Statistics of Multivariate Extremes

Alec Stephenson

August 28, 2012

Summary

This vignette uses the **evd** package to reproduce the figures, tables and analysis in Chapter 9 of Beirlant et al. (2001). The chapter was written by Segers and Vandewalle (2004). The code reproduces all figures, but for space reasons only some figures are shown. Deviations from the book are given as footnotes. Differences will inevitably exist due to numerical optimization and random number generation.

1 Introduction

The methods used here are illustrated using the `lossalae` dataset, which contains observations on 1500 liability claims. The indemnity payment (loss) and the allocated loss adjustment expense (ALAE) is recorded in USD for each claim. The ALAE is the additional expenses associated with the settlement of the claim (e.g. claims investigation expenses and legal fees). The dataset also has an attribute called `capped`, which gives the row names of the indemnity payments that were capped at their policy limit.

We first scale the data so that one unit corresponds to 100 000 USD. Putting the data on a sensible scale assists with the numerical optimization involved in maximum likelihood estimation[1]. The code below plots the raw data using the log scale for both axes (see Figure 1), and plots the data transformed to uniform $(0, 1)$ margins using an empirical transform.

```
> options(show.signif.stars=FALSE)
> library(evd); nn <- nrow(lossalae)
> loss <- lossalae/1e+05; lts <- c(1e-04, 100)
> plot(loss, log = "xy", xlim = lts, ylim = lts)

> ula <- apply(loss, 2, rank)/(nn + 1)
> plot(ula)
```

[1]The book reports an unsatisfactory fit of the GEV model to the margins. It therefore uses only empirical marginal distributions. This was perhaps due to not scaling the data. In this document we use either fully nonparametric or fully parametric methods.

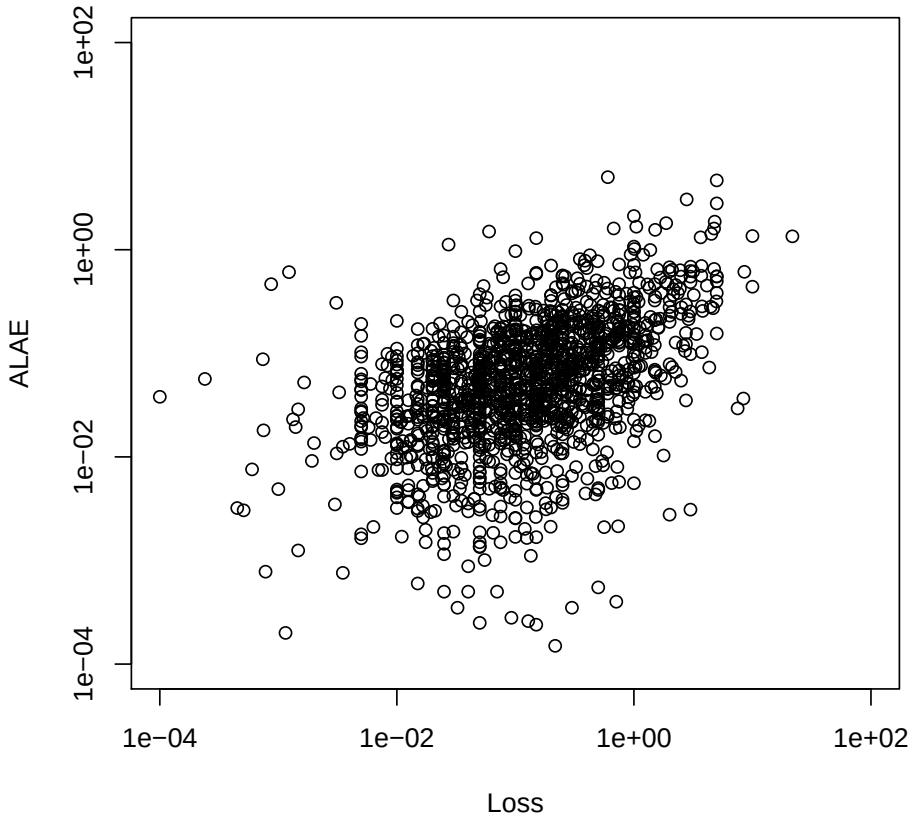

Figure 1: Scatterplot of ALAE verses Loss: original data (log-scale).

2 Parametric Models

Any bivariate extreme value distribution function can be represented in the form

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

where

$$y_j = y_j(z_j) = \{1 + \xi_j(z_j - \mu_j)/\sigma_j\}_+^{-1/\xi_j}$$

for $\sigma_j > 0$ and $j = 1, 2$, and where

$$A(\omega) = -\log\{G(y_1^{-1}(\omega), y_2^{-1}(1-\omega))\},$$

defined on $0 \leq \omega \leq 1$ is called the dependence function[2]. The marginal distributions are generalized extreme value (GEV), given by $G_j(z_j) = \exp(-y_j)$. It follows that $A(0) = A(1) = 1$, and that $A(\cdot)$ is a convex function with $\max(\omega, 1-\omega) \leq A(\omega) \leq 1$ for all $0 \leq \omega \leq 1$. At independence $A(1/2) = 1$. At complete dependence $A(1/2) = 0.5$.

