Quantile regression for count data: jittering versus regression coefficients modelling in the analysis of credits earned by university students after remote teaching. *Statistical Methods & Applications*, 1-22.

Carcaiso, V., Grilli, L. (2022). Quantile regression coefficient modeling for counts to evaluate the productivity of university students. *Book of short papers SIS 2022*, 1333-1338.

Carcaiso, V., Grilli, L. (2022). Analysis of count data by quantile regression coefficient modelling: student's gained credits after online teaching. *Proceedings of the 36th International Workshop on Statistical Modelling*, 113-116.

Conference presentations

Carcaiso, V., de Carvalho, M., Prosdocimi, I., Antoniano-Villalobos, I. (2024). Bayesian mixture models for heterogeneous extremes. (contributed talk) *6th International Conference on Advances in Extreme Value Analysis and Application to Natural Hazards (EVAN)*, Venice, Italy, 16-19 July 2024.

Carcaiso, V., de Carvalho, M., Prosdocimi, I., Antoniano-Villalobos, I. (2024). Bayesian mixture models for heterogeneous extremes. (poster presentation) *2024 ISBA World Meeting*, Venice, Italy, 1-7 July 2024.

Carcaiso, V., Prosdocimi, I., Antoniano-Villalobos, I. (2023). Regression for mixture models for extremes. (poster presentation) *Centre for Statistics Annual Conference 2024*, Edinburgh, UK, 18 June 2024.

Carcaiso, V., Prosdocimi, I., Antoniano-Villalobos, I. (2023). Regression for mixture models for extremes. (poster presentation) *Centre for Statistics Early Career Researchers Day 2024*, Edinburgh, UK, 17 June 2024.

Carcaiso, V., Prosdocimi, I., Antoniano-Villalobos, I. (2023). Where do extremes come from? Dependent mixtures for block maxima. (invited talk) *16th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2023)*, Berlin, Germany, 16-18 December 2023.

Carcaiso, V., Prosdocimi, I., Antoniano-Villalobos, I. (2023). Where do extremes come from? Dependent mixtures for block maxima. (contributed talk) *STOR-i Extremes Workshop (STEW)*, Lancaster, UK, 20-22 September 2023.

Carcaiso, V., Prosdocimi, I., Antoniano-Villalobos, I. (2023). Regression for mixture models for extremes. (poster presentation) *Extreme Value Analysis (EVA) 2023*, Milan, Italy, 26-30 June 2023.

Carcaiso, V., Prosdocimi, I., Antoniano-Villalobos, I. (2023). Regression for mixture models for extremes. (contributed talk and poster presentation) *SIS 2023 - Statistical Learning, Sustainability and Impact Evaluation*, Ancona, Italy, 21-23 June 2023.

Carcaiso, V., Grilli, L. (2022). Analysis of count data by quantile regression coefficient modelling:

student's gained credits after online teaching. (contributed talk) *36th International Workshop on Statistical Modelling*, Trieste, Italy, 18-22 July 2023.

Other Interests

Baking
Running

References

Prof. Miguel de Carvalho
School of Maths, University of Edinburgh
Edinburgh, UK
e-mail: miguel.decarvalho@ed.ac.uk

Prof. Ilaria Prosdocimi
Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice
Venice, Italy
e-mail: ilaria.prosdocimi@unive.it