MIXTURE MODELS FOR MULTIVARIATE EXTREMES

THÈSE N° 3098 (2004)

PRÉSENTÉE À LA FACULTÉ SCIENCES DE BASE

Institut de mathématiques

SECTION DE MATHÉMATIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Marc-Olivier BOLDI

ingénieur mathématicien diplômé EPF
et de nationalité française

acceptée sur proposition du jury:

Prof. A.C. Davison, directeur de thèse
Prof. S. Coles, rapporteur
Prof. A. Ledford, rapporteur
Prof. S. Morgenthaler, rapporteur

Lausanne, EPFL
2004

Contents

Remerciements

Ceux qui me connaissent savent la place qu'a pris cette thèse dans ma vie. Ils ne savent peut-être pas suffisament la place qu'ils ont eux-même occupée. Que les oubliés me pardonnent, je leur dois sans doute beaucoup aussi.

I would like to thank my supervisor Anthony Davison for the opportunity he gave me and above all the patience he had with me. I would like to thank Anthony Ledford, Stuart Coles and Stephan Morgenthaler for their interest in this work and the improvements they brought to it. A mes parents pour leur patience, leur soutien, leur amour. A Ingrid et Pierre-Philippe pour l'exemple qu'ils me donnent et à Marine et Florian pour m'avoir rendu marteau. A Thomas pour être un deuxième frère et à Aude pour sa complicité et son accueil. A Hélène et Jean-Claude pour leur porte ouverte, et aux tatas et tontons également: une deuxième famille. A Maiko, je ne sais pas pourquoi mais c'est cool de t'avoir comme pote. A Alex, Loric, Cédric, Jackson, Joëlle, Annick, Philippe et Claude-Alain pour m'avoir offert la plus extraordinaire expérience de ma vie. A Sandrine et Lionel que je ne réduirai pas en une phrase. A Roland, Andrei et Sandro pour le quotidien. A Jérome pour les parties d'échecs. A Ludo pour les envies. A Armelle qui m'a relancé au bon moment. A Hugo et Gérard pour les cafés à sat. A Réda et Chéraz pour m'avoir sauvé la vie. A Sylvain et Valérie... et vive la France!!! A Christophe qui m'a supporté, bravo, belle performance. A Baptiste et Thomas aussi. A Chloé parce que tu es belle et pas que. A Benjamin, le roi de la tarte aux prunes. A Diego qui me donne du boulot. A Rodrigue et Benoît, pour les soirées de très grands cinéphiles. A Alexandre, Texas. Revient dont, on a les mêmes à la maison! Et puis aux picolo-foots, aux immenses soirées, à la Martinique, aux Led Zeppelin, au théatre, aux blagues sur l'internet, aux mariages, aux restos, ah oui les restos, et à demain.

Version abrégée

Cette thèse est une contribution à la statistique des valeurs extrêmes. La queue d'une fonction de répartition multivariée est charactérisée par sa distribution spectrale. Nous proposons un nouveau modèle semi-paramétrique constitué d'un mélange de distributions de Dirichlet. Pour l'estimation de ses composants, un algorithme à sauts réversibles par chaînes de Markov et un algorithme EM sont développés. Leurs performances sont illustrées sur des données réelles et simulées. Ces simulations sont obtenues grâce à une nouvelle représentation des modèles logistiques et de Dirichlet. En parallèle à l'estimation de la loi spectrale, la statistique des valeurs extrêmes requière la selection d'un seuil permettant de classer les données comme extrêmes ou non. Cette sélection est obtenue grâce à une nouvelle méthode, basée sur des arguments heuristiques, qui permet une sélection indépendante de la dimension des données. Ses performances sont illustrées sur des données réelles et simulées.

L'intérêt premier d'une analyse des extrêmes réside dans l'estimation de quantiles d'événements rares et dans l'exploration de la structure de dépendance, pour lesquelles l'estimation de la mesure spectrale est un moyen plutôt qu'un but. Ces deux questions sont abordées. Pour la première, une méthode de Monte Carlo par simulation d'extrêmes est développée. Elle est comparée avec des méthodes classiques et nouvelles de la littérature. Pour la seconde, une analyse de dépendance conditionnelle originale est proposée. Elle consiste en une série de graphiques représentant des coupes de la fonction de densité spectrale. Elle éclaire différents aspects de la structure de dépendance des données. Des exemples sur des données réelles illustrent l'analyse.

Dans la dernière partie, le modèle semi-paramétrique et les méthodes présentées sont étendues au cas spatial. Cela est rendu possible en considérant la distribution spectrale comme la loi d'une probabilité aléatoire, un point de vue adopté tout au long de cette thèse. Le cas des extrêmes multivariés s'étend alors au cas d'extrêmes de mesures aléatoires. L'application est illustrée sur un jeu de données de précipitations en Chine.

Abstract

This thesis is a contribution to multivariate extreme value statistics. The tail of a multivariate distribution function is characterized by its spectral distribution, for which we propose a new semi-parametric model based on mixtures of Dirichlet distributions. To estimate the components of this model, reversible jump Monte Carlo Markov chain and EM algorithms are developed. Their performances are illustrated on real and simulated data, obtained using new representations of the extremal logistic and Dirichlet models. In parallel with the estimation of the spectral distribution, extreme value statistic machinery requires the selection of a threshold in order to classify data as extreme or not. This selection is achieved by a new method based on heuristic arguments. It allows a selection independent of the dimension of the data. Its performance is illustrated on real and simulated data.

Primal scientific interests behind a multivariate extreme value analysis reside in the estimation of quantiles of rare events and in the exploration of the dependence structure, for which the estimation of the spectral measure is a means rather than an end. These two issues are addressed. For the first, a Monte Carlo method is developed based on simulation of extremes. It is compared with classical and new methods of the literature. For the second one, an original conditional dependence analysis is proposed, which enlightens various aspects of the dependence structure of the data. Examples using real data sets are given.

In the last part, the semi-parametric model and the presented methods are extended to spatial extremes. It is made possible by considering the spectral distribution as the distribution of a random probability, an original viewpoint adopted throughout this thesis. Classical multivariate extremes are extended to extremes of random measures. The application is illustrated on rainfall data in China.

Introduction

When considering the influence through time of a phenomenon on a system, this influence is often cumulative. Effects of similar amplitude concentrate and entail an ageing of the system, like a pot filled little by little with drops. From a mathematical viewpoint, the state of the system is determined by the sum of little quantities X_i, $i = 1, 2, \ldots$. It is therefore proper to study $\sum_i X_i$. This is the ergodic theory, whose most famous result is the central limit theorem. Unlike this approach, statistics of extremes studies accidents rather than ageing. In many cases, the state of systems is due to one particular event rather than to the accumulation of many: when the pot is broken, the number of drops falling into it is no longer relevant. From a mathematical viewpoint, the sum $\sum_i X_i$ is almost equal to the maximum $\max_i X_i$. It is therefore proper to study the most extreme values of the phenomenon. This is the basis of extreme value theory.

The main principle of this theory is to rarefy the X_i by rescaling. If this is done properly, universal behavior of the most extreme values appears. This approach is similar in principle to the central limit theorem although the limit distribution thus obtained is less tightly determined. Extreme value statistics fit this limit distribution to the most extreme among the available data. By nature, these extremes are sparse, so that information brought by the dataset may be small. Models link the most extreme behavior (maybe never observed) with that in the dataset (that is moderately extreme). This is a particularity in statistics since behavior is extrapolated from few data whereas typically it is interpolated from many.

In practice, univariate and multivariate statistics of extremes bring complementary information. Univariate statistics quantify the size of the extreme while multivariate statistics detail their dependence structure. The two following examples are enlightening:

- let us consider two nearby villages in mountains with rainfall measuring systems. One being next to the other, their rainfall is likely to be dependent. However, local storms may occur during the year only at one village at a time. Therefore, rainfall during

these storm periods are likely to be independent between the two villages. However the general levels of extreme rainfall at the two villages are likely to be comparable;

- let us consider two firms belonging to two different but linked economic sectors. Management being independent a priori, their returns are likely to be independent also. However, if the stock exchanges in these two sectors are perturbed by similar crashes, we would observe a strong dependence in the extreme returns of these two firms. However these returns are not comparable a priori.

Although maybe unrealistic, these two examples illustrate the following important ideas:

- two dependent phenomena may be independent in the extremes and vice versa;
- being extreme is not in relation with the size of the observation but with its scarcity, in particular for multivariate data.

In general, a multivariate analysis of extremes is preceded by a standardization of the margins of the data on a common scale. We afterward fit a probability measure, called the spectral measure or spectral distribution, which characterizes the dependence structure of the data. This measure is a central element of the multivariate extreme statistics. This thesis addresses the fitting of this distribution, its use and its generalization to spatial extreme statistics. Chapter 1 is a review of the univariate and multivariate extremes. Chapter 2 presents a new semi-parametric model for the spectral distribution and develops two algorithms in order to fit it. Chapter 3 applies the model with the two algorithms to simulated and real data and studies their performance. Chapter 4 deals with the use of the spectral distribution function. Two issues are addressed: the estimation of the probability of rare events using Monte Carlo methods and the analysis of the dependence structure of the extremes. Chapter 5 presents a generalization of the multivariate extremes to spatial extremes.

Chapter 1

Univariate and Multivariate extremes

1.1 History

An historical survey on univariate extreme value distributions can be found in Kotz & Nadarajah (2000). The history goes back to 1709 and Nicolas Bernoulli discussing the mean of the largest distance among points lying at random on a line. The notion of the distribution of the largest value is more modern and was first introduced by von Bortkiewicz (1922). Fréchet (1927) identified one possible limit distribution for largest order statistics and, in the next year, Fisher & Tippett (1928) showed that these distributions can only be of three types. von Mises (1936) presented sufficient conditions for the convergence toward each of these types and Gnedenko (1943) gave a rigorous foundation of extreme value theory with necessary and sufficient conditions for weak convergence. The late 1930s and 1940s were marked by a number of papers dealing with practical applications of extreme value theory, among which are Weibull (1939) studying strength of materials and Gumbel with a large number of papers culminating with his book, Gumbel (1958). As pointed out by Kotz & Nadarajah (2000), the literature in extreme value analysis is now enormous and growing very quickly. To the authors, "while this extensive literature serves as a testimony to the great vitality and applicability of the extreme value distributions and processes, it also unfortunately reflects on the lack of coordination between researchers and the inevitable duplication [...] of results appearing in a wide range of diverse publications". This lack of unification was already mentioned by Pickands (1971) where the author links extreme value theory with the convergence

of point processes. Statistical inference is developed in Pickands (1975) which justifies the use of the generalized Pareto distribution in threshold methods, commonly used by hydrologists. In parallel, methods based on several largest order statistics were proposed by Weissman (1978). These methods were developed afterward by several contributors; see Davison & Smith (1990). Galambos's (1978) monograph is one of the first reference books specifically dedicated to statistical models and treating also multivariate extremes. It is followed by Leadbetter, Lindgren & Rootzén (1983), a key reference, in which is formally presented extreme value theory for stationary sequences.

The multivariate theory is naturally more recent. Surveys of the literature can be found in Galambos (1978) and Coles (2001). It goes back to, once more, a Russian contribution by Finkelstein (1953). Later, independently of one another, three works, Geoffroy (1958/1959), Tiago de Oliveira (1958) and Sibuya (1960), appeared on bivariate extremes and gave a representation of the max-stable limit distribution of standardized componentwise maxima. The first point process argument goes back to de Haan & Resnick (1977) and Pickands (1981) who gave an equivalent representation. Resnick (1987) is a key book for the point process theory applied to extreme value analysis. The development of parametric families for the componentwise maximum approach is due to Tawn (1988) and the use of the point process approach is due to Coles & Tawn (1991) and Joe, Smith & Weissman (1992). Development of non-parametric estimation goes back to Deheuvels & Tiago de Oliveira (1989). Works on stationary multivariate extremes have been mainly due to Hsing (1989), generalizing Leadbetter et al.'s (1983) conditions under which stationary series behave like independent ones.

Below are presented some main concepts in extreme value theory and its applications. These concepts can be found in the numerous reference books available, among which are Leadbetter et al. (1983), Tiago de Oliveira (1984), Resnick (1987), Embrechts, Klüppelberg & Mikosch (1997), Kotz & Nadarajah (2000), Coles (2001), Reiss & Thomas (2001) and Finkenstädt & Rootzén (2004). The next section presents a limited background on point processes useful for univariate and multivariate cases. Section 1.3 details univariate and Section 1.4 multivariate extremes.

1.2 Point processes for extremes

Below is given very brief background and results of the point process theory useful for extremes. Chapter 4 contains a more general and detailed overview than the one below.

It is extracted from Jagers (1974), Kallenberg (1983) and Resnick (1987) and serves the application for spatial extremes.

The basic working space is $\mathbb{R}^d$ equipped with $\mathcal{S}$, the σ-algebra generated by bounded rectangles. Let $\mathcal{T}$ be the collection of finite unions of these bounded rectangles. The simplest probability measure in $\mathbb{R}^d$ is the Dirac mass at a point x,

$$\delta_x(A) = \mathbb{1}_A(x), \quad A \in S,$$

where $\mathbb{1}_A$ is the indicator function of the set A. Consequently, the simplest random probability measure is the Dirac mass at a random point $X \in \mathbb{R}^d$. This is a basic unit for construction of point processes. Let $X_1, \ldots, X_n$ be an independent sample from F, then its sample point process is $\sum_{i=1}^n \delta_{X_i}$. It is an example of a finite random measure, provided that n is finite. In general a point process N is a random measure such that $N(A)$ is $\mathbb{Z}_+$-valued (possibly ∞) for any $A \in \mathcal{S}$. A classical issue in point process theory is the convergence of sample point processes as $n \to \infty$. A possible limit is the Poisson process N, defined as a point process such that for any finite disjoint collection $A_1, \ldots, A_p \in \mathcal{T}$, the vector $\{N(A_1), \ldots, N(A_p)\}$ has independent components distributed according to Poisson distributions with parameters $\lambda(A_1), \ldots, \lambda(A_p)$, respectively, where λ is a Radon measure, that is a measure finite on every bounded set in $\mathbb{R}^d$. In general, for a point process N, the family of distributions of $\{N(A_1), \ldots, N(A_p)\}$ for every finite collection $A_1, \ldots, A_p \in \mathcal{T}$ is called the finite dimensional distribution of N. It uniquely defines the point process distribution. In the same vein, a sequence of point process $\{N_n\}$ is said to converge in distribution to N, if, for any finite collection $A_1, \ldots, A_p \in \mathcal{T}$,

$$\{N_n(A_1), \ldots, N_n(A_p)\} \xrightarrow{d} \{N(A_1), \ldots, N(A_p)\},$$

where d is the classical convergence in distribution of random vectors.

The following result is a direct consequence of Jagers (1974, p.233) or can be found in Resnick (1987, p.154) in a slightly different formulation. Below, ∂A denotes the boundary of A.

Theorem 1.1

For each n, let $\{X_{n,i}\}_{i=1}^n$ be an independent and stationary sample in $\mathbb{R}^d$. Then the sample point process

$$N_n = \sum_{i=1}^n \delta_{X_{n,i}}$$

converges in distribution to the Poisson process N with intensity λ if and only if, for every $A \in \mathcal{A}$,

$$nP\left(X_{n,1} \in A\right) \overset{n\to\infty}{\longrightarrow} \lambda(A),$$

where $\mathcal{A} \subset \mathcal{S}$ is an algebra containing some basis and such that $\lambda(\partial A) = 0$.

In $\mathbb{R}^d$, $\mathcal{A}$ can be the collection of finite unions and intersections of rectangles $(x_1, \infty) \times \cdots \times (x_d, \infty)$. This key result is usually used as a corollary presented below. Therein, for any $a \in \mathbb{R}^d_+$ and $b \in \mathbb{R}^d$, the set $aA + b$ equals $\{x \in A : a^{-1}(x - b) \in A\}$, where additions, multiplications, inverses and comparisons are done componentwise.

Theorem 1.2

Let $X, X_1, X_2, \ldots,$ be an independent and stationary sample in $\mathbb{R}^d$. Then the following statements are equivalent:

(i) there exist sequences $\{a_n\} \subset \mathbb{R}^d_+$ and $\{b_n\} \subset \mathbb{R}^d$ such that for any $A \in \mathcal{A}$,

$$nP\left(X \in a_n A + b_n\right) \quad \overset{n\to\infty}{\longrightarrow} \quad \lambda(A), \tag{1.1}$$

where $\mathcal{A}$ is as in Theorem 1.1;

(ii) the sample point process

$$N_n = \sum_{i=1}^{n} \delta_{a_n^{-1}(X_i - b_n)}$$

converges in distribution to a Poisson process N with intensity λ.

The next section presents consequences of this result for univariate extremes when $d = 1$. A central issue is the study of sequences $\{a_n\}$ and $\{b_n\}$ that provide limit distributions useful for statistical inference.

1.3 Univariate extremes

In the univariate case, $\mathcal{A}$ can be reduced to the collection of intervals of the form (x, ∞), $x \in \mathbb{R}$, so that condition (1.1) reduces to

$$n\left\{1 - F(a_n x + b_n)\right\} \overset{n\to\infty}{\longrightarrow} \tau(x),$$

where F is the distribution function of X and τ is some positive function. Possible forms of τ are given by the following argument:

$$\begin{aligned} n\{1 - F(a_n x + b_n)\} \quad &= \quad n\{1 - F(a_n + b_n)\}\frac{n\{1 - F(a_n x + b_n)\}}{n\{1 - F(a_n + b_n)\}} \\ &\overset{n\to\infty}{\longrightarrow} \quad \tau(1) \lim_{n\to\infty} \frac{1 - F(a_n x + b_n)}{1 - F(a_n + b_n)}, \end{aligned}$$

that is

$$\tau(x) = \tau(1) \lim_{n\to\infty} \frac{1 - F(a_n x + b_n)}{1 - F(a_n + b_n)}, \quad \forall x.$$

A simple example is when F is the unit Fréchet distribution, $F(x) = e^{-1/x}$, $x > 0$, for which one can take $a_n = n$ and $b_n = 0$. Then

$$\tau(x) = \lim_{n\to\infty} \frac{1 - F(nx)}{1 - F(n)} = x^{-1},$$

and hence

$$\tau(x) = \tau(1)/x.$$

In general, the existence of $\{a_n > 0\}$ and $\{b_n\}$ such that a non-degenerate limit exists is not guaranteed. For discrete F, the Poisson or the geometric distributions are classical examples. For continuous F, an example is $F(x) = 1 - 1/\log x$, $x \geq e$, but most classical distribution functions admit such sequences. Three forms are possible and the necessary and sufficient conditions under which each of them holds can be found in Leadbetter et al. (1983, pp. 17–19), from which the following lines are extracted but with the change of notation $a_n = a_n^{-1}$. In each case, τ is given by the right part of the limit equation. The end-point of the distribution function F is denoted $x_F = \sup\{x : F(x) < 1\}$.

Theorem 1.3 (Extremal Types Theorem)

(*i*) *(Type I or Gumbel) There exists some strictly positive function* $g(t)$ *such that*

$$\lim_{t\uparrow x_F} \frac{1 - F\{t + xg(t)\}}{1 - F(t)} = e^{-x}, \quad \textit{for all } x \in \mathbb{R}.$$

In this case, $\int_0^\infty \{1 - F(u)\}du < \infty$ *and a possible choice for* g *is* $g(t) = \int_t^{x_F} \{1 - F(u)\}du/\{1 - F(t)\}$.

(*ii*) *(Type II or Fréchet)* $x_F = \infty$ *and there exists* $\alpha > 0$ *such that*

$$\lim_{t\uparrow\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-\alpha}, \quad \textit{for all } x > 0.$$

(*iii*) *(Type III or Weibull)* $x_F < \infty$ *and there exists* $\alpha > 0$ *such that*

$$\lim_{h\downarrow 0} \frac{1 - F(x_F - hx)}{1 - F(x_F - h)} = x^{\alpha}, \quad \textit{for all } x > 0.$$

In each case, let $\gamma_n = F^{-1}(1 - 1/n) = \inf\{x : F(x) \geq 1 - 1/n\}$. *Then the sequences* $\{a_n > 0\}$ *and* $\{b_n\}$ *can be chosen to be*

(*i*) *(Type I)* $a_n = g(\gamma_n)$ *and* $b_n = \gamma_n$;

(*ii*) *(Type II)* $a_n = \gamma_n$ *and* $b_n = 0$;

(*iii*) *(Type III)* $a_n = x_F - \gamma_n$ *and* $b_n = x_F$.

Distribution of the maximum

The distribution of the standardized maximum of a stationary and independent sample, $\{X_i\}_{i=1}^n$, is characterized by the Poisson limit result. Let N_n be the sample process of $\{X_i\}_{i=1}^n$, let N be a Poisson process with intensity τ and let $M_n = \max\{X_1, \ldots, X_n\}$. Applying Theorem 1.2 in the univariate case, if there exist sequences $\{a_n\} \subset \mathbb{R}_+$ and $\{b_n\} \subset \mathbb{R}$ such that

$$n\{1 - F(a_n x + b_n)\} \overset{n\to\infty}{\longrightarrow} \tau(x),$$

then

$$P\{a_n^{-1}(M_n - b_n) \leq x\} = P\{N_n(x) = 0\} \overset{n\to\infty}{\longrightarrow} P\{N(x) = 0\} = \exp\{-\tau(x)\}.$$

This result can also be seen in the following way,

$$\log P\{a_n^{-1}(M_n - b_n) \leq x\} = n \log F(a_n x + b_n) \approx -n\{1 - F(a_n x + b_n)\} \overset{n\to\infty}{\longrightarrow} -\tau(x).$$

Therefore, the possible limit distribution function for the standardized maxima is $\exp\{-\tau(x)\}$. Replacing $\tau(x)$ by its possible shapes gives the three types of extreme value distribution function. In practice, one never knows F, $\{a_n\}$, or $\{b_n\}$. Assuming that the asymptotic regime has been reached, one observes the maximum M_n and fits the distribution function

$$P(M_n \leq x) = \exp\left\{-\tau(a_n x + b_n)\right\}.$$

The three forms of τ can be embedded into the Generalized Extreme Value distribution.

Definition 1.4 (Generalized Extreme Value Distribution)
The generalized extreme value distribution is

$$G(x;\, \mu, \sigma, \kappa) = \begin{cases} \exp\left[-\left\{1 + \dfrac{\kappa}{\sigma}(x-\mu)\right\}^{-1/\kappa}\right], & 1 + \dfrac{\kappa}{\sigma}(x-\mu) > 0, \quad \kappa \neq 0, \\ \exp\left[-\exp\left\{-\dfrac{(x-\mu)}{\sigma}\right\}\right], & x \in \mathbb{R}, \qquad\qquad \kappa = 0. \end{cases}$$

The location, scale and shape parameters are respectively μ, σ and κ.

Thus $\tau(x)$ admits a generalized form. In practice, the estimation of μ, σ and κ requires one to decompose the sample into blocks and take the blockwise maxima. A drawback with this approach is the loss of data. An improvement is to consider the asymptotic simultaneous distribution of the highest order statistics.

Distribution of the highest order statistics

Let $X_1, \ldots, X_n$ be an independent and stationary sample, $\tilde{X}_1, \ldots, \tilde{X}_n$ the standardized sample, where $\tilde{X}_i = a_n^{-1}(X_i - b_n)$, and suppose that the Poisson limit applies. Furthermore, let $\tilde{X}_{(n)} \leq \cdots \leq \tilde{X}_{(1)}$ be the order statistics of the standardized sample. Then, for any fixed r,

$$P\{\tilde{X}_{(n-r)} < u\} = \exp\{-\tau(u)\} \sum_{k=0}^{r-1} \frac{\tau(u)^k}{k!}.$$

This gives the distribution function of the r-th largest order statistic. Furthermore, for $x_n > \cdots > x_{n-r+1} > u$, there is an $h > 0$ sufficiently small that intervals $(-\infty,\ u], (x_{n-r+1} - h,\ x_{n-r+1}],\ \ldots,\ (x_n - h,\ x_n]$ are disjoint. In this case, the Poisson limit implies that

$$\begin{aligned} P\Big\{\tilde{X}_{(n)} \in (x_n - h,\ x_n], \ldots,\ \tilde{X}_{(n-r+1)} \in (x_{n-r+1} - h,\ x_{n-r+1}],\ \tilde{X}_{(n-r)} < u\Big\} \\ = \{\tau(x_n) - \tau(x_n - h)\} \cdots \{\tau(x_{n-r+1}) - \tau(x_{n-r+1} - h)\} \exp\{-\tau(u)\}, \end{aligned}$$

so

$$\begin{aligned} P\Big\{x_n - h < \tilde{X}_{(n)} \leq x_n, \ldots, x_{n-r+1} - h < \tilde{X}_{(n-r+1)} \leq x_{n-r+1} \mid \tilde{X}_{(n-r)} < u\Big\} \\ \propto \{\tau(x_n) - \tau(x_n - h)\} \cdots \{\tau(x_{n-r+1}) - \tau(x_{n-r+1} - h)\} \exp\{-\tau(u)\}. \end{aligned}$$

Dividing by h^r and letting $h \to 0$ we obtain the simultaneous density of the r-th largest order statistics given that there is no other value above u, that is,

$$\tau'(x_n) \cdots \tau'(x_{n-r+1}) \exp\{-\tau(u)\}.$$

Setting $u = x_{n-r+1}$ and using the generalized form of τ, the following log-likelihood can be built: if $\kappa \neq 0$, then

$$\ell(\mu, \sigma, \kappa) = -n \log \sigma - \left(1 + \kappa \frac{x_{n-r+1} - \mu}{\sigma}\right)^{-\frac{1}{\kappa}} - \sum_{j=1}^{r} (1 + 1/\kappa) \log\left(1 + \kappa \frac{x_{n-j+1} - \mu}{\sigma}\right),$$

or, in the case $\kappa = 0$,

$$\ell(\mu, \sigma) = -n \log \sigma - \exp\left(-\frac{x_{n-r+1} - \mu}{\sigma}\right) - \sum_{j=1}^{r} \frac{x_{n-j+1} - \mu}{\sigma}.$$

The domain of the parameters is $\sigma > 0$, $1 + \kappa(x_{n-j+1} - \mu)/\sigma > 0$ for $j = 1, \ldots, r$.

In order to obtain good inferences, r must be selected appropriately. If r is too large the inference is only based on few observations, but if r is too small, the Poisson process may be a poor approximation of reality. Another equivalent possibility is to choose a threshold u and work with the excesses over u. This is the Peaks Over Threshold method.

Peaks Over Threshold

If X is distributed according to F and $\{a_n > 0\}$ and $\{b_n\}$ are its standardizing sequences, then for any $y > 0$ and any u,

$$P\left(\frac{X - b_n}{a_n} > y + u \,\middle|\, \frac{X - b_n}{a_n} > u\right) = \frac{1 - F\{a_n(y+u) + b_n\}}{1 - F(a_n u + b_n)} \overset{n\to\infty}{\longrightarrow} \frac{\tau(y+u)}{\tau(u)}.$$

In order to estimate a_n and b_n, one uses the generalized form of τ and obtains

$$\frac{\tau(y+u)}{\tau(u)} = \left(\frac{1 + \kappa\sigma^{-1}(y + u - \mu)}{1 + \kappa\sigma^{-1}(u - \mu)}\right)^{-1/\kappa} = \left(1 + \frac{\kappa}{\tilde{\sigma}} y\right)^{-1/\kappa},$$

where $\tilde{\sigma} = \sigma + \kappa(u - \mu)$.

Definition 1.5 (Generalized Pareto Distribution)

The Generalized Pareto distribution is

$$H(y;\, \sigma,\, \kappa) \;=\; \begin{cases} 1 - \left(1 + \dfrac{\kappa}{\sigma} y\right)^{-1/\kappa}, & 1 + \dfrac{\kappa}{\sigma} y > 0, \quad \kappa \neq 0, \quad y \geq 0 \\ \\ 1 - \exp\left(-\dfrac{y}{\sigma}\right), & \kappa = 0, \quad y \geq 0. \end{cases} \tag{1.2}$$

The parameters σ and κ are respectively scale and shape parameters.

In practice, to fit this, a high threshold u is selected and a generalized Pareto distribution is fitted to excesses of data above u. If u is sufficiently high that the Poisson limit is valid, then N_u, the number of excesses over u among $\{X_1, \ldots, X_n\}$, is approximately Poisson with parameter λ. The observed n_u estimates λ and $n^{-1}n_u$ estimates $P(X > u)$. The corresponding log-likelihood separates into two parts,

$$\ell(\lambda, \sigma, \kappa) = \ell(\lambda) + \ell(\sigma, \kappa),$$

which allows separate inferences on the parameters λ and $(\sigma,\, \kappa)$. In detail, each part of the log-likelihood is

$$\ell(\lambda) = -\lambda + n_u \log \lambda - \log n_u!,$$

and

$$\ell(\sigma, \kappa) = \begin{cases} -n_u \log \sigma - (1 + 1/\kappa) \displaystyle\sum_{i=1}^{n_u} \log(1 + \kappa\sigma^{-1} y_i), & \text{if } \kappa \neq 0, \\ \\ -n_u \log \sigma - \displaystyle\sum_{i=1}^{n_u} \sigma^{-1} y_i, & \text{if } \kappa = 0. \end{cases}$$

The Peaks Over Threshold method can be represented as a semi-parametric model. The excesses above a high threshold u are distributed according to a generalized Pareto distribution while the empirical distribution function $\hat{F}$, or any other appropriate model, is

used under u. This is the semi-parametric extremal model, see for example Coles & Tawn (1991).

Definition 1.6 (Semi-parametric extremal model)

Let $X, X_1, \ldots, X_n$ be independent and identically distributed according to F. Let $\hat{F}$ be the empirical distribution function of $X_1, \ldots, X_n$ and u a high threshold such that the Peaks Over Threshold model applies. Then the semi-parametric extremal model is the distribution function

$$\tilde{F}(x) = \begin{cases} \hat{F}(x), & \text{for } x \leq u, \\ (1-p_u) + p_u \left[1 - \left\{1 + \frac{\kappa}{\sigma}(x-u)^{-1/\kappa}\right\}\right], & \text{for } x > u, \end{cases}$$

where $p_u = P(X > u)$.

The choice of the threshold u involves a bias-variance trade-off. If the threshold is too high, estimation is based on very few data and is unlikely to be reliable, whereas if it is too low, the Pareto model is unlikely to be true and, although numerous, the excesses are not representative of the asymptotic behavior of X. For this reason, selection of the threshold is often based on choosing the lowest u such that the Pareto hypothesis seems reliable. If excesses of X above u_0 are Pareto with parameters σ and κ, then, for $u > u_0$,

$$\begin{aligned} P(X - u > y \mid X > u) &= \frac{P(X > y + u \mid X > u_0)}{P(X > u \mid X > u_0)} \\ &= \frac{\{1 + \sigma^{-1}\kappa(y - u_0 + u)\}^{-1/\kappa}}{\{1 + \sigma^{-1}\kappa(u - u_0)\}^{-1/\kappa}} \\ &= [1 + \{\sigma + \kappa(u - u_0)\}^{-1}\kappa y]^{-1/\kappa}. \end{aligned}$$

Therefore, excesses of X above $u > u_0$ are generalized Pareto with parameters $\sigma + \kappa(u - u_0)$ and κ. The mean of (1.2) is $(1+\kappa)^{-1}\sigma$, if $\kappa < 1$, and infinite otherwise. Hence, for any $u > u_0$,

$$E(X - u \mid X > u) = \frac{\sigma + \kappa(u - u_0)}{1 + \kappa},$$

and $E(X - u \mid X > u)$ is a linear function of u. This result is the basis for a graphical diagnostic known as the Mean Residual Life Plot.

Definition 1.7 (Mean Residual Life Plot)

The Mean Residual Life Plot is the graph of an empirical estimator of $E(X - u \mid X > u)$ versus u,

$$\left(u \, , \; n_u^{-1} \sum_{i : x_i > u} x_i - u \right).$$

Above a good threshold u_0, the graph should be linear with slope $(1+\kappa)^{-1}\kappa$. As the threshold increases, the empirical estimate of $E(X-u \mid X>u)$ becomes more and more variable so that the graph jitters and inference becomes difficult. Furthermore, such a diagnostic should not be used when $\kappa \geq 1$.

Extreme quantiles and return levels

In practice, scientific interest is typically focused not directly on the parameters (μ, σ, κ) or $(\lambda, \sigma, \kappa)$ but rather on the question 'what level will be exceeded with a given low probability?' or equivalently 'what is x_β, the solution of $\beta = P(X > x_\beta)$, for a very small probability β?'

If β is a low probability, then the solution to $G(x_\beta) = 1-\beta$ for the Generalized Extreme Value distribution is

$$x_\beta = \begin{cases} \mu - \kappa^{-1}\sigma[1-\{\log(1-\beta)\}^{-\kappa}], & \text{if } \kappa \neq 0, \\ \mu - \sigma\log\{-\log(1-\beta)\}, & \text{if } \kappa = 0. \end{cases}$$

For the Peaks Over Threshold method, let p_u be the probability of being above a high threshold u. Then x_β satisfies

$$\beta = P(X > x_\beta) = P(X > x_\beta \mid X \leq u)(1-p_u) + P(X > x_\beta \mid X > u)p_u.$$

If β is smaller than p_u, then $x_\beta > u$ and $P(X > x_\beta \mid X \leq u) = 0$. Consequently,

$$p_u^{-1}\beta = P(X - u > x_\beta - u \mid X > u).$$

Hence, solving $1 - G(y_\beta) = p_u^{-1}\beta$ for the generalized Pareto distribution, one obtains

$$y_\beta = \begin{cases} \left\{\left(p_u^{-1}\beta\right)^{-\kappa} - 1\right\}\kappa^{-1}\sigma, & \text{if } \kappa \neq 0, \\ -\sigma\log(p_u^{-1}\beta), & \text{if } \kappa = 0, \end{cases}$$

and the return level is $x_\beta = y_\beta + u$. Estimates of y_β can be obtained by substituting estimates of σ and κ, the estimate of p_u being $n^{-1}n_u$, the number of excesses over the number of data.

Statistical inference

Methodologies for statistical inference have been addressed in numerous places; see for example Davison & Smith (1990) and Coles (2001).

Once the threshold is selected, estimation can be based on the likelihood. The shape parameter κ is the same whether the Peaks Over Threshold or the order statistic method is used, whereas the scale parameters are linked by

$$\sigma_{\mathrm{POT}} = \sigma_{\mathrm{GEV}} + \kappa(u - \mu).$$

Here σ_{POT} is the scale parameter of the generalized Pareto distribution fitted in a Peaks over Threshold method with a threshold u and σ_{GEV} is the scale parameter of the generalized extreme value distribution fitted to the maxima of the data. In both cases, the maximum likelihood estimator is consistent provided that $\kappa > -1$ and is asymptotically normal and efficient only for $\kappa > -1/2$ (Smith 1985). This failure of likelihood based methods can be illustrated by the following fact (Embrechts et al. 1997, p.357). Let $\hat{\sigma}$ and $\hat{\kappa}$ be the maximum likelihood estimators of σ and κ respectively from an independent and identically distributed sample of size n, then one can show that

$$n^{1/2}\left(\hat{\kappa} - \kappa,\, \frac{\hat{\sigma}}{\sigma} - 1\right) \xrightarrow{d} \mathcal{N}(0, M^{-1}), \quad n \to \infty,$$

where

$$M^{-1} = (1+\kappa)\begin{pmatrix} 1+\kappa & -1 \\ -1 & 2 \end{pmatrix}.$$

Then as κ tends to $-1/2$, M^{-1} becomes singular and the limit normal distribution becomes degenerate. In general, having $\kappa \leq -1/2$ requires other methods, like Bayesian procedures, which have been studied in Coles & Powell (1996). If one does not have any prior information on the parameters, scientific experts may be able to quantify prior information on return levels. The idea is to parametrize the likelihood as a function of return levels and compute the posterior distribution. An advantage of this approach, beyond the use of prior information, is that posterior inference does not suffer from any limitation on κ.

Non-parametric methods have also been developed and represent an important part of the literature. For example, one can find the Hill-plot and probability weighted moment estimators in Embrechts et al. (1997). We do not detail them since they are not central to this work.

Dependent stationary sequences and cluster analysis

Stationary sequences have been extensively studied with a particular attention to conditions under which the limit distribution of the maximum remains a generalized extreme

value distribution and the Poisson process remains valid. Let $X_1, \ldots, X_n$ be a stationary sequence of random variables with common marginal distribution function F. Let F_I denote the simultaneous distribution function of $\{X_i\}_{i \in I}$, where I is a subset of $\{1, \ldots, n\}$, and let $F_I(x) = F_I(x, \ldots, x)$. Now define the following mixing condition:

Definition 1.8 (Condition D($\mathbf{u_n}$))

Condition $D(u_n)$ holds for the sequence u_n if for any integers

$$1 \leq i_1 \leq \cdots \leq i_p \leq j_1 \leq \cdots \leq j_{p'} \leq n$$

for which $j_1 - i_p \geq l$,

$$\left| F_{i_1,\ldots,i_p,j_1,\ldots,j_{p'}}(u_n) - F_{i_1,\ldots,i_p}(u_n) F_{j_1,\ldots,j_{p'}}(u_n) \right| \leq \alpha_{n,l},$$

where $\alpha_{n,l_n} \to 0$ as $n \to \infty$ for some sequence $l_n = o(n)$.

Leadbetter (1974) showed the following result:

Theorem 1.9

Let $\{X_i\}_{i=1}^n$ be a stationary sequence, let $M_n = \max\{X_1, \ldots, X_n\}$, and $\{a_n > 0\}$ and $\{b_n\}$ be such that $P\left\{a_n^{-1}(M_n - b_n) \leq x\right\}$ converges in distribution to a non-degenerate distribution function $G(x)$. Suppose that $D(u_n)$ is satisfied for $u_n = a_n x + b_n$, for each x such that $G(x) > 0$. Then $G(x)$ is a generalized extreme value distribution.

The effect of dependence in the sequence is detailed by Leadbetter (1983). Below, let $\{\hat{X}_i\}_{i=1}^n$ be an independent sequence with the same marginal distribution as $\{X_i\}_{i=1}^n$ and denote $\hat{M}_n = \max\{X_1, \ldots, X_n\}$.

Theorem 1.10

Suppose that there exist sequences $\{a_n > 0\}$ and $\{b_n\}$ such that $P\left\{a_n^{-1}\left(\hat{M}_n - b_n\right) \leq x\right\}$ converges in distribution to a non-degenerate distribution function $G(x)$. Suppose that $\{X_i\}_{i=1}^n$ satisfies $D(u_n)$ for $u_n = a_n x + b_n$, for each x such that $G(x) > 0$. Then there exists $0 \leq \theta \leq 1$ such that $P\left\{a_n^{-1}\left(\hat{M}_n - b_n\right) \leq x\right\}$ converges in distribution to $G^\theta(x)$.

The parameter θ is termed the extremal index. It varies from zero to one according to the strength of the dependence, the case $\theta = 0$ being degenerate. The case $\theta = 1$ corresponds to very weak dependence and is ensured by the following cluster condition.

Definition 1.11 (Condition D′($\mathbf{u_n}$))

Condition $D'(u_n)$ holds for a stationary sequence $\{X_j\}$ if

$$\limsup_{n \to \infty} n \sum_{j=2}^{[n/k]} P\{X_1 > u_n,\, X_j > u_n\} \longrightarrow 0, \quad \text{as } k \to \infty.$$

The intuitive interpretation of the extremal index is as the cluster rate or the inverse of the mean cluster size. In the absence of condition D', the dependence structure is such that a large value has a greater chance of being followed by another one. If the time between two consecutive such values is small relative to n, the passage to the limit will merge those two extremes onto the same time. The limit process is then not a Poisson process but a compound Poisson process: any occurrence can be multiple rather than single. The multiplicity is usually random and is called the cluster size distribution. Under certain conditions (Hsing, Hüsler & Leadbetter 1988) the extremal index is the inverse of the mean cluster size, that is, the rate of arrival of clusters.