The dependence function represents only the dependence structure of the distribution, and hence only the dependence parameters of parametric models need to be specified in order to produce dependence function plots. The code below plots dependence functions for four different parametric models. The first of these is given in Figure 2.

[2]The book uses the definition $B(\omega) = A(1-\omega)$.

```
> abvevd(dep = 0.5, asy = c(1,1), model = "alog", plot = TRUE)
> abvevd(dep = 0.5, asy = c(0.6,0.9), model = "alog", add = TRUE, lty = 2)
> abvevd(dep = 0.5, asy = c(0.8,0.5), model = "alog", add = TRUE, lty = 3)

> abvevd(dep = -1/(-2), model = "neglog", plot = TRUE)
> abvevd(dep = -1/(-1), model = "neglog", add = TRUE, lty = 2)
> abvevd(dep = -1/(-0.5), model = "neglog", add = TRUE, lty = 3)

> abvevd(alpha = 1, beta = -0.2, model = "amix", plot = TRUE)
> abvevd(alpha = 0.6, beta = 0.1, model = "amix", add = TRUE, lty = 2)
> abvevd(alpha = 0.2, beta = 0.2, model = "amix", add = TRUE, lty = 3)

> abvevd(dep = 1/1.25, model = "hr", plot = TRUE)
> abvevd(dep = 1/0.83, model = "hr", add = TRUE, lty = 2)
> abvevd(dep = 1/0.5, model = "hr", add = TRUE, lty = 3)
```

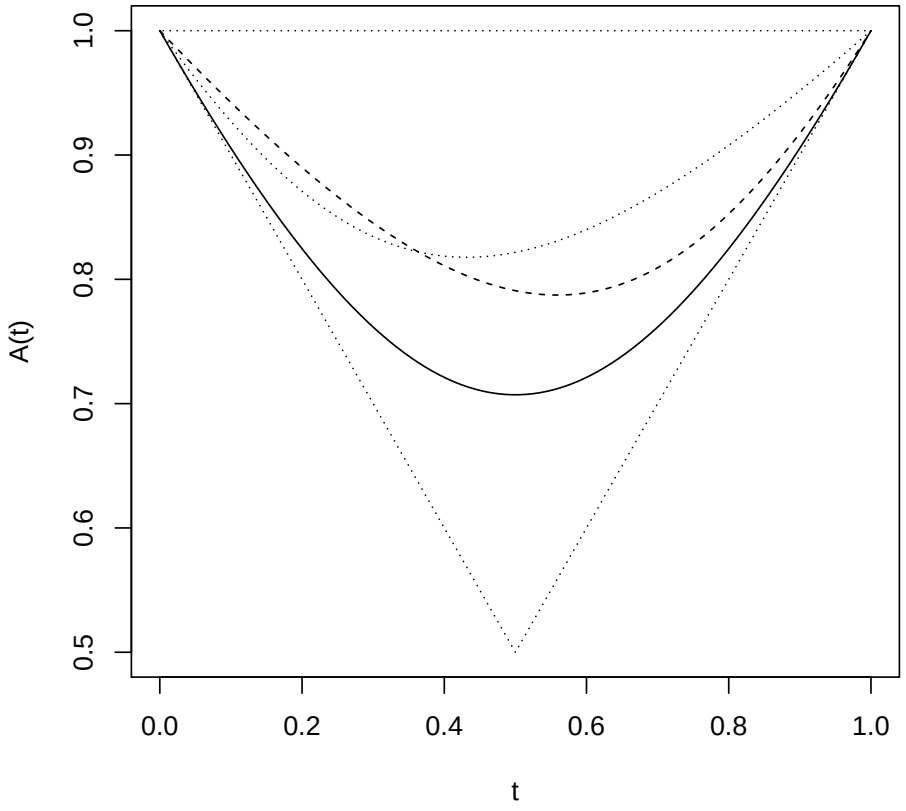

Figure 2: Dependence functions: asymmetric logistic model.

