A User's Guide to the POT Package (Version 1.1)

Mathieu Ribatet

Department of Hydrological Statistic, INRS,
University of Québec, 490 de la Couronne, G1K 9A9 Canada

Cemagref UR HH, 3bis quai Chauveau
69336 Lyon Cedex 09 France

E-mail: ribatet@hotmail.com
17th October 2006

1 Introduction

1.1 Why the POT package?

The **POT** package is an add-on package for the R statistical software (R Development Core Team, 2006). The main goal of this package is to develop tools to perform stastical analyses of Peaks Over a Threshold (**POT**).

Most of functions are related to the Extreme Value Theory (**EVT**). Coles (2001) gives a comprehensive introduction to the EVT, while Kluppelberg and Mikosch (1997) present advanced results.

1.2 Obtaining the package/guide

The package can be downloaded from CRAN (The Comprehensive R Archive Network) at `http://cran.r-project.org/`. This guide (in pdf) will be in the directory `POT/doc/` underneath wherever the package is installed.

1.3 Contents

To help users to use properly the **POT** package, this guide contains practical examples on the use of this package. Section 2 introduce quickly the Extreme Value Theory (**EVT**). Some basic examples are described in section 3, while section 4 gives a concrete statistical analysis of extreme value for river Adieéres at Beaujeu (FRANCE).

1.4 Citing the package/guide

To cite this guide or the package in publications please use the following bibliographic database entry.

```
@Manual{key,
  title = {A User's Guide to the POT Package (Version 1.0)},
  author = {Ribatet, M. A.},
  year = {2006},
  month = {August},
  url = {http://cran.r-project.org/}
}
```

1.5 Caveat

I have checked these functions as best I can but, as ever, they may contain bugs. If you find a bug or suspected bug in the code or the documentation please report it to me at `ribatet@hotmail.com`. Please include an appropriate subject line.

1.6 Legalese

2 An Introduction the EVT

2.1 The univariate case

Even if this package is only related to peaks over a threshold, a classical introduction to the EVT must deal with "block maxima". Let $X_1, \ldots, X_n$ be a series of independent and identically distributed random variables with commom distribution function F. Let $M_n = \max(X_1, \ldots, X_n)$.

Suppose there exists normalizing constants $a_n > 0$ and b_n such that:

$$\Pr\left[\frac{M_n - b_n}{a_n} \leq y\right] = F^n(a_n y + b_n) \longrightarrow G(y), \qquad n \to +\infty \tag{2.1}$$

for all $y \in \mathbb{R}$, where G is a non-degenerate distribution function. According to the Extremal Types Theorem (Fisher and Tippett, 1928), G must be either Fréchet, Gumbel or negative Weibull. Jenkinson (1955) noted that these three distributions can be merged into a single parametric family: the Generalized Extreme Value (**GEV**) distribution. The GEV has a distribution function defined by:

$$G(y) = \exp\left[-\left(1 + \xi\frac{y-\mu}{\sigma}\right)_+^{-1/\xi}\right], \tag{2.2}$$

where (μ, σ, ξ) are the location, scale and shape parameters respectively, $\sigma > 0$ and $z_+ = \max(z, 0)$.

The Fréchet case is obtained when $\xi > 0$, the negative Weibull when $\xi < 0$ while the Gumbel case is defined by continuity when $\xi \to 0$.

From this result, Pickands (1975) showed that the limiting distribution of normalized excesses of a threshold μ as the threshold approaches the endpoint μ_{end} of the variable of interest is the Generalized Pareto Distribution (**GPD**). That is, if X is a random variable which holds (2.1), then:

$$\Pr\left[X \leq y | X > \mu\right] \longrightarrow H(y), \qquad \mu \to \mu_{\text{end}} \tag{2.3}$$

with

$$H(y) = 1 - \left(1 + \xi \frac{y - \mu}{\sigma}\right)_+^{-1/\xi}, \tag{2.4}$$

where (μ, σ, ξ) are the location, scale and shape parameters respectively, $\sigma > 0$ and $z_+ = \max(z, 0)$. Note that the Exponential distribution is obtained by continuity as $\xi \to 0$.

In practice, these two asymptotical results motivated modelling block maxima with a GEV, while peaks over threshold with a GPD.

2.2 The multivariate case

When dealing with multivariate extremes, it is usual to transform data to a particular distribution. For example, Falk and Reiss (2005) used the inverted standard exponential distribution – $\Pr[Z \leq z] = \exp(z)$, $z \leq 0$, Coles et al. (1999) use the uniform distribution on $[0, 1]$. However, the most common distribution seems to be the standard Fréchet one – $\Pr[Z \leq z] = \exp(-1/z)$ (Smith, 1994; Smith et al., 1997; Bortot and Coles, 2000). Thus, in the following, we will only consider this case. For this purpose, margins are transformed according to:

$$Z_j = -\frac{1}{\log F_j\left(Y_j\right)}$$

where F_j is the distribution of the j-th margin.

Obviously, in practice, the margins F_j are unknown. When dealing with extremes, the univariate EVT tells us what to do. Thus, if block maxima or peaks over a threshold are of interest, we must replace F_j with GEV or GPD respectively.

Definition 2.2.1. A multivariate extreme value distribution in dimension d has representation:

$$G\left(y_1, \ldots, y_d\right) = \exp\left[-V\left(z_1, \ldots, z_d\right)\right] \tag{2.5}$$

with

$$V(z_1, \ldots, z_d) = \int_{T_p} \max_{j=1,\ldots,d} \left(\frac{q_j}{z_j}\right) \mathrm{dH}\left(q_1, \ldots, q_d\right)$$

where H is a measure with mass 2 called *spectral density* defined on the set

$$T_p = \left\{ (q_1, \ldots, q_d) : q_j \geq 0, \sum_{j=1}^{d} q_j^2 = 1 \right\}$$

with the constraint

$$\int_{T_p} q_j dH(q_j) = 1, \qquad \forall j \in \{1, \ldots, d\}$$

The V function is often called *exponential measure* (Klüppelberg and May, 2006) and is an homogeneous function of order -1.

Contrary to the univariate case, there is an infinity of functions V for $d > 1$. Thus, it is usual to used specific parametric families for V. Several examples for these families are given in Annexe A.

Another representation for a multivariate extreme value distribution is the Pickands' representation (Pickands, 1981). We give here only the bivariate case.

Definition 2.2.2. A bivariate extreme value distribution has the Pickands' representation:

$$G(y_1, y_2) = \exp\left[-\left(\frac{1}{z_1} + \frac{1}{z_2}\right) A\left(\frac{z_2}{z_1 + z_2}\right)\right] \tag{2.6}$$

with

$$\begin{aligned} A : [0,1] &\longrightarrow [0,1] \\ w &\longmapsto A(w) = \int_0^1 \max\left\{w(1-q), (1-w)q\right\} dH(q) \end{aligned}$$

In particular, the functions V and A are linked by the relation:

$$A(w) = \frac{V(z_1, z_2)}{z_1^{-1} + z_2^{-1}}, \qquad w = \frac{z_2}{z_1 + z_2}$$

The dependence function A holds:

1. $A(0) = A(1) = 1$;
2. $\max(w, 1-w) \leq A(w) \leq 1$, $\forall w$;
3. A is convex;
4. Two random variables (with unit Fréchet margins) are independent if $A(w) = 1$, $\forall w$;
5. Two random variables (with unit Fréchet margins) are perfectly dependent if $A(w) = \max(w, 1-w)$, $\forall w$.

We define the multivariate extreme value distributions which are identical to the block maxima approach in higher dimensions. We now establish the multivariate theory for peaks over threshold.

According to Resnick (1987, Prop. 5.15), multivariate peaks over thresholds u_j has the same representation than for block maxima. Only the margins F_j must be replaced by GPD instead of GEV. Thus,

$$F(y_1, \ldots, y_d) = \exp\left[-V\left(-\frac{1}{\log F_1(y_1)}, \ldots, -\frac{1}{\log F_d(y_d)}\right)\right], \qquad y_j > u_j \tag{2.7}$$

3 Basic Use

3.1 Random Numbers and Distribution Functions

First of all, lets start with basic stuffs. The **POT** package uses the R convention for random numbers generation and distribution function features.

```
> ##Random number generation
> rgpd(5, loc = 1, scale = 2, shape = -0.2)
[1] 4.672547 2.365295 1.899087 1.577886 2.409450
> ##Varying threshold can be performed also
> rgpd(6, c(1, -5), 2, -0.2)
[1]  2.424368 -3.389774  3.965086 -3.332016  4.707819 -4.985408
> ##The same but with a varying scale parameter
```

```
> rgpd(6, 0, c(2, 3), 0)
[1] 2.9850740 3.1486256 1.0705649 0.7401753 3.1231517 2.3994109
> ##Probability of non exceedence
> pgpd(c(9, 15, 20), 1, 2, 0.25)
[1] 0.9375000 0.9825149 0.9922927
> ##Quantile associated to probability of non exceedence
> qgpd(c(.25, .5, .75), 1, 2, 0)
[1] 1.575364 2.386294 3.772589
> ##Evaluate the density at point...
> dgpd(c(9, 15, 20), 1, 2, 0.25)
[1] 0.015625000 0.003179117 0.001141829
```

Several options can be passed to three of these four functions. In particular:

- for "pgpd", user can specify if non exceedence or exceedence probability should be computed with option `lower.tail = TRUE` or `lower.tail = FALSE` respectively;
- for "qgpd", user can specify if quantile is related to non exceedence or exceedence probability with option `lower.tail = TRUE` or `lower.tail = FALSE` respectively;
- for "dgpd", user can specify if the density or the log-density should be computed with option `log = FALSE` or `log = TRUE` respectively.