This approach is a basis for an estimation of the cluster size distribution. Hsing et al. (1988) defined

Definition 1.12 (Cluster Size Distribution)

Let

$$\pi_n(j) = P\left\{ \sum_{i=1}^{r_n} \mathbb{1}_{\{X_i > u_n\}} = j \;\middle|\; \sum_{i=1}^{r_n} \mathbb{1}_{\{X_i > u_n\}} > 0 \right\}, \quad \textit{for } j = 1, 2, \ldots,$$

for a sequence $r_n = o(n)$. Under conditions guaranteeing the compound Poisson model, there exists a sequence $k_n \to \infty$ such that if $r_n = [n/k_n]$ then

$$\pi(j) = \lim_{n\to\infty} \pi_n(j), \quad \textit{for } j = 1, 2, \ldots,$$

is the cluster size distribution.

Under some further conditions on π, the authors show that

$$\theta = \sum_{j \geq 1} j\pi(j).$$

A natural way to estimate π and θ is to choose r, group the data into blocks and count the number of excesses over u_n in each group. An estimator of $\pi(j)$ is simply the number of blocks that have j excesses over the number of blocks that have one or more excesses. The mean of the resulting law is an estimator for θ^{-1}. For the estimation of θ alone, a method directly based on the compound Poisson limit has been developed by Ferro & Segers (2003) in order to avoid an arbitrary choice of r.

1.4 Multivariate extremes

As in the univariate case, Theorem 1.1 is one basis of statistical techniques but naturally a supplementary aspect arises in the multivariate case. In general, data are standardized

to be marginally Fréchet so that $a_n = n$ and $b_n = 0$. The standardization is done for example using the semi-parametric extremal model on every margin. Let $X_1, \ldots, X_n$ be an independent and stationary sample in $\mathbb{R}^d$, marginally distributed according to a unit Fréchet distribution, $\exp(-1/x)$, $x > 0$. Then

$$N_n = \sum_{i=1}^{n} \delta_{n^{-1}X_i} \xrightarrow{d} N,$$

a Poisson process with intensity λ, if and only if

$$nP\left(X \in nA\right) \overset{n\to\infty}{\longrightarrow} \lambda(A),$$

for any $A \in \mathcal{A}$. Hence, for any $t > 0$,

$$\lambda(tA) = \lim_{n\to\infty} nP\left\{X \in n(tA)\right\} = \lim_{nt\to\infty} \frac{nt}{t} P\left\{X \in (nt)A\right\} = t^{-1}\lambda(A).$$

In other words, λ is homogeneous of degree -1. This implies that the image of λ through the transformation

$$x \longmapsto \begin{cases} (r,\, w)\,, & x \neq 0, \\ (0,\, 0)\,, & x = 0, \end{cases}$$

where $r = \|x\|$ and $w = x/\|x\|$, is the product

$$\lambda(dx) = \frac{1}{r^2} dr \times \tilde{H}(dw),$$

for a positive Radon measure $\tilde{H}$ on the simplex $\mathcal{S}_d = \left\{w \in [0,1]^d : \|w\| = 1\right\}$. This measure $\tilde{H}$ is called the spectral measure. A popular choice of $\|\cdot\|$ is the pseudo-polar scale,

$$r = \sum_{j=1}^{d} x^{(j)} \quad \text{and} \quad w^{(j)} = x^{(j)}/r, \quad j = 1, \ldots, d,$$

and $\mathcal{S}_d = \left\{w \in [0,1]^d : \sum_{j=1}^{d} w^{(j)} = 1\right\}$; see for example Coles & Tawn (1994).

The spectral measure has total mass equal to d and satisfies a mean condition

$$\int_{\mathcal{S}_d} w^{(j)} \tilde{H}(dw) \;\; = \;\; 1, \quad j = 1, \ldots, d. \tag{1.3}$$

This is due to the Fréchet margin requirement and can be deduced from the multivariate distribution of the componentwise maximum presented below.

Distribution of the componentwise maximum

Let M_n be the componentwise maximum of $n^{-1}X_1, \ldots, n^{-1}X_n$. Then

$$P\{M_n \leq x\} = P\{n^{-1}X_1 \leq x, \ldots n^{-1}X_n \leq x\} = P\{N_n(A) = 0\}.$$

By passage to the limit,

$$P\{N_n(A) = 0\} \overset{n\to\infty}{\longrightarrow} P\{N(A) = 0\} = \exp\{-\lambda(A)\},$$

where

$$A = \{rw \leq x\}^c = \left\{\exists j : rw^{(j)} > x^{(j)}\right\} = \left\{r > \min_{j=1,\ldots,d} \frac{x^{(j)}}{w^{(j)}}\right\}.$$

Therefore

$$\begin{aligned}
\exp\{-\lambda(A)\} &= \exp\left\{-\int_A \frac{1}{r^2} dr\tilde{H}(dw)\right\} \\
&= \exp\left\{-\int_{\mathcal{S}_d}\int_{r>\min\{x^{(j)}/w^{(j)}\}} \frac{1}{r^2} dr\tilde{H}(dw)\right\} \\
&= \exp\left\{-\int_{\mathcal{S}_d} \max_{j=1,\ldots,d} \frac{w^{(j)}}{x^{(j)}} \tilde{H}(dw)\right\}.
\end{aligned}$$

This result is known as the multivariate extreme value theorem and this distribution is termed the multivariate extreme value distribution.

As a by-product, the asymptotic jth marginal distribution of M_n is

$$\lim_{n\to\infty} P\{M_n^{(j)} \leq x^{(j)}\} = \exp\left\{-\frac{1}{x^{(j)}}\int_{\mathcal{S}_p} w^{(j)}\tilde{H}(dw)\right\}.$$

The data are marginally Fréchet, so $M_n^{(j)}$ is itself asymptotically distributed according to a unit Fréchet, that is

$$\exp\left\{-\frac{1}{x^{(j)}}\right\} = \exp\left\{-\frac{1}{x^{(j)}}\int_{\mathcal{S}_d} w^{(j)}\tilde{H}(dw)\right\}.$$

This clearly implies conditions (1.3) and summing them all gives that the total mass, $\tilde{H}(S_d)$, equals d. Therefore, the spectral measure can be rescaled into the spectral probability measure $H(dw) = d^{-1}\tilde{H}(dw)$. We summarize these results in a theorem.

Theorem 1.13 (Multivariate Extreme Value Theorem)
Let $X_1, \ldots, X_p$ be an independent and stationary sample in $\mathbb{R}^d$ with Fréchet margins and let $M_n = \max\{n^{-1}X_1, \ldots, n^{-1}X_n\}$. If there exists a positive finite measure λ such that, for any $A \in \mathcal{A}$,

$$nP(X \in nA) \overset{n\to\infty}{\longrightarrow} \lambda(A),$$

then

$$P(M_n \leq x) \overset{n\to\infty}{\longrightarrow} \exp\left\{-d\int_{\mathcal{S}_d} \max\left(\frac{w^{(1)}}{x^{(1)}}, \ldots, \frac{w^{(d)}}{x^{(d)}}\right) H(dw)\right\},$$

where the spectral probability measure H can be any distribution on $\mathcal{S}_d$ satisfying

$$\int_{\mathcal{S}_d} w^{(i)} H(dw) = d^{-1}, \quad j = 1, \ldots d.$$

The function

$$V(x) = d \int_{S_d} \max\left(\frac{w^{(1)}}{x^{(1)}}, \ldots, \frac{w^{(d)}}{x^{(d)}}\right) H(dw)$$

is called the dependence measure. Some models for H are defined through the corresponding dependence measure, which is homogeneous of order -1, $V(tx) = t^{-1}V(x)$, for all $t > 0$, so that the multivariate extreme value distribution $G(x) = \exp\{-V(x)\}$ is a simple max-stable distribution, that is $G(x/n) = G^n(x)$. In general, the distribution G is said to be max-stable if, for every $n \in \mathbb{N}$, there exist constants $a_n \in \mathbb{R}_+^d$ and $b_n \in \mathbb{R}^d$ such that

$$G^n(a_n x + b_n) = G(x).$$

It can be shown that the class of max-stable distributions with non-degenerate marginals coincides with the class of multivariate extreme value distributions (Resnick 1987, p.264).

Threshold method

As in the univariate case, inference based on componentwise maxima would represent a loss of data. The equivalent of the Peaks Over Threshold approach requires the choice of a sufficiently high threshold r_0. Then those values among $X_1, \ldots, X_n$ whose norm, $R = \|X\|$, exceeds r_0, are supposed to form a Poisson process. This provides a semi-parametric model for the joint distribution function F of X. In the set $\left\{x \in \mathbb{R}^d : \|x\| \leq r_0\right\}$, the empirical distribution function is reliable, while in the set $\left\{x \in \mathbb{R}^d : \|x\| > r_0\right\}$, the Poisson process is used. In general, norms other than $\|x\| = \sum_{i=1}^p x^{(i)}$ can be used. The threshold as well as spectral probability measure will be adapted according to this norm. For example, de Haan & de Ronde (1998) use the Euclidean norm $\|x\|_2^2 = \sum_{i=1}^p (x^{(i)})^2$ or $\|x\| = \max\left(x^{(1)}, \ldots, x^{(p)}\right)$. In fact any $\|x\|_a = (\sum_{i=1}^p |x^{(i)}|^a)^{1/a}$, for $1 \leq a \leq \infty$, can be used; the notation $a = \infty$ refers to the sup norm. An appropriate choice of the norm may be useful in practice. By varying the norm, one varies the shape of sets on which the probability is straightforwardly obtained after the selection of the threshold r_0.

As in the univariate case, the selection of the threshold involves a bias-variance trade-off. Too high a threshold gives unreliable inference and too a low threshold makes the Poisson process incompatible with the data. Therefore, the threshold r_0 should be the smallest value such that the Poisson process is valid. To our knowledge very few procedures have been proposed to select r_0. Coles & Tawn (1994) advise checking the independence of W and R by choosing r_0 such that the histogram of W stabilizes. However

this method is limited to $d = 2$. In Section 4.1 we present new graphical diagnostics to aid the selection of r_0 in any dimension.

Parametric inference on the spectral measure can be based on the likelihood function. The transformation to the pseudo-polar scale is

$$r_i = \sum_{j=1}^{d} x_i^{(j)}, \quad w_i^{(j)} = x_i^{(j)}/r_i, \quad j = 1, \ldots, d, \; i = 1, \ldots, n.$$

In the case where H has density h, the Poisson likelihood for observed data in $\{r > nr_0\}$ is

$$L(\alpha) \propto \prod_{i \in I_0} h(w_i),$$

where I_0 is the set of those $1 \leq i \leq n$ such that $r_i > r_0$. Appropriate maximum-likelihood estimation as well as Bayesian methods can be used in a standard way.

In the case where marginal parameters and spectral distribution are estimated together, the likelihood function becomes more complicated; see Coles & Tawn (1991). The semi-parametric extremal model transforms margins of original data $Y_1, \ldots, Y_n$ to the Fréchet scale by the transformation

$$X_i^{(j)} = \begin{cases} -1/\log\left[1 - p_j\left\{1 + \kappa_j \sigma_j^{-1}\left(Y_i^{(j)} - u_j\right)\right\}^{-1/\kappa_j}\right], & Y_i^{(j)} > u_j, \\ -1/\log\left\{\text{rank}\left(Y_i^{(j)}\right)/(n+1)\right\}, & Y_i^{(j)} \leq u_j, \end{cases}$$

where u_j is the threshold of the jth margin, p_j is the proportion of $Y_i^{(j)}$ exceeding u_j, σ_j and κ_j are the parameters corresponding to jth margin and rank $\left(Y_i^{(j)}\right)$ is the rank of $Y_i^{(j)}$ among $Y_1^{(j)}, \ldots, Y_n^{(j)}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, d$. Incorporating this transformation, the likelihood function of observations in a set A is

$$\exp\left\{-d \int_{A \cap A_0} \frac{r_0}{r} H(dw)\right\} \prod_{i \in I_0} h(w_i) r_i^{-(d+1)} \times \prod_{j \in I_0^{(i)}} \sigma_j^{-1} p_j^{-\kappa_j} \left(X_i^{(j)}\right)^2 \exp\left(1/X_i^{(j)}\right)\left\{1 - \exp\left(-1/X_i^{(j)}\right)\right\}^{1+\kappa_j},$$

where $A_0 = \{r > r_0\}$, I_0 is the set of $1 \leq i \leq n$ such that $r_i \in A \cap A_0$ and $I_0^{(i)}$ is the set of $1 \leq j \leq d$ such that $Y_i^{(j)} > u_j$, for $i = 1, \ldots, n$. Coles & Tawn (1991) use $A = \mathbb{R}_+^d \setminus \{(0, \nu_1) \times \cdots \times (0, \nu_d)\}$, for ν_j corresponding to u_j on the Fréchet scale. In Coles & Tawn (1994), $A = A_0$ so that the likelihood function is based on more data. Furthermore, they choose the marginal threshold to be the backward image of r_0 on each margin so that the exponential term is constant.

Models for the spectral measure

The spectral probability measure H is the distribution function of pseudo-polar angles of the most extreme data. When H gives probability one to the point $(d^{-1}, \ldots, d^{-1})$, the data are said to be asymptotically perfectly dependent. When H gives probability d^{-1} to each point of the form $(0, \ldots, 0, 1, 0, \ldots, 0)$, data are said to be asymptotically independent. Intermediate situations are called asymptotically dependent. Asymptotic independence has received particular attention in the literature and is treated later. Below we present models for asymptotic dependence.

A large list of parametric models can be found in Kotz & Nadarajah (2000). As they remark, there is some chaos there: some models are given by the density of h, others by their dependence measure, some are given in any dimension d, others specifically for $d = 2$. In this thesis the Poisson process approach is favored, so we concentrate on models for which h is available. In this category the most used models, without minimizing the importance of others, are the extremal logistic model and the extremal Dirichlet model.

The extremal logistic model has density function

$$h(w) = \frac{1}{2}(1-\alpha)w_1^{-2}w_2^{-1}u_1^{1-\alpha}u_2\left\{\alpha u_2 + \beta u_1\right\}^{-1}, \quad w = (w_1, w_2) \in \mathcal{S}_2, \tag{1.4}$$

where $0 < \alpha, \beta < 1$ and $u = (u_1, u_2) \in S_2$ is the solution to

$$(1-\alpha)w_2u_2^{\beta} - (1-\beta)w_1u_1^{\alpha} = 0.$$

No explicit form exists except if $\alpha = \beta$, in which case

$$h(w) = \frac{1}{2}(\alpha^{-1} - 1)(w_1w_2)^{-1-1/\alpha}\left\{w_1^{-1/\alpha} + w_2^{-1/\alpha}\right\}^{\alpha-2}.$$

This is the symmetric logistic model. As α tends to 0 or 1 the symmetric logistic density tends to asymptotic perfect dependence or independence, respectively. The dependence measure of the extremal logistic model is

$$V(x) = \left(x_1^{-1/\alpha} + x_2^{-1/\alpha}\right)^{\alpha}, \quad x \in \mathbb{R}^2.$$

The symmetric logistic model has been generalized by Tawn (1990) to the asymmetric logistic model,

$$V(x) = \sum_{c \in C}\left\{\sum_{j \in c}\left(\frac{\theta_{j,c}}{x_j}\right)^{r_c}\right\}^{1/r_c}, \quad x \in \mathbb{R}^d,$$

where C is the set of all non-empty subsets of $\{1, \ldots, d\}$ and the parameters are constrained by $r_c \geq 1$ for all $c \in C$, $\theta_{j,c} = 0$ if $j \notin c$, $\theta_{j,c} \geq 0$, $j = 1, \ldots, d$ and $\sum_{c \in C}\theta_{j,c} = 1$.

This generalization is a mixture of extremal logistic models on each subspace of $\mathcal{S}_d$. In Chapter 2 a new representation of the extremal logistic model is given. It allows exact simulation and a different kind of generalization in any dimension d.

The extremal Dirichlet model has density function, for $w \in \mathcal{S}_d$,

$$h(w) = \frac{1}{d} \frac{\Gamma\left(1 + \sum_{j=1}^{d} \alpha_j\right)}{\prod_{j=1}^{d} \Gamma(\alpha_j)} \left(\sum_{j=1}^{d} \alpha_j w_j\right)^{-(d+1)} \prod_{j=1}^{d} \alpha_j \prod_{j=1}^{d} \left(\frac{\alpha_j w_j}{\sum_{j=1}^{d} \alpha_j w_j}\right)^{\alpha_j - 1} \tag{1.5}$$

where $\alpha_j > 0$, $j = 1, \ldots, d$ and Γ is the gamma function

$$\Gamma(t) = \int_0^\infty y^{t-1} e^{-y} dy, \quad t > 0.$$

In dimension two, the dependence measure is

$$V(x) = \frac{1}{x_1} \text{Beta}\left(x^*,\, 1;\, \alpha_1 + 1,\, \alpha_2\right) + \frac{1}{x_2} \text{Beta}\left(0,\, x^*;\, \alpha_1,\, \alpha_2 + 1\right), \quad x \in \mathbb{R}^2,$$

where $x^* = \alpha_1 x_1 / (\alpha_1 x_1 + \alpha_2 x_2)$ and $\text{Beta}\,(x,\, y;\, \alpha,\, \beta)$ is the incomplete Beta function,

$$\text{Beta}\,(x,\, y;\, \alpha,\, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_x^y s^{\alpha - 1} (1 - s)^{\beta - 1} ds.$$

The case $d > 2$ is complicated to express but numerically feasible. The extremal Dirichlet model is a particular case of a theorem due to Coles & Tawn (1991); see also Appendix A.2. In Chapter 2 we give a new representation of the extremal Dirichlet model allowing exact simulation.

Asymptotic dependence and independence

The strength of tail dependence between two variables U_1, U_2 with uniform margins can be summarized by

$$\chi = \lim_{u \to 1} P\left(U_2 > u \mid U_1 > u\right),$$

if the limit exists. In the case of multivariate extreme value distributions $\chi = 2 \int_{\mathcal{S}_2} \min(w_1, w_2) H(dw)$, so that, if H does not put any mass in the interior of $\mathcal{S}_2$, $\chi = 0$. In general χ varies between zero, when data are asymptotically independent, and one, for asymptotically perfectly dependent data, growing with the strength of dependence (Sibuya 1960). Defining

$$\chi(u) = 2 - \frac{\log P(U_1 < u, U_2 < u)}{\log P(U_1 < u)}, \quad 0 \le u \le 1,$$

it follows that $\chi = \lim_{u \to 1} \chi(u)$. The function $\chi(u)$ can be empirically estimated for increasing u which provides an exploratory means to analyze strength of dependence; see for example Coles, Heffernan & Tawn (1999).

The discrimination of the strength of dependence within the class of asymptotic independence was first addressed by Ledford & Tawn (1996); see also Ledford & Tawn (1997). Observing that, for two Fréchet variables X_1, X_2 and large z,

$$P(X_1 > z, X_2 > z) \sim \begin{cases} z^{-1}, & \text{for perfect dependence,} \\ z^{-2}, & \text{for exact independence,} \end{cases}$$

they smoothly linked these bounding cases with the model

$$P(X_1 > z, X_2 > z) \quad \sim \quad \mathcal{L}(z) z^{-1/\eta}, \quad z \to \infty, \tag{1.6}$$

where η is the coefficient of tail dependence and $\mathcal{L}$ is a slowly varying function, that is $\mathcal{L}(tz)/\mathcal{L}(z) \to 1$ as $z \to \infty$, for any $t > 0$. The case of asymptotically dependent variables gives $\eta = 1$ and $\mathcal{L}(z) \not\to 0$. Asymptotically independent variables are distinguished into three cases:

1) positive association, $\frac{1}{2} < \eta < 1$, or $\eta = 1$ and $\mathcal{L}(z) \to 0$;

2) near independence, $\eta = \frac{1}{2}$;

3) negative association, $0 < \eta < \frac{1}{2}$.

For two variables U_1, U_2 with uniform margins, define

$$\bar{\chi}(u) = \frac{2 \log P(U_1 > u)}{\log P(U_1 > u, U_2 > u)}, \quad 0 \le u \le 1.$$

The limit $\bar{\chi} = \lim_{u \to 1} \bar{\chi}(u)$ equals $2\eta - 1$ if model (1.6) is valid. Empirical estimates of $\bar{\chi}(u)$ provide exploratory means to analyze the strength of dependence within the class of asymptotically independent variables (Coles et al. 1999).

Parametric estimates of η can be based on the structural variable $T = \min(X_1, X_2)$ by assuming model (1.6) exact above a selected threshold z_0 and approximating the slowly varying function $\mathcal{L}(z)$ by a constant (Ledford & Tawn 1996). See also Peng (1999) for a non-parametric estimator.

Ledford & Tawn (1997) also explored the relation with the point process viewpoint. Let M_n be the componentwise maximum of an independent bivariate sample $X_1, \ldots, X_n$ with common distribution function $F_{1,2}$ and unit Fréchet margins, F_1. Consider also $\bar{F}_{1,2}$, the survivor function of $F_{1,2}$, $F_{1,2}(z_1, z_2) = 1 - \bar{F}_1(z_1) - \bar{F}_2(z_2) + \bar{F}_{1,2}(z_1, z_2)$. Then, for any $z = (z_1, z_2) \in \mathbb{R}^2_+$,

$$P(M_n \le nz) \sim \left\{ 1 - \frac{1}{n} \left(\frac{1}{z_1} + \frac{1}{z_2} \right) + \bar{F}_{1,2}(nz) \right\}^n \overset{n \to \infty}{\longrightarrow} G(z),$$

where G is a bivariate extreme value distribution. The asymptotic independent case occurs when $n\bar{F}_{1,2}(nz) \overset{n\to\infty}{\longrightarrow} 0$, that is the mass of the spectral measure is exclusively on the border of $\mathcal{S}_2$. As knowledge of the limit does not of itself give information on how quickly the components become independent, the authors justify the approximate model, for large z_1 and z_2,

$$\bar{F}_{1,2}(z_1, z_2) = \mathcal{L}_{1,2}(z) z_1^{-c_1} z_2^{-c_2},$$

where $c_1 + c_2 = 1/\eta$ and $\mathcal{L}_{1,2}$ is bivariate slowly varying, that is $\mathcal{L}_{1,2}(tz_1, tz_2)/\mathcal{L}_{1,2}(t,t) \to g_*\{z_1/(z_1+z_2)\}$, $t \to \infty$. For statistical purposes $\mathcal{L}$ is considered as exactly ray dependent above high thresholds u_1 and u_2 so that the model becomes

$$\bar{F}_{1,2}(z_1, z_2) = K z_1^{-c_1} z_2^{-c_2} g_*\{z_1/(z_1+z_2)\}, \quad z_1 > u_1, z_2 > u_2,$$

where $K > 0$ is a constant and g_* a density on $\mathcal{S}_2$. On the pseudo-polar scale, $r = z_1 + z_2$, $w = z/r$, this model becomes

$$\bar{F}_{1,2}(z_1, z_2) = r^{-\eta} w_1^{-c_1} w_2^{-\eta+c_1} K g_*(w).$$

Models for g_* have been developed.

Models for asymptotic independence

For its simplicity and the wide range of dependence it can represent, the Gaussian model was made popular in Bortot, Coles & Tawn (2000). Let Y be a d-variate normal variable with correlation matrix $\{\rho_{ij}\}$, whose margins are transformed to the unit Fréchet scale. Then the pairwise coefficient of tail dependence is $\eta_{ij} = (1+\rho_{ij})/2$, $i, j = 1, \ldots, d$. Hence, for ρ_{ij} varying (strictly) between 0 and 1 every value of η_{ij} in range 1/2 and 1 can be achieved. Data with standard Gaussian margins in a suitable tail region $(u_1, \infty) \times \cdots \times (u_d, \infty)$ are fitted to a multivariate Gaussian model and inference on the tail dependence is readily obtained from the correlation matrix.

The inverted multivariate extreme value distribution exhibits positive association; see for example Heffernan & Tawn (2004). Let Y have unit Fréchet margins and survivor function

$$P(Y > y) = \exp\{-V(z)\},$$

where V is a dependence measure and $z_j = -1/\log \bar{F}_j(y_j)$, $j = 1, \ldots, d$. Then Y is distributed according to an inverted multivariate extreme value distribution and has pairwise tail dependence coefficients

$$\eta_{ij} = 1/V(\infty, \ldots, \infty, 1, \infty, \ldots, \infty, 1, \infty, \ldots, \infty),$$

where the ones are at the ith and jth places.

Chapter 2

Mixtures and Multivariate extremes

The previous chapter emphasized the central role of the spectral distribution function for asymptotically dependent extremes. In the first section of this chapter new representations of the extremal logistic and extremal Dirichlet spectral measure are developed. In the second section a new model is developed based on a constrained mixture of Dirichlet distributions. Its theoretical properties are discussed. In Section 2.3, two algorithms for fitting the models are developed.

2.1 Representations of extremal logistic and extremal Dirichlet models

These representations were originally done in order to simulate exactly from density functions (1.4) and (1.5). In the extremal logistic case, it turns out that the random variable u has a very simple distribution function and that, although u is not an explicit function of w, w is readily obtained from u. A similar approach is used for the extremal Dirichlet model.

2.1.1 The extremal logistic model

Let W be distributed according to the density function

$$h_W(w) = \frac{1}{2}(1-\alpha)w_1^{-2}w_2^{-1}u_1^{1-\alpha}u_2\left\{\alpha u_2 + \beta u_1\right\}^{-1}, \quad w = (w_1, w_2) \in \mathcal{S}_2,$$

with $u = (u_1, u_2) \in S_2$ the solution to

$$(1-\alpha)w_2 u_2^{\beta} - (1-\beta)w_1 u_1^{\alpha} = 0.$$

Then the density function of U is

$$h_U(u) = \frac{1}{2}\left\{\frac{1-\alpha}{u_1^{\alpha}} + \frac{1-\beta}{u_2^{\beta}}\right\}, \quad u \in \mathcal{S}_2.$$

In other words h_U is a balanced mixture of Dirichlet densities with parameters $(1-\alpha, 1)$ and $(1, 1-\beta)$. Extending this model to dimension d, the following definition is proposed:

Definition 2.1 (Extremal logistic model)

The random variable W follows the extremal logistic model if $W \in \mathcal{S}_d$ is the solution to

$$\frac{W_j}{W_d} = \frac{C_j U_j^{-\alpha_j}}{C_d U_d^{-\alpha_d}}, \quad j = 1, \ldots, d,$$

where

$$C_j = \Gamma(d-\alpha_j)/\Gamma(1-\alpha_j), \quad j = 1, \ldots, d,$$

and $U \in \mathcal{S}_d$ is distributed according to the density function

$$h_U(u) = d^{-1} \sum_{j=1}^{d} \frac{\Gamma(d-\alpha_j)}{\Gamma(1-\alpha_j)} u_j^{-\alpha_j}, \quad 0 < \alpha_j < 1, \quad j = 1, \ldots, d.$$

One can show that

$$h_W(w) \;=\; d^{-1}\left(\sum_{j=1}^{d} \alpha_j u_j\right)^{-1} \left(\prod_{i=1}^{d} \alpha_i u_i\right) \left(\sum_{j=1}^{d} C_j u_j^{-\alpha_j}\right) \prod_{j=1}^{d} w_j^{-1}, \tag{2.1}$$

which reduces in the symmetric case $\alpha_1 = \cdots = \alpha_d = \alpha$ to

$$h_W(w) = \frac{\Gamma(d-\alpha)}{\Gamma(1-\alpha)} \frac{\alpha^{d-1}}{d} \prod_{i=1}^{d} w_i^{-1/\alpha - 1} \left(\sum_{i=1}^{d} w_i^{-1/\alpha}\right)^{\alpha - d},$$

and by construction coincides with the classical bivariate extremal logistic model when $d = 2$. Furthermore, the constraints

$$\int_{S_d} w_j h(w) dw \;=\; \frac{1}{d}, \quad j = 1, \ldots, d, \tag{2.2}$$

are satisfied. The proof is given in Appendix A.1. A consequence of this representation is the following simulation algorithm:

1) choose j among $1, \ldots, d$ with probability $1/d$;

2) simulate independent $Z_j \sim \mathcal{G}(1-\alpha_j)$ and $Z_l \sim \mathcal{G}(1)$, $l = 1, \ldots, d$, $l \neq j$;

3) set $U_l = Z_l / \sum_{m=1}^d Z_m$, $l = 1, \ldots, d$;

4) set $W_l = C_l U_l^{-\alpha_l} / \sum_{m=1}^d C_m U_m^{-\alpha_m}$, $l = 1, \ldots, d$.

Here $\mathcal{G}$ is the gamma distribution and $C_j = \Gamma(d - \alpha_j)/\Gamma(1 - \alpha_j)$, $\quad j = 1, \ldots, d$. The algorithm uses the fact that the vector of ratios $Z_j / \sum_{m=1}^d Z_m$, $j = 1, \ldots, d$, where Z_j are independent and $\mathcal{G}(\alpha_j)$ distributed, is a Dirichlet $(\alpha_1, \ldots, \alpha_d)$ random variable; see for example Wilks (1962, p.177–182).

2.1.2 The extremal Dirichlet model

Let $U \in \mathcal{S}_d$ be distributed according to the mixture of Dirichlet densities

$$h_U(u) = d^{-1} \sum_{j=1}^d \frac{\Gamma\left(1 + \sum_{i=1}^d \alpha_i\right)}{\prod_{i=1}^d \Gamma\left(\alpha_i + \delta_{ij}\right)} \prod_{i=1}^d u_i^{\alpha_i - 1 + \delta_{ij}},$$

where δ_{ij} is the Kronecker symbol equal to one if $i = j$ and zero otherwise. Then $W \in \mathcal{S}_p$ with components

$$W_j = \frac{\alpha_j^{-1} U_j}{\sum_{i=1}^d \alpha_i^{-1} U_i}, \quad j = 1, \ldots, d,$$

is distributed according to the extremal Dirichlet model with parameters $\alpha_1, \ldots, \alpha_p$. This representation provides the following simulation algorithm:

1) choose j among $1, \ldots, d$ with probability $1/d$;

2) simulate independent $Z_l \sim \mathcal{G}(\alpha_l + \delta_{lj})$, $l = 1, \ldots, d$;

3) set $U_l = Z_l / \sum_{m=1}^d Z_m$, $l = 1, \ldots, d$;

4) set $W_l = \alpha_l^{-1} U_l / \sum_{i=1}^d \alpha_i^{-1} U_i$, $l = 1, \ldots, d$.

This representation applies not only to the extremal Dirichlet model but to a whole class of distributions described below. The extremal Dirichlet model was first defined by Coles & Tawn (1991). Its construction is based on the following theorem. Below $a \cdot b$ is the scalar product between a and b.

Theorem 2.2

If h^* is any positive function on S_d, with finite first moments, then

$$\tilde{h}(w) = (m \cdot w)^{-(d+1)} \prod_{j=1}^d m_j h^* \left(\frac{m_1 w_1}{m \cdot w}, \ldots, \frac{m_d w_d}{m \cdot w} \right),$$

where

$$m_j = \int_{S_p} u_j h^*(u)du, \quad j = 1, \ldots, d,$$

satisfies constraints

$$\int_{S_p} w_j \tilde{h}(w)dw = 1, \quad j = 1, \ldots, d,$$

and is therefore the density of a valid spectral measure $\tilde{H}$*.*

By letting h^* be a Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_d$, one obtains

$$\tilde{h}(w) = \frac{\Gamma(\alpha \cdot 1 + 1)}{\prod_{j=1}^d \Gamma(\alpha_j)} \prod_{j=1}^d \alpha_j^{\alpha_j} \frac{\prod_{j=1}^d w_j^{\alpha_j - 1}}{(\alpha \cdot w)^{\alpha \cdot 1 + 1}}.$$

The corresponding spectral density is divided by a factor d. Now letting

$$U_i = \frac{m_i W_i}{\sum_{j=1}^d m_j W_j}, \quad i = 1, \ldots, d,$$

it can be shown that

$$h_U(u) = p^{-1} \sum_{j=1}^d h^*(u) u_j / m_j. \tag{2.3}$$

By definition of m_j, $h^*(u)u_j/m_j$ is a density function and, therefore, $h_U(u)$ is a balanced mixture of $h^*(u)u_j/m_j$, $j = 1, \ldots, d$. Should it be easy to simulate from $h^*(u)u_j/m_j$, the simulation algorithm is straightforwardly adapted. The proof of the representation is given in Appendix A.2.

2.2 The extremal mixture model

This section investigates the use of mixture of Dirichlet distributions as a model for the spectral probability measure H. This family turns out to be very rich and hence appropriate for semi-parametric inference. In general, a random variable W distributed according to H can be viewed as a random probability vector or, equivalently, a random probability function on the finite set $\{1, \ldots, d\}$. Therefore, the spectral probability measure is in fact the distribution of a random probability measure. A well studied family of distributions of probability measures is the mixture of Dirichlet processes, presented below.

2.2.1 Random probability measures and mixtures of Dirichlet processes

Let $\mathfrak{S}$ be a topological space with some good properties that we do not detail now and $\mathcal{S}$ be its Borel algebra. Then, loosely, a random probability measure P on $(\mathfrak{S}, \mathcal{S})$ is a random

measure that integrates to one. Therefore the distribution of P can be represented by its finite dimensional distributions or, in other words, by the family of distributions of vectors $\{P(A_1), \ldots, P(A_d)\}$ for any $d \in \mathbb{N}$ and measurable partition $A_1, \ldots, A_d$ of $\mathfrak{S}$. A finite dimensional distribution family must satisfy Kolmogorov's consistency conditions (Kallenberg 1983, p.41) in order to represent a random probability measure. In particular, some natural conditions are satisfied like

1. $P(\mathfrak{S}) = 1$, almost surely,
2. for any $A \in \mathcal{S}$, $P(A) \geq 0$, almost surely,
3. for any finite sequence $A_1, \ldots, A_d, B_1, \ldots, B_d$ in $\mathcal{S}$ such that $A_i \cap B_i = \emptyset$, $i = 1, \ldots, d$,

$$\{P(A_1 \cup B_1), \ldots, P(A_d \cup B_d)\} \stackrel{d}{=} \{P(A_1) + P(B_1), \ldots, P(A_d) + P(B_d)\}.$$

A popular model for random probabilities is the Dirichlet process introduced by Ferguson (1973). Let α be a Radon measure on $\mathfrak{S}$. For any finite measurable partition sequence $A_1, \ldots, A_d$, the distribution of $\{P(A_1), \ldots, P(A_d)\}$ is Dirichlet with parameters $\alpha(A_1), \ldots, \alpha(A_d)$. A natural extension is the mixture of Dirichlet processes introduced by Antoniak (1974); see also Dalal (1978) and Dalal & Hall (1980). The parameter measure α is itself random, for example if α_u is a Radon measure and if U is a random variable with distribution function π, then the density function of $\{P(A_1), \ldots, P(A_d)\}$ is

$$h(w) = \int \frac{\Gamma\{\alpha_u(\mathfrak{S})\}}{\prod_{j=1}^d \Gamma\{\alpha_u(A_j)\}} \prod_{j=1}^d w_j^{\alpha_u(A_j)-1} \pi(du), \quad w \in S_d.$$

We talk about finite mixtures of Dirichlet processes when U takes only a finite number of values, that is

$$h(w) = \sum_{m=1}^k \pi_m \frac{\Gamma\{\alpha_m(\mathfrak{S})\}}{\prod_{j=1}^d \Gamma\{\alpha_m(A_j)\}} \prod_{j=1}^d w_j^{\alpha_m(A_j)-1}, \quad w \in S_d.$$

Finite mixtures are useful in practice since no parametric hypothesis for the mixing distribution is needed. The drawback is that this kind of model has a large number of parameters.

2.2.2 Adequacy

Theoretical properties of mixtures of Dirichlet processes have been intensively studied as they can approximate any prior distribution function and lead to analytic posteriors.

In particular, a large literature has been devoted to the study of posterior properties of a Bayes model with Dirichlet prior. This is outside the scope of this work. Indeed the richness of the mixtures of Dirichlet processes for approximation of priors is enough to define a rich class of models of spectral distributions. This property, known as adequacy, is studied in Dalal (1978) and Dalal & Hall (1980), who show that any random probability measure can be approximated in the weak sense by a finite mixture of Dirichlet processes. For details about the weak topology, see Section 5.3.1.

Let $\mathfrak{M}_1(\mathfrak{S})$ be the set of probability measures on $\mathfrak{S}$ and $\mathcal{F}(\mathfrak{S})$ be the class of all atomic probability measures with finite support on $\mathfrak{S}$, that is

$$\mathcal{F}(\mathfrak{S}) = \left\{ \sum_{i=1}^{n} a_i \delta_{s_i} \ : \ \sum_{i=1}^{n} a_i = 1, a_i \geq 0, s_i \in \mathfrak{S}, i = 1, \ldots, n, n \geq 1 \right\},$$

and let $\overline{\mathcal{F}(\mathfrak{S})}$ be the closure of $\mathcal{F}(\mathfrak{S})$ in the weak topology. Furthermore, let MDP be the class of mixtures of Dirichlet processes. Dalal & Hall (1980) show the two following results, among others:

Lemma 2.3

If $\mathfrak{S}$ is compact Hausdorff or Polish, then $\overline{\mathcal{F}(\mathfrak{M}_1(\mathfrak{S}))} = \mathfrak{M}_1(\mathfrak{M}_1(\mathfrak{S}))$.

Theorem 2.4

If $\mathfrak{S}$ is compact Hausdorff or Polish, then $\overline{\text{MDP}} = \mathfrak{M}_1(\mathfrak{M}_1(\mathfrak{S}))$.

The last result is the key result since it tells us that the class of mixtures of Dirichlet processes is weak dense in the class of random probabilities. However, the mixing distribution could be anything. For applications, it is essential that the set of *finite* mixtures of Dirichlet distributions is weak dense. This fact is true and emphasized in Dalal (1978) where the same results are presented in another language. The idea is that if the distribution of any random probability can be approximated by elements of $\mathcal{F}(\mathfrak{M}(\mathfrak{S}))$ (Lemma 2.3) then each of these elements may be approximated by finite mixtures of Dirichlet processes.