3 Componentwise Maxima

For demonstration purposes we use the data introduced in Section 1 to create a dataset of componentwise block maxima by randomly taking $k = 50$ groups of size $m = 30$, producing k

componentwise maxima taken over m observations[3]. Bivariate extreme value distributions are typically used to model data of this type. The code below creates the componentwise maxima data `cml` and produces two data plots, the first showing the original data and the componentwise maxima, and the second showing the componentwise maxima data transformed to standard exponential margins.

```
> set.seed(131); cml <- loss[sample(nn),]
> xx <- rep(1:50, each = 30); lts <- c(1e-04, 100)
> cml <- cbind(tapply(cml[,1], xx, max), tapply(cml[,2], xx, max))
> colnames(cml) <- colnames(loss)
> plot(loss, log = "xy", xlim = lts, ylim = lts, col = "grey")
> points(cml)
> ecml <- -log(apply(cml,2,rank)/51)
> plot(ecml)
```

The following code estimates and plots the dependence function $A(\cdot)$ from the componentwise maxima data. The first code chunk uses various nonparametric estimates of the dependence function, and also uses empirical (i.e. nonparametric) estimation of the margins, as specified by `epmar = TRUE`. The four different estimates are shown in Figure 3. The second code chunk uses maximum likelihood estimation for parametric models. The call to `fbvevd` fits the model, and the call to `plot` plots the parametric dependence function estimates. The argument specification `asy1 = 1` in the first call to `fbvevd` constrains the model fit so that the first asymmetry parameter of the model is fixed at the value one.

```
> pp <- "pickands"; cc <- "cfg"
> abvnonpar(data = cml, epmar = TRUE, method = pp, plot = TRUE, lty = 3)
> abvnonpar(data = cml, epmar = TRUE, method = pp, add = TRUE, madj = 1, lty = 2)
> abvnonpar(data = cml, epmar = TRUE, method = pp, add = TRUE, madj = 2, lty = 4)
> abvnonpar(data = cml, epmar = TRUE, method = cc, add = TRUE, lty = 1)
```

```
> m1 <- fbvevd(cml, asy1 = 1, model = "alog")
> m2 <- fbvevd(cml, model = "log")
> m3 <- fbvevd(cml, model = "bilog")
> plot(m1, which = 4, nplty = 3)
> plot(m2, which = 4, nplty = 3, lty = 2, add = TRUE)
> plot(m3, which = 4, nplty = 3, lty = 4, add = TRUE)
```

The objects produced by `fbvevd` contain information about the parametric fit of the bivariate extreme value distribution. For example, `m2` contains information on the fit of a (symmetric) logistic extreme value distribution, which has a single dependence parameter and three parameters on each of the GEV margins. Using `plot(m2)` produces several diagnostic plots, including quantile curves and spectral densities. Using `deviance(m2)` produces the deviance, which is equal to twice the negative log-likelihood. The following shows the parameter estimates and their standard errors, and gives an analysis of deviance table for testing `m2` verses `m3`, which is possible since the models are nested, with `m3` having one additional dependence parameter. The call to `exind.test` produces a score test for independence, following Tawn (1988). Omitting the `method` argument gives a likelihood ratio test, also from Tawn (1988), which is typically more accurate.

[3]The data may be completely different to the book due to random selection.