3.2 Threshold Selection

The location for the GPD or equivalently the threshold is a particular parameter as must often it is not estimated as the other ones. All methods to define a suitable threshold use the asymptotic approximation defined by equation (2.3). In other words, we select a threshold for which the asymptotic distribution H in equation (2.4) is a good approximation.

The **POT** package has several tools to define a reasonable threshold. For this purpose, the user must use `tcplot, mrlplot, lmomplot, exiplot` and `diplot` functions.

The main goal of threshold selection is to selects enough events to reduce the variance; but not too much as we could select events coming from the central part of the distribution[1] and induce bias.

3.2.1 Threshold Choice plot: *tcplot*

Let $X \sim GP(\mu_0, \sigma_0, \xi_0)$. Let μ_1 be a another threshold as $\mu_1 > \mu_0$. The random variable $X|X > \mu_1$ is also GPD with updated parameters $\sigma_1 = \sigma_0 + \xi_0(\mu_1 - \mu_0)$ and $\xi_1 = \xi_0$. Let

$$\sigma_* = \sigma_1 - \xi_1 \mu_1 \tag{3.1}$$

With this new parametrization, σ_* is independent of μ_1. Thus, estimates of σ_* and ξ_1 are constant for all $\mu_1 > \mu_0$ if μ_0 is a suitable threshold for the asymptotic approximation.

Threshold choice plots represent the points defined by:

$$\{(\mu_1, \sigma_*) : \mu_1 \leq x_{\max}\} \quad \text{and} \quad \{(\mu_1, \xi_1) : \mu_1 \leq x_{\max}\} \tag{3.2}$$

where $x_{\max}$ is the maximum of the observations $\mathbf{x}$.

Moreover, confidence intervals can be computed using Fisher information.

Here is an application.

[1] i.e. not extreme events.

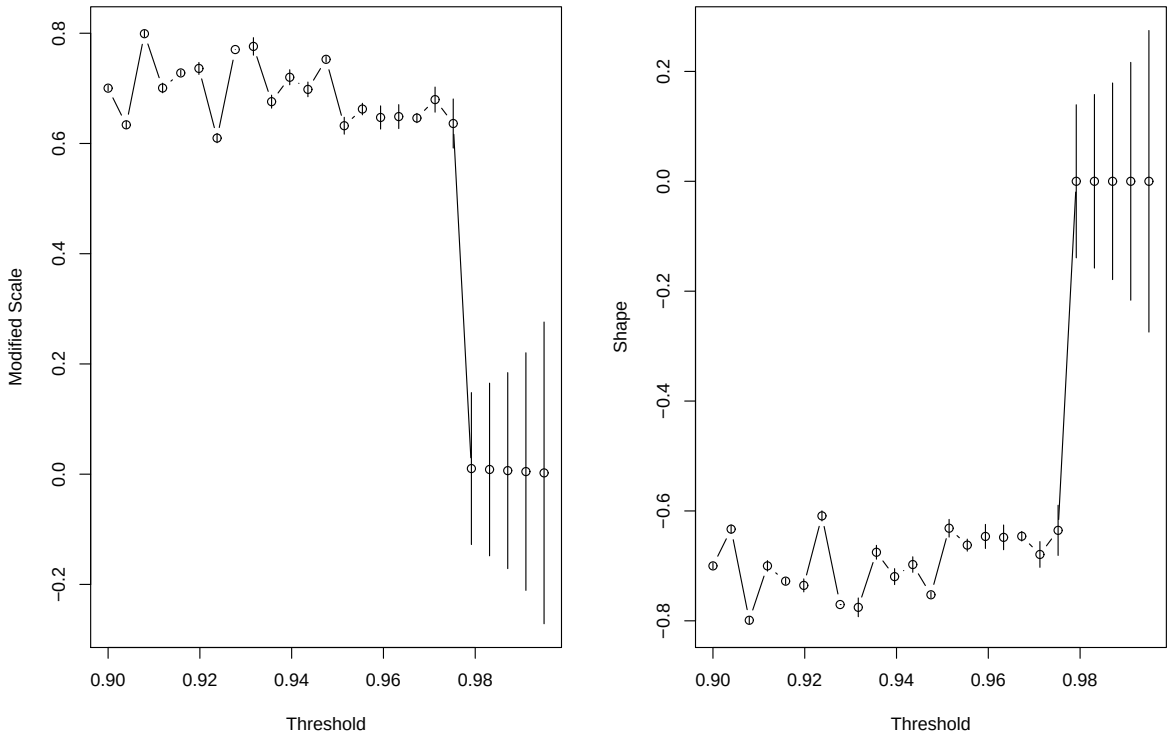

Figure 1: Threshold Choice plot on synthetic data

```
> x <- runif(10000)
> par(mfrow=c(1,2))
> tcplot(x, u.range = c(0.9, 0.995))
```

Results of the `tcplot` function is displayed in Figure 1. We can see clearly that a threshold around 0.98 is a reasonable choice. However, in practice decision are not so clear-cut as for this synthetic example.

3.2.2 Mean Residual Life Plot: *mrlplot*

The **mean residual life plot** is based on the theoretical mean of the GPD. Let X be a *r.v.* distributed as $GPD(\mu, \sigma, \xi)$. Then, theoretically we have:

$$\mathbb{E}[X] = \mu + \frac{\sigma}{1-\xi}, \qquad \text{for } \xi < 1 \tag{3.3}$$

When $\xi \geq 1$, the theoretical mean is infinite.

In practice, if X represents excess over a threshold μ_0, and if the approximation by a GPD is good enough, we have:

$$\mathbb{E}[X - \mu_0 | X > \mu_0] = \frac{\sigma_{\mu_0}}{1-\xi} \tag{3.4}$$

For all new threshold μ_1 such as $\mu_1 > \mu_0$, excesses above the new threshold are also approximate by a GPD with updated parameters - see section 3.2.1. Thus,

$$\mathbb{E}[X - \mu_1 | X > \mu_1] = \frac{\sigma_{\mu_1}}{1-\xi} = \frac{\sigma_{\mu_0} + \xi\mu_1}{1-\xi} \tag{3.5}$$

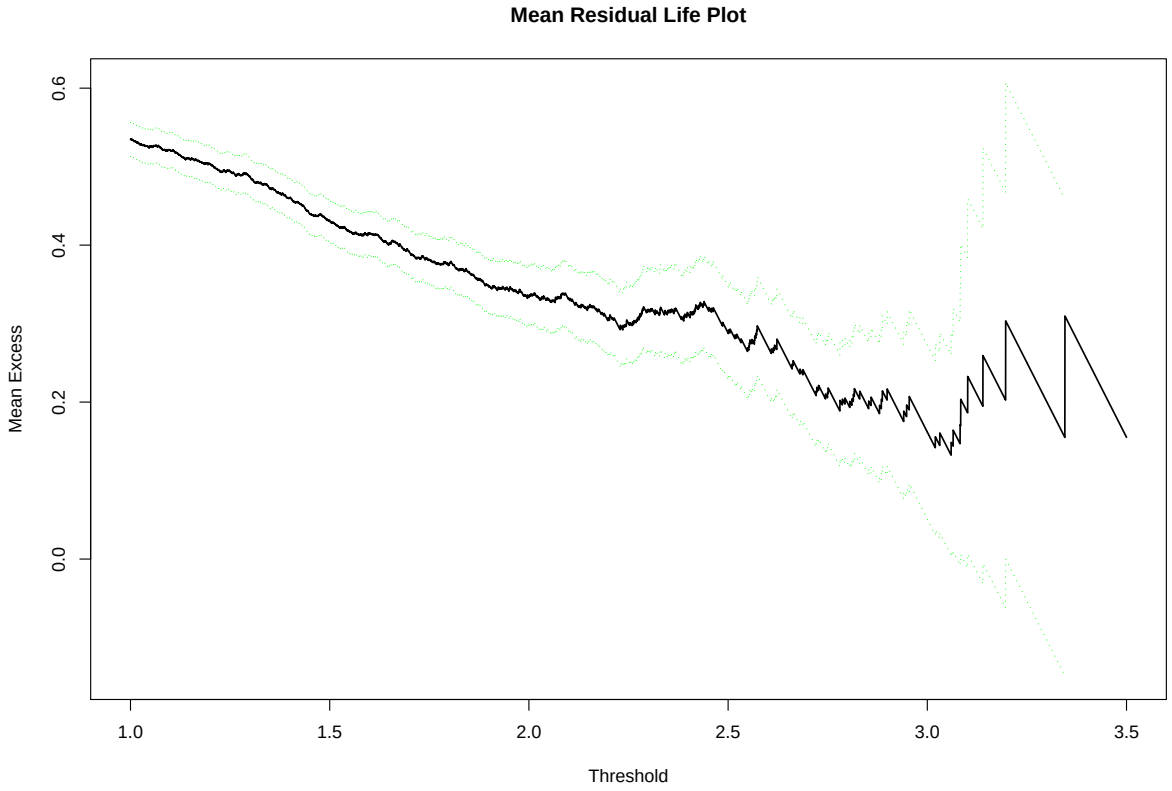

Figure 2: Mean residual life plot on synthetic data

The quantity $\mathbb{E}[X - \mu_1 | X > \mu_1]$ is linear in μ_1. Or, $\mathbb{E}[X - \mu_1 | X > \mu_1]$ is simply the mean of excesses above the threshold μ_1 which can easily be estimated using the empirical mean.