2.2.3 Application to multivariate extremes

If $\mathfrak{S}$ is the finite set $\{1, \ldots, d\}$, then the finest partition of $\mathfrak{S}$ is $\{1\}, \ldots, \{d\}$. Therefore, a probability function is simply a probability vector $(P\{1\}, \ldots, P\{d\}) \in \mathcal{S}_d$ and a random probability function is a random vector in S_d. Consequently, a Dirichlet process is a Dirichlet distribution and a finite mixture of Dirichlet processes is a finite mixture of

Dirichlet distributions,

$$h(w) = \sum_{m=1}^{k} \pi_m \frac{\Gamma\left(\sum_{j=1}^{d} \alpha_m^{(j)}\right)}{\prod_{j=1}^{d} \Gamma\left(\alpha_m^{(j)}\right)} \prod_{j=1}^{d} w_j^{\alpha_m^{(j)}-1}, \quad w \in S_d.$$

The adequacy of mixtures of Dirichlet processes implies adequacy of mixtures of Dirichlet distributions which hence generates a very rich family of spectral densities. For the mixture of Dirichlet distributions to be a valid spectral probability measure, one must impose the mean constraint for Fréchet margins, that is

$$\sum_{m=1}^{k} \pi_m \frac{\alpha_m^{(i)}}{\sum_{j=1}^{p} \alpha_j^{(m)}} = d^{-1}, \quad \text{for } i = 1, \ldots, d.$$

Sometimes it is more convenient to use the mean-scale parametrization of the Dirichlet,

$$\mu_j = \frac{\alpha_j}{\sum_{i=1}^{d} \alpha_i} \quad \text{and} \quad \nu = \sum_{i=1}^{d} \alpha_i,$$

with $\sum_{j=1}^{d} \mu_j = 1$. In order to avoid confusion with the extremal Dirichlet model the mixture of Dirichlet models will be termed the extremal mixture model. This gives the following definition:

Definition 2.5 (Extremal mixture model)

The random variable W follows the extremal mixture model if it is distributed according to the density function

$$h(w) = \sum_{m=1}^{k} \pi_m \frac{\Gamma(\nu_m)}{\prod_{j=1}^{d} \Gamma\left(\nu_m \mu_j^{(m)}\right)} \prod_{j=1}^{d} w_j^{\nu_m \mu_j^{(m)}-1}, \quad w \in S_d,$$

where, for $j = 1, \ldots, d$ and $m = 1, \ldots, k$,

$$\pi_m \geq 0, \ \sum_{m=1}^{k} \pi_m = 1, \quad \nu_m > 0, \quad \mu_j^{(m)} \geq 0, \ \sum_{m=1}^{k} \pi_m \mu_j^{(m)} = d^{-1}, \ \sum_{i=1}^{d} \mu_i^{(m)} = 1.$$

2.2.4 Discussion

The consistency of the Dirichlet process as a random probability brings a further understanding to the implicit nature of the spectral probability distribution. An extreme event occurs when the sum R of the d components of X exceeds the selected threshold r_0. The distribution of R among components of X, W, is distributed according to H. The constraints on H mean that W, seen as a random distribution on the set $\{1, \ldots, d\}$, is expected to be uniform, as each margin is on the unit Fréchet scale. Furthermore, Kolmogorov's consistency conditions ensure that the distribution of any subgroup of

components is coherent with the model. For example, let $d = 3$. If H is a mixture of Dirichlet distributions with parameters α_m and π_m, $m = 1, \ldots, k$, then, writing 'Dir' for the Dirichlet distribution,

$$\left(W^{(1)}, W^{(2)}, W^{(3)}\right) \sim \sum_{m=1}^{k} \pi_m \mathrm{Dir}\left\{\alpha_m^{(1)}, \alpha_m^{(2)}, \alpha_m^{(3)}\right\},$$

but also

$$\left(W^{(1)} + W^{(2)}, W^{(3)}\right) \sim \sum_{m=1}^{k} \pi_m \mathrm{Dir}\left\{\alpha_m^{(1,2)}, \alpha_m^{(3)}\right\}.$$

where $\alpha_m^{(1,2)} = \alpha_m^{(1)} + \alpha_m^{(2)}$. Accordingly, the mean constraints

$$\sum_{m=1}^{k} \pi_m \frac{\alpha_m^{(j)}}{\alpha_m^{(1)} + \alpha_m^{(2)} + \alpha_m^{(3)}} = 1/3, \quad j = 1, 2, 3,$$

become

$$\sum_{m=1}^{k} \pi_m \frac{\alpha_m^{(1,2)}}{\alpha_m^{(1,2)} + \alpha_m^{(3)}} = 2/3 \quad \text{and} \quad \sum_{m=1}^{k} \pi_m \frac{\alpha_m^{(3)}}{\alpha_m^{(1,2)} + \alpha_m^{(3)}} = 1/3.$$

This consistency, achieved by every distribution of random probabilities, is unfortunately not clear for the extremal logistic and extremal Dirichlet models. In particular, although they can be generalized easily to any kind of mean constraints, we have not succeeded in writing them as a distribution of random probabilities. For multivariate extremes, this is of secondary importance since, as $\{\{1\}, \ldots, \{d\}\}$ is the finest partition of $\mathfrak{S} = \{1, \ldots, d\}$, the distribution of $(W^{(1)}, \ldots, W^{(d)})$ gives implicitly the distribution on every decomposition of $\mathfrak{S}$. But on continuous space $\mathfrak{S}$, where no such finest decomposition exists, they are difficult to extend. In Chapter 5, this consistency is used to extend the extremal mixture to the spatial context, in other words, to mixtures of Dirichlet processes.

The adequacy of the extremal mixture model is not guaranteed under constraints. In absence of theoretical results, the richness of this family will be investigated by experiment. The next section is devoted to fitting procedures.

2.3 Fitting the extremal mixture model

This section presents the use of two algorithms for fitting the extremal mixture model: the EM algorithm and the reversible jump Markov chain Monte Carlo algorithm.

A review of the literature about the use of mixture densities can be found in Redner & Walker (1984), and reference books are Titterington, Smith & Makov (1985) and McLachlan & Peel (2000). This subject goes back to Pearson (1894) where a mixture of two Gaussian densities is fitted by the method of moments. In the 1960's, the increasing

capacity of computers allowed mixtures to be fitted with maximum likelihood methods. The complexity of likelihood functions, even for simple mixtures, explains the success of the EM algorithm that simplifies computations. The choice of the number of components in the mixtures is a challenging problem. Historically it is addressed separately from the estimation of the parameters (Lindsay 1995, p.74). In this context, the Bayesian formulation allowed flexible models and estimation procedures to be developed: the unknown number of components, k, is not selected but the posterior density is a mixture over k using the reversible jump Markov chain Monte Carlo algorithm to explore every possible dimension (Richardson & Green 1997).

2.3.1 The EM algorithm applied to the extremal mixture model

The use of the EM algorithm for incomplete data problems was formalized by Dempster, Laird & Rubin (1977). Since then a very large literature is devoted to its applications and improvements (Meng & Pedlow 1992). Two common criticisms are the difficulty of obtaining confidence intervals and its slow convergence. The first issue is discussed, for example, in Tanner (1996, p.74–80) or Oakes (1999) and the second in, for example, Meng & van Dyk (1997). A reference book is McLachlan & Krishnan (1997).

The EM algorithm

Apart from minor details, the following lines follow the development in Davison (2003, p.210). An independent and stationary sample $y_1, \ldots, y_n$ from a finite mixture with k components can be viewed as a partial observation of a complete dataset, $(y_1, u_1), \ldots, (y_n, u_n)$, where u_j takes values in $\{1, \ldots k\}$ and indicates which component y_j comes from. The complete data log-likelihood is

$$\log f(y, u; \theta) = \log f(y; \theta) + \log f(u \mid y; \theta),$$

where θ are the parameters and $\log f(y; \theta) = \ell(\theta)$ is the log-likelihood. Taking the expectation with respect to U after conditioning on $Y = y$, at θ', yields

$$E\left\{\log f(Y, U; \theta) \mid Y = y; \theta'\right\} = \ell(\theta) + E\left\{\log f(U \mid Y; \theta) \mid Y = y; \theta'\right\},$$

which can be written as

$$Q(\theta; \theta') = \ell(\theta) + C(\theta; \theta').$$

Therefore, for a fixed θ',

$$Q(\theta; \theta') \geq Q(\theta'; \theta') \quad \text{implies} \quad \ell(\theta) - \ell(\theta') \geq C(\theta'; \theta') - C(\theta; \theta').$$

A further argument using Jensen's inequality applied to $f(y \mid u;\, \theta)$ shows that if $f(u \mid y;\, \theta)$ is non-degenerate and no two values of θ give the same model, then $C(\theta';\, \theta') \geq C(\theta;\, \theta')$, with equality only when $\theta = \theta'$. Therefore, increasing the value of $Q(\theta;\, \theta')$ increases $\ell(\theta)$. Furthermore, under appropriate smoothness conditions, $C(\theta;\, \theta')$ is stationary at $\theta = \theta'$. Hence, if $Q(\theta;\, \theta')$ is stationary at $\theta = \theta'$, so too is $\ell(\theta)$. The result is the EM algorithm.

Definition 2.6 (EM Algorithm)

Starting at θ'*,*

1. *compute* $Q(\theta;\, \theta')$*,*

2. *for fixed* θ'*, maximize* $Q(\theta;\, \theta')$ *over* θ*, giving* $\theta^\dagger$*,*

3. *check convergence. If not converged, set* $\theta' := \theta^\dagger$ *and go to step* 1.

Convergence can be checked by criteria like

$$|\ell(\theta^\dagger) - \ell(\theta')|/|\ell(\theta')| \leq \varepsilon \quad \text{or} \quad \|\theta^\dagger - \theta'\|/\|\theta'\| \leq \varepsilon,$$

for some small ε.

Confidence intervals can be based on the information matrix via the missing information principle. Indeed,

$$\log f(y_i;\, \theta) = \log f(y_i, u_i;\, \theta) - \log f(u_i \mid y_i;\, \theta),$$

so that, at $\theta' = \theta$,

$$\frac{-\partial^2 \log f(y;\, \theta)}{\partial \theta^2} = \frac{-\partial^2 Q(\theta,\, \theta')}{\partial \theta^2} - \frac{-\partial^2 H(\theta,\, \theta')}{\partial \theta^2}, \tag{2.4}$$

where $H(\theta,\, \theta') = \sum_{i=1}^n \int \log p(u_i \mid y_i; \theta) p(u_i \mid y_i; \theta') du_i$. The first term of the right side of (2.4) is termed the complete information and is in general numerically available from the M-step. The second term is named the missing information and has to be algebraically calculated or approximated at the maximum likelihood estimator $\theta = \theta' = \hat{\theta}$ by

$$\frac{-\partial^2 H(\theta,\, \theta')}{\partial \theta^2} = \sum_{i=1}^n \operatorname{var}\left(\frac{\partial \log f(y_i, u_i;\, \theta)}{\partial \theta} \right), \tag{2.5}$$

where the variances are taken with respect to $p(u_i \mid y_i, \theta)$, $i = 1, \ldots, n$. For sake of clarity, the proof of the approximation of the variance is given in Appendix A.3. It can also be found in Tanner (1996, p.74–78) but the notation adopted by him makes the link with the present work not straightforward. This proof also enlightens a link with Oakes (1999) that we have not found explicitly in the literature.

In the case where the complete-data (y_i, u_i), $i = 1, \ldots, n$, come from a regular exponential family,

$$\log f(y_i, u_i;\, \theta) = \log b(y_i, u_i) + \theta^T s(y_i, u_i) - \log a(\theta),$$

then

$$Q(\theta,\, \theta') = \sum_{i=1}^{n} \log b(y_i, u_i) + \theta^T \sum_{i=1}^{n} \int s(y_i, u_i) p(u_i \mid y_i;\, \theta') du_i - n \log a(\theta),$$

where $p(u_i \mid y_i;\, \theta)$ denotes the distribution of U_i given $Y_i = y_i$ at θ and the subscript T is the transposition. Therefore, maximizing $Q(\theta,\, \theta')$ with respect to θ is equivalent to solving

$$\frac{\partial \log a(\theta)}{\partial \theta} = \frac{1}{n} \sum_{i=1}^{n} \int s(y_i, u_i) p(u_i \mid y_i;\, \theta') du_i.$$

The confidence intervals are obtained from

$$\frac{\partial \log f(y_i, u_i;\, \theta)}{\partial \theta} = s(y_i, u_i) - \frac{\partial \log a(\theta)}{\partial \theta},$$

so that

$$\sum_{i=1}^{n} \operatorname{var}\left(\frac{\partial \log f(y_i, u_i;\, \theta)}{\partial \theta} \right) = \sum_{i=1}^{n} \operatorname{var}\left\{ s(y_i, u_i) \right\},$$

where the variances are taken with respect to $p(u_i \mid y_i, \theta)$, $i = 1, \ldots, n$.

In the case where all components of a mixture come from the same exponential family and are equal up to the parameter, that is

$$\log f_m(y_i;\, \theta) = \log b(y_i) + \theta_m^T s(y_i) \ - \log a(\theta_m), \quad m = 1, \ldots, k,$$

where θ is the concatenation of the vectors $\theta_1, \ldots, \theta_k$, then the complete-data log-likelihood takes also the form of an exponential family:

$$\begin{aligned} \log f(y_i, u_i;\, \theta) &= \sum_{m=1}^{k} \delta_{mu_i} \left\{ \log b(y_i) + \theta_m^T s(y_i) - \log a(\theta_m) + \log \pi_m \right\}, \\ &= \log b(y_i) + \sum_{m=1}^{k-1} \delta_{mu_i} \left\{ \theta_m^T s(y_i) - \log \frac{a(\theta_m)}{a(\theta_k)} + \log \frac{\pi_m}{\pi_k} \right\} \\ &\quad + \delta_{ku_i} \theta_k^T s(y_i) + \log \pi_k - \log a(\theta_k). \end{aligned}$$

Therefore, the complete-data likelihood function comes from an exponential family by setting, with a little abuse of notation,

$$s(y_i, u_i) = \begin{pmatrix} \delta_{u_i 1} \\ \vdots \\ \delta_{u_i (k-1)} \\ \delta_{u_i 1} s(y_i) \\ \vdots \\ \delta_{u_i k} s(y_i) \end{pmatrix}, \quad \phi = \begin{pmatrix} \log \frac{\pi_1}{\pi_k} - \log \frac{a(\theta_1)}{a(\theta_k)} \\ \vdots \\ \log \frac{\pi_{k-1}}{\pi_k} - \log \frac{a(\theta_{k-1})}{a(\theta_k)} \\ \theta_1 \\ \vdots \\ \theta_k \end{pmatrix},$$

$\log b(y_i, u_i) = \log b(y_i)$ and $\log a(\phi) = \log a(\theta_k) - \log \pi_k$. In the case of a finite mixture, $p(u_i \mid y_i; \theta)$ is a discrete distribution,

$$p(u_i = m \mid y_i; \phi) = \frac{f_m(y_i; \theta)\pi_m}{\sum_{l=1}^k f_l(y_i; \theta)\pi_l} = \frac{\exp\{s(y_i)^T \theta_m + \delta_{mk}\phi_m\}}{\sum_{l=1}^k \exp\{s(y_i)^T \theta_m + \delta_{mk}\phi_m\}}, \quad m = 1, \ldots, k.$$

The function Q simplifies accordingly. Unfortunately, although mixtures of Dirichlet distributions fit into this framework, the constraints imposed by the extremal mixture model link the parameters in such a way that the complete-data likelihood function cannot be written as an exponential family anymore. In particular, standard errors obtained from a blind application of (2.4) and (2.5) would not be correct since the constrained maximum of the likelihood is not equal to the overall maximum. Therefore, we do not pursue this direction here.

Application to the extremal mixture model

The complete-data log-likelihood contributions are, for $i = 1, \ldots, n$,

$$\log f(y_i, u_i; \theta) = \sum_{m=1}^k \delta_{u_i m} \left\{ \log \Gamma(\nu_m) - \sum_{j=1}^d \log \Gamma(\nu_m \mu_m^{(j)}) + \sum_{j=1}^d \left(\nu_m \mu_m^{(j)} - 1 \right) \log w_i^{(j)} + \log \pi_m \right\},$$

where θ is the vector containing $\pi_1, \ldots, \pi_{k-1}$, $\nu_1, \ldots, \nu_k$ and $\mu_m^{(j)}$, $m = 1, \ldots, k-1$, $j = 1, \ldots, d-1$ and where

$$\begin{aligned} \pi_k &= 1 - \sum_{m=1}^{k-1} \pi_m, \\ \mu_k^{(j)} &= \frac{1}{\pi_k} \left(\frac{1}{d} - \sum_{m=1}^{k-1} \pi_m \mu_m^{(j)} \right), \quad j = 1, \ldots, d-1, \\ \mu_m^{(d)} &= 1 - \sum_{j=1}^{d-1} \mu_m^{(j)}, \quad m = 1, \ldots, k-1, \\ \mu_k^{(d)} &= 1 - \frac{1}{\pi_k} \sum_{j=1}^{d-1} \left(\frac{1}{d} - \sum_{m=1}^{k-1} \pi_m \mu_m^{(j)} \right), \end{aligned}$$

with the constraints

$$\begin{aligned}
&0 < \pi_m < 1, \quad m = 1, \ldots, k-1,\\
&1 - \sum_{m=1}^{k-1} \pi_m > 0,\\
&0 < \mu_m^{(j)} < 1, \quad m = 1, \ldots, k-1, \quad j = 1, \ldots, d-1,\\
&1 - \sum_{j=1}^{d-1} \mu_m^{(j)} > 0,\\
&0 < \frac{1}{\pi_k}\left(\frac{1}{d} - \sum_{m=1}^{k-1} \pi_m \mu_m^{(j)}\right) < 1,\\
&\nu_m > 0, \quad m = 1, \ldots, k.
\end{aligned}$$

Optimizing this kind of function is very technical but can be performed for example with `matlab`'s function `fmincon`. The constraints are difficult to simplify because any change in the parametrization may make the calculus of $\partial \log f(y_i, u_i; \theta)/\partial\theta$ very awkward. This makes the use of `R`'s function `nlmin` inadequate if confidence intervals are of interest.

Now $\partial \log f(y_i, u_i; \theta)/\partial\theta$ is a vector composed of the following elements:

$$\begin{aligned}
\frac{\partial \log f(y_i, u_i; \theta)}{\partial \pi_m} &= \frac{\delta_{u_i m}}{\pi_m} + \delta_{u_i k}\left[\nu_k\left\{\sum_{j=1}^{d}\left(\frac{\mu_k^j}{\pi_k} - \frac{\mu_m^j}{\pi_m} - \log\Gamma'(\nu_k\mu_k^{(j)})\right)\log w_i^{(j)}\right\} - \frac{1}{\pi_k}\right],\\
\frac{\partial \log f(y_i, u_i; \theta)}{\partial \nu_m} &= \delta_{u_i m}\left\{\sum_{j=1}^{d}\mu_m^{(j)}\log w_i^{(j)} + \log\Gamma'(\nu_m) - \sum_{j=1}^{d}\log\Gamma'(\nu_m\mu_m^{(j)})\right\},\\
\frac{\partial \log f(y_i, u_i; \theta)}{\partial \nu_m} &= \delta_{u_i m}\nu_m\left\{\log w_i^{(j)} - \log w_i^{(d)} - \log\Gamma'(\nu_m\mu_m^{(j)}) + \log\Gamma'(\nu_m\mu_m^{(d)})\right\}\\
&\quad -\delta_{u_i k}\nu_k\frac{\pi_m}{\pi_k}\left\{\log w_i^{(j)} - \log w_i^{(d)} - \log\Gamma'(\nu_k\mu_k^{(j)}) + \log\Gamma'(\nu_k\mu_k^{(d)})\right\},
\end{aligned}$$

where $\log\Gamma'(x)$ is the digamma function. Therefore, for the appropriate matrix A,

$$\frac{\partial \log f(y_i, u_i; \theta)}{\partial \theta} = A\begin{pmatrix}\delta_{u_i 1}\\ \vdots \\ \delta_{u_i k}\end{pmatrix},$$

so that with $B = \operatorname{var}\left\{(\delta_{u_i 1}, \ldots, \delta_{u_i k})^T\right\}$,

$$\operatorname{var}\left(\frac{\partial \log f(y_i, u_i; \theta)}{\partial \theta}\right) = ABA^T.$$

Furthermore, one can calculate that

$$C_{ml} = \delta_{mk}p(u_i = m \mid y_i) - p(u_i = m \mid y_i)p(u_i = l \mid y_i), \quad m, l = 1, \ldots, k.$$

This gives an explicit formula for the second term of the right hand side of (2.4), while the first term is obtained numerically from the optimizer.

Selection of the number of components

Akaike's Information Criterion (AIC) is often used to select models in the regression context and naturally some authors have extended it to the selection of the number of components k in finite mixtures (McLachlan & Peel 2000, p.203). The AIC is known to overfit in the regression context, so that it may overestimate k. Therefore small-sample corrected versions like the AIC_c (Hurvich & Tsai 1989) may be preferable in particular for extremes where datasets are often small, although they have no theoretical basis for non-Gaussian data. A reference book on model selection is McQuarrie & Tsai (1998) and a practically-oriented treatment can be found in Burnham & Anderson (2002).

Because of the constraints, the number of parameters is $p = 2k - 1 + (k-1)(d-1)$ for the extremal mixture model with k components. Akaike's Information Criterion is therefore

$$\mathrm{AIC}(k) = -2\hat{\ell}_k + 2p,$$

where $\hat{\ell}_k$ is the maximized likelihood. The second-order modified version of the AIC, the AIC_c, is

$$\mathrm{AIC}_c(k) = \mathrm{AIC}(k) + \frac{2p(p+1)}{n-p-1},$$

where n is the sample size. The selected k minimizes $\mathrm{AIC}_c(k)$.

2.3.2 Reversible jump Markov chain Monte Carlo

Green (1995) is a founding work for the reversible jump Markov chain Monte Carlo algorithm. This algorithm is particularly suited for mixture models (Richardson & Green 1997) where the chain varies across parameter spaces with varying dimensions. A detailed treatment of Markov chain Monte Carlo methods can be found for example in Gilks, Richardson & Spiegelhalter (1996). Below, we briefly present the Metropolis–Hastings and the reversible jump Markov chain Monte Carlo algorithms.

Generalities

The observed data $y_1, \ldots, y_n$ are treated as an incomplete data set $(y_1, u_1), \ldots, (y_n, u_n)$, where u_j indicates the component from which y_j comes. The sample $u_1, \ldots, u_n$ is supposed independent and identically distributed according to the mixing distribution $(\pi_1, \ldots, \pi_k)$. A prior density is given to every parameter, including k, according to the directed acyclic graph of Figure 2.1. With generic notation, the distribution of all variables is

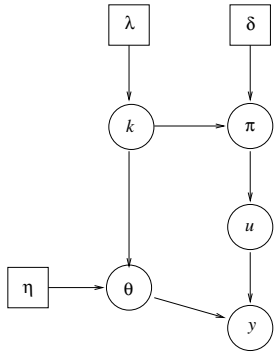


Figure 2.1: Directed acyclic graph of the hierarchical scheme for Bayesian mixture analysis.

$$p(\lambda, \delta, \eta, k, \pi, u, \theta, y) = p(y \mid \theta, u)p(\theta \mid k, \eta)p(u \mid \pi, k)p(\pi \mid k, \delta)p(k \mid \lambda)p(\delta)p(\lambda).$$

In order to satisfy the Bayesian paradigm, one must compute the posterior distribution $p(\lambda, \delta, \eta, k, \pi, u, \theta \mid y)$. This awkward task may be achieved by using a Markov chain Monte Carlo algorithm. The aim is to build a recurrent Markov chain whose stationary distribution is the target one but avoiding the calculation of its normalizing constant. After a burn-in period to ensure convergence, the output of the chain is used as a sample from the target distribution. The Metropolis–Hastings algorithm is a popular version of such algorithms.

Let $\pi(\cdot)$ be the target density to simulate from. From the state x, the next state is x' simulated from a proposal density $q(\cdot \mid x)$. The jump from x to x' is accepted with probability $\alpha(x \mid x')$. Hastings (1970) proposed using

$$\alpha(x' \mid x) = \min\left\{1\,,\, \frac{\pi(x')q(x \mid x')}{\pi(x)q(x' \mid x)}\right\}.$$

This ensures that the chain of successive x so generated is reversible and converges to the unique stationary distribution $\pi(\cdot)$. The proposal density can be arbitrary, as long as $q(x \mid x') > 0$ whenever $q(x' \mid x) > 0$. In the Bayesian context, the target distribution is the posterior distribution. The ratios in $\alpha(\cdot \mid \cdot)$ avoid calculation of the constant normalizing π.

Green (1995) adapted the Metropolis–Hastings algorithm to varying dimension problems, allowing the study of mixture models from a fully Bayesian approach (Richardson

& Green 1997). In this context the number of components in a parameter vector is conditional on k, which is itself random. In order to jump from one dimension k to another, a countable family of move types m is defined. From the current state x, a move type m with destination x' with joint distribution $q_m(dx' \mid x)$ is proposed, and is accepted with probability

$$\alpha(x' \mid x) = \min\left\{1\,,\, \frac{\pi(dx')q_m(dx \mid x')}{\pi(dx)q_m(dx' \mid x)}\right\}.$$

Each move has to be reversible, in the sense that $q_m(dx \mid x')$ and $q_m(dx' \mid x)$ are positive together. In application, each move type m consists in a move forward, increasing the dimension, and a move backward, decreasing the dimension and reversing the move forward. A move type m that does not change the dimension is a classical Metropolis–Hastings jump. If the move type m changes the dimension and the destination x' is in higher dimension, a new random variable v is proposed and x' is set to $x'(x,\, v)$. The dimension of v is the increase of dimension due to the jump forward. Due to dimension reduction, the backward jump is deterministic. Thus the acceptance probability turns out to be

$$\alpha(x' \mid x) = \min\left\{1\,,\, \frac{\pi(x')r_m(x')}{\pi(x)r_m(x)q(v \mid x,\, m)}\left|\frac{\partial x'}{\partial(x,\, v)}\right|\right\},$$

where $r_m(x)$ is the probability to select a move type m when in state x and $q(v \mid x,\, m)$ is the density proposing v from state x when the move type is m. The acceptance probability of the backward move is

$$\alpha(x \mid x') = \min\left\{1\,,\, \frac{\pi(x)r_m(x)}{\pi(x')r_m(x')q(v \mid x',\, m)}\left|\frac{\partial x}{\partial(x',\, v)}\right|\right\}.$$

A particularly useful case is when v is chosen independently of the current state x. In this case, the forward move acceptance probability is

$$\alpha(x' \mid x) = \min\left\{1\,,\, \frac{\pi(x')r_m(x')}{\pi(x)r_m(x)q(v)}\left|\frac{\partial x'}{\partial(x,\, v)}\right|\right\}.$$

Output analysis

The interest of Bayesian methods lies in an analysis based on the whole posterior distribution of the parameters and not only in point estimates. In the case of mixture analysis with an unknown number of components, k, it may be advantageous to analyze a quantity whose dimension does not depend on k. The density of the data is such a parameter, furthermore, it contains most information of interest. Each iteration of the chain corresponds to a simulated density

$$f(\cdot \mid k,\, \pi, \theta) = \sum_{m=1}^{k} \pi_m f(\cdot \mid \theta_m).$$

The posterior distribution of this density may be studied via standard summaries, such as the posterior quantiles, median and mean. A conditional or an overall approach can be followed based on simulated $f(\cdot \mid k, \pi, \theta)$ for a fixed k or $f(\cdot \mid k, \pi, \theta)$ for every k, respectively. Richardson & Green (1997, p.745) advise against plug-in estimates of the form $f(\cdot \mid k, \hat{\pi}, \hat{\theta})$ which potentially 'give a poor over-smooth approximation of the predictive density'.

The posterior means $E\{f(\cdot \mid k, \pi, \theta) \mid k, y\}$ and $E\{f(\cdot \mid k, \pi, \theta) \mid y\}$ are not themselves finite mixtures of distributions in general, but their Monte Carlo estimates,

$$\begin{aligned} \hat{E}\{f(\cdot \mid k, \pi, \theta) \mid k, y\} &= |R_k|^{-1} \sum_{r \in R_k} \pi_m^{(r)} f(\cdot \mid \theta_m^{(r)}), \\ \hat{E}\{f(\cdot \mid k, \pi, \theta) \mid y\} &= |R|^{-1} \sum_{r=1}^{|R|} \sum_{m=1}^{k_r} \pi_m^{(r)} f(\cdot \mid \theta_m^{(r)}), \end{aligned}$$

are. Here, for every k, R_k is the set of indexes for which $k_r = k$ and $R = \cup_k R_k$. One can hence simulate data from the posterior mean estimate by uniformly resampling some r's in R_k or in R and simulating data from $f(\cdot \mid k_r, \pi^{(r)}, \theta^{(r)})$, if possible. This strategy is useful to obtain Monte Carlo estimates of some awkward characteristics of the density of the data. It is used in Chapter 3.

The output analysis should allow a diagnostic of (non-)convergence. General practice is based on the determination of a burn-in period during which the chain should have reached the stationary distribution and the burn-out period for the analysis, R according to our previous notation. The inspection of stationarity is often based on heuristic rules, mainly graphical. The output is plotted and stationarity is declared when each of them seems to have stabilized. In the case of varying dimension, those diagnostics should be done conditionally on k. As a consequence, it may be advantageous to make diagnostics on k and on parameters independent of k. However, one should be aware that reducing convergence to one or two graphics may lead one to miss a non-convergent aspect of the output.

Brooks & Giudici (2000) proposed a three-plot diagnostic based on a scalar summary quantity that should be independent of k. Several independent chains with dispersed starting points are launched in parallel. The diagnostic compares estimates of the global variance, the mean variance within models and the variance between models. If the chain has reached convergence, then overall and within-chain estimates should have converged toward the same quantities. Let c be the chain index, for $c = 1, \ldots, C$, m the model index, for $m = 1, \ldots, M$, and θ the summary statistic. Then the three diagnostics are

built according to the following principles:

a) Let V be the variance estimator based on the θ's in every chain and let VW_c be the variance estimator based on the θ's in chain c. Then V and $C^{-1}\sum_{c=1}^{C} VW_c$ should both have converged toward the same quantity, the global variance.

b) Let W_m be the variance estimator based on the θ's in model m in every chain and let W_mW_c be the variance estimator based on the θ's in model m in chain c. Then $M^{-1}\sum_{m=1}^{M} W_m$ and $C^{-1}\sum_{c=1}^{C} M^{-1}\sum_{m=1}^{M} W_mW_c$ should both have converged toward the same quantity, the mean variance within models.

c) Let B_m be the variance between models based on the θ's in every chain, that is

$$B_m = \frac{1}{M-1}\sum_{m=1}^{M}(\bar{\theta}_m - \bar{\theta})^2,$$

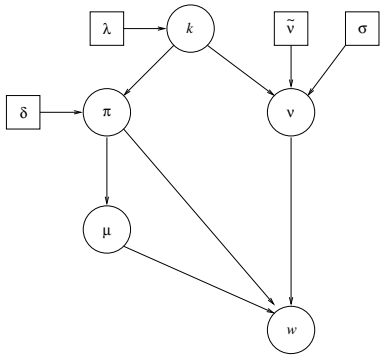
with an obvious notation. Let B_mW_c be the variance between model in chain c, then B_m and $C^{-1}\sum_{c=1}^{C} B_mW_c$ should both have converged toward the same quantity, the variance between models.

The fact that one must launch several chains in parallel makes these diagnostics heavy to compute. If the complexity is too high, then only one chain can be used. In this case, diagnostics consider convergence of the variance estimator without comparing global and within-chain variance estimators. In such case, the user may not be sure that the convergence is not around a local mode instead of a global one.

The choice of the scalar θ is critical since it must contain enough information to infer convergence. These diagnostics cannot be based on k, for example. Nevertheless, simple summaries on k, such as a histogram, should also be performed.

Application to extremal mixture models

For the extremal mixture model, the hierarchical scheme is given in Figure 2.2, in which F_μ is some complicated distribution function incorporating the constraints linking π with μ. Its construction can be found in Appendix A.5. The h_m are the Dirichlet components of the mixture, each with parameters $\left\{\mu_m^{(j)}\right\}_{j=1}^{d}$ and ν_m, for $m = 1, \ldots, k$. Comparing with Figure 2.1, the index variables u have been integrated out. The conditioning of μ given π due to the model constraint is a novel aspect compared to Figure 2.1. This adds technical difficulties, among which are the awkward form of F_μ and the choice of move types, which must ensure that the acceptance probability is not too low.



$$k \sim \text{Poisson}(\lambda),$$

$$\log \nu \mid k \sim \mathcal{N}_k\left(\log \tilde{\nu},\, \sigma^2 I_k\right)$$

$$\pi \mid k \sim \text{Dirichlet}(\delta_1, \ldots, \delta_k),$$

$$\mu \mid \pi,\, k \sim F_\mu,$$

$$w \mid \nu,\, \pi,\, \mu,\, k \sim \sum_{m=1}^{k} \pi_m h_m,$$

Figure 2.2: Hierarchical scheme for the extremal mixture model.

The hyperprior parameters must be specified by the user. If no prior information on a parameter is available, the prior distribution should be as vague as possible. The choice of λ may be guided by an inspection of the histograms of various components of w. Setting $\delta_m = 1$, for $m = 1, \ldots, k$, makes π uniformly distributed on $\mathcal{S}_k$ avoiding misleading prior information. Hyperparameters on $\log \nu$ are more sensitive since any choice implies prior information. An approach adapted to the data should be used, with a sensitivity analysis in case of doubt.

For the proposal, two move types have been defined. Suppose that current state of the algorithm is k, $\pi_1, \ldots, \pi_k$, $\mu_1, \ldots, \mu_k$ and $\nu_1, \ldots, \nu_k$. Then either one of the two following move types is proposed.

i) A 'SPLIT/COMBINE' move type: the forward step, 'SPLIT', divides one random component m_0 into m_1 and m_2. The backward step, 'COMBINE', merges m_1 and m_2 into m_0. Updates are done in such a way that constraints are preserved locally, that is

$$\pi_{m_0} = \pi_{m_1} + \pi_{m_2} \quad \text{and} \quad \pi_{m_0}\mu_{m_0} = \pi_{m_1}\mu_{m_1} + \pi_{m_2}\mu_{m_2}.$$

For the 'SPLIT', a random variable $v \in (0, 1)$ is simulated according to a Beta distribution and then $\pi_{m_1} = v\pi_{m_0}$ and $\pi_{m_2} = (1 - v)\pi_{m_0}$. Next μ_{m_2} is simulated according to a Dirichlet distribution on S_d and μ_{m_1} is set to $\pi_{m_1}^{-1}(\pi_{m_0}\mu_{m_0} - \pi_{m_2}\mu_{m_2})$. The scales $\log \nu_{m_1}$ and $\log \nu_{m_2}$ are independently simulated according to normal variables with mean $\log \nu_0$. The component m_0 is selected at random, uniformly. The 'COMBINE' sets $\pi_{m_0} = \pi_{m_1} + \pi_{m_2}$ and $\mu_{m_0} = \pi_{m_0}^{-1}(\pi_{m_1}\mu_{m_1} + \pi_{m_2}\mu_{m_2})$. The scale

$\log \nu_{m_0}$ is simulated according to a normal with mean $\log \nu_{m_1} + \log \nu_{m_2}$. The couple (m_1, m_2) is selected at random, uniformly. The Jacobian of the transformation

$$(\pi_{m_0}, \mu_{m_0}, v, \mu_{m_2}) \longmapsto (\pi_{m_1}, \pi_{m_2}, \mu_{m_1}, \mu_{m_2})$$

is $|\pi_0/v^d|$, see Appendix A.5;

ii) A 'MCMC' move type: this updates parameters π, μ and ν without changing k. This operation is done by randomly building a series of couples (m_{i_1}, m_{i_2}), with $1 \leq i_1 < i_2 \leq k$, then apply successively a 'COMBINE' move then a 'SPLIT'. The update hence obtained preserves the constraints.

The proposal distributions are defined according to a supplementary level of randomization. At each iteration, the size of the forthcoming move is simulated among 'BIG', 'MEDIUM' and 'SMALL'. If 'BIG' is selected and if the forthcoming move is 'SPLIT', then v is proposed according to a uniform on $(0, 1)$ and μ_{m_2} is proposed according to a uniform distribution on the simplex S_d, so that the proposition is far from the current state. If 'MEDIUM' or 'SMALL' is selected, then v is proposed according to a Beta distribution and μ_{m_2} according to a Dirichlet distribution. The parameters of those proposals are fixed by the user. Logically, a sharp shape of the proposal density should be attributed to 'SMALL' and a smoother shape for 'MEDIUM'. The scale parameters for the normal proposal of $\log \nu_{m_0}$ are defined by the user, the higher the variance, the bigger the move size. Finally, the selection of the move size is independent of the current state and its distribution vector is specified by the user.

The selection of the move type is independent of the move size and its distribution is to be fixed by the user. Once the move type is selected, the selection of the move is independent of the current state of the chain, except if $k = 1$, because in this case a 'COMBINE' is impossible. In details, the user specifies $p_c = P(\text{'COMBINE'})$ and $p_s = P(\text{'SPLIT'})$. The move type 'SPLIT/COMBINE' is selected with probability $p_{sc} = p_s + p_c$ and 'MCMC' with probability $p_m = 1 - p_{sc}$. If 'SPLIT/COMBINE' is selected, then a 'SPLIT' is done with probability $1 \cdot \mathbb{1}_{\{k=1\}} + \mathbb{1}_{\{k\neq 1\}} \cdot p_s/p_{sc}$ and 'COMBINE' with the complementary probability.

2.3.3 Discussion

The two algorithms are compared on simulated and real data in the next chapter. Nevertheless, their respective advantages and drawbacks can already be discussed. The back-

ground knowledge for applying the EM algorithm is simpler to acquire than that for the reversible jump, which needs more time and practice. The EM algorithm is in a frequentist framework which seems more realistic for multivariate extremes, where often no prior information is available. One the other hand, the EM algorithm is a non-stochastic algorithm and may encounter the curse of dimensionality when used with such a non-parsimonious model as the extremal mixture model, though a prior on k may be a good way to impose parsimony. The use of the EM algorithm becomes very complicated when confidence intervals are of interest, while uncertainty assessment is straightforward for the reversible jump algorithm. In general, the output of the reversible jump algorithm provides a complete characterization of the posterior distribution while the EM algorithm solves only a part of the problem.

Chapter 3

Algorithms and model performance

The first section presents an analysis of data simulated to assess practical performance of the EM and reversible jump algorithms. In Section 3.2 the performance of the extremal mixture model is assessed from simulated and real data.

3.1 Algorithm performance

The two fitting procedures are applied to a dataset simulated from the extremal mixture model described in Appendix A.4.

3.1.1 The EM algorithm

Figure 3.1 details the various fitted densities obtained for $k = 1, \ldots, 6$. Each row shows the components of the fitted density and the left-hand plot shows a histogram of data and the true and fitted density. The fit looks good for $k = 4$ and, for $k > 4$, additional components are either redundant or negligible. The AIC_c criterion, shown in Figure 3.2, selected the correct number of components. The parameter estimates are

$$\hat{\pi} = \begin{pmatrix} 0.4019_{(1.1\cdot 10^{-4})} \\ 0.2721_{(5.4\cdot 10^{-5})} \\ 0.1877_{(1.9\cdot 10^{-4})} \\ 0.1383 \end{pmatrix}, \quad \hat{\mu} = \begin{pmatrix} 0.5093_{(6.9\cdot 10^{-5})} & 0.4907 \\ 0.7971_{(2.7\cdot 10^{-5})} & 0.2029 \\ 0.2097_{(1.6\cdot 10^{-4})} & 0.7903 \\ 0.2828 & 0.7172 \end{pmatrix}, \quad \hat{\nu} = \begin{pmatrix} 0.9598_{(6.1\cdot 10^{-4})} \\ 20.186_{(4.5\cdot 10^{-4})} \\ 0.6018_{(6.1\cdot 10^{-5})} \\ 53.048_{(1.1)} \end{pmatrix},$$

with standard errors based on the information matrix indicated in brackets. Hence, the

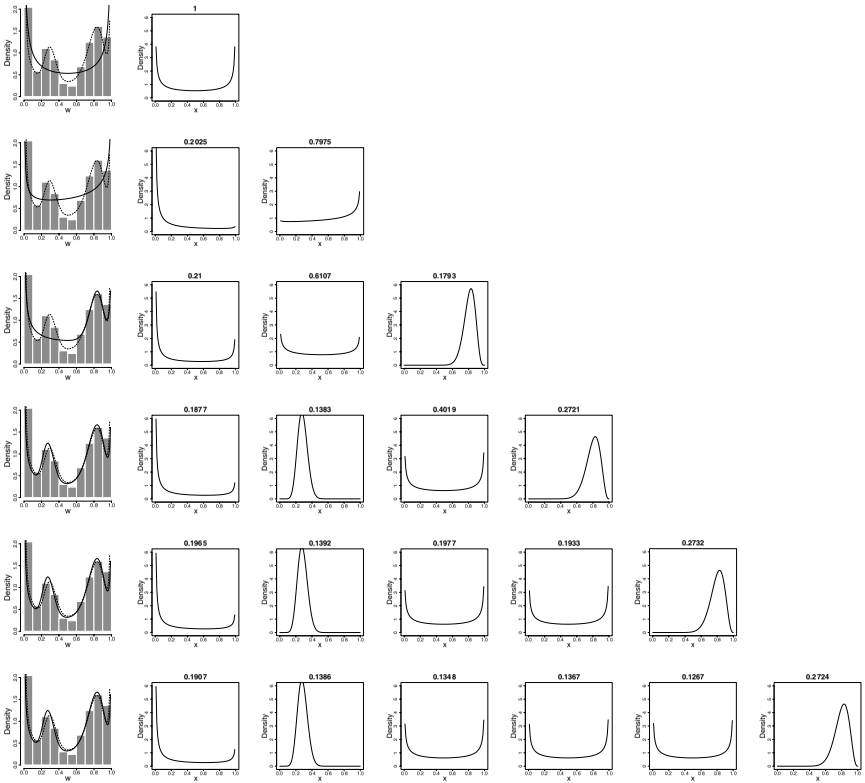


Figure 3.1: Results of the EM algorithm applied to data from the extremal mixture model. Left hand side plot: histogram of data, superimposed with the true density (dashed) and fitted density (full). From left to right: density functions of each component. Top to bottom: $k = 1$ to $k = 6$.