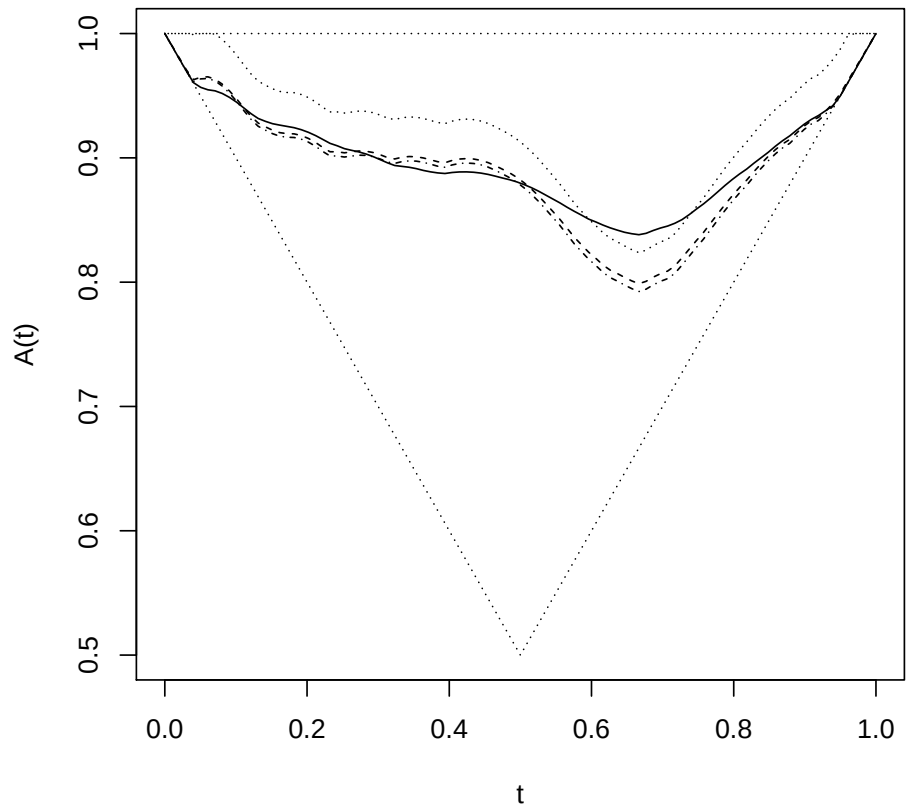

Figure 3: Nonparametric dependence function estimates by Pickands (dotted line), Deheuvels (dashed line), Hall-Tajvidi (dot-dashed line) and Capéràa-Fougères-Genest (solid line) based on componentwise block maxima data and using empirical marginal estimation.

```
> round(rbind(fitted(m2), std.errors(m2)), 3)

      loc1 scale1 shape1  loc2 scale2 shape2   dep
[1,] 2.688  1.662  0.263 0.616  0.367  0.545 0.821
[2,] 0.270  0.218  0.125 0.060  0.059  0.157 0.097

> anova(m3, m2)

Analysis of Deviance Table

   M.Df Deviance Df  Chisq Pr(>chisq)
m3    8   300.44
m2    7   302.99  1 2.5415     0.1109

> evind.test(cml, method = "score")

        Score Test Of Independence

data:  cml
```

```
norm.score = -1.8308, p-value = 0.03357
alternative hypothesis: true dependence is greater than independence
sample estimates:
      dep
0.8205622
```

The code below uses the function `qcbvnonpar` to plot quantile curves using nonparametric dependence function estimates. Quantile curves are defined as

$$Q(F, p) = \{(z_1, z_2) : F(z_1, z_2) = p\},$$

where F is a distribution function and p is a probability. We use the default nonparametric estimation method and we again use empirical estimation of the margins[4], as specified by `epmar = TRUE`. For parametric dependence models similar plots can be produced using e.g. `plot(m2, which = 5)`. Note that because we plot curves corresponding to the distribution of the original dataset rather than the componentwise maxima, we pass the argument `mint = 30`.

```
> lts <- c(0.01,100)
> plot(loss, log = "xy", col = "grey", xlim = lts, ylim = lts)
> points(cml); pp <- c(0.98,0.99,0.995)
> qcbvnonpar(pp, data = cml, epmar = TRUE, mint = 30, add = TRUE)
```

4 Excesses Over A Threshold

We now consider all the 1500 observations on liability claims. We assume that the data are distributed according to the distribution function F, and we are interested in $F(z)$ where $z = (z_1, z_2)$ is in some sense large. The methods we use assume that F is in the domain of attraction of some bivariate extreme value distribution G, and we focus on large data points to estimate features of G, and hence of $F(z)$ for large z.

Typically we focus on points z that lie above a certain threshold. The functions `tcplot` and `mrlplot` can be used for producing plots on each margin to help determine thresholds u_1 and u_2 for methods that focus primarily on points z such that $z_1 > u_1$ and $z_2 > u_2$. Alternatively, the function `bvtcplot` can be used to help determine a single threshold u^* for methods that focus on points z such that $r(z) > u^*$, where $r(z) = x_1(z_1) + x_2(z_2)$, and $x_j(z_j) = -1/\log \hat{F}_j(z_j)$ for $j = 1, 2$ where F_j is estimated empirically.

Following Segers and Vandewalle (2004), a sensible choice for threshold u^* might be found from Figure 5 by taking the kth largest $r(z)$, where k is the largest value for which the y-axis is close to two. Figure 5 is plotted below using `bvtcplot`. The value of k is returned invisibly. Setting `spectral = TRUE` uses the kth largest points to plot a nonparametric estimate of $H([0, \omega])$ where H is the spectral measure of G.

```
> k0 <- bvtcplot(loss)$k0
> bvtcplot(loss, spectral = TRUE)
```

[4]Using parametric marginal estimates tends to produce more sensible quantile curve plots, but we follow the book here. Unlike the book, the quantile curves in Figure 4 are not step functions because the empirical marginal transforms include interpolation.