A mean residual life plot consists in representing points:

$$\left\{ \left(\mu, \frac{1}{n_\mu} \sum_{i=1}^{n_\mu} x_{i,n_\mu} - \mu \right) : \mu \leq x_{\max} \right\} \tag{3.6}$$

where n_μ is the number of observations $\mathbf{x}$ above the threshold μ, x_{i,n_μ} is the i-th observation above the threshold μ and $x_{\max}$ is the maximum of the observations $\mathbf{x}$.

Confidence intervals can be added to this plot as the empirical mean can be supposed to be normally distributed (Central Limit Theorem). However, normality doesn't hold anymore for high threshold as there are less and less excesses. Moreover, by construction, this plot always converge to the point $(x_{\max}, 0)$.

Here is another synthetic example.

```
> x <- rnorm(10000)
 mrlplot(x, u.range = c(1, 3.5), col = c("green", "black", "green"))
```

Figure 2 displays the mean residual life plot. A threshold around 2.5 should be reasonable.

3.2.3 L-Moments plot: *lmomplot*

L-moments are summary statistics for probability distributions and data samples. They are analogous to ordinary moments – they provide measures of location, dispersion, skewness, kurtosis,

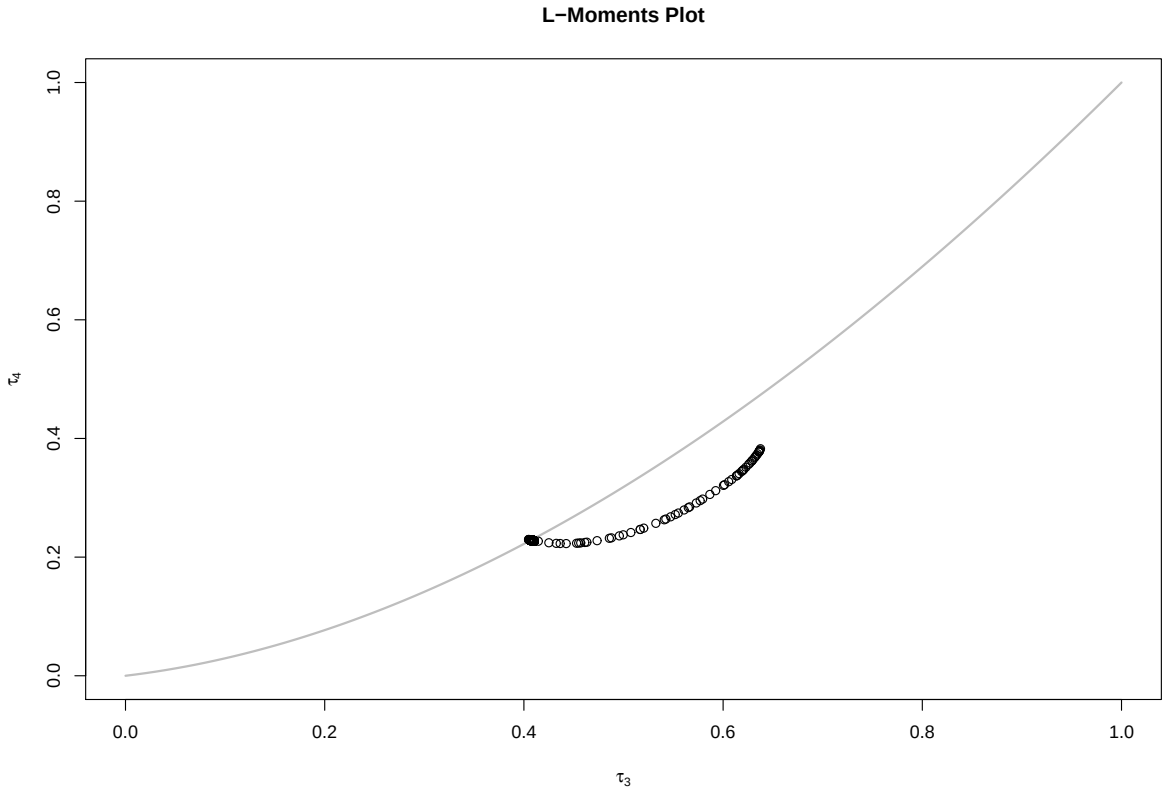

Figure 3: L-Moment plot on synthetic data

and other aspects of the shape of probability distributions or data samples – but are computed from linear combinations of the ordered data values (hence the prefix L).

For the GPD, the following relation holds:

$$\tau_4 = \tau_3 \frac{1 + 5\tau_3}{5 + \tau_3} \tag{3.7}$$

where τ_4 is the **L-Kurtosis** and τ_3 is the **L-Skewness**.

The **L-Moment** plot represents points defined by:

$$\{(\hat{\tau}_{3,u}, \hat{\tau}_{4,u}) : u \leq x_{\max}\} \tag{3.8}$$

where $\hat{\tau}_{3,u}$ and $\hat{\tau}_{4,u}$ are estimations of the L-Kurtosis and L-Skewness based on excesses over threshold u and $x_{\max}$ is the maximum of the observations $\mathbf{x}$. The theoretical curve defined by equation (3.7) is traced as a guideline.

Here is a trivial example.

```
> x <- c(1 - abs(rnorm(200, 0, 0.2)), rgpd(100, 1, 2, 0.25))
> lmomplot(x, u.range = c(0.9, 2), identify = FALSE)
```

Figure 3 displays the L-Moment plot. By passing option `identiy = TRUE` user can click on the graphic to identify the threshold related to the point selected.

We found that this graphic has often poor performance on real data.

3.2.4 Dispersion Index Plot: *diplot*

The **Dispersion Index plot** is particularly useful when dealing with time series. The EVT states that excesses over a threshold can be approximated by a GPD. However, the EVT also states that the occurrences of these excesses must be represented by a Poisson process.

Let X be a *r.v.* distributed as a Poisson distribution with parameter λ. That is:

$$\Pr[X = k] = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k \in \mathbb{N}. \tag{3.9}$$

Thus, we have $\mathbb{E}[X] = Var[X]$. Cunnane (1979) introduced a **Dispersion Index** statistic defined by:

$$DI = \frac{s^2}{\lambda} \tag{3.10}$$

where s^2 is the intensity of the Poisson process and λ the mean number of events in a block - most often this is a year. Moreover, a confidence interval can be computed by using a χ^2 test:

$$I_\alpha = \left[\frac{\chi^2_{(1-\alpha)/2,M-1}}{M-1}, \frac{\chi^2_{1-(1-\alpha)/2,M-1}}{M-1}\right] \tag{3.11}$$

where $\Pr[DI \in I_\alpha] = \alpha$.

For the next example, we use the data set *ardieres* included in the **POT** package. Moreover, as *ardieres* is a time series, and thus strongly auto-correlated, we must "extract" extreme events while preserving independence between events. This is achieved using function **clust**[2].

```
> data(ardieres)
> events <- clust(ardieres, u = 2, tim.cond = 8 / 365,
+ clust.max = TRUE)
> diplot(events, u.range = c(2, 20))
```

The Dispersion Index plot is presented in Figure 4. From this figure, a threshold around 5 should be reasonable.

3.3 Fitting the GPD

3.3.1 The univariate case

The main function to fit the GPD is called **fitgpd**. This is a generic function which can fit the GPD according several estimators. There are currently 7 estimators available: method of moments `moments`, maximum likelihood `mle`, biased and unbiased probability weighted moments `pwmb, pwmu`, mean power density divergence `mdpd`, median `med` and pickands' `pickands` estimators. Details for these estimators can be found in (Coles, 2001), (Hosking and Wallis, 1987), (Juárez and Schucany, 2004), (Peng and Welsh, 2001) and (Pickands, 1975).

The MLE is a particular case as it is the only one which allows varying threshold. Moreover, two types of standard errors are available: "expected" or "observed" information of Fisher. The option `obs.fish` specifies if we want observed (`obs.fish = TRUE`) or expected (`obs.fish = FALSE`).

As Pickands' estimator is not always feasible, user must check the message of feasibility return by function `fitgpd`.

We give here several didactic examples.

[2]The clust function will be presented later in section 3.6.