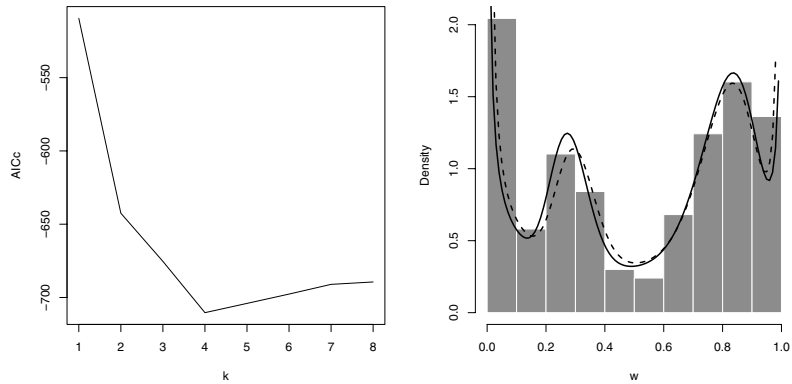


Figure 3.2: EM algorithm with AIC_c selection. Left: AIC_c score as a function of k. Right: histogram of simulated data, fitted density for $k = 4$ (full), true density (dashed).

EM algorithm and AIC_c criterion both seem to provide good fitting procedures. However this conclusion is brought up again in the discussion.

3.1.2 Reversible jump algorithm

The algorithm was run starting from $k = 1$, $\pi = 1$, $\nu = 1.18$ and $\mu = [0.5, 0.5]$. The value of ν was determined by the method of moments. After a $30,000$ iteration burn-in period, the final estimate based on $20,000$ iterations is shown in Figure 3.3. The fit looks good as the posterior median almost covers the true density. The posterior mean, not shown here, does the same. The credibility interval covers the true density function except maybe for the first mode that seems to be more on the right than the final estimate. The posterior distribution of k strongly favors $k = 4$.

Convergence diagnostics

Two other chains of $50,000$ iterations were run from starting points $k = 4$,

$$\pi = \begin{pmatrix} 0.5 \\ 0.25 \\ 0.125 \\ 0.125 \end{pmatrix}, \quad \mu = \begin{pmatrix} 0.5 & 0.5 \\ 0.8 & 0.2 \\ 0.1 & 0.9 \\ 0.3 & 0.7 \end{pmatrix}, \quad \nu = \begin{pmatrix} 0.9 \\ 20 \\ 1 \\ 5 \end{pmatrix},$$

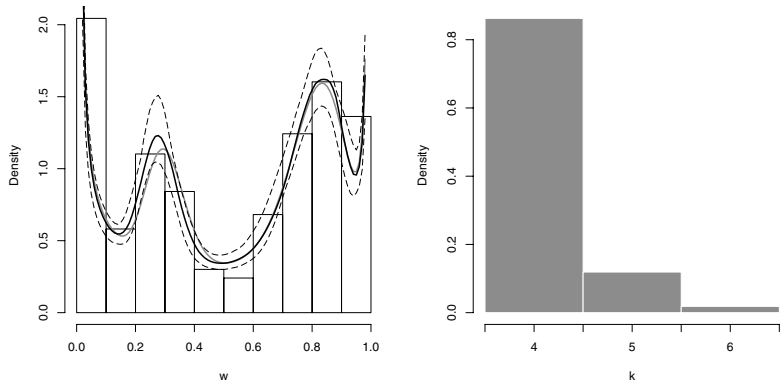


Figure 3.3: Reversible jump algorithm result. Left: histogram of the data, the true density (full gray), the posterior median (full black), 90% credibility interval (dashed). Right: posterior distribution of k.

which corresponds to the real density, and $k = 6$,

$$\pi = \begin{pmatrix} 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \end{pmatrix}, \quad \mu = \begin{pmatrix} 0.5 & 0.5 \\ 0.85 & 0.15 \\ 0.15 & 0.85 \\ 0.35 & 0.65 \\ 0.65 & 0.35 \\ 0.5 & 0.5 \end{pmatrix}, \quad \nu = \begin{pmatrix} 100 \\ 4 \\ 4 \\ 10 \\ 10 \\ 2 \end{pmatrix}.$$

The scalar used to assess the convergence according to Brooks & Giudici (2000)'s plots, described in Section 2.3.2, is

$$\begin{aligned} \theta &= \int_0^1 \min\{w, 1-w\} H(dw) = \int_0^{1/2} w H(dw) + \int_{1/2}^1 (1-w) H(dw), \\ &= \sum_{m=1}^{k} \pi_m \left\{ \mu_1^{(m)} \text{Beta}\left(0,\, 1/2;\, \alpha_1^{(m)} + 1,\, \alpha_2^{(m)}\right) + \mu_2^{(m)} \text{Beta}\left(1/2,\, 1;\, \alpha_1^{(m)},\, \alpha_2^{(m)} + 1\right) \right\}, \end{aligned}$$

where $\alpha_j^{(m)} = \nu_m \mu_j^{(m)}$. The interpretation of θ is the same as that of χ in Section 1.4. In the case of asymptotic independence, H gives mass $1/2$ at points 0 and 1 so that θ is 0, and in the case of asymptotic perfect dependence, H gives mass 1 at point $1/2$ so that θ is $1/2$. For intermediate situations, θ varies between 0 and $1/2$ indicating the strength of dependence.

The quantity θ was computed for the three chains every ten iterations to obtain $\theta_r^{(1)}$, $\theta_r^{(2)}$ and $\theta_r^{(3)}$ for $r = 1, \ldots, 5000$. The diagnostic is shown in Figure 3.4.

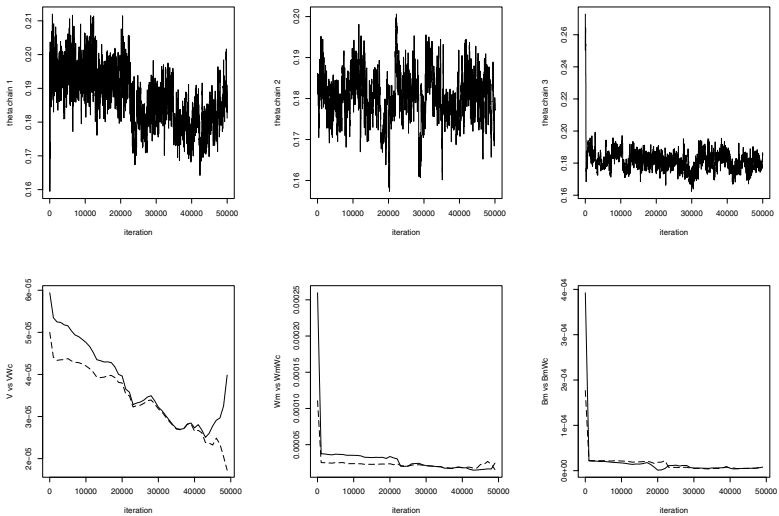


Figure 3.4: Reversible jump algorithm convergence diagnostics. Top line: parameter θ's path in chains 1,2 and 3. Bottom line: global variance (full) versus within chain variance (dashed); the left plot shows the total variance; the middle plot shows the intra-model variance; the right plot shows inter-model variance.

The intra-model and inter-model variances indicate convergence from the viewpoint of global estimation or mean chain estimation, though the upper panels do not seem fully satisfactory. The total variance does not seem to have stabilized. The global and the mean chain estimation have met, which is a good sign of convergence, but they are still decreasing. The series of $\theta^{(1)}$ shows a clear non-stationarity at the end of the series, which may be due to the imprecision of the estimation of the variance at the end of the series. This imprecision may be due to the sparsity of data there. This feature is also present in the two other plots but is hidden because of the scale. A lack of stationarity in the explored models for chain 1 is revealed by an inspection of the series of k for the three chains, shown in Figure 3.5. Chains 2 and 3 look more stable around $k = 4$.

This study of the non-convergence of the chain reveals the importance of the starting point of the algorithm. Indeed, the first chain has a starting points with $k = 1$, the furthest from the true density from the viewpoint of its shape. Although the first chain gives a good fit to the true density, its lack of convergence suggests it needs more iterations. The second chain shows no evidence of non-convergence and it is natural considering that its

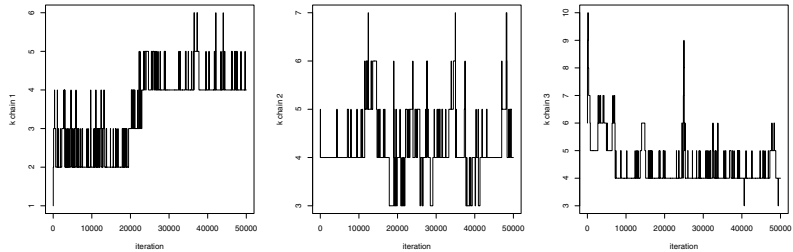


Figure 3.5: Reversible jump algorithm convergence diagnostics on k. From the left plot to the right: the path of k for chains $1, 2$ and 3.

starting point is the true density. The last chain also indicates convergence. Its starting point is a more complex model with $k = 6$ and potentially closer to the true model than the starting point of chain 1. From a practical viewpoint the first chain was in fact the fairest among the three, in the sense that a practitioner would naturally use such a generic starting point. Nevertheless, the non-convergence would have been surely detected with diagnostics based on only one chain.

Sensitivity analysis

Two causes to the sensitivity to prior specification can be distinguished: the prior specification on k and the prior specification on the other parameters conditional on k. This latter aspect is difficult to quantify since the influence of the prior varies as the chain moves from one state to another. For example, the influence of the prior on π may inflate as k grows. Below we study the sensitivity of the Poisson prior distribution on k to the hyperparameter λ and the sensitivity of the prior normal on $\log \nu$ to the hyperparameters $\log \tilde{\nu}$ and σ. The influence of the hyperparameter on π is not addressed, because with $\delta = 1$, it is the uniform distribution for any k. The prior is therefore uninformative and should always be used in practice.

To assess the sensitivity to the prior specification of k, three chains were ran for $20,000$ iterations, after a $30,000$ iteration burn-in period. The results are shown in Figure 3.6. The hyperpriors on $\log \nu$ were fixed to $\log \tilde{\nu} = 0$ and $\sigma = 10$ for the three chains. The final estimate when $\lambda = 1$ is not good and misses one component, though the credibility intervals cover the true density. When $\lambda = 2$ or 8 the fit is good. Furthermore, the histogram of k shows even when $\lambda = 8$ the number of components remains around four and five. This means that taking a high λ does not involve an artificial explosion of the

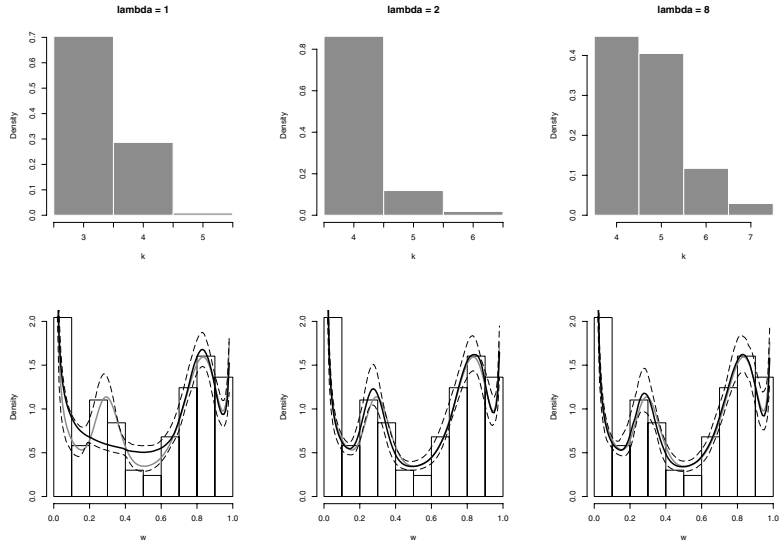


Figure 3.6: Sensitivity to k prior specification. From left to right $\lambda = 1, 2, 8$. Top: posterior histogram of k. Bottom: histogram, true density (full gray), posterior median (full black), 90% credibility intervals (dashed).

number of components.

For assessment of the sensitivity to the prior specification $\log \nu$, three chains were ran for $20,000$ iterations, after a $30,000$ iteration burn-in period. The results are shown in Figure 3.7. From the left to the right, the hyperpriors on $\log \nu$ were fixed to $\log \tilde{\nu} = 0$ and $\sigma = 1$, $\sigma = 10$ and $\sigma = 100$. The hyperprior parameter on k was fixed to $\lambda = 2$ for the three chains. The conclusion is dramatic for small σ. The influence on the algorithm is crucial. In detail, for this data set, the range of $\log \nu$ is $\log 0.9 = -0.105$ up to $\log 50 = 3.91$. With a mean 0 and a standard deviation 1, the prior of $\log \nu$ misses $\log 20 = 2.996$ and $\log 50 = 3.91$, corresponding to the two central bumps, with a probability higher than 95%. Furthermore the chain was ran for another $50,000$ iterations without any sign of improvement. This effect is serious and may compromise an analysis. Naturally, one should try to be uninformative, which means taking σ as large as possible. In practice however, the algorithm may then encounter numerical problems.

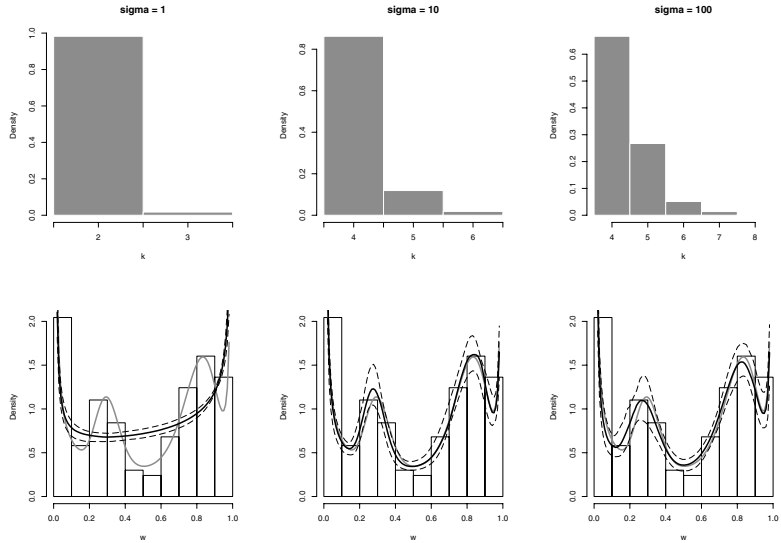


Figure 3.7: Sensitivity to $\log \nu$ prior specification. From left to right $\sigma = 1, 10, 100$. Top: posterior histogram of k. Bottom: histogram, true density (full gray), posterior median (full black), 90% credibility intervals (dashed).

3.1.3 Discussion

Both methods of estimation have advantages and drawbacks. The EM algorithm is easy to implement, in principle at least, if a numerical optimizer is available. However confidence intervals are very hard or even impossible to obtain in particular for parameters that are not in the original parameterization. This is a serious drawback since in general the parameters of the extremal mixture models are not of interest compared to the dependence measure or return levels. Using the EM algorithm in such a context would involve developing methods for uncertainty assessment other than those presented in this work. One possibility is a bootstrap procedure but the time required would make such a method impracticable. As a second drawback, the EM algorithm is implemented for incomplete data coming from a mixture distribution. In particular, it is impossible to use it for componentwise maxima, the multivariate extreme value distribution not being itself a mixture since

$$P\{M_n \leq x\} = \exp\left\{-d \int_{S_d} \max_{j=1,\ldots,d} \left(\frac{w_j}{x_j}\right) H(dw)\right\}.$$

A third drawback is the slowness of the EM algorithm. Procedures for speeding it up have been proposed in the literature (Meng & van Dyk 1997), but they have not been tried here. In this context, the constraints on the parameters present a technical challenge.

The reversible jump algorithm is more complex in its theory and implementation but the application looks more natural than the EM algorithm. As a first drawback, the reversible jump algorithm needs a working prior specification although in general no prior information on the parameters of an extremal mixture model is available. Furthermore, convergence is not guaranteed and must be part of the inference. Nevertheless, the fit provided by the posterior mean or median turned out to be satisfactory. The algorithm provides uncertainty assessment on k as well as on any scalar or vector quantity whose size is independent of k, such as θ. Uncertainty assessment for π, μ and ν is more difficult because of labeling issues. This problem may be solved by imposing further constraints on the parameters in order to follow the path of each component. One possibility would be to sort π but this extra complexity may dramatically slow down the convergence of the algorithm. In our case, we use the mixture as a semi-parametric model, following the discussion of Richardson & Green (1997, p.785): 'in this case, no interpretation should be given to the components and [...] only summaries which are invariant to the labeling, like density estimates [...], should be produced.'. As a final comment, the algorithm is computationally intensive and so may be slow. However, in our experience, it usually turns out to be faster than EM algorithm. Furthermore, it offers a more complete exploration of the posterior density than does maximum likelihood estimation.

In view of its greater flexibility, the reversible jump algorithm is favored in the rest of this thesis.

3.2 Model performance

Below we study the performance of the extremal model itself. This section is divided into two parts, the first a simulation study, the second an analysis of oceanographic data.

3.2.1 Simulated data analysis

Datasets of length 500 from classical spectral density functions were simulated and the extremal mixture model was fitted using a reversible jump algorithm. The initial point is always $k = 1$ and ν is chosen by the method of moments. The analysis is based on $30,000$ iterations after a $50,000$ iteration burn-in period. The three first datasets were

simulated from the extremal logistic model in dimension $d = 2$ with parameters $(0.2, 0.2)$, $(0.8, 0.8)$ and $(0.6, 0.6)$. These parameters were chosen for the three types of density shapes they provide. The results are shown in Figure 3.8. For parameters $(0.2, 0.2)$ and $(0.8, 0.8)$, the fits look good. In the first case, the true density falls into the credibility intervals and in the second case it is visually indistinguishable from the posterior median. In the case of parameter $(0.6, 0.6)$, the final estimate is quite far from the true density. The shape of the credibility interval indicates the path of the algorithm with almost fixed points around $w = 0.15$ and $w = 0.85$. One possible reason is the great sensitivity of the shape of the symmetric extremal logistic model density around $\alpha = 0.5$, that is small changes in α change the density quite dramatically. The histogram itself is not representative of the density. To explore this phenomenon, an unrealistic experiment with $10,000$ data simulated from a symmetric extremal logistic model with parameter $(0.6, 0.6)$ was conducted. The results are shown in Figure 3.9. The first line shows various shapes of the symmetric density for parameters around 0.5. This illustrates the great variability encountered. The second line shows the results of the reversible jump algorithm and the visited k. The histogram, and hence the data, is more representative of the density. Here, the fit looks quite good although the convergence could be improved, in view of the path of k.

The fourth to fifth datasets are simulated from asymmetric extremal logistic models with parameters $(0.2, 0.8)$ and $(0.6, 0.4)$. Results are given in Figure 3.10. Here again, the model seems to capture the main part of the true density. The gray lines representing the true density are cut at 0.1 and 0.9 because of the implicit form of the density. Solving the required equation is numerically difficult beyond these limits. The extremal mixture model does not suffer from this drawback.

The next three datasets were simulated from the extremal Dirichlet model with parameters (α, β) equal to $(1, 0.2)$, $(3, 0.4)$ and $(10, 2)$, respectively. Results are presented in Figure 3.11 in a similar way as before. The fit is fairly good and this experiment confirms that the final fit is really close to the histogram.

In order to illustrate the performance in dimension $d = 3$, two datasets were simulated from an extremal Dirichlet model with parameters (0.5, 1, 30) and an extremal logistic model with parameters (0.2, 0.5, 0.8). The goodness of fit is judged on the bivariate density function of the two first components of the data. Results are given in Figure 3.12. The conclusion is the same as in dimension $d = 2$. The fits looks adequate in view of the

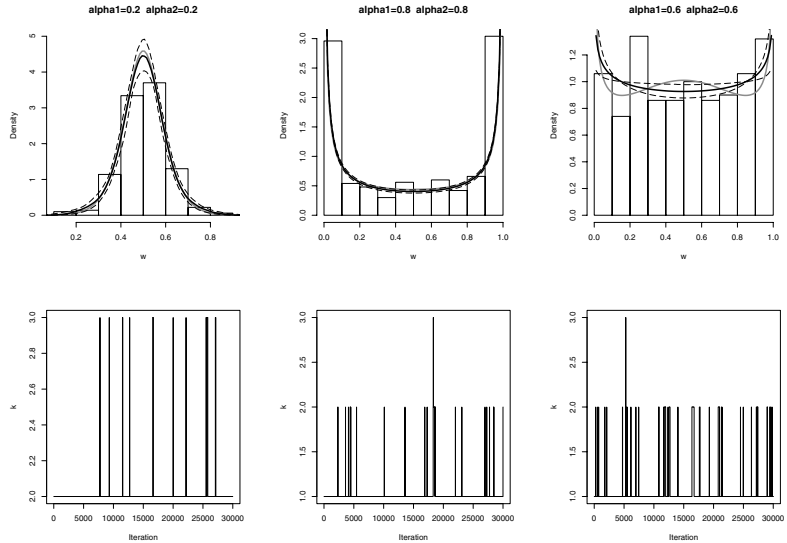


Figure 3.8: Fitting the symmetric extremal logistic model with the extremal mixture model. Top: histograms superimposed by the true density (gray), the posterior median estimates (black) and the 90% credibility interval (dash). Bottom: the path of k.

data, which are more representative of the true density in the extremal Dirichlet model case than in the extremal logistic model. As in the case $d = 2$, an unrealistic simulation with $5,000$ data from the same extremal logistic model was conducted. Results are shown in Figure 3.13. Here again, the extremal mixture model is close to the data.

3.2.2 Real data analysis

In order to illustrate performance on real data, the model was fitted to the dataset from Coles & Tawn (1994) described in Appendix A.4.

In their analysis, Coles & Tawn (1994) used the semi-parametric extremal model for the margins and the extremal Dirichlet model for the dependence structure. The multivariate threshold r_0 was selected to be exp(3.3), giving 222 joint excesses. The marginal thresholds have been chosen in a compatible manner, that is $p_k = 222/2895$ and u_k is the $(1 - p_k)$-empirical quantile of $X_1^{(k)}, \ldots, X_n^{(k)}$. Although not necessary, this choice simplifies the application. The authors obtain $u_1 = 6.59$, $u_2 = 11.6$ and $u_3 = 0.351$. The marginal parameters and the parameters of dependence were simultaneously

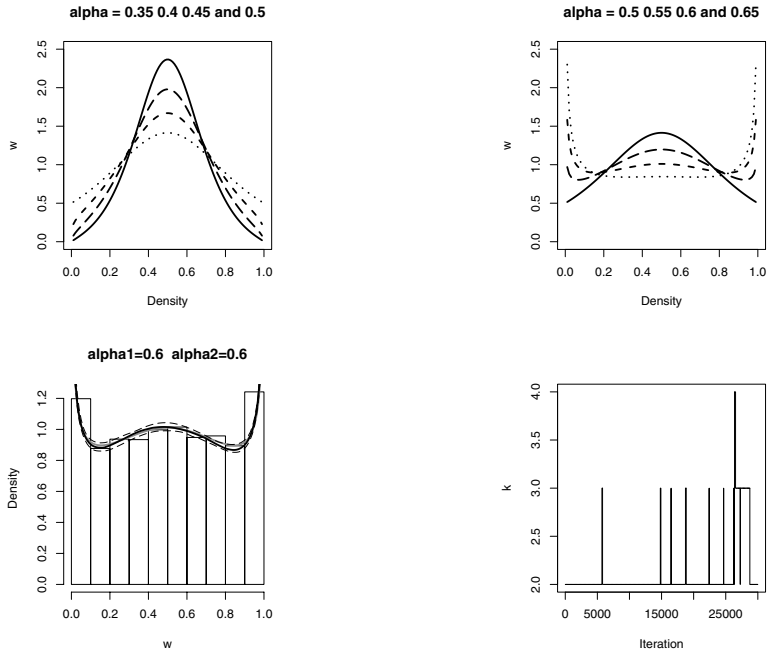


Figure 3.9: The sensitivity of the symmetric extremal logistic model for α around 0.5. Top: true h for $\alpha = 0.35$ up to 0.65; the more continuous the line, the lower α. Bottom left: histogram of $10,000$ data with $\alpha = 0.6$ superimposed by the posterior median estimates (black) and the 90% credibility interval (dash); bottom right: the path of k.

estimated by maximum likelihood. The fitted value for the dependence parameter is $\hat{\alpha} = (0.497, 0.985, 0.338)$. Coles & Tawn (1994) did not indicate the values of the marginal parameters.

An extremal mixture model is fitted with the reversible jump algorithm modified to take into account the marginal parameters and the parameters of dependence simultaneously. A central interest in Coles & Tawn (1994) is estimation of the return level. This subject is addressed in the next chapter. Technical aspects of the reversible jump algorithm for the dependence and the margin parameters are explained in Appendix A.5.3.

The results are presented in Figure 3.14, which shows the spectral density contours. The gray dots are the pseudo-angles obtained by non-parametric transformation. The estimate with the extremal mixture model seems to offer a better fit to the data than the extreme Dirichlet model. The fit looks better in the center of the simplex, which

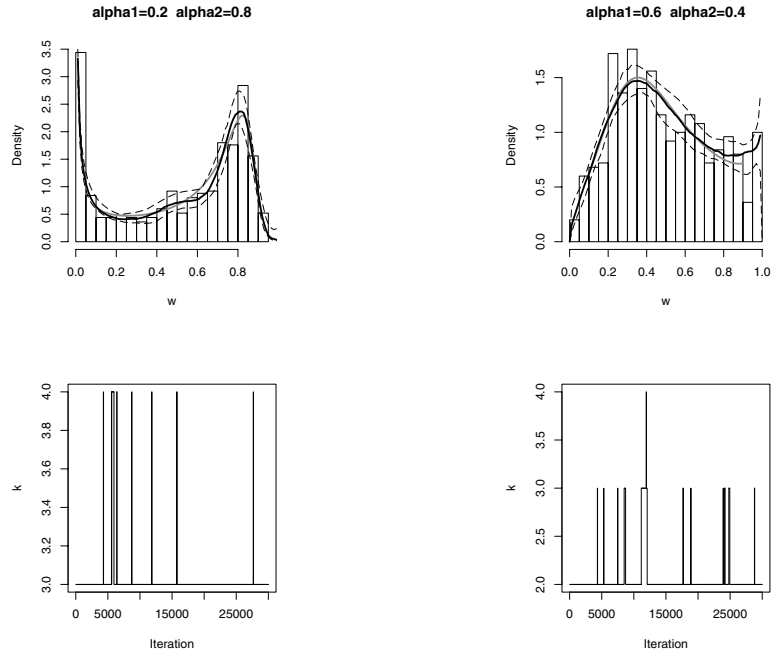


Figure 3.10: Fitting the asymmetric extremal logistic model with the extremal mixture model. Top line: histograms of simulated data superimposed with the true density (gray), the posterior median estimates (black) and the 90% credibility interval (dash). Bottom line: the values of k visited by the algorithm.

indicates that the model takes the dependence of the pseudo-angles into account. Secondly, the fit looks better at the edges showing that the model is sensitive to some of the asymptotic independence structure in the data. Most of the density is concentrated at the left corners of the simplex and along the diagonal edge. In view of this picture, one could suspect asymptotic independence of the couple (X_2, X_3), that is, the surge and the period. This important characteristic of the dataset is more difficult to detect with the extreme Dirichlet model. For example, looking at the fitted curves at edges $\{w_1 = 0\}$ and $\{w_2 = 0\}$, we see that the extremal Dirichlet model attributes stronger asymptotic dependence to the first edge than to the second, in contradiction with the data, which are more closely fitted by the extremal mixture model. Here we see that the smoothness of the extremal Dirichlet model is perhaps a little excessive and therefore that estimation at the edges is influenced by the data in the center of the simplex.

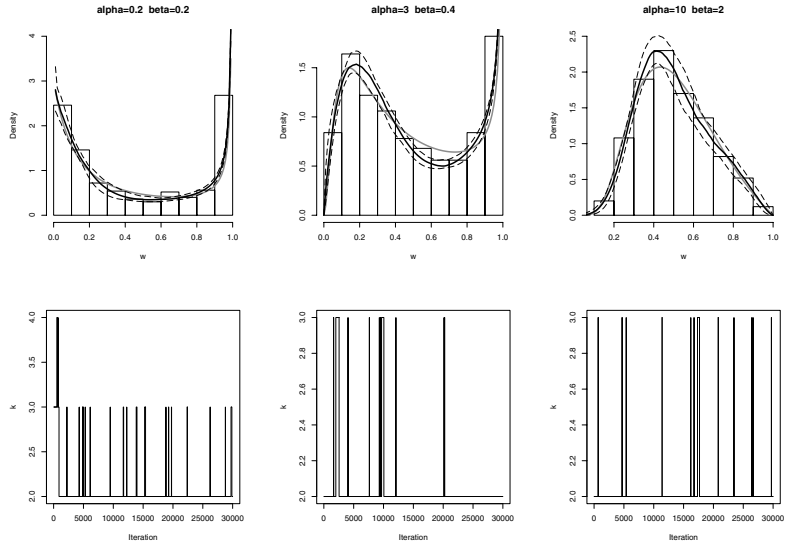


Figure 3.11: Fitting the symmetric extremal Dirichlet model with the extremal mixture model. Top line: histograms of simulated data superimposed with the true density (gray), the posterior median estimates (black) and the 90% credibility interval (dash). On the bottom line: the path of k.

3.2.3 Discussion

In experiments, the extremal mixture model has demonstrated its flexibility, allowing for a large variety of shapes. In particular, this model fits the data very closely when the true density and the histograms are not close to one another. This situation was observed for data simulated from an extremal logistic model with one parameter close to 0.5. The extremal mixture model thus appears to be a good model for spectral densities, at least as good as the extremal logistic and the extremal Dirichlet models. In particular, it is easier to use than the extremal logistic, which has no explicit likelihood, except in its symmetric form. Furthermore, the reversible jump algorithm turns out to be a good procedure even if it could not be used in a fully automatic way for these experiments.

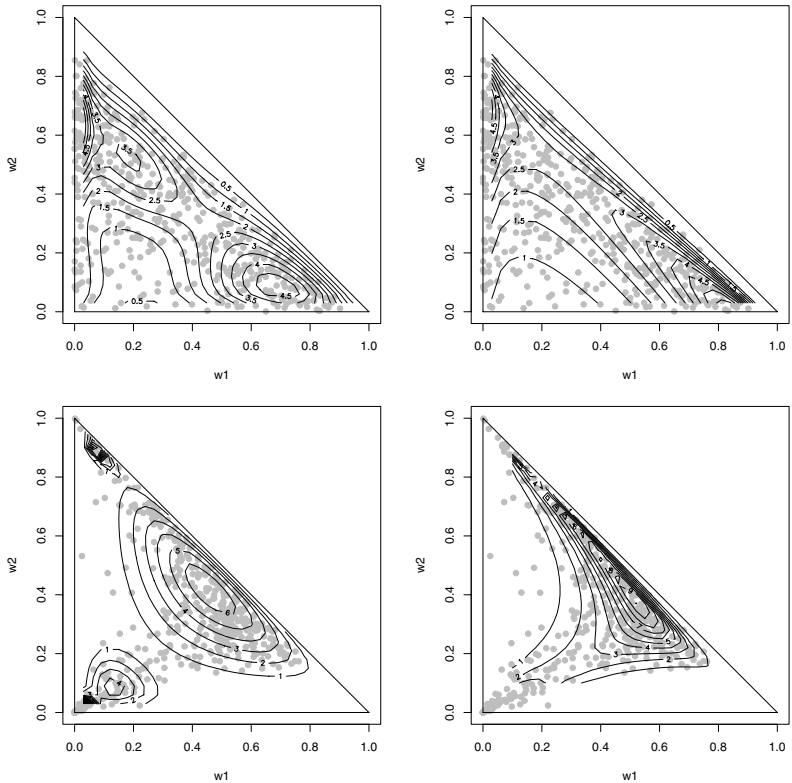


Figure 3.12: Extremal mixture model performance in three dimensions. Top left: extremal Dirichlet model data fitted by extremal mixture model; top right: true density. Bottom left: extremal logistic model data fitted by extremal mixture model; bottom right: true density.

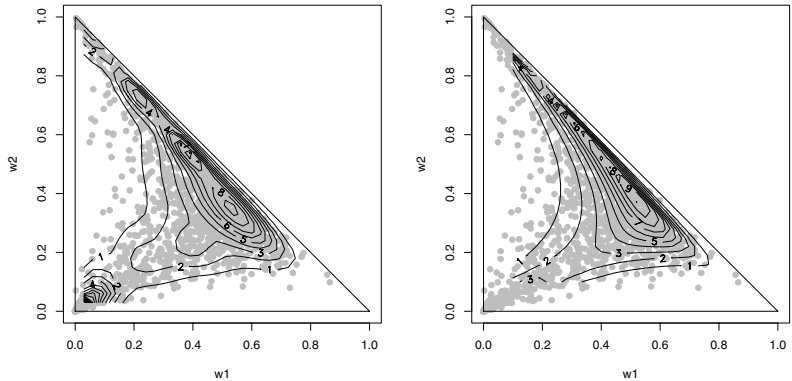


Figure 3.13: Sensitivity of the extremal logistic model. Left: the extremal mixture model fit. Right: the true density.

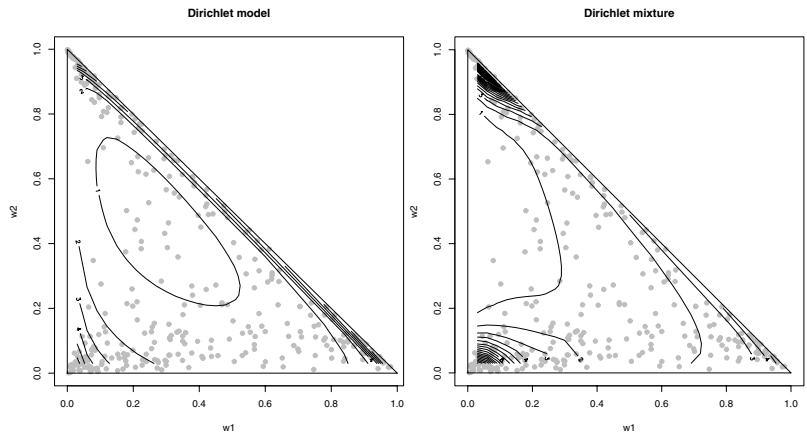


Figure 3.14: Spectral density contours on oceanographic data. Left: the original extremal Dirichlet estimate. Right: the extremal mixture model posterior mean estimate.

Chapter 4

Return level estimation and dependence analysis

This chapter deals with practical aspects of extreme value analysis using the extremal mixture model. It is divided into three sections. The first presents diagnostics for the selection of the multivariate threshold for the Poisson process approach. It is not directly linked with the extremal mixture model but is used in the following sections. In Section 4.2, the estimation of return levels via Monte Carlo estimates is addressed. Section 4.3 presents an original dependence analysis of a real dataset using the extremal mixture model.

4.1 Selection of the multivariate threshold

In this section we propose three diagnostics to choose the threshold in the multivariate case. The justification is based on the Poisson limit.

4.1.1 Three diagnostic plots

Let $X_1, \ldots, X_n$ be an independent and stationary sample in $\mathbb{R}^d$ with Fréchet margins, let $R_i = \sum_{j=1}^{d} = X_i^{(j)}$ and let $W_i = X_i/R_i$. For a threshold r_0, let $I_0 = \{i : R_i > r_0\}$. Then three features (null hypothesis) are tested:

i) for $i \in I_0$, R_i is distributed according to the distribution function $F(r) = 1 - r_0/r$, $r > r_0$;

ii) for $i \in I_0$, W_i has mean vector $(d^{-1}, \ldots, d^{-1}) \in \mathbb{R}^d$;

iii) for $i \in I_0$, R_i and W_i are linearly independent.

The p-value of each test is plotted versus $\log r_0$. The threshold is selected when the p-value is acceptable. Three classical tests are used:

i) the Kolmogorov–Smirnov test (Davison 2003, p.328). This is based on the statistic

$$\max_{i \in I_0} \left\{ i/n - U_{(i)}, U_{(i)} - (i-1)/n \right\},$$

where $U_i = 1 - r_0/R_i$, $i \in I_0$, and $U_{(i)}$ is the ith order statistic of $\{U_l\}_{l \in I_0}$. The p-value has to be calculated numerically and is available from standard statistical software, for example in `R` by function `ks.test`.

ii) the multivariate T-test. It is based on Hotelling's T^2 statistic (Davison 2003, p.260)

$$T^2 = n_0 \left(\bar{W} - \mu_0\right)^T S^{-1} \left(\bar{W} - \mu_0\right) \sim \frac{(d-1)(n_0-1)}{n_0-(d-1)} F_{d-1, n_0-(d-1)},$$

where $\bar{W}$ is the sample mean of $\left(W_i^{(1)}, \ldots, W_i^{(d-1)}\right)_{i \in I_0}$, $\mu_0 = (d^{-1}, \ldots, d^{-1}) \in \mathbb{R}^{d-1}$, n_0 is the number of excesses and $F_{\alpha,\beta}$ is the F distribution with parameter α and β. The F distribution is available from most statistical software.

iii) the linear dependence test between $\log R$ and W. A normal linear model is fitted,

$$\log(R_i/r_0) = \left[1, W_i^{(1)}, \ldots, W_i^{(d-1)}\right] \beta + \varepsilon_i, \quad i \in I_0,$$

with $\beta = (\beta_0, \ldots, \beta_{d-1})^T$ and $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. The linear dependence test

$$\mathcal{H}_0 : \beta_1 = \cdots = \beta_{d-1} = 0 \quad \text{vs} \quad \mathcal{H}_1 : \exists j \in (1, \ldots, d-1) \, : \, \beta_j \neq 0,$$

is based on the statistics

$$\frac{\{\mathrm{SS}(\hat{\beta}_0) - \mathrm{SS}(\hat{\beta})\}/(d-1)}{\mathrm{SS}(\hat{\beta})/(n_0-d)} \sim F_{d-1, n_0-d}.$$

where $\mathrm{SS}(\hat{\beta}_0)$ is the residual sum of squares under $\mathcal{H}_0$ and $\mathrm{SS}(\hat{\beta})$ is the residual sum of squares under the full model (Davison 2003, p.379). The p-value of this test is available from statistical software.