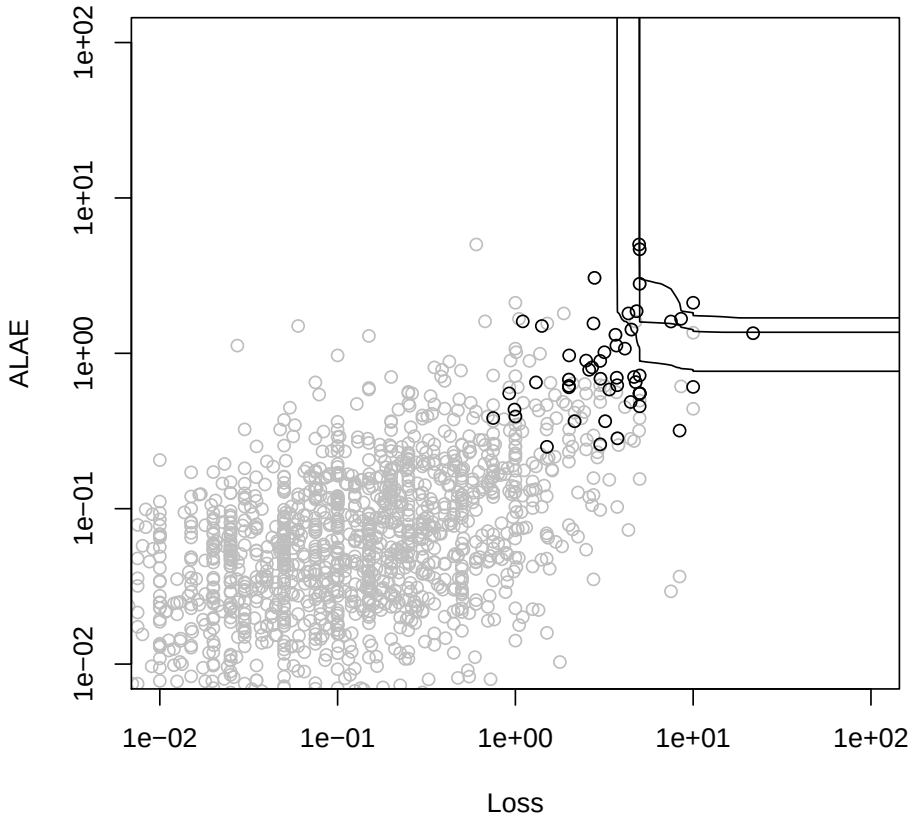

Figure 4: Estimated quantile curves $Q(\hat{F}, p)$ for $p = 0.98, 0.99, 0.995$ based on the componentwise block maxima data shown as black circles, using the Capéràa-Fougères-Genest nonparametric estimate of the dependence function and using empirical marginal estimation.

The parametric approach to the problem can employ models similar to those used for bivariate extreme value distributions. We first consider the margins separately by fitting a univariate generalized Pareto distribution to the excesses over the threshold u_j on each margin $j = 1, 2$. We choose the thresholds so that the number of exceedances is roughly[5] half of the value `k0`.

```
> thresh <- apply(loss, 2, sort, decreasing = TRUE)[(k0+5)/2,]
> mar1 <- fitted(fpot(loss[,1], thresh[1]))
> mar2 <- fitted(fpot(loss[,2], thresh[2]))
> rbind(mar1,mar2)

         scale     shape
mar1 0.8313558 0.4562441
mar2 0.2148189 0.4455429
```

Parametric threshold models can be fitted using the function `fbvpot`, with the parametric model specified using the `model` argument. The default approach uses censored likelihood methodology, where a bivariate extreme value dependence structure is fitted to the data censored at the marginal thresholds u_1 and u_2. Alternatively, a Poisson process model can be employed using the `likelihood` argument. **SENTENCE**. Some examples of parametric fits are given below. Diagnostic plots for the fitted models can be produced using e.g. `plot(m2)`.

[5]The value is chosen so that the thresholds match exactly with those used in the book.