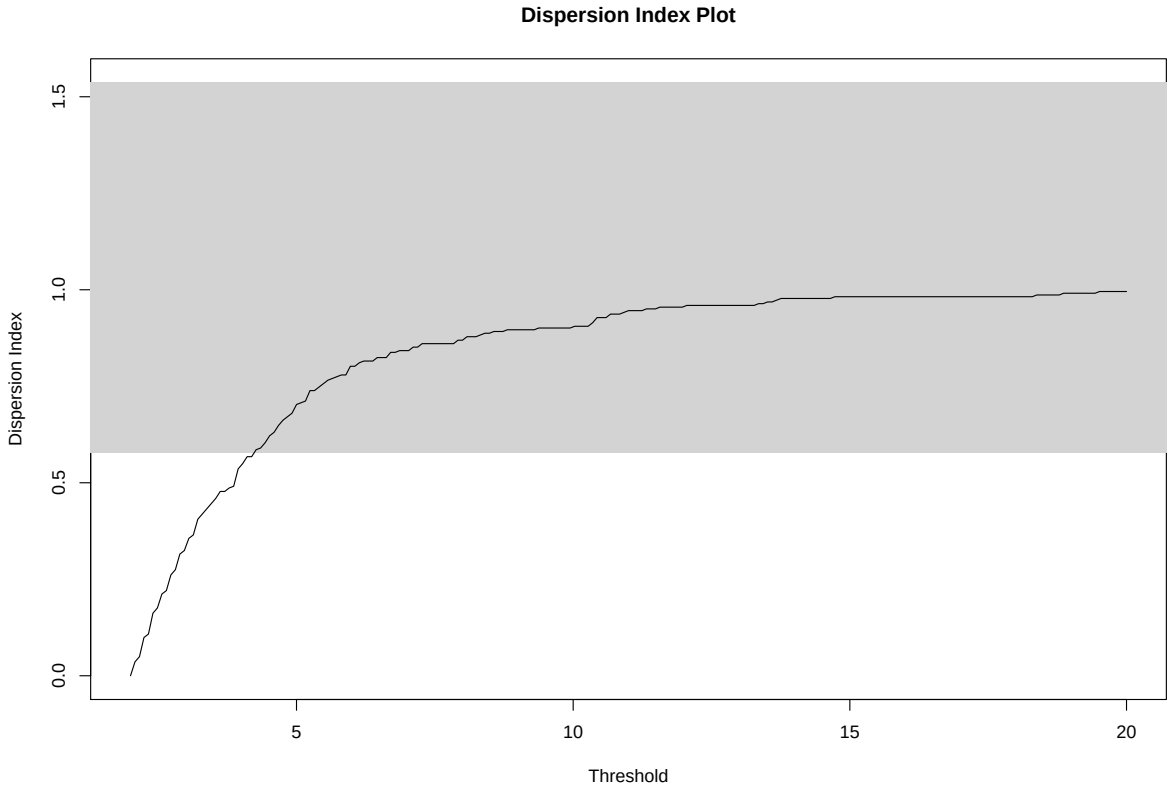

Figure 4: Dispersion index plot for the dataset *ardieres*

```
> x <- rgpd(200, 1, 2, 0.25)
> mom <- fitgpd(x, 1, "moments")$param
> mle <- fitgpd(x, 1, "mle")$param
> pwmu <- fitgpd(x, 1, "pwmu")$param
> pwmb <- fitgpd(x, 1, "pwmb")$param
> pickands <- fitgpd(x, 1, "pickands")$param
> med <- fitgpd(x, 1, "med", start = mle)$param
> mdpd <- fitgpd(x, 1, "mdpd")$param
> print(rbind(mom, mle, pwmu, pwmb, pickands, med, mdpd))
            scale      shape
mom      1.693222 0.22258795
mle      1.758304 0.18825359
pwmu     1.774835 0.18511688
pwmb     1.783859 0.18097373
pickands 1.867132 0.02210745
med      1.826378 0.08445534 ##Convergence: iteration limit reached
mdpd     1.788432 0.16608265
```

The MLE method allows to fix either the scale or the shape parameter. For example, if we want to fit a Exponential distribution, just do:

```
> x <- rgpd(100, 1, 2, 0)
> fitgpd(x, thresh = 1, shape = 0, method = "mle")
> ##The same but with a fixed scale value
> fitgpd(x, thresh = 1, scale = 2, method = "mle")
```

If now, we want to fit a GPD with a varying threshold, just do:

```
> x <- rgpd(500, 1:2, 0.3, 0.01)
> fitgpd(x, 1:2, method = "mle")
```

Note that the varying threshold is repeated cyclically until it matches the length of object `x`.

3.3.2 The bivariate case

The generic function to fit bivariate POTs is **fitbvgpd**. There is currently 6 models for the bivariates GPD – see Annexe A. All of these models are fitted using maximum likelihood estimator. Moreover, the approach uses *censored* likelihood – see (Smith et al., 1997).

```
> x <- rgpd(500, 0, 1, 0.25)
> y <- rgpd(500, 2, 0.5, -0.25)
> Mlog <- fitbvgpd(cbind(x,y), c(0,2), model = "log")
> Mlog

Call: fitbvgpd(data = cbind(x, y), threshold = c(0, 2), model = "log")
Estimator: MLE
Dependence Model and Strenght:
Model : Logistic
lim_u Pr[ X_1 > u | X_2 > u] = 0.02

Deviance: 1397.460
     AIC: 1407.460

Marginal Threshold: 0 2
Marginal Number Above: 500 500
Marginal Proportion Above: 1 1
Joint Number Above: 500
Joint Proportion Above: 1
Number of events such as {Y1 > u1} U {Y2 > u2}: 500

Estimates
 scale1   scale2   shape1   shape2    alpha
 1.0912   0.5155   0.2241  -0.2500   0.9853

Standard Errors
 scale1   scale2   shape1   shape2    alpha
0.07892  0.03026  0.05767  0.03960  0.02398

Asymptotic Variance Covariance
        scale1      scale2      shape1      shape2      alpha
scale1   6.228e-03   3.624e-05  -3.033e-03  -3.347e-05   1.561e-05
scale2   3.624e-05   9.157e-04  -7.282e-07  -1.002e-03   1.856e-05
shape1  -3.033e-03  -7.282e-07   3.326e-03   2.136e-05  -9.574e-05
shape2  -3.347e-05  -1.002e-03   2.136e-05   1.568e-03  -5.620e-05
alpha    1.561e-05   1.856e-05  -9.574e-05  -5.620e-05   5.750e-04

Optimization Information
```

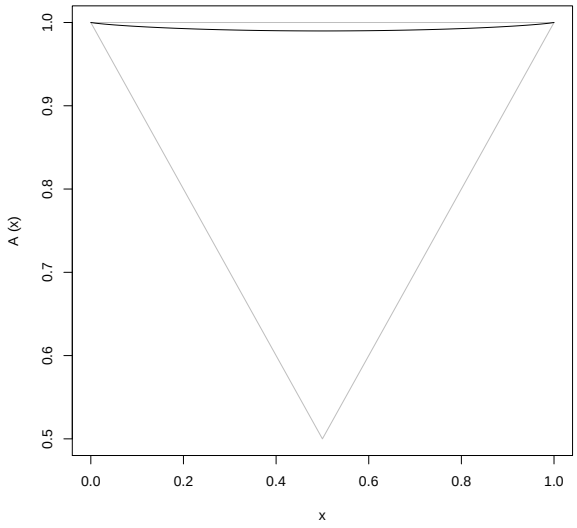

Figure 5: The Pickands' dependence function.

```
Convergence: successful
Function Evaluations: 47
Gradient Evaluations: 10
```

In the summary, we can see `lim_u Pr[ X_1 > u | X_2 > u] = 0.02` . This is the χ statistics of Coles et al. (1999). For the parametric model, we have:

$$\chi = 2 - V(1,1) = 2\left(1 - A(0.5)\right)$$

For independent variables, $\chi = 0$ while for perfect dependence, $\chi = 1$. In our application, the value 0.02 indicates that the variables are independent – which is obvious. In this perspective, it is possible to fixed some parameters. For our purpose of independence, we can run:

```
> fitbvgpd(cbind(x,y), c(0,2), model = "log", alpha = 1)
> ##This is equivalent to fit x and y separately of course.
```

Note that as all bivariate extreme value distributions are asymptotically dependent, the $\overline{\chi}$ statistic of Coles et al. (1999) is always equal to 1.

Another way to detect the strength of dependence is to plot the Pickands' dependence function. This is simply done with the **pickdep** function.

```
> pickdep(Mlog)
```

The horizontal line corresponds to independence while the other ones corresponds to perfect dependence. Please note that by construction, the *mixed* and *asymetric mixed* models can not model perfect dependence variables.

3.3.3 Markov Chains for Exceedances

The classical way to perform an analysis of peaks over a threshold is to fit the GPD to cluster maxima. However, there is a waste of data as only the cluster maxima is considered. On the

contrary, if we fit the GPD to all exceedances, standard errors are underestimated as we consider independence for dependent observations. Here is where Markov Chains can help us. The main idea is to model the dependence structure using a Markov Chains while the joint distribution is obviously a multivariate extreme value distribution. This idea was first introduces by Smith et al. (1997).

In the remainder of this section, we will only focus with first order Markov Chains. Thus, the likelihood for all exceedances is:

$$L(y_1, \ldots, y_n; \theta, \psi) = \frac{\prod_{i=2}^{n} f(y_{i-1}, y_i; \theta, \psi)}{\prod_{i=2}^{n-1} f(y_i; \theta)} \tag{3.12}$$

where $f(y_{i-1}, y_i; \theta, \psi)$ is the joint density, $f(y_i; \theta)$ is the marginal density, θ is the marginal GPD parameters and ψ is the dependence parameter. The marginals are modelled using a GPD, while the joint distribution is a bivariate extreme value distribution.

For our application, we use the **simmc** function which simulate a first order Markov chain with extreme value dependence structure.

```
> ## First simulate a first order Markov Chain with
> ## uniform(0,1) marge.
> mc <- simmc(1000, alpha = 0.5, model = "log")
> ## Transform it to a GPD
> mc <- qgpd(mc, 2, 1, 0.15)
> fitmcgpd(mc, 2, "log")

Call: fitmcgpd(data = mc, threshold = 2, model = "log")
Estimator: MLE
Dependence Model and Strenght:
Model : Logistic
lim_u Pr[ X_1 > u | X_2 > u] = 0.571
Deviance: 1448.343
     AIC: 1454.343

Threshold: 2
Number Above: 998
Proportion Above: 1

Estimates
 scale   shape   alpha
0.9453  0.1682  0.5146

Standard Error Type:

Standard Errors
  scale    shape    alpha
0.09219  0.04742  0.02237

Asymptotic Variance Covariance
       scale       shape       alpha
scale   0.0084994  -0.0018018  -0.0010052
shape  -0.0018018   0.0022488  -0.0004062
alpha  -0.0010052  -0.0004062   0.0005005
```

```
Optimization Information
  Convergence: successful
  Function Evaluations: 72
  Gradient Evaluations: 13
```

3.4 Confidence Intervals

Once a statistical model is fitted, it is usual to gives confidence intervals. Currently, only `mle, pwmu, pwmb, moments` estimators can computed confidence intervals. Moreover, for method `mle`, "standard" and "profile" confidence intervals are available.