4.1.2 Performance on simulated and real data

In order to illustrate its performance on real data, the diagnostics were applied to Newlyn dataset, from Coles & Tawn (1994), described in Appendix A.4. The margins are transformed to the unit Fréchet using the empirical distribution function. The result of the

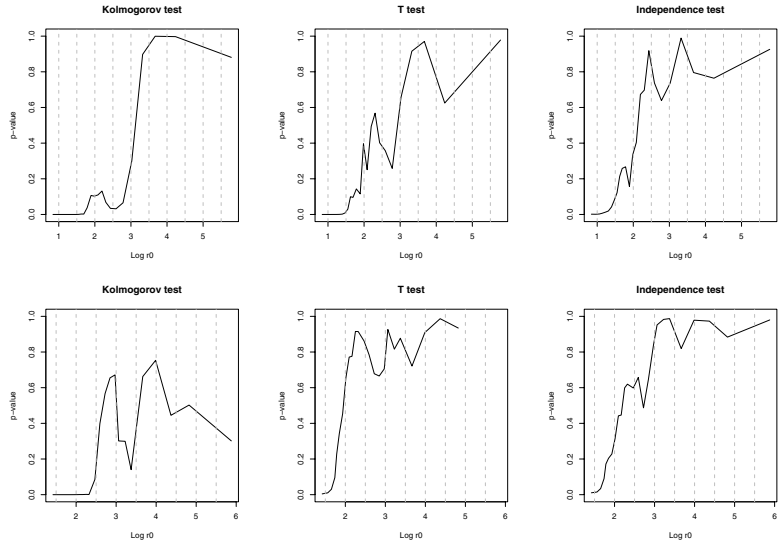


Figure 4.1: Selection of the threshold for real data. Top: Newlyn data. Bottom: air quality data.

diagnostics is shown in the top line of Figure 4.1. It is hard to draw a rigorous conclusion, but the threshold exp(3.3) seems reasonable. Coles & Tawn (1994) chose this value using a visual selection procedure based on the stabilization of the histograms of W_i as r_0 increases. Such a procedure is much more difficult in dimension greater than two, unlike this diagnostic which does not depend on the dimension. The bottom line of Figure 4.1 shows the result when applied to air quality data, from Heffernan & Tawn (2004), see Appendix A.4. In this example the selection is less obvious; we take $r_0 = \exp(3.6)$ below.

For the simulation study, four bivariate datasets were simulated. The size of each dataset is 5000. The margins are transformed to the Fréchet scale using empirical transformation.

1. A mixture of 4800 standard normal data with correlation $\rho = 0.5$, margins transformed to unit Fréchet, restricted to the set $\{r \leq r_0\}$ and 200 bivariate data from the extremal Poisson model whose spectral density being asymmetric logistic with parameters $(0.2, 0.8)$ and R restricted to $\{r > r_0\}$. The threshold r_0 equals $\exp(3.86)$.

2. The same mixture but the spectral measure of the Poisson model is an extremal

Dirichlet with $\alpha = (0.4, 0.6)$.

3. The inverted multivariate extreme value distribution with a logistic spectral function with parameters $\alpha = (0.42, 0.42)$.

4. A normal distribution with correlation $\rho = 0.5$.

The two first datasets are constructed in order to build an artificial threshold r_0. The two last datasets were chosen because they exhibit asymptotic independence so that the classical extremal Poisson characterization is inappropriate. The results are shown in Figure 4.2. For the first two datasets, the artificial threshold appears appropriate on the diagnostic plots. In fact it seems in both cases that an even lower threshold would be appropriate. For the two last datasets, the diagnostic plot does not reveal any appropriate threshold. In particular, the three tests show strong incoherence; a good threshold for the Kolmogorov's test on R turns out to be inappropriate for Hotelling's test on the mean of W.

4.1.3 Discussion

The three-diagnostic plots have both positive and negative points. Firstly, they are independent of the dimension of the data. Three graphs are enough to make a first selection of the threshold even if a more detailed selection may be needed afterward. Secondly, they only require standard routines available in most software packages and are very easy to compute. However, this selection remains mainly heuristic. Firstly, the hypothesis of normality for two among the three tests is not satisfied. This deficiency is worsened by the fact that the region of interest in the plots corresponds to p-value of tests done with very few data. The power of such a procedure is likely to be very low. Secondly, there is no uncertainty assessment in such an approach. There is no guide to what kind of p-value is high enough for the threshold to be appropriate. Furthermore, as the threshold is not a parameter, there is no true threshold for a model and hence no numerical study of the performance of these diagnostics can be done. We conclude that these diagnostics are a step toward a more rigorous threshold selection but they remain heuristic and should not be used in any automatic procedure.

As a final comment, the test on the mean of W could be replaced by an empirical likelihood test. This would relax the hypothesis of normality but, on the other hand, would imply an explosion in the computation time. There must be other directions in which these diagnostics can be improved.

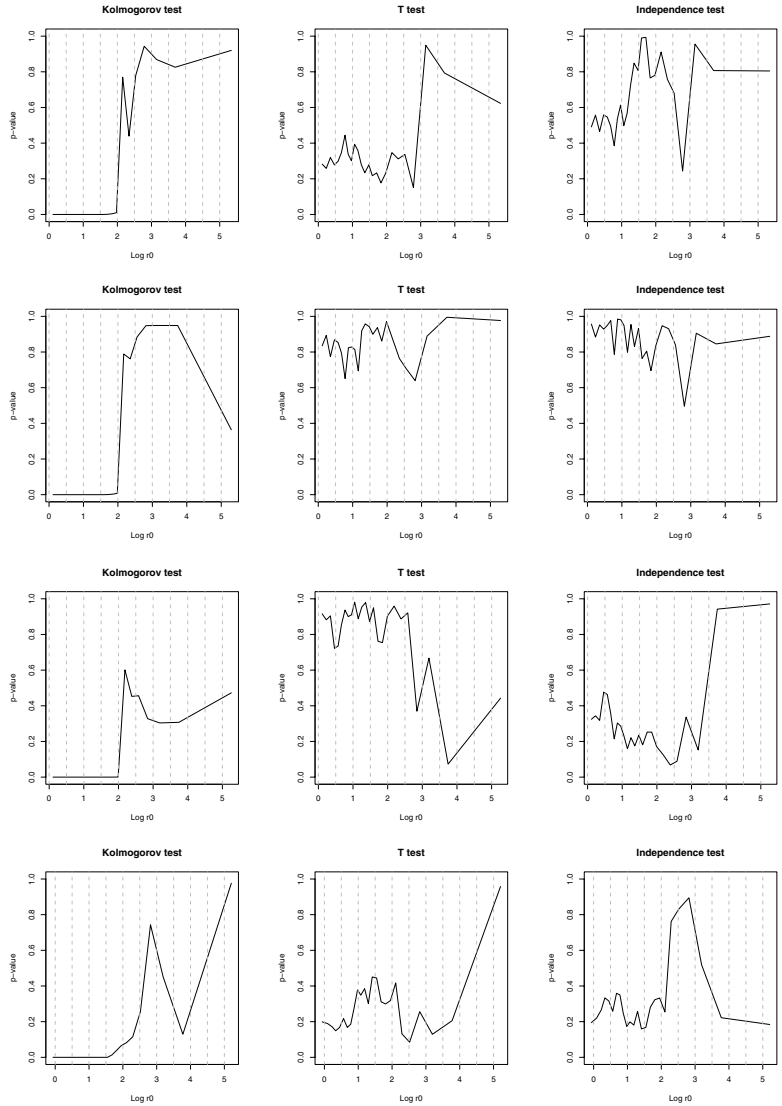


Figure 4.2: Selection of the threshold for simulated data. From top to bottom: normal and Poisson with logistic H, normal and Poisson with Dirichlet H, inverted multivariate extreme value distribution and normal dataset.

4.2 Simulation of multivariate extremes

Simulation of multivariate extremes refers to simulation either from a multivariate extreme value distribution or from the extremal point process. This section concentrates on the second case. The first case is treated for logistic type distributions in Stephenson (2003). A general method, although difficult to put in practice, can be found for the bivariate case in Kotz & Nadarajah (2000, p.142–143), where the simulation from the extremal point process is also presented with a focus on the use of the rejection algorithm for simulating from the spectral measure. Their idea only requires simulation from a Poisson process, but the marginalization problem is not addressed, and methods and applications in the multivariate case are only sketched. We here present more general methods for simulation of point processes for multivariate extremes; despite their simplicity we have not found them in explicit form in the literature, though there is related work by Bruun & Tawn (1998).

4.2.1 Generalities

The simulation is based on the semi-parametric model implied by the Poisson process limit. In other words, for a given selected threshold r_0, a simulated datum is either in the set $\{r \leq r_0\}$ with probability $1 - p_0$, in which case it is not extreme and is distributed according to the empirical distribution function, or in the set $\{r > r_0\}$, in which case it is extreme and distributed according to the Poisson process with intensity measure

$$d\frac{r_0}{r^2}dr \times H(dw), \quad w \in S_d, \quad r > r_0$$

where H is the spectral probability measure. In order to simulate a fixed number n_0 of points from a Poisson process in a compact set A, one has to simulate an independent and identically distributed sample of size n_0 according to the intensity measure restricted to A and properly normalized. Therefore, the simulation scheme of N extreme events in the set $\{r_1 < r \leq r_2\}$, $r_1 \geq r_0$, is obtained by repeating the following algorithm N times:

1) Simulate $U \sim \mathcal{U}(0,1)$ and set $R := r_1 \{1 - (1 - r_1/r_2)U\}^{-1}$;

2) Simulate $(W^{(1)}, \ldots, W^{(p)}) \sim H(dw)$;

3) Set $X^{(j)} := RW^{(j)}$, $j = 1, \ldots, p$.

The simulation of R comes from the fact that

$$P\{R \leq r \mid r_1 < R \leq r_2\} = \frac{P\{r_1 < R \leq r \mid R > r_0\}}{P\{r_1 < R \leq r_2 \mid R > r_0\}} = \frac{1 - r_1/r}{1 - r_1/r_2},$$

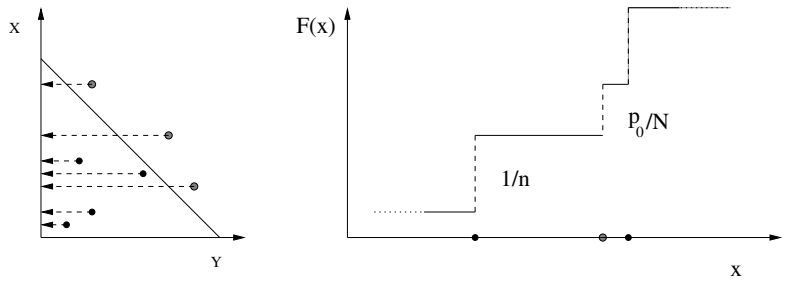


Figure 4.3: Illustration of the mixed empirical distribution function. Left: simulated extremes are in gray, observations in black. Right: the empirical cumulative distribution function increases by $1/n$ at each observation and by p_0/N at each simulated extreme.

for $r_1 < r \leq r_2$. Extreme events can hence be simulated at arbitrarily high levels. In dimension two, this algorithm is described in Kotz & Nadarajah (2000, p. 143) where they take $r_1 = r_0$ and $r_2 \to \infty$ and simulate from H using a rejection sampling algorithm.

In practice, it can be useful to transform the simulated data back to the original scale of the observations. The data simulated in $\{r > r_0\}$ have not the same weight as the original data in $\{r \leq r_0\}$. Let n be the total number of data, n_0 the number of data in $\{r > r_0\}$, and N the number of data simulated in $\{r > r_0\}$. Then, ignoring observed data in $\{r > r_0\}$, the empirical distribution function of $X^{(j)}$ increases by $1/n$ at each $x_i^{(j)}$ observed in $\{r \leq r_0\}$ and increases by $(n^{-1}n_0)N^{-1}$ at each datum simulated in $\{r > r_0\}$. Formalizing, let I_0 be the index set such that $i \in I_0$ means $x_i \in \{r > r_0\}$. Then the empirical distribution function of the j-th margin is

$$F_j(x) = \frac{1}{n} \sum_{i \in I_0^c} \mathbb{1}_{\{x_i^{(j)} \geq x\}} + \frac{n_0}{n} \frac{1}{N} \sum_{i \in I_0} \mathbb{1}_{\{x_i^{(j)} \geq x\}}.$$

In other words, the final empirical distribution function mixes the original empirical distribution function restricted to $\{r \leq r_0\}$ with the empirical distribution function of data simulated in $\{r > r_0\}$. The weights of this mixture are respectively $1 - p_0$ and p_0, where p_0 is the probability of being in $\{r > r_0\}$. Those $x_i^{(j)}$ from $\{r > r_0\}$ are not necessarily themselves greater than r_0. Figure 4.3 gives a schematic illustration. This mixed empirical distribution function can then be used to transform the data to the uniform scale, then back to the original scale, using for example the semi-parametric extremal model.

An example

This principle is illustrated on real data. The original data are an air quality measurement series during 1994–1998, extracted from Heffernan & Tawn (2004), see Appendix A.4. For a purpose of illustration, only summer series of ozone and nitrogen dioxide were considered and we ignore potential time dependence. The data are shown in Figure 4.4, on the original scale and on the Fréchet scale. The semi-parametric extremal model has been used. Threshold selection gives $u_1 = 57.8$ and $u_2 = 52$, while the marginal parameter maximum likelihood estimates are $\hat{\sigma}_1 = 8.05$, $\hat{\sigma}_2 = 16.0$, $\hat{\kappa}_1 = 0.230$ and $\hat{\kappa}_2 = -0.436$. A multivariate threshold is selected at $r_0 = 20.4$ and the spectral density is estimated using various models: a symmetric extremal logistic model with maximum likelihood estimate $\hat{\alpha} = 0.688$, an extremal Dirichlet model with maximum likelihood estimate $\hat{\alpha} = (0.469, 0.665)$ and a posterior mean of an extremal mixture model fitted with the reversible jump algorithm. From those estimates, $N = 1000$ extremal events are simulated then transformed back to the original scale. Figure 4.5 shows the result. In each panel, the original data below the multivariate threshold are in black, the gray points are the extremal events. The top left panel refers to an empirical simulation scheme using the pseudo-angle histogram for the spectral distribution.

Various features of the data appear clearly, such as the finite end-point of ozone series, because $\hat{\kappa}_2$ is negative, and the very low probability of sets like $[100, 150] \times [20, 40]$. The empirical model ignores the possibility of independence at moderate extreme levels as it simulates no data in $[50, 100] \times [0, 20]$ or in $[0, 25] \times [60, 80]$. In the area of highly extreme events, $[100, 150] \times [60, 80]$, the extremal Dirichlet and the extremal logistic seem to be more or less the same while the extremal mixture model is more spread out and so is closer to the empirical model. No clear cut distinction appears between the three parametric models.

Extreme probability estimates and return levels

A direct application is the estimation of the probability of extreme sets and return levels. The proportion of data from the estimated model that fall into the extreme set gives an unbiased estimate of the probability of the extreme set under the estimated model. Let A be the set of interest. Then

$$P(A) = P(A \mid R \leq r_0)P(R \leq r_0) + P(A \mid R > r_0)P(R > r_0).$$

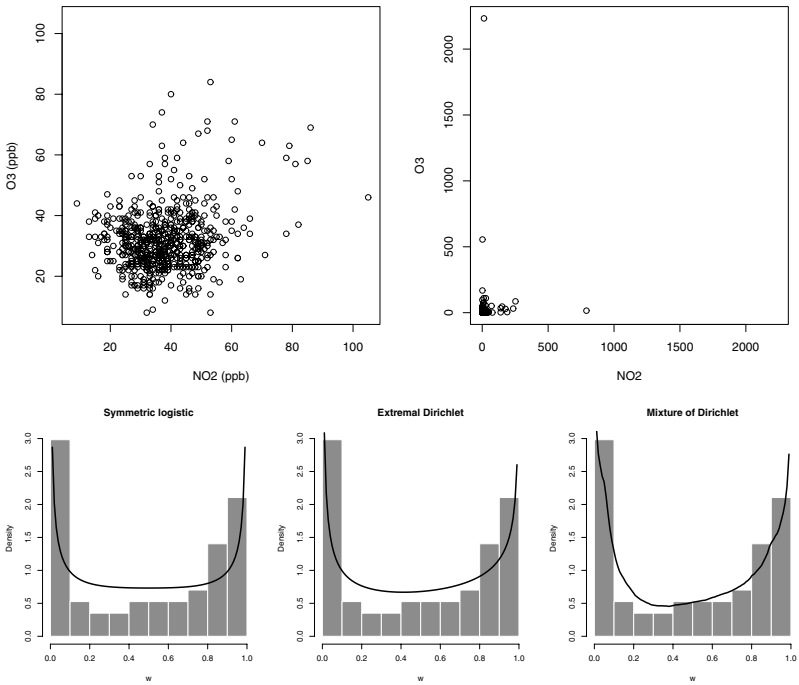


Figure 4.4: Ozone and nitrogen dioxide extremal analysis. Top left: original data set. Top right: data on the Fréchet scale. Bottom, from left to right: histogram of w with fitted extremal logistic, extremal Dirichlet and extremal mixtures.

If A is an extreme set, then $P(A \mid R \leq r_0)$ is zero, otherwise this quantity is estimated empirically. Since the multivariate threshold selection defines $p_0 = P(R > r_0)$, it is sufficient to concentrate on estimation of $P(A \mid R > r_0)$, which we write as $P_0(A)$ for convenience. Then, for any $r_s > r_0$,

$$P_0(A) = P_0(A,\, R \leq r_s) + P_0(A,\, R > r_s).$$

In order to achieve a precision ε in the estimation of $P_0(A)$, it is not necessary to look at the domain further than an r_s such that $P_0(R > r_s) < \varepsilon$, since $P_0(A,\, R > r_s) < P_0(R > r_s)$. It is therefore enough to restrict the simulation to the set $\{r_0 < r \leq r_s\}$ or, in other words, to r_s such that $r_0/r_s = \varepsilon$, that is $r_s = r_0/\varepsilon$. Now

$$P_0(A,\, R \leq r_s) = P(A \mid r_0 < R \leq r_s) P_0(R \leq r_s) = P(A \mid r_0 < R \leq r_s)(1 - r_0/r_s),$$

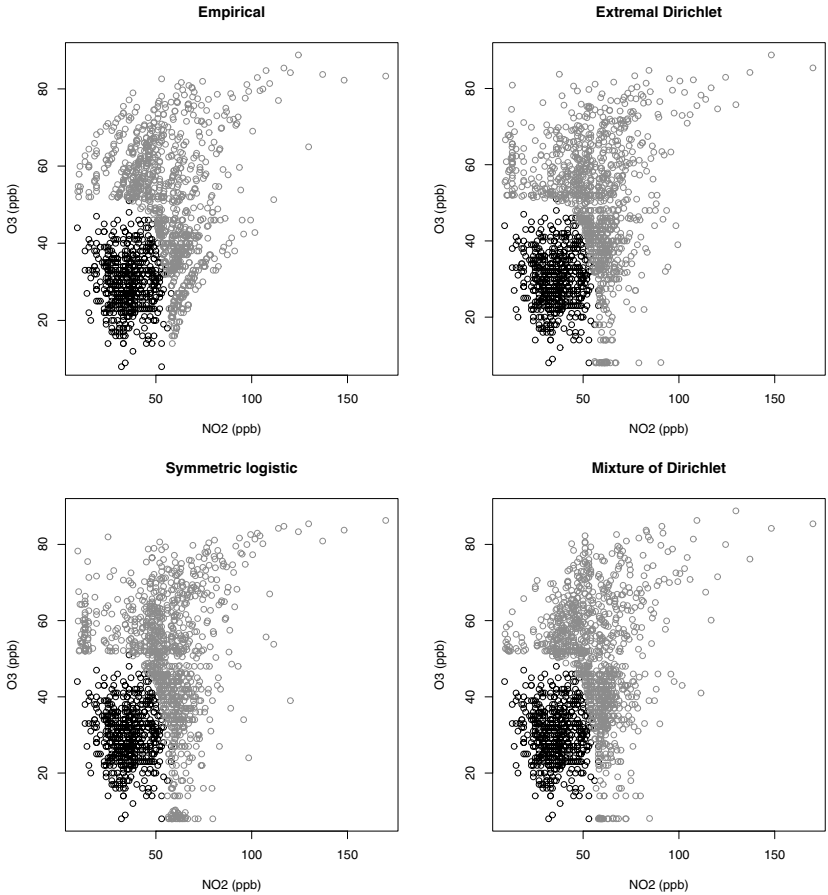


Figure 4.5: Simulation of extreme events. Black points are original non-extreme data. Gray points are simulated from the limit Poisson process with various spectral probability measures.

where $P(A \mid r_0 < R \leq r_s)$ is estimated by the proportion of simulated points falling in A. An extension of this principle is to decompose the set $[r_0\,,\, r_s)$ into an arbitrary number of sets, $[r_i\,,\, r_{i+1})$, and simulate the same amount of data in each set so that the space is more uniformly explored. This approach is straightforward in any dimension since it is only applied to the coordinate r.

The estimation of the return level of a set is of most interest. In the multivariate case,

a set A_v is parameterized by a scalar v and the estimation of v_p such that

$$p = P\left(X \in A_{v_p}\right)$$

is of interest. When p is small and so A_v is away from the central part of the data, then previous methods can be used in a numerical optimization procedure minimizing an objective function, for example,

$$f(v) = \|P\left(X \in A_v\right) - p\|.$$

4.2.2 Simulation study

Mimicking the simulation study of Heffernan & Tawn (2004), the return levels of distributions A, B, C and D, described in Appendix A.4, are estimated. Two kinds of events are considered:

1) simultaneously extremal events. The return level is v such that $p_v = P(Y > v)$;

2) unilaterally extremal events. The return level is v such that $P(Y_1 > r, Y_2 < v) = p$ where r is such that $P(Y_1 > r) = p/q$.

Here, comparisons are done componentwise. True values of v are given in the Appendix A.4. In Heffernan & Tawn (2004), a new approach is developed. The idea is to fit the distributions of the data given that there are extremes in one component, that is the distributions of $X_{-i} \mid X_i > u_i$, for $i = 1, \ldots, d$. The asymptotic characterization is similar to the multivariate extreme distribution. The extremal events considered are perfectly designed for this kind of method, which reveals very good performance in particular for models C and D which are asymptotically independent. The estimation of return levels is also based on Monte Carlo integration by simulating from the fitted model. Their results are compared with classical methods for simultaneous extremal events but not for unilaterally extremal events for which classical methods are inappropriate. These results are reproduced in Tables 4.1 and 4.2. Therein NP1 refers to a non-parametric approach with the coefficient tail of dependence, η, fixed to 1, NP refers to a non-parametric approach with estimated η, and HT refers to the method developed in Heffernan & Tawn (2004).

Method MD refers to the Monte Carlo procedure developed in this thesis. From each distribution, a sample of 5000 data is simulated, a threshold is selected, and the spectral probability distribution is estimated using the extremal mixture model with a reversible

Dist.	Method	p		
		10^{-4}	10^{-6}	10^{-8}
A	MD	0.04 $(-0.37, 0.49)$	-0.01 $(-0.26, 0.28)$	0.02 $(-0.19, 0.23)$
	NP1	-0.1 $(-1.0, 0.8)$	-0.1 $(-0.7, 0.5)$	-0.0 $(-0.5, 0.4)$
	NP	-0.8 $(-15, 0.6)$	-0.8 $(-16, 0.4)$	-0.6 $(-17, 0.3)$
	HT	-1.4 $(-4.0, 0.8)$	-1.6 $(-4.1, 0.5)$	-1.6 $(-5.0, 0.4)$
B	MD	4.53 $(4.00, 5.11)$	2.09 $(2.58, 3.23)$	1.98 $(1.73, 2.24)$
	NP1	4.6 $(3.7, 5.6)$	3.0 $(2.4, 3.6)$	2.1 $(1.7, 2.5)$
	NP	-1.0 $(-14, 5.3)$	-2.9 $(-17, 3.4)$	-3.9 $(-19, 2.4)$
	HT	-4.0 $(-12, 4.2)$	-5.7 $(-15, 0.5)$	-6.1 $(-17, 0.0)$
C	MD	29.4 $(28.4, 30.2)$	34.1 $(33.4, 34.7)$	53.5 $(52.6, 54.3)$
	NP1	23 $(22, 24)$	26 $(26, 28)$	29 $(28, 29)$
	NP	-0.6 $(-16, 14)$	-0.1 $(-18, 17)$	0.2 $(-18, 18)$
	HT	-0.6 $(-8.6, 5.3)$	0.6 $(-13, 8.2)$	0.8 $(-18, 9.8)$
D	MD	20.3 $(19.5, 21.2)$	26.3 $(25.7, 26.8)$	26.9 $(26.5, 27.3)$
	NP1	28 $(27, 29)$	31 $(30, 32)$	32 $(32, 33)$
	NP	-1.9 $(-15, 14)$	-1.9 $(-17, 16)$	-2.2 $(-18, 17)$
	HT	-0.6 $(-10, 7.3)$	-0.1 $(-15, 9.2)$	-0.1 $(-25, 12)$

Table 4.1: Median (2.5 and 97.5 percentiles)($\times 100$) of the posterior distribution of relative errors of v_p for simultaneously extremal events.

jump algorithm. The posterior distribution of the relative error $(\hat{v}-v)/v$ is obtained from the chain. To do so, a subset $I = \{r_1, \ldots, r_l\} \subset \{1, \ldots, R\}$, where R is the number of iterations of the chain, is drawn at random. Then for each $r \in I$, a sample of extremes is simulated and the corresponding v_r is computed. The resulting chain of relative errors $(v_{r_1} - v)/v, \ldots, (v_{r_l} - v)/v$ is analyzed. This choice of a $l \leq R$ is made because each step of this algorithm can be quite long, but, in principle, the whole chain $\{1, \ldots, R\}$ could be taken.

For simultaneously extremal events, method MD performs similarly to method NP1, which is also based on the Poisson process approach. In particular, it breaks down for distributions C and D that exhibit asymptotic independence. For distribution A it is the best method, from the viewpoint of bias and uncertainty, even though the uncertainty is underestimated here. Indeed, these values are the credibility intervals for v, which

Dist.	q	p			Methods
		10^{-4}	10^{-6}	10^{-8}	
A	0.2	1.86 $(-0.33, 3.86)$	1.17 $(-0.05, 2.25)$	0.90 $(-0.02, 1.69)$	MD
	0.2	-3.1 $(-13, 2.7)$	-4.7 $(-15, 1.6)$	-5.1 $(-16, 1.0)$	HT
	0.5	0.33 $(-0.41, 1.08)$	0.27 $(-0.30, 0.75)$	0.21 $(-0.15, 0.59)$	MD
	0.5	-2.0 $(-9.4, 1.2)$	-2.5 $(-11, 0.8)$	-2.6 $(-12, 0.5)$	HT
	0.8	-0.05 $(-0.37, 0.22)$	-0.00 $(-0.22, 0.20)$	0.03 $(-0.15, 0.17)$	MD
	0.8	-0.8 $(-6.7, 3.8)$	-0.9 $(-7.8, 3.5)$	-1.0 $(-9.2, 2.9)$	HT
B	0.2	3.39 $(2.18, 4.60)$	1.82 $(1.14, 2.40)$	1.54 $(0.99, 2.02)$	MD
	0.2	-15 $(-36, 0.7)$	-17 $(-43, -1.1)$	-16 $(-47, -2.0)$	HT
	0.5	2.99 $(2.17, 3.71)$	2.03 $(1.52, 2.48)$	1.40 $(1.04, 1.73)$	MD
	0.5	-11 $(-25, -0.9)$	-12 $(-29, -1.9)$	-12 $(-32, -2.2)$	HT
	0.8	6.79 $(6.18, 7.50)$	4.40 $(3.87, 4.87)$	3.33 $(2.94, 3.67)$	MD
	0.8	-8.4 $(-19, -0.3)$	-9.1 $(-21, -1.6)$	-9.2 $(-23, -1.6)$	HT

Table 4.2: Median (2.5 and 97.5 percentiles)($\times$100) of the posterior distribution of relative errors of v_p for non-simultaneously extremal events.

do not take into account the uncertainty of estimation of the spectral probability. As a conclusion, the method MD does not have any clear advantage over method NP1 that is far simpler. In the case of asymptotic independence, method MD and NP1 should not be used.

For non-simultaneous extremal events, methods NP and NP1 are not applicable so method HT has no rival other than method MD. For distribution A, both methods are statistically unbiased but while method HT seems to underestimate v with very large uncertainty, method MD seems to overestimate v with a smaller uncertainty. However, the same comments as for simultaneous extremal events about the underestimation of this uncertainty for MD apply here. For distribution B, method HT has a significant negative bias with very large uncertainty while method MD is closer to the true return level from a point estimate viewpoint but reveals significant positive bias. In fact, method MD shows similar performance for simultaneous and non-simultaneous extremal events. Results for distributions C and D are not shown in Table 4.2 because they are not relevant. The Monte Carlo scheme for method MD hardly simulates any points in the set of interest so that the numerical estimation of v remains at the initial points. Therefore method HT is

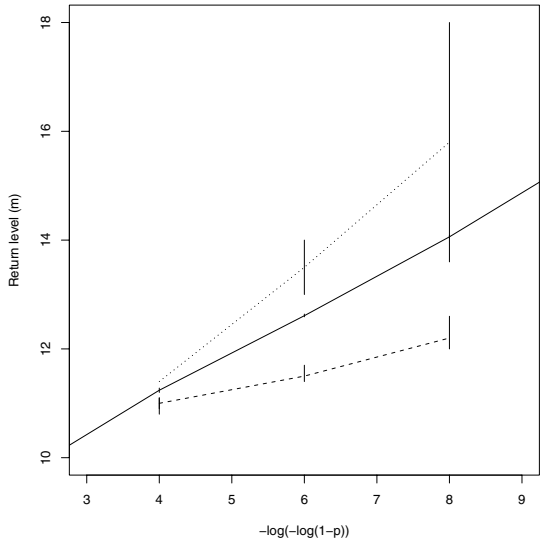


Figure 4.6: Newlyn return level estimates. Unbroken line: extremal mixture model. Dashed line: the structural variable. Dotted line: the extremal Dirichlet model. Vertical lines indicate confidence intervals for the structural variable and the Dirichlet model and credibility intervals for the mixture model.

the only one applicable in this context.

The conclusions of these experiments are that

- method MD can be used for general set shape as long as no asymptotic independence is suspected while method NP1 is to be preferred for simultaneously extremal events because of its simplicity;
- under asymptotic independence, method HT can be used generically while method NP may be preferable for simultaneous extremal events because of its simplicity. Method MD should not be used in this context.

4.2.3 Real data analysis

An illustration using real data is done on the Newlyn data; see Appendix A.4. The failure region, $Q(v, X) \geq 0.002$, is defined as a function of the design parameter v, corresponding to the sea-wall height. It is to be estimated for $-\log\{-\log(p_v)\} = i$, where $p_v = P\{Q(v, X) \geq 0.002\}$ and $i = 4, 6, 8$. The results are shown in Figure 4.6. Credibility

intervals are almost invisible, indicating that in this case, they do not represent the overall uncertainty correctly, perhaps because the uncertainty on the marginal parameters is not taken into account here.

For comparison, the results obtained in Coles & Tawn (1994) from the structural variable approach and the extremal Dirichlet model have been reproduced. A problem raised by Coles & Tawn (1994) was the inconsistency between the multivariate and the structure variable approaches. The return level curve from the extremal mixture model seems to be a compromise between the two approaches and is more consistent, in some sense. This shows that a part of the problem was a lack of fit due to the extremal Dirichlet model. However, the return level curve of the extremal mixture model is not within the confidence band of the structural variable approach. This can be explained by the fact that the asymptotic independence of the data cannot be taken into account by the Poisson process, even if the extremal mixture model fits the pseudo-polar angles closely. The application of the Gaussian tail model to this dataset brings an even more consistent result and is discussed in Bortot et al. (2000).

4.3 Dependence analysis

This section develops a dependence analysis based on Dirichlet distribution properties given in Section 2.2.4. It is illustrated on real data.

4.3.1 Generalities

In the case of dimension $d > 2$, suppose that the pseudo-angles W follow an extremal mixture model,

$$W \sim \sum_{m=1}^{k} \pi_m \mathrm{Dir}\left(\alpha^{(m)}\right).$$

In such case, classical properties of the Dirichlet distribution apply; see Wilks (1962), for example. Let $\mathrm{Dir}(\alpha)$ denote a Dirichlet distribution with parameter vector α and $\mathrm{dDir}(\alpha)$ be the corresponding Dirichlet density. Also, if $\mu = (\mu_1, \ldots, \mu_d)$ is a vector then $\mu_{(1,2)}$ denotes the vector (μ_1, μ_2) and $\mu_{-(1,2)}$ denotes the vector $(\mu_3, \ldots, \mu_d)$. Generalization to μ_J and μ_{-J} for any subset $J \subset \{1, \ldots, d\}$ is obvious. Also $\sum \mu_J$ is the sum of every element of μ_J.

Let I be an indicator variable taking value in $(1, \ldots, k)$ with probability $(\pi_1, \ldots, \pi_k)$

such that

$$W \mid I = m \sim \mathrm{Dir}\left(\alpha^{(m)}\right), \quad m = 1, \ldots, k.$$

Then

$$\left(\frac{W_1}{W_1+W_2}, \frac{W_2}{W_1+W_2}\right) \Bigg| \left(W_{-(1,2)} = w_{-(1,2)}, I = m\right) \sim \mathrm{Dir}\left(\alpha^{(m)}_{(1,2)}\right)$$

so that

$$\left(\frac{W_1}{W_1+W_2}, \frac{W_2}{W_1+W_2}\right) \Bigg| \left(W_{-(1,2)} = w_{-(1,2)}\right) \sim \sum_{m=1}^{k} \pi_m(w_{-(1,2)}) \mathrm{Dir}\left(\alpha^{(m)}_{(1,2)}\right),$$

where

$$\pi_m(w_{-(1,2)}) = \frac{\pi_m \mathrm{dDir}\left(\alpha^{(m)}_{-(1,2)}; \sum \alpha^{(l)}_{(1,2)}\right)\left\{w_{-(1,2)}; \sum w_{(1,2)}\right\}}{\sum_{l=1}^{k} \pi_l \mathrm{dDir}\left(\alpha^{(l)}_{-(1,2)}; \sum \alpha^{(l)}_{(1,2)}\right)\left\{w_{-(1,2);w_{(1,2)}}\right\}}.$$

Using similar arguments, we have also

$$\left(\frac{W_1}{W_1+W_2}, \frac{W_2}{W_1+W_2}\right) \Bigg| \left(\sum W_{-(1,2)} = \sum w_{-(1,2)}\right) \sim \sum_{m=1}^{k} \pi_m\left(\sum w_{-(1,2)}\right) \mathrm{Dir}\left(\alpha^{(m)}_{(1,2)}\right),$$

where

$$\pi_m\left(\sum w_{-(1,2)}\right) = \frac{\pi_m \mathrm{dDir}\left(\sum \alpha^{(m)}_{-(1,2)}; \sum \alpha^{(m)}_{(1,2)}\right)\left\{\sum w_{-(1,2)}; \sum w_{(1,2)}\right\}}{\sum_{l=1}^{k} \pi_l \mathrm{dDir}\left(\sum \alpha^{(l)}_{-(1,2)}; \sum \alpha^{(l)}_{(1,2)}\right)\left\{\sum w_{-(1,2)}; \sum w_{(1,2)}\right\}}.$$

Therefore, conditional distributions of an extreme in any subgroup of $\{1, \ldots, d\}$ can be easily obtained from the global spectral measure.

Summary statistics can help to get useful information from these conditional distributions without plotting them all. In the following the conditional expectation is used,

$$E\left(\frac{W_1}{W_1+W_2} \Bigg| \sum W_{-(1,2)} = \sum w_{-(1,2)}\right) = \sum_{m=1}^{k} \pi_m\left(\sum w_{-(1,2)}\right) \frac{\alpha_1^{(m)}}{\alpha_1^{(m)} + \alpha_2^{(m)}}, \quad (4.1)$$

and the dependence measure,

$$\begin{aligned} E\left\{\min\left(\frac{W_1}{W_1+W_2}, \frac{W_2}{W_1+W_2}\right) \Bigg| \sum W_{-(1,2)} = \sum w_{-(1,2)}\right\} \\ = \sum_{m=1}^{k} \frac{\pi_m\left(\sum w_{-(1,2)}\right)}{\alpha_1^{(m)} + \alpha_2^{(m)}} \Big\{\alpha_1^{(m)} \mathrm{Beta}\left(0,\, 1/2;\, \alpha_1^{(m)} + 1,\, \alpha_2^{(m)}\right) \\ + \alpha_2^{(m)} \mathrm{Beta}\left(1/2,\, 1;\, \alpha_1^{(m)},\, \alpha_2^{(m)} + 1\right)\Big\}. \quad (4.2) \end{aligned}$$

As references, the dependence measure is 0 for distributions concentrated on 0 and 1, it is 1/4 for the uniform distribution and it is 1/2 for the distribution concentrated on 1/2.

Finally, the unconditional distribution and associated statistics may also be used. Let $i_1, \ldots, i_p$ be disjoint subsets of indices of $\{1, \ldots, d\}$ with $p \leq d$. Then, by integrating out previous formulas, it is straightforward to see that any normalized subgroup

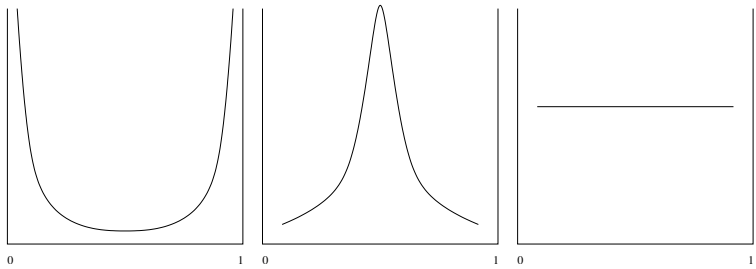


Figure 4.7: Illustration of typical density shapes. From left to right: near-independence, near-dependence, uniform.

$\left(W_{i_1}/\sum_{j=1}^{p} W_{i_j}, \ldots, W_{i_p}/\sum_{j=1}^{p} W_{i_j}\right)$ is distributed according to the mixture of Dirichlet distribution with parameters $\alpha_{i_j}^{(m)}$ and mixing distribution $\{\pi_m\}_{m=1}^{k}$, $j = 1, \ldots, p$, $m = 1, \ldots, k$. This is equivalent to restricting the analysis to a subset of indices and then ignoring a part of the dependence, so we do not use them below.

4.3.2 Application

Without loss of generality, we give the interpretation of these distributions in dimension three. An observation $X = (X_1, X_2, X_3)$ is extreme if $R = \sum_{j=1}^{3} X_j > r_0$. In such case, the split of R among components $1, 2$ and 3 is given by $W = (W_1, W_2, W_3)$. The density $h_3^{1,2}$ of

$$\frac{W_1}{W_1 + W_2} \middle| W_3 = w_3$$

describes the split of $R - Rw_3$ among components 1 and 2 given that a proportion w_3 of R comes from component 3. For example, given that 10% of this extreme is due to component 3, $h_3^{1,2}$ shows how the remaining 90% distribute themselves among components 1 and 2. It is clear that the split of those 90% tells us almost everything about the behavior of the extreme. On the contrary, if only 10% of the extreme event is due to components 1 and 2, the split of those 10% is less important than the remaining 90%.

The evolution of $h_3^{1,2}$ as w_3 varies from 0 to 1 gives information on the dependence between the three components. In the dependence analysis, we are looking for the kind of shapes shown in Figure 4.7. The left shape indicates independence: the extreme is due either to component 1 or 2. The middle shape indicates dependence: the extreme is due to components 1 and 2 in equal proportion. The right shape indicates a random situation: the extreme is proportioned uniformly. For example, as w_3 goes from 0 to 1, an

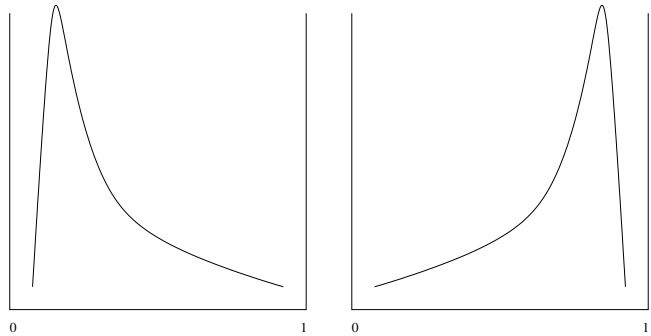


Figure 4.8: Illustration of typical skewed density shapes. Left: skewed to the right. Right: skewed to the left.

evolution of the graph of $h_3^{1,2}$ from the left type of shape to the middle one means that the dependence between 1 and 2 increases as their proportion in the extreme diminishes. This may be interpreted as an impact on the concentration of component 3 on components 1 and 2, a phenomenon that may be relevant in chemical applications, for example.