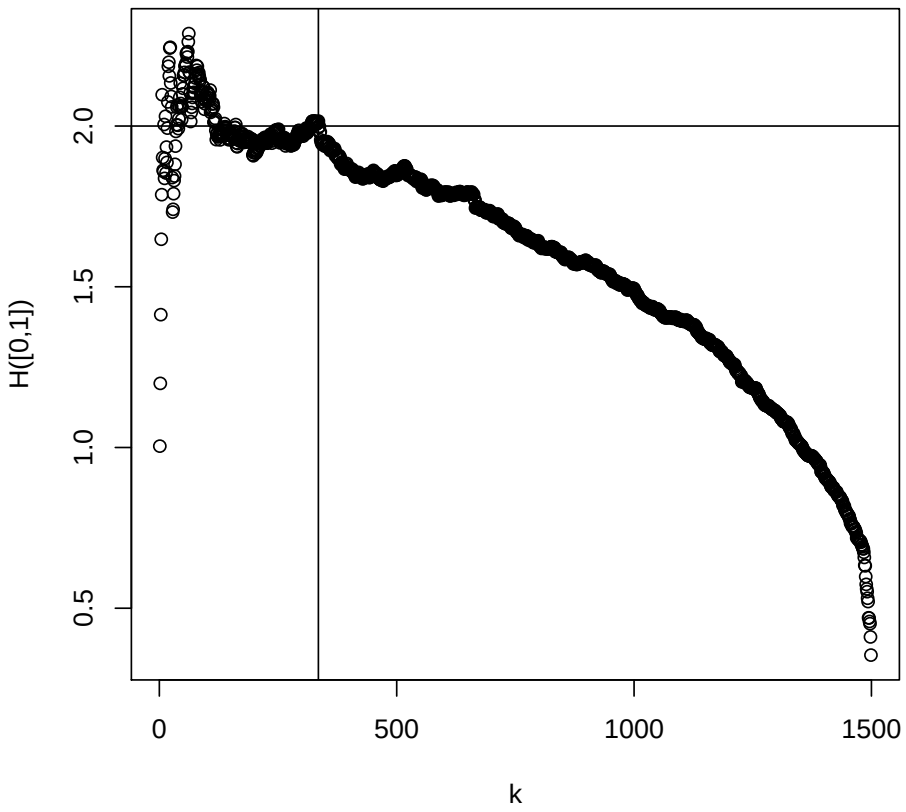

Figure 5: A plot of $(k/n)r_{(n-k)}$ as a function of k, where $r_{(1)} \leq \cdots \leq r_{(n)}$ are the ordered values of r. The y-axis provides an estimate of $H([0,1]) = 2$ for the spectral measure H of G.

```
> m1 <- fbvpot(loss, thresh, model = "alog", asy1 = 1)
> m2 <- fbvpot(loss, thresh, model = "bilog")
> m3 <- fbvpot(loss, thresh, model = "bilog", likelihood = "poisson")
> round(rbind(fitted(m2), std.errors(m2)), 3)

     scale1 shape1 scale2 shape2 alpha  beta
[1,]  0.780  0.601  0.205  0.556 0.579 0.760
[2,]  0.113  0.133  0.027  0.118 0.086 0.047
```

The following code plots parametric and nonparametric estimates for the bivariate extreme value dependence structure fitted to the upper tail of F. The parametric estimates use the previously fitted models. The nonparametric estimate can be plotted using the `"pot"` method and takes the value `k0` to specify the threshold.

```
> abvnonpar(data = loss, method = "pot", k = k0, epmar = TRUE,
+   plot = TRUE, lty = 3)
> plot(m1, which = 2, add = TRUE)
> plot(m2, which = 2, add = TRUE, lty = 4)
> plot(m3, which = 2, add = TRUE, lty = 2)
```

Figure 6 uses our fitted asymmetric logistic model `m1` to plot quantile curves at probabilities $p = 0.98, 0.99, 0.995$. The thresholds used for the censored likelihood model fit are also added to the plot.

```
> lts <- c(1e-04, 100)
> plot(loss, log = "xy", col = "grey", xlim = lts, ylim = lts)
> plot(m1, which = 3, p = c(0.95,0.975,0.99), tlty = 0, add = TRUE)
> abline(v=thresh[1], h=thresh[2])
```

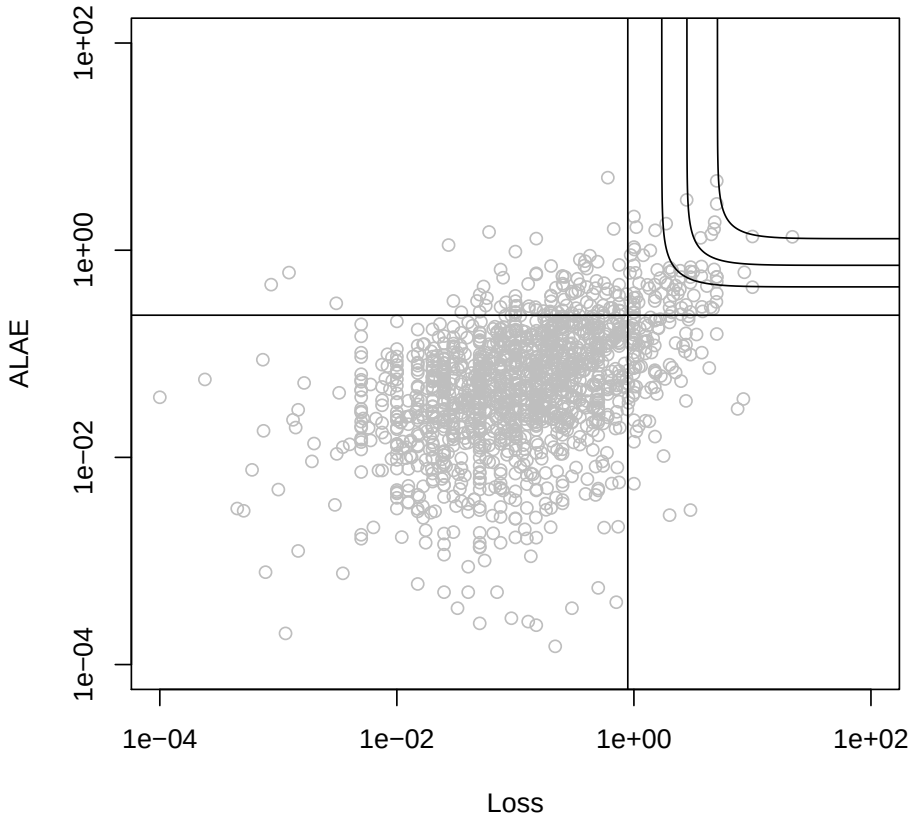

Figure 6: Quantile curves for probabilities $p = 0.98, 0.99, 0.995$ for an asymmetric logistic model fit using censored likelihood estimation, with censoring at marginal thresholds given by the vertical and horizontal lines.