If we want confidence intervals for the scale parameters:

```
> x <- rgpd(100, 1, 2, 0.25)
> mle <- fitgpd(x, 1, method = "mle")
> mom <- fitgpd(x, 1, method = "moments")
> pwmb <- fitgpd(x, 1, method = "pwmb")
> pwmu <- fitgpd(x, 1, method = "pwmu")
> gpd.fiscale(mle, conf = 0.9)
> gpd.fiscale(mom, conf = 0.9)
> gpd.fiscale(pwmu, conf = 0.9)
> gpd.fiscale(pwmb, conf = 0.9)
```

For shape parameter confidence intervals, simply use function `gpd.fishape` instead of `gpd.fiscale`. Note that the *fi* stands for "Fisher Information".

Thus, if we want profile confidence intervals, we must use functions `gpd.pfscale` and `gpd.pfshape`. The *pf* stands for "profile". These functions are only available with a model fitted with MLE.

```
> gpd.pfscale(mle, range = c(1, 2.5), conf = 0.9)
> gpd.pfshape(mle, range = c(-0.1, 0.6), conf = 0.95)
```

Confidence interval for quantiles - or return levels - are also available. This is achieved using: (a) the Delta method or (b) profile likelihood.

```
> gpd.firl(pwmu, prob = 0.95)
> gpd.pfrl(mle, prob = 0.95, range = c(4.8, 15))
```

The profile confidence interval functions both returns the confidence interval and plot the profile log-likelihood function. Figure 6 depicts the graphic window returned by function `gpd.pfrl` for the return level associated to non exceedence probability 0.95.

3.5 Model Checking

To check the fitted model, users must call function **plot** which has a method for the `uvpot`, `bvpot` and `mcpot` classes. For example, this is a generic function which calls functions: `pp.gpd` (probability/probability plot), `qq.gpd` (quantile/quantile plot), `dens.gpd` (density plot) and `retlev.gpd` (return level plot) for the uvpot class.

Here is a basic illustration of the function `plot`.

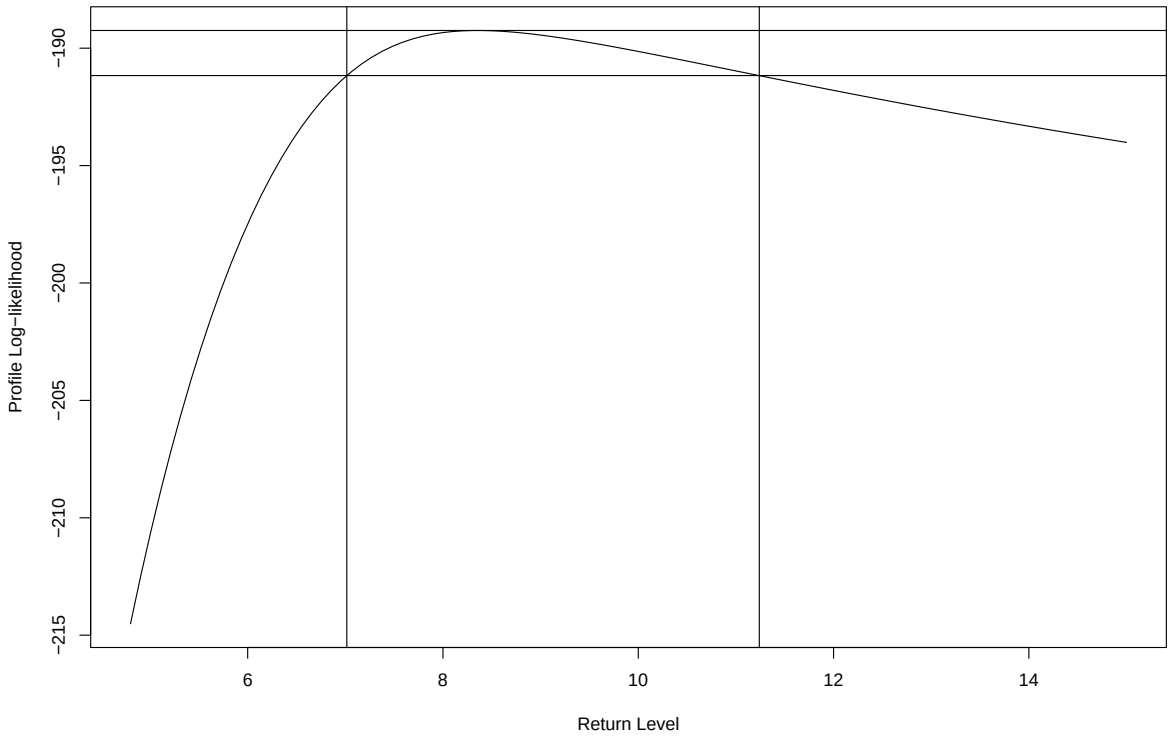

Figure 6: Profile log-likelihood function for a given return level

```
> x <- rgpd(200, 10, 0.5, -0.2)
> fitted <- fitgpd(x, 10, method = "mle")
> par(mfrow=c(2,2))
> plot(fitted, npy = 1)
```

Figure 7 displays the graphic windows obtained with the latter execution.

If one is interested in only a probability/probability plot, there is two options. We can call function `pp.gpd` or equivalently `plotgpd` with the **which** option. The "which" option select which graph you want to plot. That is:

- which = 1 for a probability/probability plot;
- which = 2 for a quantile/quantile plot;
- which = 3 for a density plot;
- which = 4 for a return level plot;

Note that "which" can be a vector like `c(1,3)` or `1:3`.

Thus, the following instruction gives the same graphic.

```
> plot(fitted, which = 1)
> pp.gpd(fitted)
```

If a return level plot is asked ($4 \in$ `which`), a value for `npy` is needed. "npy" corresponds to the *mean number of events per year*. This is required to define the "return period". If missing, the default value (i.e. 1) will be chosen.

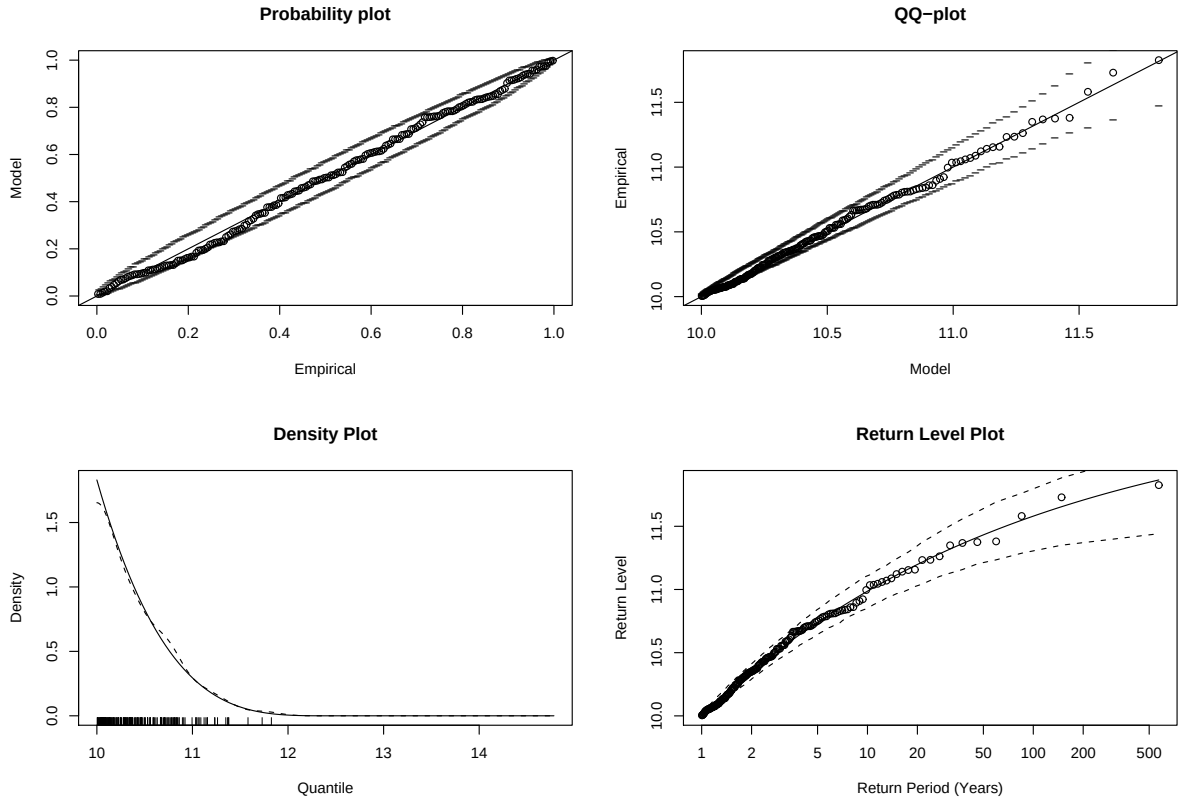

Figure 7: Checking plots from function *plotgpd*

3.6 Declustering Techniques

In opposition to block maxima, a peak over threshold can be problematic when dealing with time series. Indeed, as often time series are strongly auto-correlated, select naively events above a threshold may lead to dependent events.