When the components are numerous, it may be difficult to represent all the conditional densities. In such cases, the use of summary quantities such as (4.1) and (4.2) is an alternative. In the left shape of Figure 4.7, the conditional dependence (4.2) is close to 0, in the middle case, it is close to 1/2 and in the right case, it equals 1/4. Therefore a single plot can indicate the evolution of the graph of $h_3^{1,2}$ from the left to the middle type of shape. In the same vein, the conditional expectation quantifies the skewness of $h_3^{1,2}$. The two shapes shown by Figure 4.8 have the same conditional dependence but the left one has a higher conditional expectation than the right one. Note that the three shapes in Figure 4.7 have the same conditional expectation. This shows that the two summary statistics should be used in order to obtain an accurate picture of the evolution of the shape of $h_3^{1,2}$ as w_3 goes from 0 to 1.

The conditional dependence analysis is here done on the Newlyn data and on the air quality data; see Appendix A.4. In three dimensions the spectral measure can be completely represented, see Figure 3.14, while this is impossible in five dimensions. As conditional distributions are an alternative representation of the spectral probability measure, they are much more interesting in five dimensions than in three. However, the main ideas are well illustrated in three dimensions.

Newlyn data

The three components are the wave height, the period and the surge, which we will refer to as (w, p, s). Figure 4.9 shows the conditional distribution of $(w, p \mid s)$, $(w, s \mid p)$ and $(p, s \mid w)$ for various levels of the conditioning component. For example, the interpretation of the left plot of the middle line is the spectral density of the distribution of an extreme event among the wave and the surge, given that 5% of this extreme is due to the period. Levels above 71% are not represented for sake of legibility and because the conditional distributions do not vary much beyond. From a practical viewpoint, we conclude the following:

- From the leftmost plots of the two topmost lines, we see that the wave is rarely the only cause of an extreme event. Indeed, the level of the wave is high only when the contribution of the conditioned component is high.

- From the right plots of the two topmost lines, we see that the contribution of the wave height combined with another cause can be relatively large. This is in distinction from the two other causes, period and surge, that have an almost uniform distribution, as shown by the rightmost plots of the third line.

- The leftmost plots of the bottom line show that the period and the surge exhibit near-independence with slightly heavier weight to the surge when the wave contribution to the extreme is low. This dependence increases as the wave contribution to the extreme increases.

In absence of any knowledge about wave phenomena we do not draw any further conclusion. For example, conditional densities of $(w, p \mid s)$ and $(w, s \mid p)$ exhibit a doubtful shape at conditional component level 0.05. This could be due to a smooth propagation of the independence exhibited by $(p, s \mid w)$ at conditional component level 0.05 and 0.15. This smooth propagation may appear more clearly in three dimensions, in Figure 3.14. This shows that conclusions must be drawn with caution.

Figure 4.10 shows the conditional expectation and the conditional dependence measure. The gray vertical dashed line is at 1/3. This is a reference for the conditioned component whose expectation equals 1/3 because of the constraints on the spectral distribution. Therefore, 1/3 can be interpreted as a reference and, informally, levels to the left of this line are low values of the conditional component while levels to the right of it are high values. We can see that this limit represents also a change in the behavior of

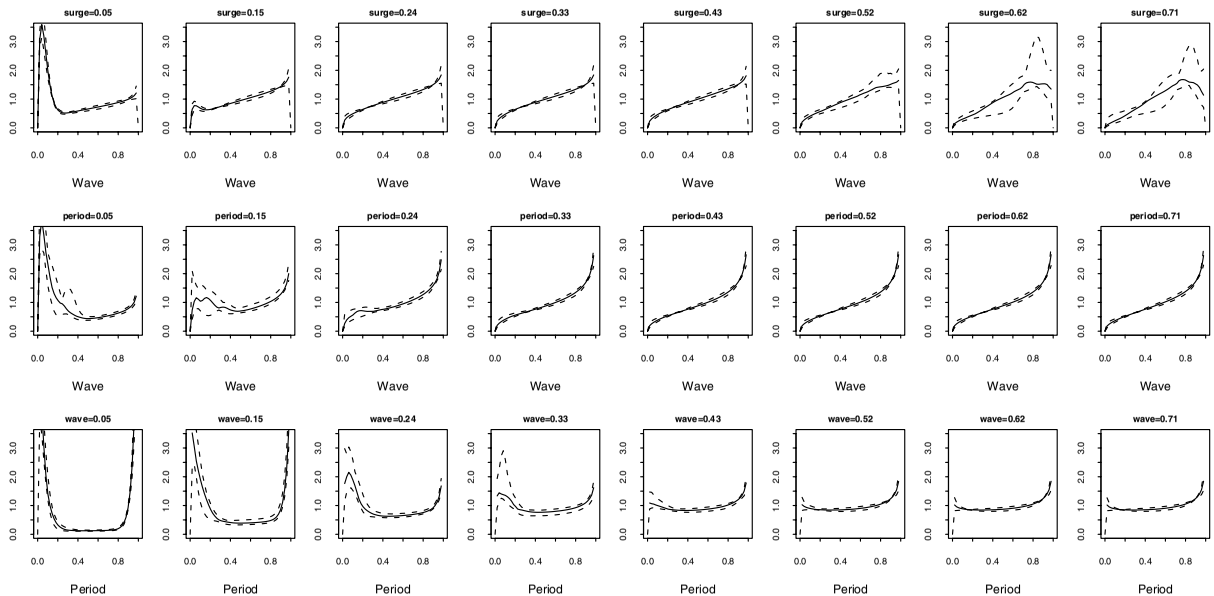


Figure 4.9: Conditional distributions of wave, period and surge. From top to bottom: $(w, p \mid s)$, $(w, s \mid p)$, $(p, s \mid w)$. From left to right: the evolution as the conditioned component goes from 0 to 0.71. The title indicates the conditioned component. The full lines are the posterior median density estimate and the dashed lines 90% credibility intervals.

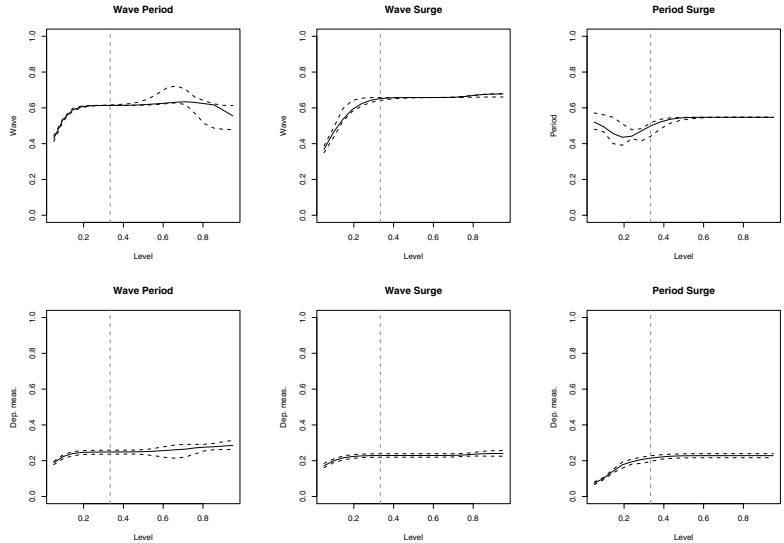


Figure 4.10: Conditional summary statistics of Newlyn data. Top: conditional expectation. Bottom: conditional dependence. The full lines are the posterior median density estimate and the dashed lines 90% credibility intervals. The vertical dashed gray line indicates 1/3. The title indicates the unconditioned pair.

the conditional density, which stabilizes, whatever the pair considered. This should be compared with Figure 4.9, in which this feature is obvious. Each conditional density tends more or less toward a uniform-like distribution. Other features like the independence of the surge and the period given a low wave level can be also seen. From the conditional expectation plots, we see that the wave level follows the conditional component level at low levels, then it stabilizes around 0.6 which is higher than the expected 0.5.

Air quality data

It is difficult to represent the conditional density for the ten pairs at every level, so we concentrate on the summary statistics. Conditional expectations are shown in Figure 4.11 and conditional dependence measures in Figure 4.12. We draw the following conclusion:

- The pair (NO_2, NO) looks close to uniform distribution, whatever the level of the rest. When paired with another component, NO behaves like NO_2; this can be seen by comparing (NO, O_3) and (NO_2, O_3), for example. This suggests these two components

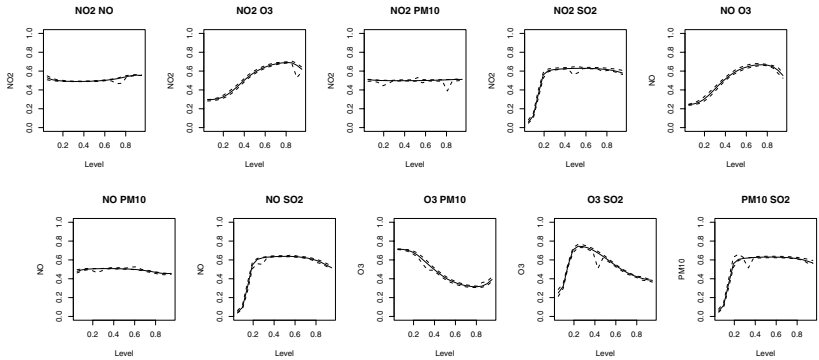


Figure 4.11: Conditional expectation for air quality data. Title indicates the unconditioned pair. Full lines show the posterior median estimates and dashed lines indicate 90% credibility intervals. The ordinate axis title indicates the name of component whose expectation is plotted.

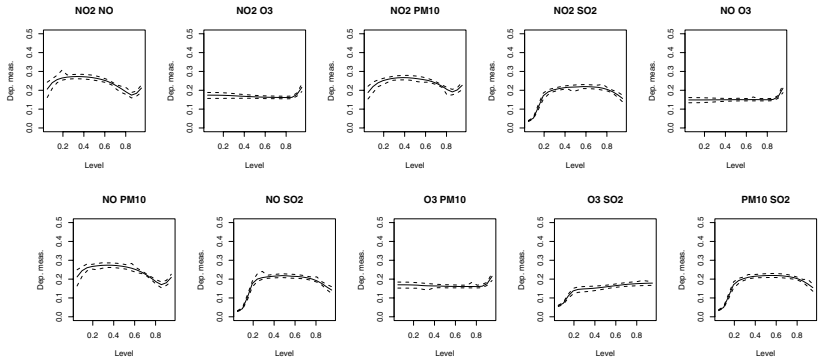


Figure 4.12: Conditional dependence measure for air quality data. Title indicates the unconditioned pair. Full lines show the posterior median estimates and dashed lines indicate 90% credibility intervals.

play the same role in the extremal behavior of the data set.

- Looking at pairs containing O_3, the conditional expectation shows that the contribution of O_3 to the extreme event is high when the level of conditioning component is low, here below 0.6, except for the pair (O_3, SO_2) that exhibits complicated behavior at low conditioning levels. According to Heffernan & Tawn (2004) and references therein, high levels of pollution in winter are mainly due to vehicle emissions trapped by cold

and stable weather conditions. Therefore pollution is mainly due to nitrogen and sulfur compounds and particulate matter. This is coherent with the observation of expectation plots since major episodes of pollution contain low levels of O_3. However an episode of pollution with low levels of three components would be mainly due to ozone and not to the last cause: for example, if a pollution episode happens and is not due to NO, NO_2 and SO_2, then it will be mainly due to ozone and not to PM_{10}. Furthermore, the level of conditional dependence between O_3 and another pair is low and stable. This reveals that ozone plays a particular role in winter, independent of the other pollution causes.

- Now, SO_2 shows strong independence with any other component at low conditioning levels and weak dependence at higher conditioning levels. From the expectation plot containing SO_2, we see that if these two components are the one cause of an extreme then the expected level of SO_2 is very high. This tendency is inverted rapidly at lowest levels of the conditioning components.

These observations allowed us to form three almost independent groups of components, O_3, SO_2, and (NO_2, NO, PM_{10}). In general pollution is due to high levels of (NO_2, NO, PM_{10}) going with varying levels of SO_2 and low levels of O_3. If pollution is caused by only one component then it is likely to be SO_2 or O_3. It is likely to be SO_2 if the level of (NO_2, NO, PM_{10}) is very low and it is likely to be O_3 if this level is moderately low.

4.3.3 Discussion

The conditional analysis may reveal some intrinsic properties of the data but turns out to be subtle, needing practice to interpret. These conditional densities provide useful means of exploration of the spectral measure even in high dimensions. They can be generalized to any spectral distribution function since the conditional density is simply a rescaled slice of the spectral density. The advantage of the extremal mixture model is that this slice is straightforwardly obtained from the parameters.

Chapter 5

Spatial extremes

Spatial extremes is a growing field. Although the literature is small compared to that for univariate and multivariate extremes, the major theoretical results exist in applicable forms. Until now statistical applications have been restricted to max-stable processes, so inference is limited to pointwise maxima. In the multivariate context the Poisson process characterization has advantages over the componentwise maximum viewpoint. Although the Poisson process characterization is well documented in the spatial context, to our knowledge, no statistical application has appeared. The aim of this chapter is to use the viewpoint of W as a random probability measure in order to investigate this.

The first section reviews literature on spatial extremes. The second section gives motivation and the main heuristic ideas. The theoretical justification of our approach is original in the literature, and exploits random measure theory. The third section presents some topological and random measure theory background, detailed treatments of which can be found in Resnick (1987) and Jagers (1974), although most of what follows is extracted from Kallenberg (1983). Section 5.4 presents the application to spatial extremes and Section 5.5 illustrates the method on a real data set.

5.1 State of the art

A class of models for the study of the extremal behavior of stochastic process is the class of simple max-stable processes. Such a process $\{Z_t\}_{t\in T}$ is defined by the property that the pointwise maximum of $\{Z_t^{(i)}\}_{t\in T}$, n independent copies of $\{Z_t\}_{t\in T}$, has the same distribution as $\{nZ_t\}_{t\in T}$, that is $\max\left\{Z_t^{(1)},\ldots,Z_t^{(n)}\right\}_{t\in T} \stackrel{d}{=} \{nZ_t\}_{t\in T}$. In the case $T=[0,1]$, de Haan (1984) characterizes every continuous simple max-stable processes by

a spectral representation. For each continuous simple max-stable process $\{Z_t\}_{t\in T}$, there exists a set S and a finite positive measure ρ on S such that

$$Y_t = \max_{k\geq 1} f_t(S_k)X_k, \quad t \in T,$$

has the same finite dimensional distributions as Z_t, where $\{X_k, S_k\}$ is the enumeration of a Poisson process on $\mathbb{R}^+ \times S$ with intensity $x^{-2}dx \times \rho(ds)$ and $\{f_t\}_{t\in T}$ is a suitably chosen family of positive L_1–functions. Conversely, every such representation is continuous simple max-stable. The functions f_t, $t \in T$, are called the spectral functions.

Marginal distribution functions of simple max-stable processes are Fréchet, that is $\exp\{-c(t)/x\}$, $x > 0$, for $c(t) \geq 0$. As a consequence of the spectral representation the finite dimensional distribution of Z is multivariate extreme: for $0 \leq t_1 < \cdots < t_d \leq 1$ and $z_1, \ldots z_d > 0$,

$$P\{Z_{t_i} \leq z_i,\, i = 1, \ldots, d\} = \exp\left\{-\int_0^1 \max_{i=1,\ldots,d} \frac{f_{t_i}(s)}{z_i} ds\right\}.$$

Furthermore, the spectral functions can be chosen continuous in L_1, that is $\|f_{t_i} - f_t\|_1 \to 0, t_i \to t$. Like the spectral functions are linked to the spectral measure in the finite dimensional case, Giné, Hahn & Vatan (1990) make the correspondence with a spectral measure on the space of density functions. For the following we need to define some notation:

C the space of continuous functions on $[0, 1]$,

$C^+ = \{f \in C : f > 0\}$,

$\bar{C}_1^+ = \{f \in C : f \geq 0 \text{ and } \|f\|_\infty = 1\}$,

$\bar{C}^+ = (0, \infty] \times \bar{C}_1^+$.

Then Z is continuous simple max-stable if and only if there exists a finite measure H on C_1^+ with $\int_{C_1^+} f(t)dH(f) = 1$, $t \in [0, 1]$, such that

$$P\{Z < f\} = \exp\left\{-\int_{C_1^+} \|g/f\|_\infty dH(g)\right\}$$

or equivalently

$$P\left\{\sup_{t\in K_i} Z(t) \leq x_i, i = 1, \ldots, d\right\} = \exp\left\{-\int_{C_1^+} \max_{i=1,\ldots,d} \frac{\sup_{t\in K_i} g(t)}{x_i} dH(g)\right\},$$

for all compact $K_1, \ldots, K_d \subset [0, 1]$ and all positive $x_1, \ldots, x_d$.

The Poisson process characterization is developed by de Haan & Lin (2001). Let $X, X^{(1)}, X^{(2)}, \ldots$ be independent and identically distributed random elements of C^+. Then the following statements are equivalent (among others):

(*i*) $n^{-1} \max_{i=1,\ldots,n} X^{(i)} \xrightarrow{d} Z$ in C^+, with Z a continuous simple max-stable process.

(*ii*) $\nu_n \xrightarrow{w} \nu$ in the space of measures on C^+ (and then ν is homogeneous of degree -1), with

$$\nu_n(E) = nP\left(n^{-1}X \in E\right), \quad E \in \mathcal{B}(\bar{C}^+).$$

(*iii*) $N_n \xrightarrow{d} N$ in the space of random measures on $\bar{C}^+$, where

$$N_n = \sum_{i=1}^{n} \delta_{\{n^{-1}X^{(i)}\}}$$

and N is a Poisson process.

Notions of random measures, Poisson process, and vague convergence are explained in the next section. Anyway, the interpretation of this result is rather clear when compared to results in the multivariate case since it is the equivalence between convergence of the componentwise maximum, convergence of the survival function and convergence to the Poisson process.

The previous characterization extends to max-stable processes with more general margins. A process Z is continuous max-stable if there exist sequences of norming functions $a_n > 0$ and b_n such that, for $n \in \mathbb{N}$,

$$a_n^{-1}\left(\max_{i=1,\ldots,n} Z^{(i)} - b_n\right) \stackrel{d}{=} Z,$$

where $Z^{(i)}$, $i = 1, \ldots, n$, are independent replicates of Z. A max-stable process can be represented as

$$a(t)\frac{Z_t^{\gamma(t)} - 1}{\gamma(t)} - b(t),$$

for $a, b, \gamma \in C$, $a > 0$, and Z continuous simple max-stable. de Haan & Lin (2001) also characterizes convergence to max-stable processes. Let $X, X^{(1)}, X^{(2)}, \ldots$ be independent and identically distributed random elements of C. Suppose that $F_t(x) = P\left(X_t \leq x\right)$ is continuous with respect to t for each x. Define

$$U_t(s) = F_t^{\leftarrow}(1 - 1/s), \quad s > 0, 0 \leq t \leq 1.$$

The following statements are equivalent (among others):

(*i*) $a_n(t)^{-1}\left\{\max_{i=1,\dots,n} X^{(i)} - b_n(t)\right\} \xrightarrow{w} Z$, where $a_n > 0$ and b_n are continuous functions, chosen in such way that, for $t \in [0,1]$,

$$P\{Z_t \leq x\} = \exp\left\{-(1+\gamma(t)x)^{-1/\gamma(t)}\right\}, \quad x > 0.$$

Then γ is continuous.

(*ii*) $a_n(t)^{-1}\left\{\max_{i=1,\dots,n} X^{(i)} - U_t(n)\right\} \xrightarrow{w} Z$.

(*iii*) $n^{-1}\max_{i=1,\dots,n}\left[1 - F_t\left\{X_t^{(i)}\right\}\right]^{-1} \xrightarrow{w} \{1+\gamma(t)Z_t\}^{1/\gamma(t)}$, and the limit is automatically simple max-stable.

This result tells us how marginal standardization can be used to obtain a simple max-stable limit and then apply the previous convergence characterization. Once more this is a perfect parallel with the multivariate case, where margins have to be put on the Fréchet scale before the spectral distribution can be estimated.

Coles (1993) exploited the spectral representation of max-stable processes. T is the area of study containing observation sites $\tilde{T} = \{t_1, \dots, t_q\}$. The spectral density function h of a selected number of tuning sites $T_1 = \{t_1, \dots, t_p\} \subset \tilde{T}$ is estimated using a parametric multivariate extremal model. We call α the parameter of h, $\mathcal{A}$ its parameter space and $\tilde{\alpha}$ the estimate of α. The spectral functions f_t, called storm profiles, are used to link the information brought by h to the whole area T. Proximity coefficients are defined as

$$a_j(t) = \frac{d_j(t)^{-\xi}}{\sum_{i=1}^{p} d_i^{-\xi}(t)}, j = 1, \dots, p,$$

where $d_j(t)$ is the distance from t to $t_j \in T_1$ and for some smoothing parameter $\xi > 0$. Parametric forms are given to storm profiles, namely

$$f_t(w) = \frac{h\{w, \phi(t)\}}{h(w, \tilde{\alpha})} \sum_{j=1}^{d} a_j(t) w_j,$$

where ϕ is a function valued in $\mathcal{A}$ with $\phi(t) = \tilde{\alpha}$ for $t \in T_1$. With this form, constraint

$$\int_{S_p} f_t(w) H(dw) = 1,$$

is guaranteed as well as the equality $f_{t_j}(w) = w_j$ for $t_j \in T_1$, $j = 1, \dots, p$. A parametric form is given to ϕ in order to perform maximum likelihood estimation from $T_2 = T \setminus T_1$. The smoothing parameter ξ is selected in order to diminish a formal multiplicative error in the empirical estimation of storm profile at sites T_2. Further diagnostics are developed in order to select an adequate partition T_1 and T_2 iteratively.

A similar procedure is used in Coles & Walshaw (1994), where the parametric form of the spectral functions accounts for the direction of wind data, and in Coles & Tawn (1996) for rainfall processes. An extreme variogram has also been developed in Ancona-Navarrete & Tawn (2002) allowing for pairwise dependence diagnostics.

Alternative approaches like that of Casson & Coles (1999) link the observations through a Bayesian structure. Extreme observations follow an extremal Poisson model given the parameters. Known link functions of these parameters are distributed according to a Gaussian random field with a mean given by a regression function, depending on the spatial location.

Except that of Casson & Coles (1999), the procedures described are based on the spectral representation of simple max-stable processes. Inference is therefore based on maxima which implies, as in the multivariate case, a loss of data that are often already sparse. Another drawback discussed in Coles (1993) is the fact that the inference is based on a limited number of observation sites, the remaining ones being used to validate the model. The author detected that the model missed a part of spatial dependence. The approach presented below is based on the point process convergence result and tries thus to solve the sparsity of data. Secondly, the coherence of the used model, namely mixtures of Dirichlet process, extends very naturally to spatial contexts and helps to consider data at every observation site. Furthermore, the extension from multivariate to spatial data is conceptually very simple.

5.2 Motivation

In this section, we illustrate the main ideas in order to motivate the next theoretical section. Let Ω be a geographical area and $X_1, \ldots, X_n$ be independent and identically distributed replicates of a random measure X defined on Ω. An example is rainfall. The aim is to analyze the stochastic behavior of X. However, we only observe X at a finite number of sites $s_1, \ldots, s_p \in \Omega$, which correspond to meteorological stations for example. These sites may have their own characteristics, such as their altitude. This situation is illustrated in Figure 5.1 where each site can be classified according to the subset A_1, A_2, A_3 of Ω to which it belongs.

A natural approach to analyze the stochastic behavior of X is to determine its distribution. Unfortunately, the distribution of a random measure is an abstract object, difficult to deal with in practice. Using the finite dimensional distributions of X is one

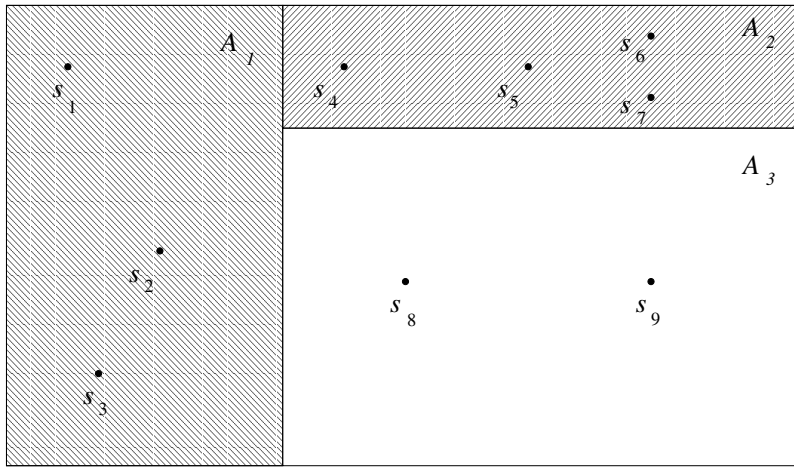


Figure 5.1: Illustration of spatial situation. Nine stations of observations $s_1, \ldots, s_9$ (black dots) on a area (rectangle), belonging to various subsets A_1, A_2 and A_3 (shaded).

possibility. Each element of this family is the distribution function $F_{B_1,\ldots,B_d}$ of the vector $\{X(B_1), \ldots, X(B_d)\}$ for any measurable and disjoint decomposition $B_1, \ldots, B_d$ of Ω. This family fully characterizes the distribution of X and conversely: a distribution F for X determines $\{F_{B_1,\ldots,B_d}\}$ uniquely and a coherent family $\{F_{B_1,\ldots,B_d}\}$ determines F. The coherence is ensured by the Kolmogorov's conditions. An toy example is the following Gaussian random field for which

$$\{X(B_1), \ldots, X(B_d)\} \sim \mathcal{N}_d \left\{ \begin{pmatrix} \mu(B_1) \\ \vdots \\ \mu(B_d) \end{pmatrix}, \begin{pmatrix} \sigma(B_1) & 0 & \cdots & 0 \\ 0 & \ddots & \vdots & \vdots \\ \vdots & \ldots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma(B_d) \end{pmatrix} \right\},$$

where μ and σ are two Radon measures, σ being positive. The observation of the vector $\{X_i(B_1), \ldots, X_i(B_d)\}$, $i = 1, \ldots, n$, allows inference on $\{\mu(B_1), \ldots, \mu(B_d)\}$, for example. If μ is parametrized, then we can draw conclusions about the entire measure μ.

For the extremes, the general behavior of X is not of interest, but only its extremal behavior. In finite dimension, the extremal paradigm assumes convergence to a Poisson process with a homogeneous intensity. This amounts to assuming that $R = \sum_{j=1}^{d} X^{(j)}$ and $W = X/R$ are independent for large R. It turns out that this result remains valid

for random measures, as will be shown. The total intensity $R = X(\Omega)$ and the random probability measure $W = X/R$ are independent for large R. Therefore, the inference is to be done on H, the distribution of W. In the following, we use the mixture of Dirichlet processes as a model, that is we suppose that

$$\{W(B_1), \ldots, W(B_d)\} \sim \sum_{m=1}^{k} \pi_m \text{Dir}\left\{\alpha_m(B_1), \ldots, \alpha_m(B_d)\right\},$$

for any disjoint and measurable decomposition $B_1, \ldots, B_d$ of Ω,where $\{\pi_m\}_{m=1}^k$ is a probability vector and α_m a Radon measure, $m = 1, \ldots, k$. The observation of W_i for extremes X_i, that is those whose R_i exceeds a high threshold r_0, allows inference on k, π and $\{\alpha_m(B_1), \ldots, \alpha_m(B_d)\}$, $m = 1, \ldots, k$.

However, we observe X and hence W only at points $s_1, \ldots, s_p$. From these observations, an estimate $\left\{\hat{W}(B_1), \ldots, \hat{W}(B_d)\right\}$ of $\{W(B_1), \ldots, W(B_d)\}$ is obtained. There are many possibilities, like the empirical estimate

$$\hat{W}(B_1) = \frac{|B_1|}{|\Omega|} \frac{1}{\#\{j : s_j \in B_1\}} \sum_{j:s_j \in B_1} W_{s_j}.$$

Another possibility is a smoothing approach: from $\{X_{s_1}, \ldots, X_{s_p}\}$ a smooth intensity function on Ω is extrapolated, say f. From it, we deduce

$$R = \int_\Omega f(s)ds, \quad \tilde{f} = f/R,$$

which leads to the numerical estimate

$$\hat{W}(B_1) = \int_{B_1} \tilde{f}(s)ds.$$

Naturally, we cannot perform the analysis for all decompositions $B_1, \ldots, B_d$ and we have to choose an appropriate one. If it is too coarse, then the resulting information about the α_m's is poor, and if it is too fine, then the problem may become infeasible if the number of observational sites p is large. In Figure 5.1, a heuristic strategy would be to choose $d = 3$ and $B_j = A_j$, $j = 1, 2, 3$. In other words, stations are regrouped by similarity. In practice, however, this selection is not easy and may be guided by an expert of the scientific domain from which the data are extracted.

Some further issues must be treated, such as the marginalization of X. In the finite dimensional case, each margin must be in the domain of attraction of a unit Fréchet distribution and this implies constraints on H. We will see that this feature can be extended in a natural way in the infinite dimensional case. Also main ideas of the choice of the threshold, simulation and conditional dependence analysis will be illustrated in an example.

5.3 Theoretical background

5.3.1 Topology and convergence

Let $\mathfrak{S}$ be a locally compact second countable Hausdorff space (that is every point has a compact neighborhood, $\mathfrak{S}$ has a countable base and distinct points may be separated by disjoint neighborhoods). A subset B is said to be bounded or relatively compact if its closure, $\bar{B}$, is compact. $\mathcal{F}_c$ denotes the set comprising every continuous function $f : \mathfrak{S} \to \mathbb{R}_+$ with compact support. The Borel algebra of $\mathfrak{S}$ is denoted $\mathcal{S}$ and $\mathcal{B}$ is the class of all bounded B in $\mathcal{S}$. A Radon or locally finite measure μ is a measure on $\mathfrak{S}$ such that $\mu B < \infty$, $\forall B \in \mathcal{B}$. The class of all Radon measures is noted $\mathfrak{M}$, and $\mathfrak{N}$ denotes the class of all Radon measures valued in $\mathbb{Z}_+ \cup \{\infty\}$. For each Radon measure on $\mathfrak{S}$ μ, $\mathcal{B}_\mu$ denotes the class of all sets $B \in \mathcal{B}$ such that $\mu \partial B = 0$, where ∂B is the boundary of B.

The vague topology on $\mathfrak{M}$ is generated by the sets $\{\mu : s < \int f d\mu < t\}$, $s, t \in \mathbb{R}$, $f \in \mathcal{F}_c$. We write $\mu_n \xrightarrow{v} \mu$ and say that a sequence $\{\mu_n\}$ vaguely converges to μ if, by definition,

$$\int f d\mu_n \overset{n\to\infty}{\longrightarrow} \int f d\mu, \quad \forall f \in \mathcal{F}_c,$$

or if one of the following equivalent statements is satisfied:

i) $\mu_n B \to \mu B, \quad \forall B \in \mathcal{B}_\mu,$

ii) $\limsup_{n\to\infty} \mu_n F \leq \mu F$ and $\liminf_{n\to\infty} \mu_n G \geq \mu G$, for all closed $F \in \mathcal{B}$ and all open $G \in \mathcal{B}$.

A subset $M \subset \mathfrak{M}$ is relatively compact in the vague topology if and only if

$$\sup_{\mu \in M} \mu B < \infty, \quad \forall B \in \mathcal{B}.$$

The weak topology on $\mathfrak{M}$ is generated by the same sets as the vague topology but replacing $\mathcal{F}_c$ by the set of continuous bounded functions $f : \mathfrak{S} \to \mathbb{R}_+$, say $\mathcal{F}_b$. A sequence $\{\mu_n\}$ weakly converges to μ if, by definition,

$$\int f d\mu_n \overset{n\to\infty}{\longrightarrow} \int f d\mu, \quad \forall f \in \mathcal{F}_b.$$

We then write $\mu_n \xrightarrow{w} \mu$. A subset $M \subset \mathfrak{M}$ is relatively compact in the weak topology if and only if

$$\sup_{\mu \in M} \mu B < \infty, \quad \forall B \in \mathcal{B}, \quad \text{and} \quad \inf_{B \in \mathcal{B}} \sup_{\mu \in M} \mu B^c = 0.$$

The following statements are equivalent:

i) $\mu_n \xrightarrow{w} \mu,$

ii) $\mu_n \xrightarrow{v} \mu$ and $\mu_n \mathfrak{S} \longrightarrow \mu \mathfrak{S}$,

iii) $\mu_n \xrightarrow{v} \mu$ and $\inf_{B \in \mathcal{B}} \limsup_{n \to \infty} \mu_n B^c = 0$.

In particular, convergence in the vague and the weak topologies coincide for probability measures. An important property is that $\mathfrak{M}$ and $\mathfrak{N}$ are Polish in the vague and in the weak topologies, that is, there exists a complete and separable metric.

5.3.2 Random measures

$\mathfrak{M}$ and $\mathfrak{N}$ are equipped with the σ-algebras, $\mathcal{M}$ and $\mathcal{N}$, generated by the vague topology. This coincides with the σ-algebras in $\mathfrak{M}$ and $\mathfrak{N}$, respectively, generated by the mappings $\mu \mapsto \mu B$, from $\mathfrak{M}$ to $\mathbb{R}$, defined for any $B \in \mathcal{B}$. A random measure or a point process on $\mathfrak{S}$ is a measurable mapping from some fixed probability space $(\Omega, \mathcal{A}, P)$ into $(\mathfrak{M}, \mathcal{M})$ or $(\mathfrak{N}, \mathcal{N})$ respectively.

The distribution of a random measure ξ is $P\xi^{-1}$, defined as

$$P\xi^{-1}(M) = P(\xi^{-1}M) = P\{\xi \in M\}, \quad M \in \mathcal{M} \text{ or } \mathcal{N}.$$

Equality in distribution $\xi \stackrel{d}{=} \eta$ means $P\xi^{-1} = P\eta^{-1}$ or one of the following equivalent statements:

(i) $\int f d\xi = \int f d\eta, \quad f \in \mathcal{F}_c,$

(ii) $L_\xi(f) = L_\eta(f), \quad f \in \mathcal{F}_c,$

(iii) $(\xi I_1, \ldots, \xi I_k) \stackrel{d}{=} (\eta I_1, \ldots, \eta I_k)$, $k \in \mathbb{N}$, $I_1, \ldots, I_k \in \mathcal{B}$.

The third equivalence could be written for a smaller class than $\mathcal{B}$ but these distinctions are outside of the scope of this work. L_μ denotes the Laplace transform of μ,

$$L_\mu(f) = E(e^{-\int f d\mu}), \quad f \in \mathcal{F}_c.$$

It can be defined for any positive measurable function $f : \mathfrak{S} \longrightarrow \mathbb{R}_+$. The distributions of each vector $(\mu I_1, \ldots, \mu I_k)$, $k \in \mathbb{N}$, $I_1, \ldots, I_k \in \mathcal{B}$, are called the finite dimensional distributions of μ. For such a family of distributions to define a random measure μ, it must be coherent in a sense made precise by Kolmogorov's consistency conditions. For further details, see Kallenberg (1983, p.41).

The convergence in distribution of the sequence of random measures $\{\mu_n\}$ to μ means the vague convergence of their distributions or one of the following equivalent statements:

(i) $\int f d\xi_n \xrightarrow{d} \int f d\xi, \quad f \in \mathcal{F}_c,$

(ii) $L_{\xi_n}(f) \xrightarrow{d} L_\xi(f), \quad f \in \mathcal{F}_c,$

(iii) $(\xi_n I_1, \ldots, \xi_n I_k) \xrightarrow{d} (\xi I_1, \ldots, \xi I_k)$, $k \in \mathbb{N}$, $I_1, \ldots, I_k \in \mathcal{B}_\xi$,

where $\mathcal{B}_\xi = \{B \in \mathcal{B} : \xi \partial B = 0, \text{ a.s.}\}$.

Convergence in distribution is derived from the vague topology, but sometimes the weak topology is used instead, in which case $\mathfrak{M}$ and $\mathfrak{N}$ are equipped with the σ-algebras generated by the weak topology. In order to make the difference between the weak and the vague convergence in distribution, we write $\xrightarrow{wd}$ or $\xrightarrow{vd}$ instead of $\xrightarrow{d}$, respectively. The following statements are equivalent:

(i) $\xi_n \xrightarrow{wd} \xi$,

(ii) $\xi_n \xrightarrow{vd} \xi$ and $\xi_n \mathfrak{S} \xrightarrow{d} \xi \mathfrak{S}$,

(iii) $\xi_n \xrightarrow{vd} \xi$ and $\inf_{B \in \mathcal{B}} \limsup_{n \to \infty} P(\xi_n B^c > \varepsilon) = 0$, $\forall \varepsilon > 0$.

For random probabilities, that is random measures such that $\mu \mathfrak{S} = 1$, almost surely, the two notions coincide.

The Dirac measure at $s \in \mathfrak{S}$, a random point, is the measure $\delta_s \in \mathfrak{N}$ such that

$$\delta_s B = \mathbb{1}_B(s), \quad B \in \mathcal{B}.$$

If $s_1, \ldots, s_n \in \mathfrak{S}$ are independent and identically distributed according to F then the point process

$$\xi = \sum_{i=1}^{n} \delta_{s_i}$$

is called a sample process. If we let $n = \nu$ be random and $\mathbb{Z}_+$-valued, then ξ is a mixed sample process. If furthermore ν is distributed according to a Poisson distribution with mean $a \geq 0$ then ξ is a Poisson process with intensity $\lambda = aF$. The L-transform of a Poisson process with intensity λ is

$$L_\xi(f) = \exp\left\{-\lambda\left(1 - e^{-f}\right)\right\}, \quad f \in \mathcal{F}.$$

By extension, any point process with a similar L-transform is called a Poisson process even if $\lambda \in \mathfrak{M}$ is unbounded. A Poisson process ξ_α with intensity $\alpha\lambda$, α an $\mathbb{R}^+$-valued random variable, $\lambda \in \mathfrak{M}$, is called a mixed Poisson process. It has L-transform

$$L_\alpha\left\{\lambda\left(1 - e^{-f}\right)\right\}, \quad f \in \mathcal{F},$$

where L_α is the L-transform of α. More generally, if $\lambda = \eta$ is a random measure on $\mathfrak{S}$, then ξ is called a doubly stochastic Poisson process or a Cox process directed by η. Its L-transform is

$$L_\eta \left(1 - e^{-f}\right), \quad f \in \mathcal{F},$$

where L_η is the L-transform of η. An array of point processes $\{\xi_{nj}\}$, $j = 1, \ldots, r_n$, $1 \leq r_n \leq \infty$, $n \in \mathbb{N}$, is called a null-array if for any fixed n, the ξ_{nj} are independent and

$$\lim_{n\to\infty} \sup_j P\left(\xi_{nj} B > 0\right) = 0, \quad B \in \mathcal{B}.$$

This notion is central in extreme value theory because of the following key result, originally due to Jagers (1972).

Theorem 5.1

Let ξ be a Poisson process on $\mathfrak{S}$ with intensity measure $\lambda \in \mathfrak{M}$ and let $\{\xi_{nj}\}$ be a null-array of point processes on $\mathfrak{S}$. Then $\sum_j \xi_{nj} \xrightarrow{d} \xi$ if and only if

(i) $\sum_j P\left(\xi_{nj} I > 0\right) \overset{n\to\infty}{\longrightarrow} \lambda I, \quad I \in \mathcal{B}_\lambda,$

(ii) $\sum_j P\left(\xi_{nj} B > 1\right) \overset{n\to\infty}{\longrightarrow} 0, \quad B \in \mathcal{B}.$

The original result remains valid if the two conditions are satisfied on smaller sets than $\mathcal{B}_\lambda$ and $\mathcal{B}$ but these distinctions lie outside the scope of this work.