Models based on bivariate extreme value distributions assume that the margins are either asymptotically dependent or are perfectly independent. They cannot account for situations where the dependence between the margins vanishes at increasingly extreme levels. The remainder of this section illustrates the estimation of dependence measures that can identify such cases.

We consider three quantities as defined in Coles *et al.* (1999). The coefficient of extremal dependence $\chi \in [0, 1]$ is the tendency for one variable to be large given that the other is large. When $\chi = 0$ the variables are asymptotically independent, and when $\chi > 0$ they are asymptotically independent. The second measure $\bar{\chi}$ identifies the strength of dependence for asymptotically independent variables. When $\bar{\chi} = 1$ the variables are asymptotically dependent, and when $-1 \leq \bar{\chi} < 1$ they are asymptotically independent. The third measure is the coefficient of tail dependence η, which satisfies $\bar{\chi} = 2\eta - 1$.

The following code produces Figure 7 which shows estimates of the functions $\chi(u)$ and $\bar{\chi}(u)$, as defined in Coles *et al.* (1999), for $0 < u < 1$. The functions are defined so that $\chi = \lim_{u\to 1} \chi(u)$ and $\bar{\chi} = \lim_{u\to 1} \bar{\chi}(u)$. In this case $\chi(u) > 0$ for all u but there is little evidence that $\bar{\chi}$ is close to one, so it is difficult to specify the form of dependence on the basis of this plot.

```
> old <- par(mfrow = c(2,1))
> chiplot(loss, ylim1 = c(-0.25,1), ylim2 = c(-0.25,1), nq = 200,
+   qlim = c(0.02,0.98), which = 1:2, spcases = TRUE)
> par(old)
```

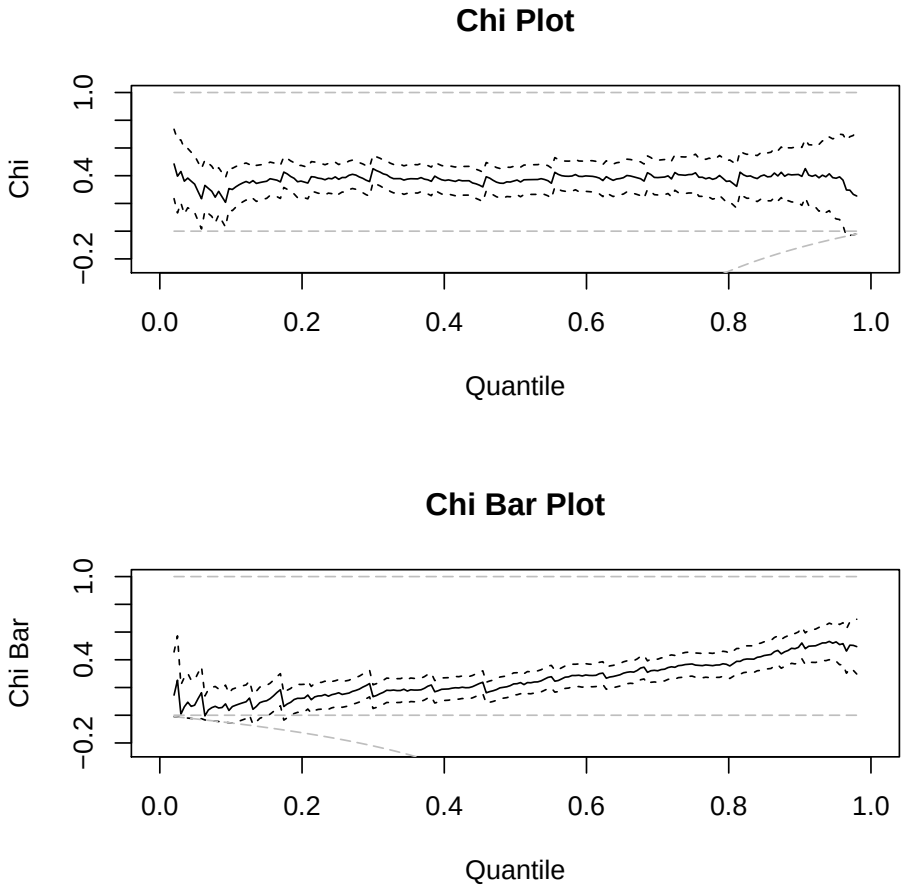

Figure 7: The dependence measures $\chi(u)$ and $\bar{\chi}(u)$. Estimates (solid line), 95% pointwise confidence intervals (dot-dashed lines). The dashed lines represent the theoretical limits of the functions and the exact independence case at zero.