The function **clust** tries to identify peaks over a threshold while meeting independence criteria. For this purpose, this function needs at least two arguments: the threshold `u` and a time condition for independence `tim.cond`. Clusters are identify as follow:

1. The first exceedence initiates the first cluster;
2. The first observation under the threshold `u` "ends" the cluster unless `tim.cond` does not hold;
3. The next exceedence which hold `tim.cond` initiates a new cluster;
4. The process is iterated as needed.

Here is an application on flood discharges for river Ardière at Beaujeu. A preliminary study shows that two flood events can be considered independent if they do not lie within a 8 days window. Note that unit to define `tim.cond` must be the same than the data analyzed.

```
> data(ardieres)
> clust(ardieres, u = 2, tim.cond = 8 / 365)
```

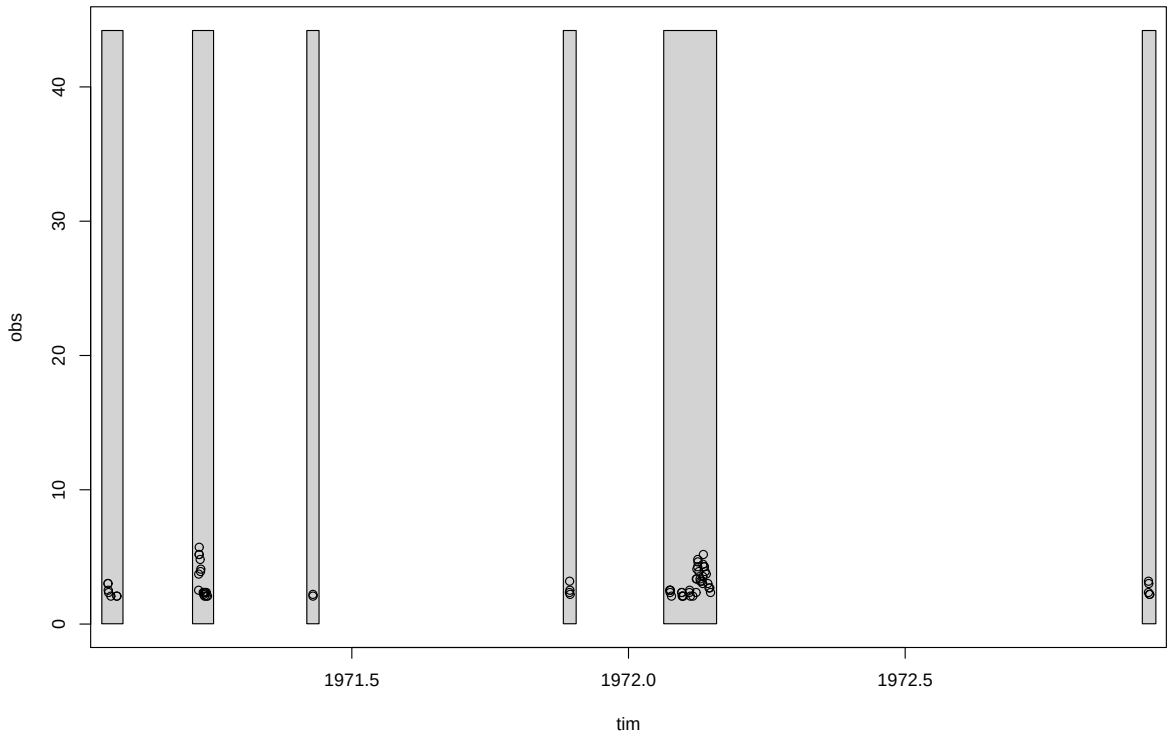

Figure 8: The identified clusters. Data Ardières, u = 2, tim.cond = 8

Several options can be passed to the "clust" function. By default, it will return a list with the identified clusters. Usually, we want only cluster maxima, this is achieved by passing option `clust.max = TRUE`. Users can also ask for a graphic representation of clusters by passing option `plot = TRUE` - see Figure 8.

```
> clustMax <- clust(ardieres, u = 2, tim.cond = 8 / 365,
+ clust.max = TRUE, plot = TRUE, xlim = c(1971.1, 1972.9))
```

3.7 Miscellaneous functions

3.7.1 Return periods: *rp2prob* and *prob2rp*

The functions **rp2prob** and **prob2rp** are useful to convert return periods to non exceedence probabilities and vice versa. It needs either a return period either a non exceedence probability. Moreover, the mean number of events per year "npy" must be specified.

```
> rp2prob(50, 1.8)
  npy retper      prob
1 1.8     50 0.9888889
> prob2rp(0.6, 2.2)
  npy   retper prob
1 2.2 1.136364  0.6
```

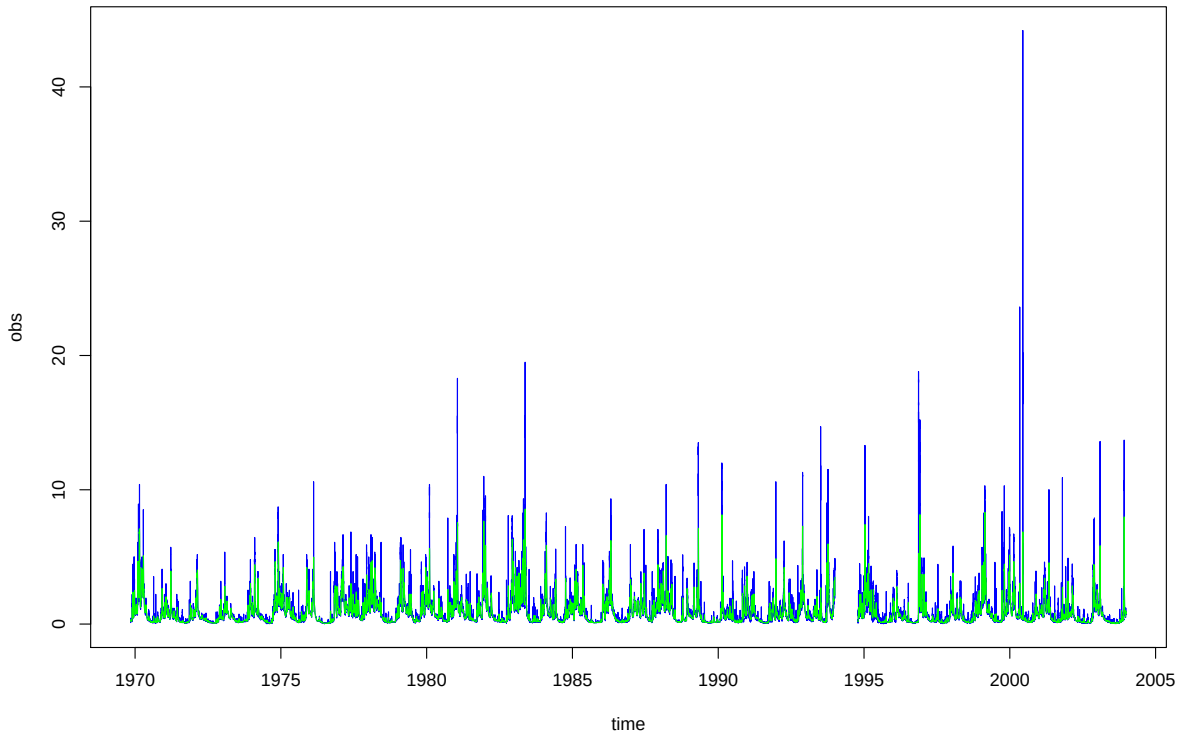

Figure 9: Instantaneous flood discharges and averaged dischaged over duration 3 days. Data ardieres

3.7.2 Unbiased Sample L-Moments: *samlmu*

The function **samlmu** computes the unbiased sample L-Moments.

```
> x <- runif(50)
> samlmu(x, nmom = 5)
        l_1         l_2         t_3         t_4         t_5
 0.53337554  0.16743489 -0.04026843  0.01243610  0.01386457
```

3.7.3 Mobile average window on time series: *ts2tsd*

The function **ts2tsd** computes an "average" time series `tsd` from the initial time series `ts`. This is achieved by using a mobile average window of length `d` on the initial time series.

```
> data(ardieres)
> tsd <- ts2tsd(ardieres, 3 / 365)
> plot(ardieres, type = "l", col = "blue")
> lines(tsd, col = "green")
```

The latter execution is depicted in Figure 9.

4 A Concrete Statistical Analysis of Peaks Over a Threshold

In this section, we provide a full and detailed analysis of peaks over a threshold for the river Ardières at Beaujeu. Figure 9 depicts instantaneous flood discharges - blue line.

As this is a time series, we must selects independent events above a threshold. First, we fix a relatively low threshold to "extract" more events. Thus, some of them are not extreme but regular events. This is necessary to select a reasonable threshold for the asymptotic approximation by a GPD - see section 2.

```
> summary(ardieres)
      time           obs
 Min.   :1970   Min.   : 0.022
 1st Qu.:1981   1st Qu.: 0.236
 Median :1991   Median : 0.542
 Mean   :1989   Mean   : 1.024
 3rd Qu.:1997   3rd Qu.: 1.230
 Max.   :2004   Max.   :44.200
                NA's   : 1.000
> events0 <- clust(ardieres, u = 1.5, tim.cond = 8/365,
+  clust.max = TRUE)
> par(mfrow=c(2,2))
> mrlplot(events0[,"obs"])
> abline( v = 6, col = "green")
> diplot(events0)
> abline( v = 6, col = "green")
> tcplot(events0[,"obs"])
```

From Figure 10, a threshold value of $6m^3/s$ should be reasonable. The Mean residual life plot - top left panel- indicates that a threshold around $10m^3/s$ should be adequate. However, the selected threshold must be low enough to have enough events above it to reduce variance while not too low as it increase the bias[3].