For application to statistical extreme value analysis, consider an array of independent random elements of $\mathfrak{S}$, $\{s_{nj}\}_{j=1}^n$, distributed according to F_n, $n \in \mathbb{N}$, and corresponding Dirac mass arrays, $\{\xi_{nj}\}_{j=1}^n$, with $\xi_{nj} = \delta_{s_{nj}}$, $j = 1, \ldots, n$, $n \in \mathbb{N}$. Then $\{\xi_{nj}\}_{j=1}^n$ is a null-array if and only if

$$\lim_{n\to\infty} P\left(s_{nj} \in B\right) = 0, \quad B \in \mathcal{B},$$

in other words, points vanish away from any bounded set. Incorporating conditions (i) and (ii) in Theorem 5.1, the first states that

$$nF_n \xrightarrow{v} \lambda,$$

while the second is trivially satisfied.

The multivariate extreme value theorem is the particular case of Theorem 5.1 where $\mathfrak{S}$ is $\mathbb{R}^d$, $s_{nj} = n^{-1} X_{nj}$, where the X_{nj} are positive independent and identically distributed according to F with unit Fréchet margins, such that

$$nF(n\cdot) \xrightarrow{v} \lambda.$$

In that case, λ has to be homogeneous of degree -1,

$$\lambda(tB) = t^{-1} \lim_{n\to\infty} ntF(ntB) = t^{-1}\lambda B, \quad B \in \mathcal{B}_\lambda.$$

Note that λ cannot place mass at the origin, since vanishing points concentrate at the origin and the limiting point process could not be Poisson because of an infinite mass at 0. In particular, sets in $\mathcal{B}_\lambda$ are bounded away from 0. Now consider the transformation $T : x \mapsto (\|x\|\,,\, x/\|x\|) = (r\,,\, w)$, and the image of λ through T is $dr/r^2 \times H(dw)$, with H the spectral measure previously introduced.

The generality of Theorem 5.1 allows us to apply the result to more general spaces and in particular to the spatial context, presented in the next section.

5.4 Spatial extremes

This section deals with applications of the previous sections to spatial extremes. The Poisson process of vectors is generalized to a Poisson process of measures, the spectral distribution is generalized to the spectral process and its application is illustrated on a real dataset.

5.4.1 The spectral process

As explained in Section 5.2, observations are not constituted of vectors but of random measures, $X_1, \ldots, X_n$, on some area Ω equipped with the Lebesgue measure. Therefore the sample space is the set of Radon measures on Ω, $\mathfrak{S} = \mathfrak{M}(\Omega)$. It is equipped with the vague topology, making it Polish, so that the previous theory applies. We do not consider pointwise measures but measures on local areas. Loosely, the typical events that can be observed are

$$\{s_1 < X(A_1) < t_1, \ldots, s_d < X(A_d) < t_d\}, \quad s_i, t_i \in \mathbb{R}, i = 1, \ldots, d,$$

for any collection of measurable sets $A_1, \ldots, A_d$. In order to do inference on the distribution of X, we suppose that X is positive and that the array $\{\delta_{n^{-1}X_{ni}}\}_{i=1}^n$, $n \in \mathbb{N}$, is a null-array, that is, n^{-1} is the correct normalizing sequence. Under appropriate conditions and applying Theorem 5.1, the sample point process $\sum_{i=1}^n \delta_{n^{-1}X_{ni}}$ converges to a Poisson process with intensity measure λ, homogeneous of degree -1. By considering the norm $\|X\| = X(\Omega)$ we have that λ decomposes on the pseudo-polar scale $R = \|X\|$ and

$W = X/\|X\|$ into a product

$$\lambda(dr, dw) = \frac{1}{r^2} dr \times \tilde{H}(dw),$$

where $\tilde{H}$ is a finite positive measure on the set of probability measures on Ω. Let us define H to be the normalized version of $\tilde{H}$. Then H is the distribution of the random probability measure W on Ω, that is, the spectral process. It is characterized by its finite dimensional distributions and a valid model for H is the mixture of Dirichlet processes.

Inference can be done from a selected decomposition $\Omega_1, \ldots, \Omega_d$ by fitting $H_{\Omega_1,\ldots,\Omega_d}$ to observation $\{W_i(\Omega_1), \ldots, W_i(\Omega_d)\}$, $i \in I_0$, where $I_0 = \{1 \leq i \leq n : X_i(\Omega) > r_0\}$, for a high threshold r_0. The physical interpretation of this model is the following: an extreme observation is defined by the fact that the total intensity on Ω, $R = X(\Omega)$, exceeds a given threshold r_0. The random distribution of this extreme on the area Ω is given by $X/X(\Omega) = W$. For example, if W follows a mixture of Dirichlet processes then the proportions of R in $\Omega_1, \ldots, \Omega_d$ are $W(\Omega_1), \ldots, W(\Omega_d)$ and this vector is distributed according to

$$\sum_{m=1}^{k} \pi_m \text{Dir}\{\alpha_m(\Omega_1), \ldots, \alpha_m(\Omega_d)\},$$

where α_m is a positive Radon measure on Ω, $m = 1, \ldots, k$.

Assuming that n^{-1} is the right standardization sequence implies there are constraints on H, as in the multivariate case. By standard arguments, we have that the asymptotic distribution function of the componentwise maximum

$$M^{(n)}_{\Omega_1,\ldots,\Omega_d} = \max_{i=1,\ldots,n} \left\{ n^{-1} X_i(\Omega_1), \ldots, n^{-1} X_i(\Omega_d) \right\}$$

is

$$\lim_{n\to\infty} P\left(M^{(n)}_{\Omega_1,\ldots,\Omega_d} \leq y \right) = \exp\left\{ -\int_{S_d} \max_{i=1,\ldots,d} \left(\frac{w_i}{y_i} \right) \tilde{H}_{\Omega_1,\ldots,\Omega_d}(dw) \right\}.$$

Without loss of generality, consider F_{Ω_1}, the first marginal of the finite dimensional distribution of F corresponding to the decomposition $\Omega_1, \ldots, \Omega_d$. The convergence requirement states that F_{Ω_1} is in the domain of attraction of a Fréchet distribution. Comparing with the marginal of the componentwise maximum distribution we have that

$$\lim_{n\to\infty} P\left(M^{(n)}_{\Omega_1} \leq y_1 \right) = \exp\left(-C_{\Omega_1}/y_1 \right),$$

where

$$C_{\Omega_1} = \int_{S_d} w_1 H_{\Omega_1,\ldots,\Omega_d}(dw).$$

A consequence is that the total mass of $\tilde{H}$ is $C(\Omega)$. Therefore $H = \tilde{H}/C(\Omega)$. For example in the case $C_{\Omega_i} = |\Omega_i|$, the marginalization consists in transforming original data, $X_i(\Omega_j)$, to data $Y_i(\Omega_j)$ in the Fréchet domain of attraction with parameters $|\Omega_i|$, $i = 1, \ldots, n$, $j = 1, \ldots, d$.

In the case where H is a mixture of Dirichlet processes, the constraints become

$$\sum_{m=1}^{k} \pi_m \frac{\alpha_m(\Omega_j)}{\alpha_m(\Omega)} = \frac{|\Omega_j|}{|\Omega|}, \quad j = 1, \ldots, d.$$

Assuming this for every decomposition $\Omega_1, \ldots, \Omega_d$, $\sum_{m=1}^{k} \pi_m \alpha_m / \alpha_k(\Omega)$ is the uniform measure on Ω. In other words, the expectation of W is the Lebesgue measure on Ω.

As a final remark, we note that it is possible to reconcile the pointwise approach with the setwise approach. Considering infinitesimal sets around each of the d observational sites, we consider the distribution of the vector $(W_1, \ldots, W_d, W_{d+1})$, where W_j is the proportion due to the infinitesimal set around site j, $j = 1, \ldots, d$, and W_{d+1} is the proportion due to the rest of the area, that is virtually all the area. Then the marginal distribution of $(W_1/(1 - W_{d+1}), \ldots, W_d/(1 - W_{d+1}))$ is a mixture of Dirichlet distributions with parameters $\alpha_j^{(m)}$, $j = 1, \ldots, d$, $m = 1, \ldots, k$, and mixing distribution $\{\pi_m\}_{m=1}^{k}$. Here $\alpha_j^{(m)}$ is the intensity of the measure $\alpha^{(m)}$ on the infinitesimal set around site j. This remark allows us to estimates the density function of $\alpha^{(m)}$ at every observational site, $m = 1, \ldots, k$, and then perhaps smooth them to obtain estimates where no point can be observed.

The setwise and the pointwise procedures give us two coherent ways to incorporate the available information in order to obtain estimates and a detailed description of the extremes of the phenomenon on the area under study.

5.4.2 Estimation

When the spectral model is a mixture of Dirichlet processes, the estimation of the mixing distribution $\{\pi_m\}_{m=1}^{k}$ and the parameter measures $\{\alpha_m\}_{m=1}^{k}$ are of interest. It can be also interesting to select k or to mix over various k by a Bayesian approach, as in the multivariate context. The estimation is based on a likelihood function obtained for a fixed measurable decomposition $\Omega_1, \ldots, \Omega_d$ of Ω.

In practice, a finite number n of data $Y_j^{(i)}$ are observed at a finite number q of sites in the area Ω, $j = 1, \ldots, q$, $i = 1, \ldots, n$. We interpolate each $Y^{(i)}$ to obtain the corresponding observed functions. Now we fix the decomposition $\Omega_1, \ldots, \Omega_d$ of Ω according to the question which the inference is to answer. For the example below we fix this decomposition according to the topography of Ω. A new dataset, $Y^{(i)}(\Omega_j)$ is built by integrating

numerically the resulting functions, $i = 1, \ldots, n$, $j = 1, \ldots, d$. It is marginalized to $X_j^{(i)}$ having Fréchet margin with parameter $|\Omega_j|$ using the extremal semi-parametric model. As a by-product, univariate thresholds u_j and parameter estimates $\{\sigma_j, \kappa_j\}$, $j = 1, \ldots, q$, are obtained.

Now the inference is similar to the multivariate case; a threshold r_0 is selected for the R_i and pseudo-polar angles $W_j^{(i)} = X_j^{(i)}/R_i$, $j = 1, \ldots, d$, $i = 1, \ldots, n$, having $R_i > r_0$ are used to fit the mixture of Dirichlet distributions with mixing distribution $\pi_1, \ldots, \pi_k$ and parameters $\alpha_1, \ldots, \alpha_m$ under constraints

$$\sum_{m=1}^{k} \pi_m \frac{\alpha_j^{(m)}}{\sum_{l=1}^{d} \alpha_l^{(m)}} = \frac{|\Omega_j|}{|\Omega|}, \quad j = 1, \ldots, d.$$

This approach has naturally the drawback that it estimates the parameter measures α_m, $m = 1, \ldots, k$, only on the coarser decomposition $\Omega_{I_1}, \ldots, \Omega_{I_d}$. For a general measurable set A, the estimate of $\alpha_m(A)$ is

$$\sum_{j=1}^{d} \frac{|A \cap \Omega_j|}{|\Omega_j|} \alpha_m(A), \quad m = 1, \ldots, k.$$

Therefore the decomposition $\Omega_1, \ldots, \Omega_d$ must be chosen with care prior to the analysis.

5.4.3 Real data analysis

The data we consider are sequences of monthly total precipitations (mm) at sixty meteorological stations in China during the period 1951–1988. They are a part of a much larger data set available at the web site of the University Corporation for Atmospheric Research, `http://dss.ucar.edu/datasets/ds578.5/`. As there are some exceptional missing data, we complete them by averaging over the immediate neighbors.

In order to remove the strong seasonality of the data, we work on their anomalies. For example, for February data, the mean of every February observation through the 38 years is subtracted and the result is divided by the standard error. At each site, the autocorrelation and partial autocorrelation functions of the resulting time series reveal weak correlation. A few sites exhibit low, almost statistically insignificant, correlation at lag one. We henceforth ignore this and assume that the data are stationary and independent in time. The altitude of each station is known. In absence of any other information on the topography of the area we assume this topography to be smooth. The resulting height profile of the area and the station positions are shown in Figure 5.2. The decomposition is obtained by thresholding the heights every 700m from 0m up to 2800m plus one area $>$ 2800m.

group (j)	Ω_j	$\lvert\Omega_j\rvert$	u_j	$\hat{\sigma}_j$	$\hat{\kappa}_j$
1	0–700m	0.460	186.1	131.6 (22.9)	−0.303 (0.103)
2	700–1400m	0.293	147.3	105.1 (20.7)	−0.105 (0.132)
3	1400–2100m	0.121	55.1	38.8 (6.4)	−0.372 (0.092)
4	2100–2800m	0.088	52.0	20.6 (4.8)	0.247 (0.184)
5	$>$ 2800m	0.038	28.5	28.9 (5.7)	−0.105 (0.132)

Table 5.1: Marginal parameters. First column is the index of the group. Second column gives the heights of station in group j, $j = 1, \ldots, 5$. Third column gives the size of the area covered by each group, relative to the total area of Ω. Columns 4, 5 and 6 gives respectively the threshold, estimates of σ and κ for the generalized Pareto tail of each group. Standard errors based on limiting covariance matrices are given in brackets.

Following the estimation procedure, we interpolate the anomalies by a function. This procedure is standard in `matlab`; function `griddata` interpolates by triangle-based linear functions on the desired grid, uniform in our case. The integration over an area Ω_j, $j = 1, \ldots, d$, is the sum of the values of the function at every observation point in Ω_j. The resulting five dimensional vector is standardized to a Fréchet scale with parameter $|\Omega_j|$. Marginal parameters, thresholds and the size of each group are shown in Table 5.1.

The choice of the multivariate threshold is made via the diagnostic set shown in Figure 5.3. A threshold of exp(3) looks appropriate but as it leaves only 27 excesses, not enough for the semi-parametric approach. We hence proceed with $r_0 = \exp(2.6)$ and 48 excesses.

The reversible jump algorithm is launched for $100,000$ iterations, and its convergence assessed using the three plot diagnostic set based on the parameter

$$\theta = \int_0^1 \min\left\{w_{(1,5)}, w_{(2,3,4)}\right\} H(dw),$$

where $w_{(1,5)}$ is the probability corresponding to groups $(1, 5)$ and $w_{(2,3,4)}$ is the probability corresponding to groups $(2, 3, 4)$. This choice was made for sake of simplicity, firstly because computation of θ is awkward for $d > 2$, and secondly because the total area of groups $(1, 5)$ is the same as the total area of groups $(2, 3, 4)$. The three plot diagnostic set is shown in Figure 5.4. Each plot has dramatic jitters at its right-hand side, due to the diminished number of observations which contribute to the variances. A check on k

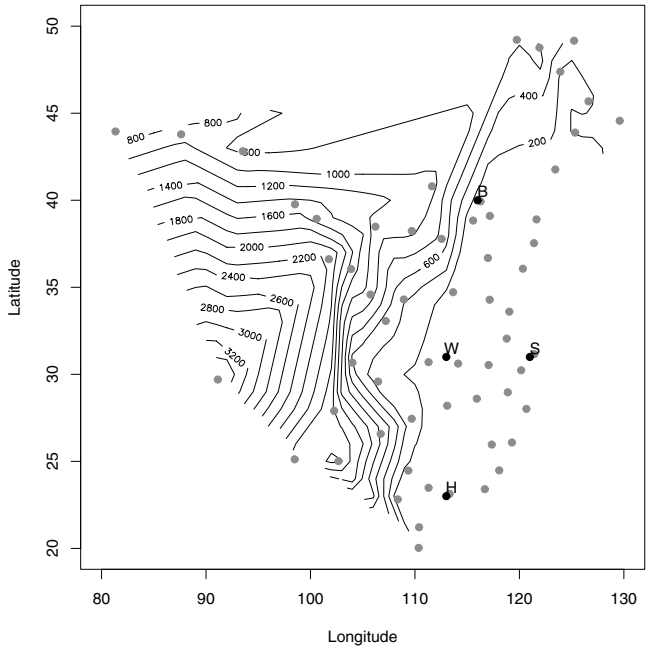


Figure 5.2: Topography of China. Level curves indicates altitude. Grey dots are the observational sites. Black dots are cities 'B' for Beijing, 'S' for Shanghai, 'W' for Wuhan, 'H' for Hong Kong.

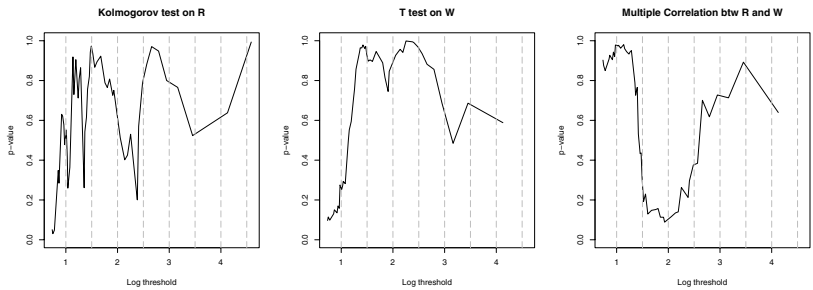


Figure 5.3: Three plot diagnostic set applied to China data.

and other parameters shows no evidence of non-convergence. Below we proceed with the last $20,000$ iterations of the chain, assuming convergence.

Figures 5.5 and 5.6 show the conditional expectation and the conditional dependence measure between groups. For the pair $(1,2)$ the construction is the following: consider

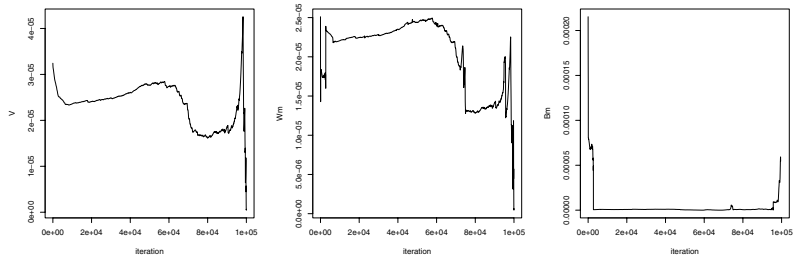


Figure 5.4: Convergence diagnostic plot for the reversible jump algorithm, based on the dependence measure between groups $(1, 5)$ and $(2, 3, 4)$.

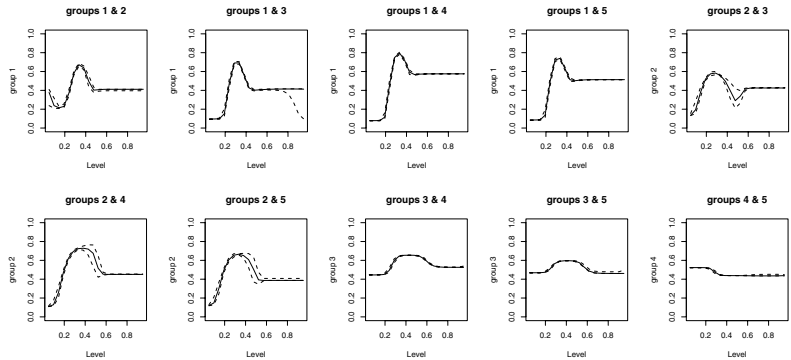


Figure 5.5: Conditional expectation of groups 1 to 5.

an area of unit size in group 1, $A_1 \subset \Omega_1$, an area of unit size in group 2, $A_2 \subset \Omega_2$, and A_0 the remaining part, $A_0 = \Omega \setminus (A_1 \cup A_2)$. Then the vector $\{W(A_1), W(A_2), W(A_0)\}$ is distributed according to the mixture of Dirichlet distributions with parameter $\alpha_m(A_1), \alpha_m(A_2)$ and $\alpha_m(A_0)$ and mixing distribution $\{\pi_m\}$, $m = 1, \ldots, k$. Since α_m is constant groupwise, for $m = 1, \ldots, k$,

$$\alpha_m(A_i) = \frac{\alpha_m(\Omega_i)}{|\Omega_i|}, \quad i = 1, 2,$$

and

$$\alpha_m(A_0) = \frac{|\Omega_1 \setminus A_1|}{|\Omega_1|}\alpha_m(A_1) + \frac{|\Omega_2 \setminus A_2|}{|\Omega_2|}\alpha_m(A_2) + \sum_{i=3}^{5} \alpha_m(\Omega_i).$$

Figures 5.5 and 5.6 show the summary statistics of the conditional distribution function of $\{W(A_1)/W(A_1 \cup A_2)\,,\, W(A_2)/W(A_1 \cup A_2)\}$ given $W(A_0)$.

The interpretation requires care. First, recall that we are studying the anomalies of precipitation data. An extreme event is an extreme anomaly, not an extreme precipitation. Secondly, it is conditional on the estimation procedure which in our case may be debated,

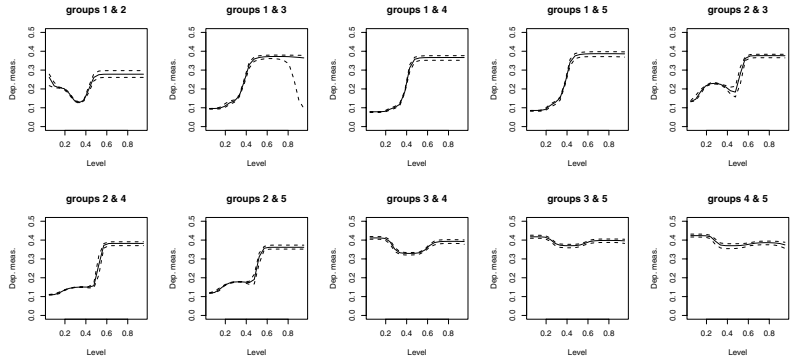


Figure 5.6: Conditional dependence measure for groups 1 to 5.

for example the choice of groups based on heights, or for the small number of data on which it is based. Disregarding these issues and taking the estimation as sure, it seems that groups 3, 4 and 5 are strongly dependent. It looks in fact as if they were one: the conditional expectation remains around 0.5 so that their expected contribution to an extreme event is equal; and the conditional dependence remains high whatever the conditioning level. This indicates their contribution to an extreme event is equal most of the time. These three groups are formed by stations above 1400m but they also represent the south-west region of China and, maybe, represent a typical meteorological region of China. Strengthening this observation we see that taking group 1 with one of the three groups 3, 4 or 5 gives the same behavior. The observation can also be done for group 2. Group 1 and group 2 seem to form two distinct groups. They exhibit moderate dependence, whatever the conditioning level of the remaining part, although the evolution of this dependence is quite difficult to interpret.

Therefore, we have isolated groups 1, 2, and $(3, 4, 5)$. Figure 5.7 shows the estimated density function of the vector $\{W(\Omega_1), W(\Omega_2), W(\Omega_3 \cup \Omega_4 \cup \Omega_5)\}$. We emphasize the fact that the expectation of this vector is not $(3^{-1}, 3^{-1}, 3^{-1})$, as in the multivariate case, but $(0.460, 0.293, 0.247)$. One can see that the three groups are well separated making us suspect asymptotic independence.

Finally, Figure 5.8 shows a simulation of W from the model on a uniform grid. The reversible jump algorithm has favored $k = 3$, each component being centered on a cluster, see Figure 5.7. The pictures given by Figure 5.8 are quite typical. When $k = 3$ is selected, either one of those typical realizations is selected with probability $\pi = (\pi_1, \pi_2, \pi_3)$. The

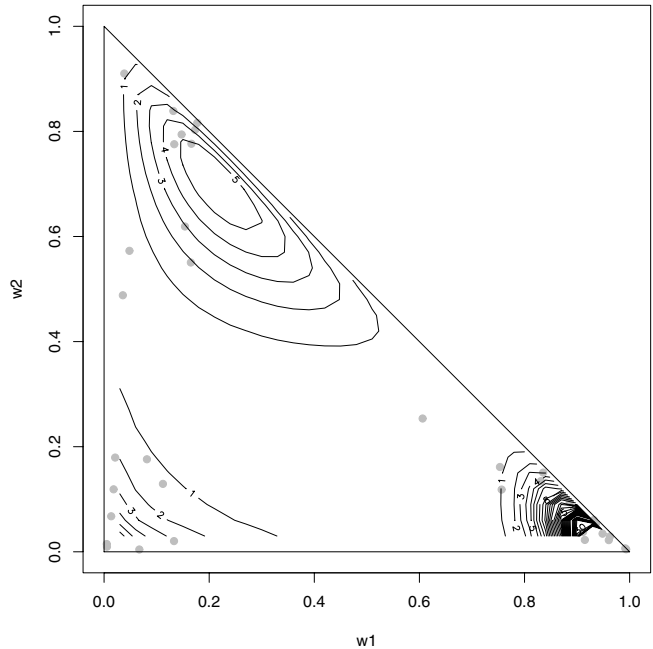


Figure 5.7: Spectral density function of groups 1, 2 and $(3, 4, 5)$.

location of groups, 1, 2, and $(3, 4, 5)$ appears clearly. Within each group, there is complete chaos, this is normal, since for a more precise image, one should have taken a thinner decomposition $\Omega_1, \ldots, \Omega_d$. Let us remark that multiplying this surface by a scalar R simulated from the distribution function $F(r) = 1 - r_0/r$, $r > r_0$, gives us a simulation scheme for the extremal curve, extremal in the sense that its total intensity exceeds r_0. Monte Carlo procedures can be developed in this direction.

5.5 Discussion and outlook

This presentation of spatial extremes emphasizes the central role of the dependence probability measure H as the distribution of a random probability. In this context, the mixture of Dirichlet processes is useful but surely other simple models exist. We tried to generalize classical extremal models, like the logistic and extremal Dirichlet, without success. It is easy to make them have constrained means but we never succeeded in keeping the coherence, to guarantee that if $(w_1, \ldots, w_d)$ is distributed according to a member of the

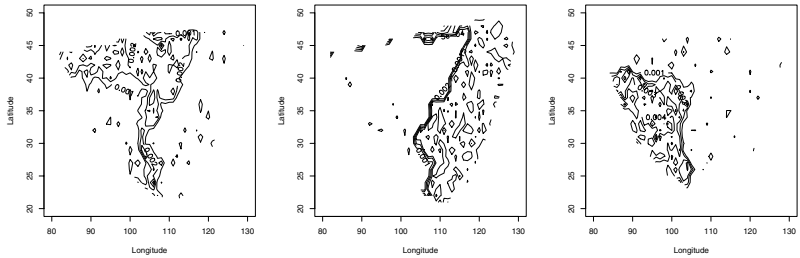


Figure 5.8: Illustration of spatial simulation. Simulated from one output with $k = 3$ of the Markov chain on groups 1, 2 and $(3, 4, 5)$. Either one of the three plots is obtained with probability π_1, π_2 and π_3 respectively.

family with mean $(\mu_1, \ldots, \mu_d)$, say, then $(w_1 + w_2, w_3, \ldots, w_d)$ will be in the same family with mean $(\mu_1 + \mu_2, \mu_3, \ldots, \mu_d)$.

A possibility to find other models is to consider a family of parametric densities $\{f_\theta : \theta \in \Theta\}$ on Ω with a prior distribution $\theta \sim p(\theta)$. Then f_θ plays the role of W while $p(\theta)$ plays the role of the spectral distribution. The constraints are that the expectation of f_θ under $p(\cdot)$ is the uniform density. Simple parametric models for the spectral distribution function may thus be found.

The choice of the decomposition $\Omega_1, \ldots, \Omega_d$ is crucial from the practical viewpoint and has not been discussed. A possible guideline is to search for groups that exhibit most independence between one another. In this context, models for asymptotic independence in the spatial context are to be developed. A simple generalization of the Gaussian tail model may be a Gaussian random field.

Conclusions

The research presented in this thesis is oriented toward the study of a new semi-parametric model for the spectral density in extremal statistics.

In Chapter 2, this model is defined in a general framework that lays the foundations of the generalization in Chapter 5. It is the occasion for a journey into non-parametric Bayesian statistics. Theory from this domain shows the richness of mixtures of Dirichlet distributions, or processes, which constitute the model. However, the spectral distribution constraints, which the model must satisfy, takes us away from the known framework and may potentially harm the usefulness of Dirichlet models. This question is hence explored by practical examples. To this end, two approaches, frequentist and Bayesian, are used: first the EM algorithm and second the reversible jump algorithm. Their theory and use is reviewed and they are adapted to our model.

In Chapter 3, the two algorithms are applied to real and simulated data in order to determine their strength and weaknesses and to study flexibility of the model. The EM algorithm turns out to be efficient in dimension two but has strong limits: the algorithm does not converge in high dimension; parameter uncertainty is difficult to assess. Furthermore the selection of the number of components cannot be done by the algorithm itself but by an information criterion. The reversible jump algorithm is efficient even in high dimensions but has also its drawbacks: convergence diagnostics have to be constructed and the convergence is never guaranteed; uncertainty seems to be underestimated at least in our example and the prior and hyperprior parameters have to be chosen properly in the absence of real prior information. The model turns out to be rich in the sense that it fits classical models of the literature, such as the logistic and the Dirichlet extremal model, rather well. This richness has its price, however, because the model is not parsimonious, a serious problem for the extremes where data are sparse. This lack of parsimony can be artificially solved by an appropriate choice of the prior distributions when using the reversible jump algorithm. However, a real incorporation of prior information in multi-

variate extremes remains unsolved. In parallel to this study, two new representations of the logistic and the Dirichlet model have been developed. Beyond the fact that they allow exact and rapid simulation of random variables, they also generalize the logistic model to any dimension.

The use of the spectral distribution is addressed in Chapter 4. Firstly, the estimation of quantiles for extreme events of general shape is partially solved thanks to Monte Carlo methods. These methods work well as long as the data are not asymptotically independent. In this latter case, the problem is beyond the model since the Poisson process limit is not valid in this case. It would be interesting to incorporate the semi-parametric model and the notion of random probability measures into the approaches specific to the asymptotic independent case. This would allow a generalization of these methods to the spatial context in the same way as in Chapter 5. Secondly, we studied a conditional analysis that gives a qualitative rather than a quantitative exploration of the dependence structure of the data. It brings to the fore hidden aspects of this dependence and shows strong links between the components of the data, as shown in several examples. These methods cannot be used in an automatic way, a practiced data analyst must be consulted. Its generalization to other models is straightforward in principle but technically difficult because of the computations involved.

The spatial approach proposed in Chapter 5 is very natural and the results are promising. We show how to use the Poisson process limit in this context, thus diminishing the loss of data. The basis of our method is the use of mixtures of Dirichlet distributions and their preponderant role in the theory of random probability measures. It would be interesting to develop simpler models for a rapid exploration of the data. This would render the choice of the decomposition of the studied area easier. It would also allow us to incorporate standard hypotheses in spatial statistics such as isotropy. The research of such models may be guided to Bayesian statistics where distributions of random probability measures, called prior distributions, are numerous. Finally, the notion of asymptotic independence is still to be addressed. One idea is to generalize the Gaussian tail model, which seems to be natural. This would almost complete the matching between multivariate and spatial extremes presented in this thesis.

Appendix

A.1 The Extremal Logistic Model

Below we prove formula (2.1) and the fact that $h_W(w)$ satisfies the constraints (2.2). The change of variable is

$$w_i = \frac{C_i u_i^{-\alpha_i}}{\sum_{j=1}^d C_j u_j^{-\alpha_j}} = \frac{C_i u_i^{q_i}}{K},$$

whose differential is

$$dw_i = \frac{C_i u_i^{q_i-1}}{K^2}\left(q_i K du_i - u_i dK\right), \quad i = 1, \ldots, d-1,$$

where $K = \sum_{j=1}^d C_j u_j^{-\alpha_j}$ and $q_i = -\alpha_i$. Therefore

$$\begin{aligned} dw_1 dw_2 &= \frac{C_1 C_2}{K^4}\left(q_1 q_2 u_1^{q_1-1} u_2^{q_2-1} K^2 du_1 du_2 - K q_1 u_1^{q_1-1} u_2^{q_2} du_1 dK - K q_2 u_1^{q_1} u_2^{q_2-1} dK du_2\right), \\ &= K^{-3} C_1 q_1 u_1^{q_1-1} C_2 q_2 u_2^{q_2-1}\left(K du_1 du_2 - q_2^{-1} u_2 du_1 dK - q_1^{-1} u_2 dK du_2\right). \end{aligned}$$

By recurrence we obtain

$$dw_1 \cdots dw_l = K^{-l-1} \prod_{i=1}^l C_i q_i u_i^{q_i-1}\left(K du_1 \cdots du_l - \sum_{i=1}^l q_i^{-1} u_i du_1 \cdots du_{i-1} dK du_{i+1} \cdots du_l\right),$$

Furthermore,

$$dK = \sum_{i=1}^{d-1} C_i q_i u_i^{q_i-1} du_i - C_d q_d u_d^{q_d-1} \sum_{i=1}^{d-1} du_i = \sum_{i=1}^{d-1}\left(C_i q_i u_i^{q_i-1} - C_d q_d u_d^{q_d-1}\right) du_i,$$

so that

$$du_1 \cdots du_{i-1} dK du_{i+1} \cdots du_{d-1} = \left(C_i q_i u_i^{q_i-1} - C_d q_d u_d^{q_d-1}\right) du_i.$$

Finally, writing $dw = dw_1 \cdots dw_{d-1}$ and $du = du_1 \cdots du_{d-1}$, one gets

$$\begin{aligned}
dw &= K^{-d} \prod_{i=1}^{d-1} C_i q_i u_i^{q_i-1} \left\{ K - \sum_{i=1}^{d-1} q_i^{-1} u_i \left(C_i q_i u_i^{q_i-1} - C_d q_d u_d^{q_d-1} \right) \right\} du \\
&= K^{-d} \prod_{i=1}^{d-1} C_i q_i u_i^{q_i-1} \left(K - \sum_{i=1}^{d-1} C_i u_i^{q_i} + C_d q_d u_d^{q_d-1} \sum_{i=1}^{d-1} q_i^{-1} u_i \right) du \\
&= K^{-d} \prod_{i=1}^{d-1} C_i q_i u_i^{q_i-1} \left(C_d u_d^{q_d} + C_d q_d u_d^{q_d-1} \sum_{i=1}^{d-1} q_i^{-1} u_i \right) du \\
&= K^{-d} \prod_{i=1}^{d} C_i q_i u_i^{q_i-1} \left(q_d^{-1} u_d + \sum_{i=1}^{d-1} q_i^{-1} u_i \right) du \\
&= K^{-d} \left(\prod_{i=1}^{d} C_i q_i u_i^{q_i-1} \right) \left(\sum_{i=1}^{d} q_i^{-1} u_i \right) du \\
&= K^{-d} \left\{ \prod_{i=1}^{d} C_i (-\alpha_i)^{-1} u_i^{-\alpha_i-1} \right\} \left\{ \sum_{i=1}^{d} (-\alpha_i) u_i \right\} du.
\end{aligned}$$

Therefore the density of W is

$$\begin{aligned}
h_W(w)dw &= h_U\{u(w)\} \left| \frac{du}{dw} \right| dw \\
&= d^{-1} \left(\sum_{i=1}^{d} \alpha_i u_i \right)^{-1} \left(\prod_{i=1}^{d} \alpha_i u_i \right) \left(\sum_{i=1}^{d} C_i u_i^{-\alpha_i} \right) \prod_{i=1}^{d} w_i^{-1} dw.
\end{aligned}$$

Furthermore, taking W_1 without loss of generality,

$$\begin{aligned}
E(W_1) &= E \left(\left\{ \sum_{i=1}^{d} C_i U_i^{-\alpha_i} \right\}^{-1} C_1 U_1^{-\alpha_1} \right) \\
&= d^{-1} \int_{S_d} \left(\sum_{i=1}^{d} C_i u_i^{-\alpha_i} \right)^{-1} \left(\sum_{i=1}^{d} C_i u_i^{-\alpha_i} \right) C_1 u_1^{-\alpha_1} du \\
&= d^{-1},
\end{aligned}$$

which proves (2.2).

A.2 The Extremal Dirichlet Model

Below we demonstrate (2.3). Let $K = \sum_{i=1}^{d} m_i w_i$. Then the change of variable is

$$u_i = m_i w_i / K, \quad i = 1, \ldots, d-1,$$

whose differential is

$$du_i = (K du_i - w_i dK) m_i / K^2, \quad i = 1, \ldots, d-1.$$

Therefore

$$du_1 du_2 = K^{-4}\left(K^2 dw_1 dw_2 - K w_2 dw_1 dK - K w_1 dK dw_2\right) m_1 m_2$$

and, by recurrence, for any $l \leq d-1$,

$$du_1 \cdots du_l = K^{-2(l+1)}\left(K^l dw_1 \cdots dw_l - K^{l-1}\sum_{j=1}^{l} w_j dw_1 \cdots dw_{j-1} dK dw_{j+1} \cdots dw_l\right)\prod_{j=1}^{l} m_j.$$

The differential of K is

$$dK = d\left(\sum_{i=1}^{d} m_i w_i\right) = \sum_{i=1}^{d-1}(m_i - m_d) dw_i,$$

so that

$$dw_1 \cdots dw_{j-1} dK dw_{j+1} \cdots dw_{d-1} = (m_j - m_d) dw_1 \cdots dw_{d-1}.$$

Noting $du = du_1 \cdots du_{d-1}$ and $dw = dw_1 \cdots dw_{d-1}$, one has

$$\begin{aligned}
du &= K^{-2(d-1)}\left\{K^{d-1} - K^{d-2}\sum_{j=1}^{d-1}(m_j - m_d) w_j\right\}\prod_{j=1}^{d-1} m_j dw, \\
&= K^{-d}\left\{K - \sum_{j=1}^{d-1} m_j w_j + m_d \sum_{j=1}^{d-1} w_j\right\}\prod_{j=1}^{d-1} m_j dw, \\
&= K^{-d}\left\{\sum_{i=1}^{d} m_i w_i - \sum_{j=1}^{d-1} m_j w_j + m_d \sum_{j=1}^{d-1} w_j\right\}\prod_{j=1}^{d-1} m_j dw, \\
&= K^{-d}\prod_{j=1}^{d} m_j dw,
\end{aligned}$$

that is

$$du = \prod_{j=1}^{d} m_j \left(\sum_{j=1}^{d} u_j/m_j\right)^{-d} dw$$

or equivalently

$$dw = \left(\prod_{j=1}^{d} m_j\right)^{-1}\left(\sum_{j=1}^{d} u_j/m_j\right)^{d} du.$$

Therefore,

$$\begin{aligned}
h_W(w) dw &= d^{-1}\left(\sum_{j=1}^{d} m_j w_j\right)^{-(d+1)} \prod_{j=1}^{d} m_j h^*\left(\frac{m_1 w_1}{m \cdot w}, \ldots, \frac{m_d w_d}{m \cdot w}\right) dw, \\
&= d^{-1}\left(\sum_{j=1}^{d} u_j/m_j\right)^{d+1} \prod_{j=1}^{d} m_j h^*(u)\left(\prod_{j=1}^{d} m_j\right)^{-1}\left(\sum_{j=1}^{d} u_j/m_j\right)^{d} du, \\
&= d^{-1} h^*(u)\sum_{j=1}^{d} u_j/m_j
\end{aligned}$$

which is the claimed formula.