We now consider the coefficient of tail dependence η. We can estimate η using univariate theory because of its relationship with $T = \min\{x_1(z_1), x_2(z_2)\}$. If we fit a generalized Pareto distribution to the data points in T that exceed a large fixed threshold, then the estimated shape parameter of the fitted distribution provides an estimate of η. The call to `tcplot` plots estimates of η at different thresholds in order to assist with threshold choice. The plot seems roughly linear after $u = 0.8$, so we take the 80th percentile of T as our threshold. Finally, we use `anova` to perform a likelihood ratio test for asymptotic dependence, with the null hypothesis $\eta = 1$ versus the alternative $\eta < 1$.

```
> fla <- apply(-1/log(ula), 1, min)
> thresh <- quantile(fla, probs = c(0.025, 0.975))
> tcplot(fla, thresh, nt = 100, pscale = TRUE, which = 2, vci = FALSE,
+   cilty = 2, type = "l", ylim = c(-0.2,1.2), ylab = "Tail Dependence")
> abline(h = c(0,1))
```

```
> thresh <- quantile(fla, probs = 0.8)
```

```
> m1 <- fpot(fla, thresh = thresh)
> cat("Tail Dependence:", fitted(m1)["shape"], "\n")

Tail Dependence: 0.7979134

> m2 <- fpot(fla, thresh = thresh, shape = 1)
> anova(m1, m2, half = TRUE)

Analysis of Deviance Table

   M.Df Deviance Df  Chisq Pr(>chisq)
m1    2   1596.5
m2    1   1599.6  1 6.2247     0.0126
```

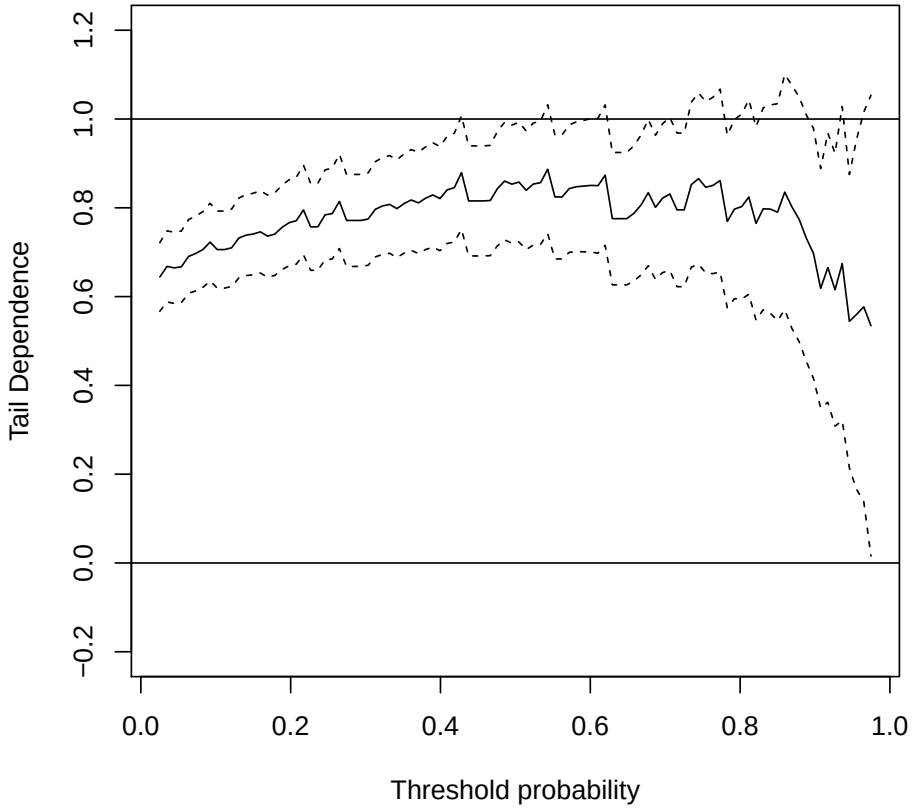

Figure 8: Maximum likelihood estimates (solid line) and 95% pointwise confidence intervals (dot-dashed lines) for η at different threshold probabilities.

POISSON LIKELIHOOD CODE AND SENTENCE TODO

Bibliography

Beirlant, J., Goegebeur, Y., Segers, J and Teugels, J. (2004) *Statistics of Extremes: Theory and Applications*. Wiley, U.K.

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

Segers, J. and Vandewalle, B. (2004). Statistics of Multivariate Extremes. In Beirlant et al. (eds.), *Statistics of Extremes: Theory and Applications*. Wiley, U.K.

Tawn, J. (1988). Bivariate extreme value theory: Models and estimation. *Biometrika*, **75**, 397–415.