Thus, we can now "re-extract" events above the threshold $6m^3/s$, obtaining object `events1`. This is necessary as sometimes `events1` is not equal to observations of `events0` greater than $6m^3/s$. We can now define the mean number of events per year "npy". Note that an estimation of the extremal index is available.

```
> events1 <- clust(ardieres, u = 6, tim.cond = 8/365,
+ clust.max = TRUE)
> npy <- length(events1[,"obs"]) / (diff(range(ardieres[,"time"],
 na.rm = TRUE)) - diff(ardieres[c(20945,20947),"time"]))
> ##Because there is a gap !!!
> print(npy)
[1] 1.677934
> attributes(events1)$exi
[1] 0.1225383
```

Let's fit the GPD.

[3] As the asymptotic approximation by a GPD is not accurate anymore.

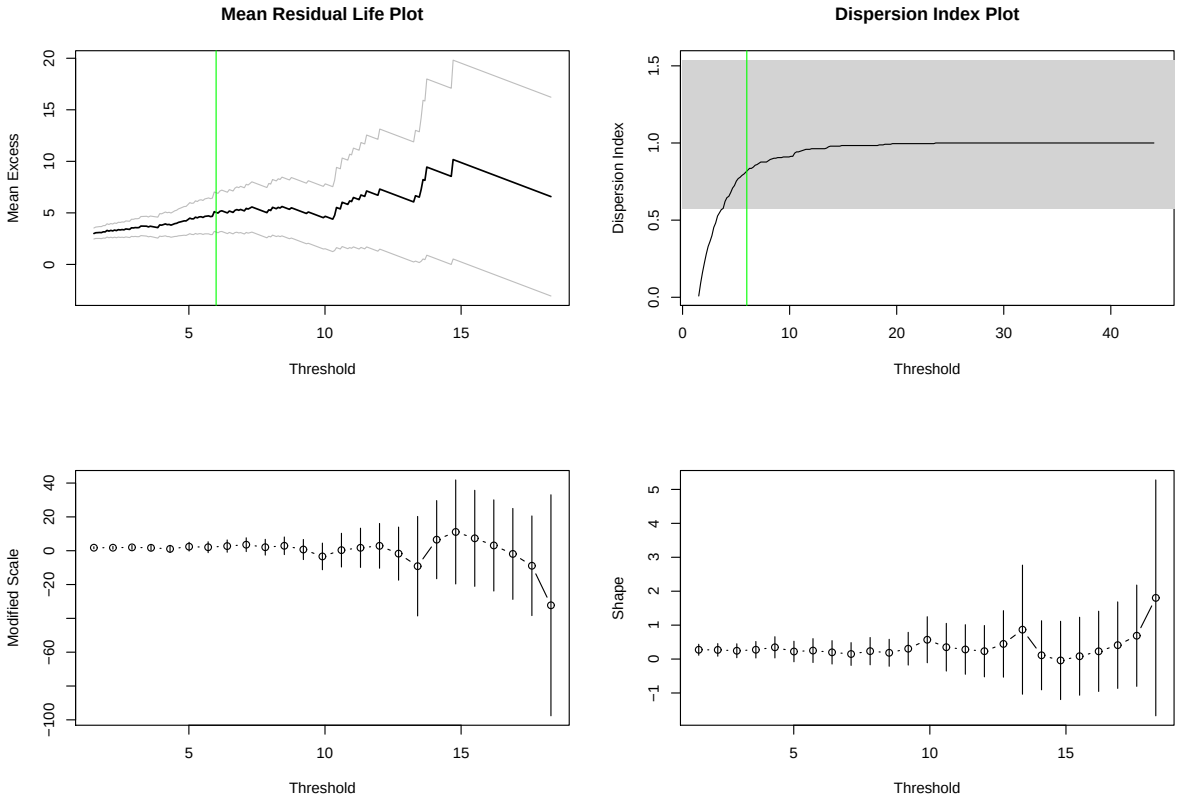

Figure 10: Threshold selection for river Ardières at Beaujeu.

```
> mle <- fitgpd(events1[,"obs"], thresh = 6, method = "mle")
Estimator: MLE

Varying Threshold: FALSE

Threshold: 6
Number Above: 56
Proportion Above: 1

Estimates
 scale   shape
3.8285  0.1579

Standard Error Type: Observed

Standard Errors
 scale   shape
0.7224  0.1349

Asymptotic Variance Covariance
       scale     shape
scale   0.52180  -0.05714
shape  -0.05714   0.01819

Optimization Information
```

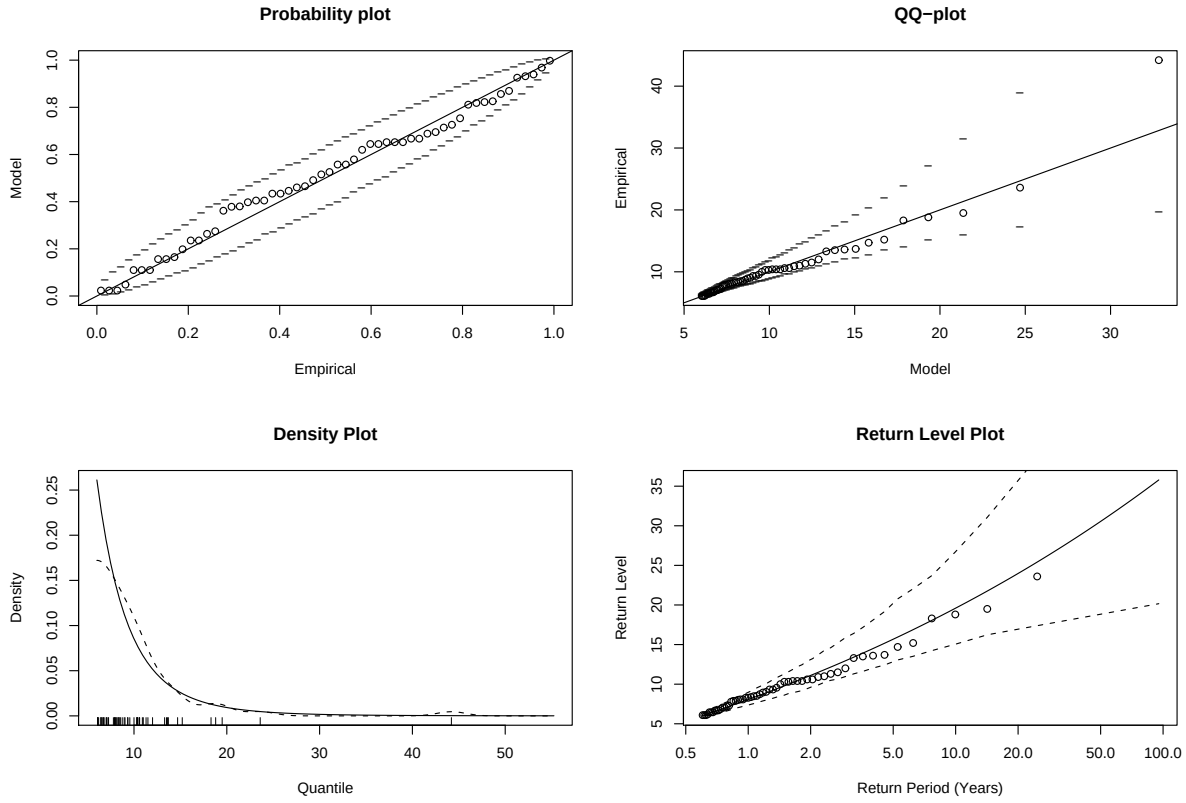

Figure 11: Graphic diagnostics for river Ardières at Beaujeu

```
  Convergence: successful
  Function Evaluations: 36
  Gradient Evaluations: 9
```

The result of function **fitgpd** gives the name of the estimator, if a varying threshold was used, the threshold value, the number and the proportion of observations above the threshold, parameter estimates, standard error estimates and type, the asymptotic variance-covariance matrix and convergence diagnostic.