A.3 Confidence intervals for the EM algorithm

Below are given the proofs of (2.4) and (2.5). According to Oakes (1999),

$$\frac{\partial^2 \log f(y,u;\,\theta)}{\partial\theta^2} = \frac{\partial^2 Q(\theta,\,\theta')}{\partial\theta^2} + \frac{\partial^2 Q(\theta,\,\theta')}{\partial\theta\partial\theta'},$$

at $\theta = \theta' = \hat{\theta}$, the maximum likelihood estimator. The first term of the right side of the equality is often numerically available from the optimizer used during the M step. In order to avoid awkward analytical calculation, the second term can be approximated by

$$\mathrm{var}\left\{\frac{\partial \log f(y,U;\,\theta)}{\partial\theta}\right\},$$

where the variance is taken with respect to U given $Y = y$, at $\theta = \hat{\theta}$. The proof uses the hypothesis that differentiation and integration can be exchanged. Indeed,

$$\begin{aligned}
\frac{\partial^2 Q(\theta,\,\theta')}{\partial\theta\partial\theta'} &= \frac{\partial^2}{\partial\theta\partial\theta'}\int \log f(y,u;\,\theta) f(u \mid y;\,\theta')du \\
&= \int \frac{\partial \log f(y,u;\,\theta)}{\partial\theta}\frac{\partial f(u \mid y;\,\theta')}{\partial\theta'}du \\
&= \int \frac{\partial \log f(y,u;\,\theta)}{\partial\theta}\frac{\partial f(u \mid y;\,\theta')}{\partial\theta'}\frac{f(u \mid y;\,\theta')}{f(u \mid y;\,\theta')}du \\
&= \int \frac{\partial \log f(y,u;\,\theta)}{\partial\theta}\frac{\partial \log f(u \mid y;\,\theta')}{\partial\theta'} f(u \mid y;\,\theta')du \\
&= \int \frac{\partial \log f(y,u;\,\theta)}{\partial\theta}\left\{\frac{\partial \log f(y,u;\,\theta')}{\partial\theta'} - \frac{\partial \log f(y;\,\theta')}{\partial\theta'}\right\} f(u \mid y;\,\theta')du \\
&= A - B,
\end{aligned}$$

with, at $\theta = \theta'$,

$$A = \int \left\{\frac{\partial \log f(y,u;\,\theta)}{\partial\theta}\right\}^2 f(u \mid y;\,\theta\}du$$

and

$$B = \frac{\partial \log f(y;\,\theta)}{\partial\theta}\int \frac{\partial \log f(y,u;\,\theta)}{\partial\theta} f(u \mid y;\,\theta)du.$$

At $\theta = \hat{\theta}$, $\frac{\partial \log f(y;\,\theta)}{\partial\theta} = 0$ so that $B = 0$. Furthermore,

$$\begin{aligned}
\int \frac{\partial \log f(y,u;\,\theta)}{\partial\theta} f(u \mid y;\,\theta)du &= \frac{\partial}{\partial\theta}\int \log f(y,u;\,\theta) f(u \mid y;\,\theta)du \\
&= \frac{\partial}{\partial\theta} Q(\theta,\,\theta')\Big|_{\theta'=\theta}, \\
&= 0,
\end{aligned}$$

at any M step so that

$$A = \mathrm{var}\left(\frac{\partial \log f(y,U;\,\theta)}{\partial\theta}\right),$$

where the variance is taken with respect to $f(u \mid y;\,\theta)$.

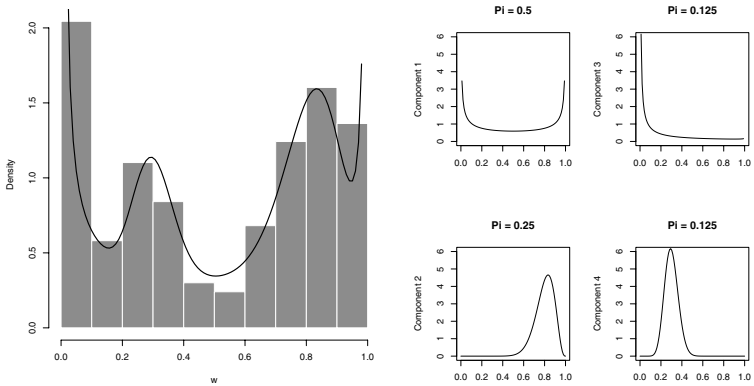


Figure A.1: Histogram of 500 data from the extremal mixture model.

A.4 Datasets

A.4.1 Simulated from the extremal mixture model

This dataset of length 500 is simulated from the extremal mixture model. The number of components is $k = 4$ and the parameters are

$$\pi = \begin{pmatrix} 0.5 \\ 0.25 \\ 0.125 \\ 0.125 \end{pmatrix}, \quad \mu = \begin{pmatrix} 0.5 & 0.5 \\ 0.8 & 0.2 \\ 0.1 & 0.9 \\ 0.3 & 0.7 \end{pmatrix}, \quad \nu = \begin{pmatrix} 0.9 \\ 20 \\ 1 \\ 50 \end{pmatrix}.$$

They have been chosen in order to obtain a multimodal shape. Figure A.1 shows the density function superimposed on a histogram of simulated data and the four components of the mixture.

A.4.2 Simulated from distributions A, B, C and D

Below, comparisons are done componentwise. In Heffernan & Tawn (2004), two kinds of events are considered:

1) simultaneously extremal event. For a fixed p_v the return level is v such that $p_v = P(Y > v)$;

2) unilaterally extremal event. For fixed p, q the return level is v such that $P(Y_1 > r,\, Y_2 < v) = p$ where r is such that $P(Y_1 > r) = p/q$;

Probabilities and return levels associated with these events are considered for four distributions A, B, C and D, given on the Gumbel scale because plots are more readable.

Distribution A: the symmetric logistic distribution with parameter $\alpha = 0.5$;

$$P(Y \leq y) = \exp\left[-V\{\exp(y)\}\right],$$

where

$$V(x) = \left(x_1^{-1/\alpha} + x_2^{-1/\alpha}\right)^{\alpha}.$$

Distribution B: the asymmetric logistic distribution with parameters $\theta_{1,\{1\}} = 0.1 = 1 - \theta_{1,\{1,2\}}$, $\theta_{2,\{2\}} = 0.75 = 1 - \theta_{2,\{1,2\}}$ and $\alpha_{\{1,2\}} = 0.2$;

$$P(Y \leq y) = \exp\left[-V\{\exp(y)\}\right],$$

where

$$V(x) = \frac{\theta_{1,\{1\}}}{x_1} + \frac{\theta_{2,\{2\}}}{x_2} + \left[\left\{\frac{\theta_{1,\{1,2\}}}{x_1}\right\}^{1/\alpha_{\{1,2\}}} + \left\{\frac{\theta_{2,\{1,2\}}}{x_2}\right\}^{1/\alpha_{\{1,2\}}}\right]^{\alpha_{\{1,2\}}}.$$

Distribution C: the inverted extreme value distribution with symmetric logistic dependence structure, with parameter $\eta = 0.75$;

$$P(Y > y) = \exp\left[-V\left\{-1/\log\left(1 - e^{-e^{-y}}\right)\right\}\right],$$

where $V(x)$ is as for distribution A with α such that $2^{-\alpha} = \eta$.

Distribution D: the bivariate normal distribution with parameter $\eta = 0.75$;

$$Y = -\log[-\log\{\Phi(Z)\}],$$

where Φ is the standard normal distribution function and Z is distributed according to a bivariate normal distribution with mean zero and unit variance and correlation $\rho = 2\eta - 1$.

On the Fréchet scale, for the bivariate extreme value distributions and large v,

$$P\left(Y > v\right) = 1 - 2\exp(-1/v) + \exp\left\{-V(v,v)\right\} \approx 2/v - V(v,v). \tag{A.1}$$

This allows us to find explicit true v for distributions A and B in the case of simultaneously extremal events. For distribution C it is trivial and for distribution D numerical procedures have to be used.

Distribution	p		
	10^{-4}	10^{-6}	10^{-8}
A	8.7	*13.3*	17.9
B	7.8	12.4	17.0
C	6.5	9.7	13.2
D	6.9	10.2	13.8

Distribution	q	p		
		10^{-4}	10^{-6}	10^{-8}
A	0.2	6.7	11.3	15.9
	0.5	8.8	13.4	18.0
	0.8	10.5	15.1	19.7
B	0.2	6.2	10.8	15.4
	0.5	7.8	12.4	17.0
	0.8	*9.2*	13.8	18.4

Table A.1: True return levels. Left table for simultaneously extremal events. Right table for unilaterally extremal events. Doubtful values indicated in italics.

Formula (A.1) extends to dimension d using

$$P(Y > v) = 1 - \sum_{i=1}^{d} P(Y_i \leq v_i) + \sum_{i_1 \neq i_2}^{d} P(Y_{i_1} \leq v_{i_1}, Y_{i_2} \leq v_{i_2}) - \cdots + (-1)^d P(Y \leq v),$$

which gives, for $v_1 = \cdots = v_d = v$ and an exchangeable model V,

$$\begin{aligned} P(Y > v) &= 1 - \binom{d}{1} P(Y_1 \leq v) + \binom{d}{2} P(Y_1 \leq v, Y_2 \leq v) - \cdots + (-1)^d P(Y \leq v) \\ &= \binom{d}{1}(1 - e^{-1/v}) + \binom{d}{2}(1 - e^{-V_2(v)}) - \cdots + (-1)^d (1 - e^{-V_d(v)}) \\ &\approx \binom{d}{1} v^{-1} - \binom{d}{2} V_2(v) + \cdots - (-1)^d V_d(v), \end{aligned}$$

where

$$V_j(v) = V(\underbrace{v, \ldots, v}_{j \text{ times}}, \infty, \ldots, \infty), \quad j = 1, \ldots, d.$$

This gives for $d = 5$ and distribution A, $v_p = 8.3, 12.9, 17.5$ for $p = 10^{-4}, 10^{-6}, 10^{-8}$, respectively. The more complex formula for non-exchangeable models like distribution B can be derived from similar arguments. It is not used in this work.

For non-simultaneous extremal events, $p/q = P(Y_1 > r) = \exp(-1/r)$ implies that $r = -1/\log(p/q)$. Then a numerical procedure must be applied in order to find the v's. Some of these are indicated in Heffernan & Tawn (2004) and we computed the others. Although we could check this numerical procedure with indicated values, we also observed lots of numerical instability so that those values should be used with some caution. It turned out in a communication with Dr Janet Heffernan that Heffernan & Tawn (2004) used the `Splus` function `uniroot` in order to find these values, while ours were obtained from the `R` function `optimize`. Results are given in Table A.1. In italics, those return

levels that we found different from Heffernan & Tawn (2004). By default and for the consistency of this presentation, we indicated our results.

For logistic type distributions A and B, simulation can be done from the `R` package `evd` (Stephenson 2003), while simulation from the multivariate normal is standard. Simulation from distribution C is obtained from distribution A with the tail inverted, that is taking $1 - u$ instead of u on the uniform scale.

A.4.3 Newlyn data

The data are a sequence from 1971 to 1977 of hourly surge records for the port of Newlyn, Cornwall, and 3-hourly wave records from a ship. A primary analysis has been done in order to obtain approximately independent vectors. The final series is of length 2894. The data are used assuming temporal independence. The three components are X_1, the inshore significant wave height, X_2, the significant wave period, and X_3, the surge. A design failure region is given by

$$Q(v, X) \geq 0.002,$$

with

$$Q(v, X) = a_1 X_1^* X_2^* \exp\left\{-a_2(v - X_3 - l)/(X_2^* \sqrt{X_1^*})\right\},$$

where $l = 4.3$m is the tidal level relative to the seabed, $a_1 = 0.25$ and $a_2 = 26.0$ are design coefficients, and X_1^* and X_2^* are the offshore significant wave height and wave period, respectively. Relationships for unobserved offshore data from observed inshore data are given by

$$X_1^* = X_1 \exp\left[1 - \exp\left\{-(l + X_3)^2/(2X_1^2)\right\}\right]^{1/2}, \quad X_2^* = X_2.$$

Their extremal dependence structure has been studied in Coles & Tawn (1994), Bortot et al. (2000) and Coles & Pauli (2002). The first article assumed extremal dependence and fitted an extremal Dirichlet spectral measure to the trivariate series. The result was compared to a structural approach in order to infer on an extremal event concerning a functional of the data. The structural approach consists in computing the functional series $Q(v, X)$ and studying its extremal structure with univariate techniques. From the viewpoint of return levels, the two methods turn out to be inconsistent for these data. In particular, probabilities of extremal events turn out to be higher under the multivariate model. The final conclusions of Coles & Tawn (1994) are that the multivariate approach

should be preferred since it is not conditioned on a particular functional of interest so that it takes into account the whole dependence structure. Bortot et al. (2000) raised doubts about these conclusions because the multivariate extremal model did not take into account potential asymptotic independence of the data. The Newlyn data were re-examined using a trivariate Gaussian tail model. The resulting return levels were lower than with an extremal Dirichlet model, thus reconciling the structural and multivariate approaches. Indeed, the return levels of the structural approach were in the confidence band around those of the Gaussian tail model except for very low probabilities, in which case the confidence intervals of the two methods still overlap. Coles & Pauli (2002) develop a model linking asymptotic dependence and asymptotic independence. The conclusions for the data are that the triple surge-wave-height exhibits asymptotic dependence while each pair involving the period exhibits asymptotic independence.

A.4.4 Air quality data

The air quality data set consists of a daily series of monitoring measurements of levels of ozone (O_3), nitrogen dioxide (NO_2), nitrogen oxides (NO) and particulate matter (PM_{10}), in Leeds city center, UK, from 1994 to 1998 inclusive. They are extracted from Heffernan & Tawn (2004) and can be downloaded from `http://www.airquality.co.uk`. Gases are recorded in parts per billion, *ppb*, and PM_{10} in μgm^{-3}. We concentrate on winter data, from November to February, inclusive. Missing values (`NA`'s) are deleted and stationarity is assumed since this work does not focus on temporal dependence.

A.5 Reversible jump Markov chain Monte Carlo algorithm

This section details prior and proposal densities and acceptance probability calculations.

A.5.1 Prior distributions

The prior distribution for k is Poisson with parameter `hyppark`, fixed by the user. The prior distribution of $\log \nu$, given k, is normal with mean and variance specified by the user into the vector `hypparlnu`. The prior distribution of π, given k, is a Dirichlet distribution in S_k with common parameters $\delta_1 = \cdots = \delta_k$, specified by the user with `hypparPi`. The default value `hypparPi = 1` leads to a uniform prior.

The prior distribution of μ, given π and k, is constructed conditionally. We write

$$\mu = \begin{bmatrix} \mu_1^{(1)} & \cdots & \mu_{d-1}^{(1)} \\ \vdots & \ddots & \vdots \\ \mu_1^{(k-1)} & \cdots & \mu_{d-1}^{(k-1)} \end{bmatrix},$$

then, writing by row, the density

$$f(\mu) = f(\mu_1^{(1)}, \ldots, \mu_{d-1}^{(1)}, \mu_1^{(2)}, \ldots, \mu_{d-1}^{(2)}, \ldots, \mu_{d-1}^{(k-1)})$$

is the product of successive conditionals, starting from the right,

$$\begin{gathered} f_{d-1}^{(k-1)} = f(\mu_{d-1}^{(k-1)} \mid \mu_1^{(1)}, \ldots, \mu_{d-1}^{(1)}, \mu_1^{(2)}, \ldots, \mu_{d-1}^{(2)}, \ldots, \mu_{d-2}^{(k-1)}) \\ f_{d-2}^{(k-1)} = f(\mu_{d-2}^{(k-1)} \mid \mu_1^{(1)}, \ldots, \mu_{d-1}^{(1)}, \mu_1^{(2)}, \ldots, \mu_{d-1}^{(2)}, \ldots, \mu_{d-3}^{(k-1)}) \\ \vdots \\ f_2^{(1)} = f(\mu_2^{(1)} \mid \mu_1^{(1)}) \\ f_1^{(1)} = f(\mu_1^{(1)}). \end{gathered}$$

Those conditionals are assumed to be uniform on the largest interval that allows constraints to be satisfied, that is

$$I_i^{(m)} = \left[0 \,,\, \min \left(1 - \sum_{j=1}^{i-1} \mu_j^{(m)} \,,\, \frac{c_i - \sum_{l=1}^{m-1} \pi_l \mu_i^{(l)}}{\pi_m} \right) \right], \quad i = 1, \ldots, d-1, \quad j = 1, \ldots, k-1,$$

where c_i is the constraint constant

$$\sum_{m=1}^{k} \pi_m \mu_i^{(m)} = c_i, \quad i = 1, \ldots, d,$$

and we recall that in the multivariate case, $c_1 = \cdots = c_d = d^{-1}$. Now the prior density on μ is the inverse of the product of the lengths of all $I_i^{(m)}$. There are no hyperparameters. A possible extra complication would be to impose a beta distribution instead of the uniform, but we have not found this useful.

A.5.2 Proposals and acceptance probability for dependence parameters

'SPLIT/COMBINE' move type

The 'SPLIT' move for component $(\pi_0 \,,\, \mu_0)$ can be summarized as

$$\begin{pmatrix} \pi_0 \\ \mu_0 \end{pmatrix} \longmapsto \begin{pmatrix} \pi_0 & v \\ \mu_0 & \mu_2 \end{pmatrix} \longmapsto \begin{pmatrix} \pi_1 = v\pi_0 & \pi_2 = (1-v)\pi_0 \\ \mu_1 = \dfrac{\mu_0 - (1-v)\mu_2}{v} & \mu_2 = \mu_2 \end{pmatrix},$$

where v is some random variable in $(0,1)$ and μ_2 in S_d. The Jacobian of this transformation is the determinant

$$\frac{\partial(\pi_1,\pi_2,\mu_1,\mu_2)}{\partial(\pi_0,v,\mu_0,\mu_2)} = \begin{array}{c|cccc} & \pi_1 & \pi_2 & \mu_1 & \mu_2 \\ \hline \pi_0 & v & 1-v & 0^T & 0^T \\ v & \pi_0 & -\pi_0 & \frac{\mu_2-\mu_0}{v} & 0^T \\ \mu_0 & 0 & 0 & \frac{1}{v}I_{d-1} & 0^T 0 \\ \mu_2 & 0 & 0 & -\frac{1-v}{v}I_{d-1} & I_{d-1} \end{array} \tag{A.2}$$

where 0 is the column vector of length $d-1$, 0^T its transpose and I_{d-1} is the identity matrix of side $d-1$. Therefore, the Jacobian is π_0/v^{d-1}.

The proposal from v, $q(v)$, depends on the jump size and on tuning parameters set by the user. The proposal for μ_2 depends on the jump size, on tuning parameters and on the current μ_0; we write $q(\mu_2 \mid \mu_0)$. The proposal ratio contribution for the 'SPLIT' of (π_0, μ_0) is then

$$\frac{1}{q(\mu_2 \mid \mu_0)q(v)}\frac{v^{d-1}}{\pi_0}. \tag{A.3}$$

For $\log\nu_0$, there is no change of variable. The two components $\log\nu_1$ and $\log\nu_2$ are proposed according to a normal distribution with mean $\log\nu_0$ and the merged component $\log\nu_0$ is proposed according to a normal distribution with mean $(\log\nu_1 + \log\nu_2)/2$. In both case, the tuning variance is determined by the user. The proposal ratio contribution for a 'SPLIT' is then

$$\frac{q(\log\nu_0 \mid \log\nu_1, \log\nu_2)}{q(\log\nu_1 \mid \log\nu_0)q(\log\nu_2 \mid \log\nu_0)}. \tag{A.4}$$

where q is the appropriate normal density.

Finally, the choice of the of component index l among $1,\ldots,k$ to be split is random, as one of the two component indexes (l_1, l_2) to be combined. The proposal ratio contribution of this choice for a 'SPLIT' is hence

$$\frac{2k}{(k+1)k}. \tag{A.5}$$

Overall, the product of (A.3), (A.4) and (A.5) gives the ratio of the proposals for a 'SPLIT'. For a 'COMBINE', it is the inverse. In Section 2.3.2, the ratio of proposals is multiplied to the ratio of posterior to calculate the acceptance probability.

'MCMC' move

The 'MCMC' move is done in two steps. The first step updates $\log\nu_1,\ldots,\log\nu_k$ according to a normal random walk with tuning variance depending on the size of the jump, fixed

by the user. The proposal ratio contribution is then 1 because of the symmetry of the normal density. The second step fixes those parameters and updates π and μ with a succession of 'COMBINE' and 'SPLIT' moves

$$\begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix} \longmapsto \pi_0 \longmapsto \begin{pmatrix} \pi_1' \\ \pi_2' \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \longmapsto \mu_0 \longmapsto \begin{pmatrix} \mu_1' \\ \mu_2' \end{pmatrix}.$$

For a generic parameter x, the move goes from x to x_0 then to x' and the backward move goes from x' to x_0 then to x. The proposal ratio contribution is therefore

$$\frac{q(x_0 \mid x')q(x \mid x_0)}{q(x_0 \mid x)q(x' \mid x_0)} = \frac{q(x_0 \mid x')}{q(x' \mid x_0)} \frac{q(x \mid x_0)}{q(x_0 \mid x)}.$$

In other words, it is the product of the ratio contribution of the two moves. They are hence readily obtained from (A.3), (A.4) and (A.5).

A.5.3 Margin parameters

When the reversible jump algorithm is used to estimate dependence and marginal structure together, then the algorithm randomly alternates two kernels, one, concerning the margins, and, the other, the dependence structure. The dependence kernel is the reversible jump part of the algorithm. The margin kernel is a standard Metropolis–Hastings kernel. For the semi-parametric extremal model of each margin, new parameters are proposed according to a proposal density, the acceptance probability ratio is computed, and then the decision to keep or reject the proposed values taken.

In this acceptance probability ratio appear proposal densities, prior densities and likelihoods. The calculation of proposal and prior densities are straightforward if properly chosen. The calculation of likelihoods is detailed in Coles & Tawn (1991) on a set $A_0 = \mathbb{R}^p \setminus \{(0, \nu_1) \times \ldots \times (0, \nu_d)\}$, where ν_i is the threshold for the i-th margin. Borrowing their notation, the likelihood in A is

$$\exp\{V(\nu)\} \prod_{i=1}^{n_A} \left(h(w_i)(nr_i)^{-(d+1)} \prod_{\substack{j=1,\ldots,d: \\ X_{i,j} > n\nu_j}} [\sigma_j^{-1} p_j^{-\kappa_j} X_{i,j}^2 \exp(1/X_{i,j})\{1 - \exp(-1/X_{i,j})\}^{1+\kappa_j}] \right),$$

where:

- $V(\nu)$ is the rate measure of the extremal Poisson model evaluated on the set $\nu = (\nu_1, \infty) \times \ldots \times (\nu_d, \infty)$;
- p_j is the probability that component j is above ν_j;

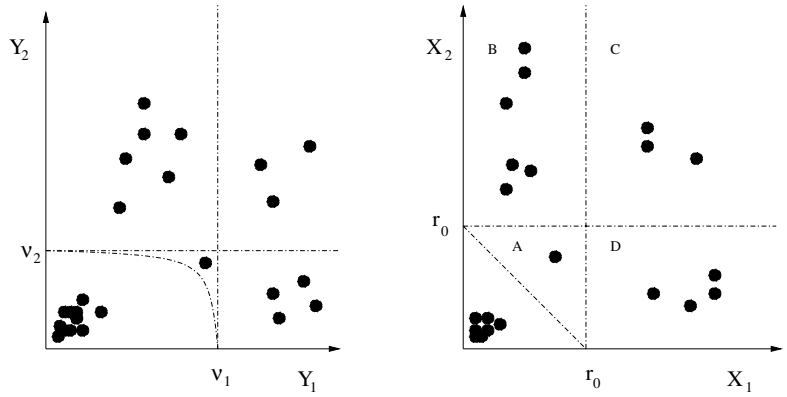


Figure A.2: The effect of the semi-parametric extremal model. On the left, the original data, on the right, data on the Fréchet scale.

- the X_i are Fréchet data obtained with the semi-parametric-extremal model, $X_{i,j}$ being the j-th component of X_i, for $j = 1, \ldots, p$ and $i = 1, \ldots, n_0$;
- w_i are the pseudo-angles of X_i and r_i is its norm;
- κ_j corresponds to $-k_j$ in Coles & Tawn (1991).

This likelihood is the Poisson likelihood of data in A: the product with the exponential and the $h(w_i)$ is the likelihood of the observed data on the pseudo-polar scale while the part corresponding to the right product is the Jacobian of the transformation of the data to the Fréchet scale with the semi-parametric extremal model.

In our case, the set A_0 is not $\mathbb{R}^p \setminus \{(0, \nu_1) \times \ldots \times (0, \nu_p)\}$ but $\{r > r_0\}$ for a selected multivariate threshold, r_0. In this case, the likelihood slightly changes because the leftmost exponential term is constant with respect to h. Furthermore, data in the set $\{r > r_0\} \setminus \nu$ are not touched by a change of margin parameters since each of their components is under the corresponding marginal threshold. Figure A.2 illustrates this in dimension $d = 2$. On the left panel, the original data Y and on the right panel, data on the Fréchet scale, X. The marginal thresholds are the backward image of the multivariate threshold r_0. The extremal Poisson model applies in sets A, B, C and D. New margin parameters for the first components influence Fréchet data in sets C and D, new margin parameters for the second components influence Fréchet data in sets C and B. Hence the global likelihood of A, B, C and D incorporates effects of data in C for both margin components and

for the dependence structure, in D for component 1 and the dependence structure, in B for component 2 and the dependence structure, and in A for the dependence structure alone. Using the likelihood on the set $\mathbb{R}^p \setminus \{(0,\, \nu_1) \times \ldots \times (0,\, \nu_p)\}$ deprives us of data in A, which brings information on the dependence structure, and forces us to calculate the measure of B, C and D, which is difficult, while the measure of A, B, C and D together is straightforward because only likelihood ratios are needed. Naturally, those comments are valid whatever the dependence structure, whether it is from one model or another.

We now discuss the choice of the prior and the proposal density, that were made for convenience. The $\log \sigma$ is a priori normal with mean zero and a large variance. The new $\log \sigma$ is proposed according to a normal distribution with mean the current $\log \sigma$ and variance defined according to the size of the jump. The generalized Pareto distribution domain,

$$1 + \sigma^{-1} \kappa y > 0,$$

implies that the proposition of a new κ must be done properly if one does not want to see every new proposition systematically rejected. For component j, one must have

$$\chi_j = \sigma_j^{-1} \kappa_j + 1/\max\{y_{i,j}\} > 0,$$

where the maximum is taken over the excesses above margin threshold u_j. Thus, new $\log \chi_j$ are proposed according to a normal random walk and one sets

$$\kappa_j = \sigma_j \left(\chi_j - 1/\max\{y_{i,j}\}\right).$$

We now calculate the proposal density. For notational convenience, we fix j, write $m = -1/\max\{y_{i,j}\}$,

$$\begin{array}{ll} x = \log(\sigma) = u, & u = \log(\sigma) = x \\ y = \kappa = e^u \left(e^v + m\right), & v = \log(\chi) = \log\left(y e^{-x} - m\right). \end{array} \quad \text{and}$$

The differentials are $dx = du$ and $dy = e^u \left(e^v + m\right) du + e^u e^v dv$ so that

$$dxdy = e^u e^v dudv \quad \text{and} \quad dudv = e^{-x} (y e^{-x} - m)^{-1} dxdy.$$

Therefore, the proposal density is

$$\begin{aligned} q(u, v \mid u', v')dudv &= (2\pi\nu_u\nu_v)^{-1} e^{-(u-u')^2/2\nu_u^2} e^{-(v-v')^2/2\nu_v^2} dudv \\ &= (2\pi\nu_u\nu_v)^{-1} e^{-(x-x')^2/2\nu_u^2} e^{-(\log(ye^{-x}-m)-\log(y'e^{-x'}-m))^2/2\nu_v^2} \frac{dxdy}{e^x\left(ye^{-x}-m\right)}, \end{aligned}$$

and the proposal density ratio is

$$\frac{q(x, y \mid x', y')}{q(x', y' \mid x, y)} = \frac{e^{x'}\left(y'e^{-x'} - m\right)}{e^{x}\left(ye^{-x} - m\right)} = \frac{\sigma'\chi'}{\sigma\chi},$$

where $'$ indicates a new proposed value.

Bibliography

Ancona-Navarrete, M. A. & Tawn, J. A. (2002), 'Diagnostics for pairwise extremal dependence in spatial processes', *Extremes* **5**(3), 271–285.

Antoniak, C. E. (1974), 'Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems', *Annals of Statistics* **2**(6), 1152–1174.

Bortot, P., Coles, S. G. & Tawn, J. A. (2000), 'The multivariate Gaussian tail model: An application to oceanographic data', *Applied Statistics* **49**, 31–49.

Brooks, S. P. & Giudici, P. (2000), 'Markov chain Monte Carlo convergence assessment via two-way analysis of variance', *Journal of Computational and Graphical Statistics* **9**(2), 266–285.

Bruun, J. T. & Tawn, J. A. (1998), 'Comparison of approaches for estimating the probability of coastal flooding', *Journal of the Royal Statistical Society Series C* **47**(3), 405–423.

Burnham, K. P. & Anderson, D. R. (2002), *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*, 2nd edn, Springer-Verlag, New-York.

Casson, E. & Coles, S. G. (1999), 'Spatial regression models for extremes', *Extremes* **1**(4), 449–468.

Coles, S. G. (1993), 'Regional modelling of extreme storms via max-stable processes', *Journal of the Royal Statistical Society series B* **55**(4), 797–816.

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

Coles, S. G., Heffernan, J. & Tawn, J. A. (1999), 'Dependence measures for extreme value analyses', *Extremes* **2**, 339–365.

Coles, S. G. & Pauli, F. (2002), 'Models and inference for uncertainty in extremal dependence', *Biometrika* **89**(1), 183–196.

Coles, S. G. & Powell, E. A. (1996), 'Bayesian methods in extreme value modelling: A review and new developments', *International Statistical Review* **64**(1), 119–136.

Coles, S. G. & Tawn, J. A. (1991), 'Modelling extreme multivariate events', *Journal of the Royal Statistical Society series B* **91**(2), 377–392.

Coles, S. G. & Tawn, J. A. (1994), 'Statistical methods for multivariate extremes: an application to structural design (with discussion)', *Applied Statistics* **43**(1), 1–48.

Coles, S. G. & Tawn, J. A. (1996), 'Modelling extremes of the areal rainfall process', *Journal of the Royal Statistical Society series B* **58**(2), 329–347.

Coles, S. G. & Walshaw, D. (1994), 'Directional modelling of extreme wind speed', *Applied Statistics* **43**(1), 139–157.

Dalal, S. R. (1978), 'A note on the adequacy of mixtures of Dirichlet processes', *Sankhyā* **40**, 185–191.

Dalal, S. R. & Hall, G. J. (1980), 'On approximating parametric Bayes models by nonparametric Bayes models', *Annals of Statistics* **8**(3), 664–672.

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

Davison, A. C. & Smith, R. L. (1990), 'Models for exceedances over high thresholds (with discussion)', *Journal of the Royal Statistical Society series B* **52**(3), 393–442.

de Haan, L. (1984), 'A spectral representation for max-stable processes', *Annals of Probability* **12**(4), 1194–1204.

de Haan, L. & de Ronde, J. (1998), 'Sea and wind: Multivariate extremes at work', *Extremes* **1**(1), 7–45.

de Haan, L. & Lin, T. (2001), 'On convergence toward an extreme value distribution in $C[0, 1]$', *Annals of Probability* **29**(1), 467–483.

de Haan, L. & Resnick, S. (1977), 'Limit theory for multivariate sample extremes', *Zeitschrift für Wahrscheinlichkeitstheorie verwandte Gebeite* **40**, 317–337.

Deheuvels, P. & Tiago de Oliveira, J. (1989), 'On the non-parametric estimation of the bivariate extreme value distribution', *Statistics and Probability Letters* **8**, 315–323.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm (with discussion)', *Journal of the Royal Statistical Society series B* **39**, 1–38.

Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997), *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, Berlin.

Ferguson, T. S. (1973), 'A Bayesian analysis of some nonparametric problems', *Annals of Statistics* **1**(2), 209–230.

Ferro, C. A. T. & Segers, J. (2003), 'Inference for clusters of extreme values', *Journal of the Royal Statistical Society series B* **62**(2), 545–556.

Finkelstein, B. V. (1953), 'On the limiting distributions of the extreme terms of a variational series of a two dimensional random quantity', *Dokllady Akademii SSSR* **91**, 209–211. (in Russian).

Finkenstädt, B. & Rootzén, H., eds (2004), *Extreme Values in Finance, Telecommunications and the Environment*, Chapman and Hall, Boca Raton.

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

Fréchet, M. (1927), 'Sur la loi de probabilité de l'écart maximum', *Annales de la Société Polonaise de Mathématiques de Cracovie* **6**, 93–116.

Galambos, J. (1978), *The Asymptotic Theory of Extreme Order Statistics*, Wiley, Malabar.

Geoffroy, J. (1958/1959), 'Contribution à la théorie des valeurs extrêmes', *Publ. Inst. Statist. Univ. Paris* **7/8**, 37–185.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J., eds (1996), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London.

Giné, E., Hahn, M. G. & Vatan, P. (1990), 'Max-infinitely divisible and max-stable sample continuous processes', *Probability Theory and Related Fields* **87**, 139–165.

Gnedenko, B. (1943), 'Sur la distribution limite du terme maximum d'une série aléatoire', *Annals of Mathematics* **44**, 423–453. Translated and reprinted in: *Breakthroughs in Statistics*, Vol. **I**, 1992, eds. S. Kotz and N. L. Johnson, Springer-Verlag, pp. 195–225.

Green, P. J. (1995), 'Reversible jump Markov chain Monte Carlo computation and Bayesian model determination', *Biometrika* **82**(4), 711–732.

Gumbel, E. J. (1958), *Statistics of Extremes*, Columbia University Press.

Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**, 97–109.

Heffernan, J. E. & Tawn, J. A. (2004), 'A conditional approach for multivariate extreme values (with discussion)', *Journal of the Royal Statistical Society Series B* **66**(3), 497–546.

Hsing, T. (1989), 'Extreme value theory for multivariate stationary sequences', *Journal of Multivariate Analysis* **29**, 274–291.

Hsing, T., Hüsler, J. & Leadbetter, M. R. (1988), 'On the exceedance point process for a stationary sequence', *Probability Theory and Related Fields* **78**, 97–112.

Hurvich, C. M. & Tsai, C.-L. (1989), 'Regression and time series model selection in small samples', *Biometrika* **76**, 297–307.

Jagers, P. (1972), 'On the weak convergence of superpositions of point processes', *Zeitschrift für Wahrscheinlichkeitstheorie verwandte Gebeite* **22**, 1–7.

Jagers, P. (1974), Aspect of random measures and point processes, *in* P. Ney & S. Port, eds, 'Advances in Probability and Related Topics', Vol. 3, Marcel Dekker, New York, pp. 179–239.

Joe, H., Smith, R. L. & Weissman, I. (1992), 'Bivariate threshold methods for extremes', *Journal of the Royal Statistical Society Series B* **54**, 171–183.

Kallenberg, O. (1983), *Random Measures*, Akademie-Verlag, Berlin.

Kotz, S. & Nadarajah, S. (2000), *Extreme Value Distributions: Theory and Applications*, Imperial College Press, London.

Leadbetter, M. R. (1974), 'On extreme values in stationary sequences', *Zeitschrift für Wahrscheinlichkeitstheorie und verswante Gebiete* **28**, 289–303.

Leadbetter, M. R. (1983), 'Extremes and local dependence of stationary sequences', *Zeitschrift für Wahrscheinlichkeitstheorie und verswante Gebiete* **65**, 291–306.

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

Ledford, A. W. & Tawn, J. A. (1996), 'Statistics for near independence in multivariate extreme values', *Biometrika* **83**, 169–187.

Ledford, A. W. & Tawn, J. A. (1997), 'Modelling dependence within joint tail regions', *Journal of the Royal Statistical Society Series B* **59**(2), 475–499.

Lindsay, B. G. (1995), *Mixture Models: Theory, Geometry and Applications*, number 5 *in* 'NSF-CBMS Regional Conference Series in Probability and Statistics', Institute of Mathematical Statistics, Hayward.

McLachlan, G. J. & Krishnan, T. (1997), *The EM Algorithm and Extensions*, Wiley, New York.

McLachlan, G. J. & Peel, D. (2000), *Finite Mixture Models*, Wiley, New York.

McQuarrie, A. D. R. & Tsai, C.-L. (1998), *Regression and Time Series Model Selection*, Singapore: World Scientific.

Meng, X.-L. & Pedlow, S. (1992), EM: A bibliographic review with missing articles, *in* 'Proceedings of the Statistical Computing Section, American Statistical Association', American Statistical Association, Alexandria, Virginia, pp. 24–27.

Meng, X.-L. & van Dyk, D. (1997), 'The EM algorithm – an old folk-song sung to a fast new tune (with discussion)', *Journal of the Royal Statistical Society Series B* **59**, 511–567.

Oakes, D. (1999), 'Direct calculation of the information matrix via the EM algorithm', *Journal of the Royal Statistical Society Series B* **61**(2), 479–482.

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

Peng, L. (1999), 'Estimation of the coefficient of tail depedence in bivariate extremes', *Statistics and Probability Letters* **43**, 399–409.

Pickands, J. (1971), 'The two-dimensional Poisson process and extremal processes', *Journal of Applied Probability* **8**, 745–756.

Pickands, J. (1975), 'Statistical inference using extreme order statistics', *Annals of Statistics* **3**, 119–131.

Pickands, J. (1981), Multivariate extreme value distributions, *in* 'Proceedings of the 43rd Session of the I.S.I.', International Statistical Institute, The Hague, pp. 859–878.

Redner, R. A. & Walker, H. F. (1984), 'Mixture densities, maximum likelihood and the EM algorithm', *SIAM Review* **26**, 195–239.

Reiss, R.-D. & Thomas, M. (2001), *Statistical Analysis of Extreme Values with Applications to Insurance Finance, Hydrology and Other Fields*, 2nd edn, Birkhaüser, Basel.

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

Richardson, S. & Green, P. J. (1997), 'On Bayesian analysis of mixtures with an unknown number of components (with discussions)', *Journal of the Royal Statistical Society Series B* **59**, 731–792.

Sibuya, M. (1960), 'Bivariate extreme statistics', *Ann. Inst. Stat. Math. Tokyo* **11**, 195–210.

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

Stephenson, A. (2003), 'Simulating multivariate extreme value distributions of logistic type', *Extremes* **6**, 49–59.

Tanner, M. A. (1996), *Tools for Statistical Inference–Methods for Exploration of Posterior Distributions and Likelihood Functions*, 3rd edn, Springer-Verlag, New York.

Tawn, J. A. (1988), 'Bivariate extreme value theory: Models and estimation', *Biometrika* **75**, 397–415.

Tawn, J. A. (1990), 'Modelling multivariate extreme value distributions', *Biometrika* **77**(2), 245–53.

Tiago de Oliveira, J. (1958), 'Extremal distributions', *Revista da Fac. Ciencias, Univ. Lisboa* **A 7**, 215–227.

Tiago de Oliveira, J., ed. (1984), *Statistical Extremes and Applications*, NATO Advanced Study Institute, Vimeiro.

Titterington, D. M., Smith, A. F. M. & Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York.

von Bortkiewicz, L. (1922), 'Variantionsbreite und mittelerer Fehler', *Sitzungsber. Berli. Math. Ges.* **21**, 3–11.

von Mises, R. (1936), 'La distribution de la plus grande de n valeurs', *Rev. Math. Union Interbalk.* **1**, 141–160. Reproduced in *Selected Papers of Richard von Mises, II* (1954), pp. 271–294, Amer. Math. Soc.

Weibull, W. (1939), 'The phenomenon of rupture in solids', *Ing. Vet. Akad. Handlingar* **153**, 2.

Weissman, I. (1978), 'Estimation of parameters and large quantiles based on the k largest observations', *Journal of the American Statistical Association* **6**, 1641–1648.

Wilks, S. S. (1962), *Mathematical Statistics*, Wiley, New York.

Curriculum Vitae

Né le 7 mars 1977 à Saint-Calais (France), je suis de nationalité française. J'ai effectué ma scolarité au lycée international de Ferney-Voltaire (France) où j'ai obtenu un baccalauréat français de type S scientifique en 1995. A la suite de mes études à l'Ecole Polytechnique Fédérale de Lausanne (Suisse), j'ai obtenu un diplôme d'ingénieur mathématicien en 2000. Depuis lors je suis assistant pour le professeur Anthony C. Davison, directeur de ma thèse de doctorat en statistiques.