Figure 11 shows graphic diagnostics for the fitted model. It can be seen that the fitted model "mle" seems to be appropriate. Suppose we want to know the return level associated to the 100-year return period.

```
> ##First convert return period in prob
> rp2prob(retper = 100, npy = npy)
       npy retper      prob
1 1.677934    100 0.9940403
> prob <- rp2prob(retper = 100, npy = npy)[,"prob"]
> qgpd(prob, loc = 6, scale = mle$param["scale"],
+  shape = mle$param["shape"])
36.19317
```

To take into account uncertainties, Figure 12 depicts the profile confidence interval for the quantile associated to the 100-year return period.

```
> gpd.pfrl(mle, prob, range = c(25, 90), nrang = 200)
```

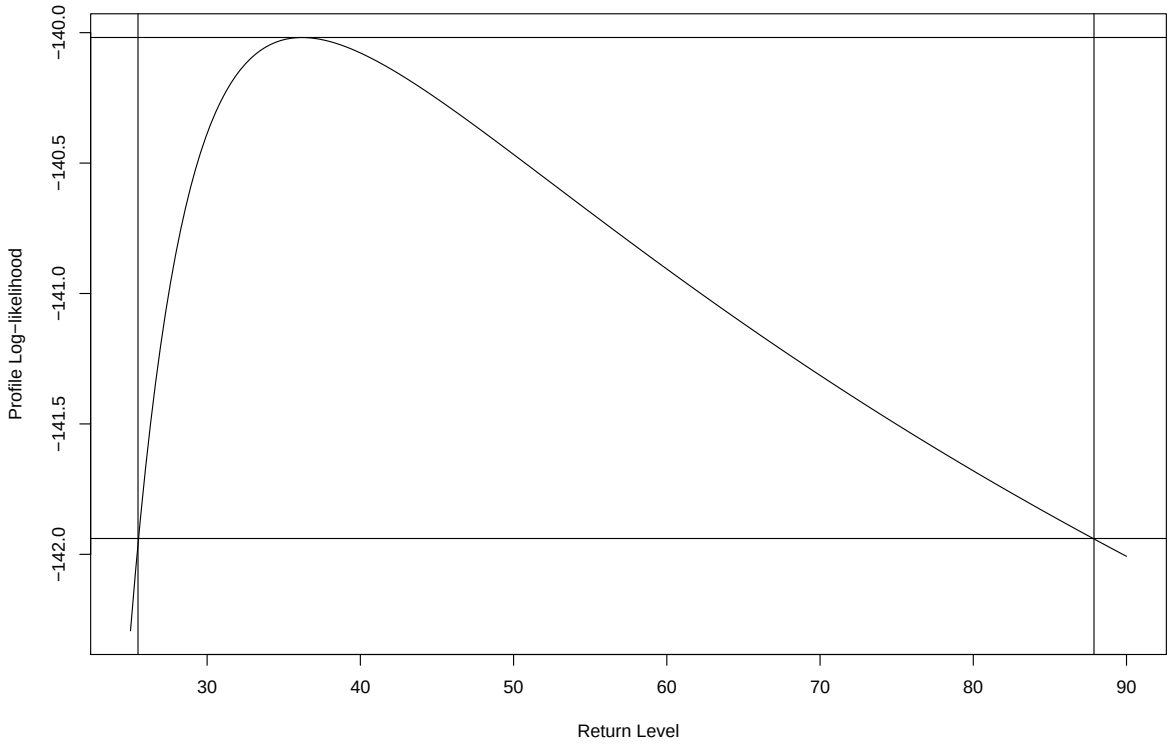

Figure 12: Profile-likelihood function for the 100-year return period quantile

```
If there is some troubles try to put vert.lines = FALSE or change
 the range...
conf.inf conf.sup
25.48995 87.87688
```

Sometimes it is necessary to know the estimated return period of a specified events. Lets do it with the larger events in "events1".

```
> maxEvent <- max(events1[,"obs"])
> print(maxEvent)
[1] 44.2
> prob <- pgpd(maxEvent, loc = 6, scale = mle$param["scale"],
+ shape = mle$param["shape"])
> print(prob)
0.997501
> prob2rp(prob, npy = npy)
       npy   retper     prob
1 1.677934 238.4804 0.997501
```

Thus, the largest events that occurs in June 2000 has approximately a return period of 240 years.

Maybe it is a good idea to fit the GPD with the other estimators available in the **POT** package.

A Dependence Models for Bivariate Extreme Value Distributions

A.1 The Logisitic model

The logisitic model is defined by:

$$V(x,y) = \left(x^{-1/\alpha} + y^{-1/\alpha}\right)^{\alpha}, \qquad 0 < \alpha \leq 1 \tag{A.1}$$

Independence is obtained when $\alpha = 1$ while total dependence for $\alpha \to 0$.

The Pickands' dependence function for the logistic model is:

$$\begin{array}{rcl} A : [0,1] & \longrightarrow & [0,1] \\ w & \longmapsto & \left[(1-w)^{\frac{1}{\alpha}} + w^{\frac{1}{\alpha}}\right]^{\alpha} \end{array}$$

A.2 The Asymetric Logistic model

The asymetric logistic model is defined by:

$$V(x,y) = \frac{1-\theta_1}{x} + \frac{1-\theta_2}{y} + \left[\left(\frac{x}{\theta_1}\right)^{-\frac{1}{\alpha}} + \left(\frac{y}{\theta_2}\right)^{-\frac{1}{\alpha}}\right]^{\alpha},$$

with $0 < \alpha \leq 1$, $0 \leq \theta_1, \theta_2 \leq 1$.

Independence is obtained when either $\alpha = 1$, $\theta_1 = 0$ or $\theta_2 = 0$. Differents limits occur when θ_1 and θ_2 are fixed and $\alpha = 1 \to 0$.

The Pickands' dependence function for the asymetric logistic model is:

$$A(w) = (1-\theta_1)(1-w) + (1-\theta_2)w + \left[(1-w)^{\frac{1}{\alpha}}\theta_1^{\frac{1}{\alpha}} + w^{\frac{1}{\alpha}}\theta_2^{\frac{1}{\alpha}}\right]^{\alpha}$$

A.3 The Negative Logistic model

The negative logistic model is defined by:

$$V(x,y) = \frac{1}{x} + \frac{1}{y} - \left(x^{\alpha} + y^{\alpha}\right)^{-\frac{1}{\alpha}}, \qquad \alpha > 0 \tag{A.2}$$

Independence is obtained when $\alpha \to 0$ while total dependence when $\alpha \to +\infty$.

The Pickands' dependence function for the negative logistic model is:

$$A(w) = 1 - \left[(1-w)^{-\alpha} + w^{-\alpha}\right]^{-\frac{1}{\alpha}}$$

A.4 The Asymetric Negative Logistic model

The asymetric negative logistic model is defined by:

$$V(x,y) = \frac{1}{x} + \frac{1}{y} - \left[\left(\frac{x}{\theta_1}\right)^{\alpha} + \left(\frac{y}{\theta_2}\right)^{\alpha}\right]^{-\frac{1}{\alpha}}, \qquad \alpha > 0, \quad 0 < \theta_1, \theta_2 \leq 1$$

Independence is obtained when either $\alpha \to 0$, $\theta_1 \to 0$ or $\theta_2 \to 0$. Different limits occur when θ_1 and θ_2 are fixed and $\alpha \to +\infty$.

The Pickands' dependence function for the asymetric negative logistic model is:

$$A(w) = 1 - \left[\left(\frac{1-w}{\theta_1}\right)^{-\alpha} + \left(\frac{w}{\theta_2}\right)^{-\alpha}\right]^{-\frac{1}{\alpha}}$$

A.5 The Mixed model

The mixed model is defined by:

$$V(x,y) = \frac{1}{x} + \frac{1}{y} - \frac{\alpha}{x+y}, \qquad 0 \leq \alpha \leq 1$$

Independence is obtained when $\alpha = 0$ while total dependence could never be reached.

The Pickands' dependence function for the mixed model is:

$$A(w) = 1 - w\,(1-w)\,\alpha$$

A.6 The Asymetric Mixed model

The asymetric mixed model is defined by:

$$V(x,y) = \frac{1}{x} + \frac{1}{y} - \frac{(\alpha+\theta)\,x + (\alpha+2\theta)\,y}{(x+y)^2}, \qquad \alpha \geq 0, \quad \alpha + 2\theta \leq 1, \quad \alpha + 3\theta \geq 0$$

Independence is obtained when $\alpha = \theta = 0$ while total dependence could never be reached.

The Pickands' dependence function for the asymetric mixed model is:

$$A(w) = \theta w^3 + \alpha w^2 - (\alpha+\theta)\,w + 1$$

References

P. Bortot and S. Coles. The multivariate gaussian tail model: An application to oceanographic data. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 49(1):31–49, 2000.

S. Coles. *An Introduction to Statistical Modelling of Extreme Values*. Springer Series in Statistics. Springers Series in Statistics, London, 2001.

S. Coles, J. Heffernan, and J. Tawn. Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365, December 1999.

C. Cunnane. Note on the poisson assumption in partial duration series model. *Water Resour Res*, 15(2):489–494, 1979.

M. Falk and R.-D. Reiss. On pickands coordinates in arbitrary dimensions. *Journal of Multivariate Analysis*, 92(2):426–453, 2005.

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

J.R.M. Hosking and J.R. Wallis. Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3):339–349, 1987.

A. F. Jenkinson. The frequency distribution of the annual maximum (or minimum) values of meteorological events. *Quaterly Journal of the Royal Meteorological Society*, 81:158–172, 1955.

S.F. Juárez and W.R. Schucany. Robust and efficient estimation for the generalized pareto distribution. *Extremes*, 7(3):237–251, 2004. ISSN 13861999 (ISSN).

C. Klüppelberg and A. May. Bivariate extreme value distributions based on polynomial dependence functions. *Math Methods Appl Sci*, 29(12):1467–1480, 2006. ISSN 01704214 (ISSN).

C. Kluppelberg and T. Mikosch. Large deviations of heavy-tailed random sums with applications in insurance and finance. *Journal of Applied Probability*, 34(2):293–308, 1997.

Liang Peng and A.H. Welsh. Robust estimation of the generalized pareto distribution. *Extremes*, 4(1):53–65, 2001.

J. Pickands. Multivariate extreme value distributions. In *Proceedings 43rd Session International Statistical Institute*, 1981.

J. III Pickands. Statistical inference using extreme order statistics. *Annals of Statistics*, 3: 119–131, 1975.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

S. I. Resnick. *Extreme Values, Regular Variation and Point Processes*. New–York: Springer–Verlag, 1987.

R. L. Smith. *Multivariate Threshold Methods*. Kluwer, Dordrecht, 1994.

R.L. Smith, J.A. Tawn, and S.G. Coles. Markov chain models for threshold exceedances. *Biometrika*, 84(2):249–268, 1997. ISSN 00063444 (ISSN).