The Extended Generalized Pareto Distribution 3 (EGP3) in texmex

Harry Southworth

12th March 2015

Contents

1 Introduction **2**
1.1 Acknowledgements 2
1.2 Software 2

2 Using the EGP3 distribution for extreme value modelling **2**
2.1 River Nidd example 2
2.2 Pharmaceutical example 4
2.3 Discussion 10

3 The EGP3 distribution: some technical details **10**
3.1 Distribution function, probability density function and random number generation 10
3.2 Return levels 12
3.2.1 Deriviatives 12
3.3 Upper endpoint 13

4 Appendix **14**
4.1 Information on the R session 14

1 Introduction

Version 2 of the the texmex [4] package for R [3] introduced the ability to add new families of extreme value distributions to the package and itself added the Generalized Extreme Value (GEV) distribution, previous releases having supported modelling only with the Generalized Pareto Distribution (GPD). We here describe the implementation of a a new extreme value family, the Extended Generalized Pareto Distribution 3 (EGP3) described by Papastathopoulos and Tawn [2].

The EGP3 family introduces a new parameter, $\kappa > 0$. For the purposes of numerical stability and avoiding negative values, when modelling data we work with $\lambda = \log \kappa$.

1.1 Acknowledgements

This work was partly funded by AstraZeneca. I'm also grateful to Yiannis Papastathopoulos and Paul Metcalfe for various comments and corrections.

1.2 Software

R version 3.1.1 (2014-07-10) [3] will be used for all analyses. A summary of the R session appears in the Appendix, in the interests of reproducibility.

2 Using the EGP3 distribution for extreme value modelling

The additional parameter, κ in the EGP models is allowed to vary over the positive real line, and in each case a value of $\kappa = 1$ results in a distribution identical to the GPD. This property suggests a new diagnostic plot to aid threshold selection: plot the estimated value of κ with a confidence interval over a range of thresholds and select the lowest threshold which contains the value $\hat{\kappa} = 1$. GPD modelling can then be performed on values above this threshold. This diagnostic plot provides a useful addition to the standard methods of examining the values of $\hat{\sigma}_*$ and $\hat{\xi}$ over a range of thresholds, and examining the mean residual life plot as described by Coles [1].

2.1 River Nidd example

Following Papastathopoulos and Tawn, we work with the River Nidd data, producing the standard threshold selection plots and the new plot based on EGP3.

```
library(texmex)
par(mfrow=c(2, 2))
plot(gpdRangeFit(nidd, cov="numeric", umax=90, umin=65, nint=20))
plot(mrl(nidd))
plot(egp3RangeFit(nidd, umax=90, umin=65, nint=20))
```

Figure 1 displays the results. The lowest threshold for which $\hat{\kappa}$ is similar to 1 is at about 75, suggesting that GPD models can be used above this level.

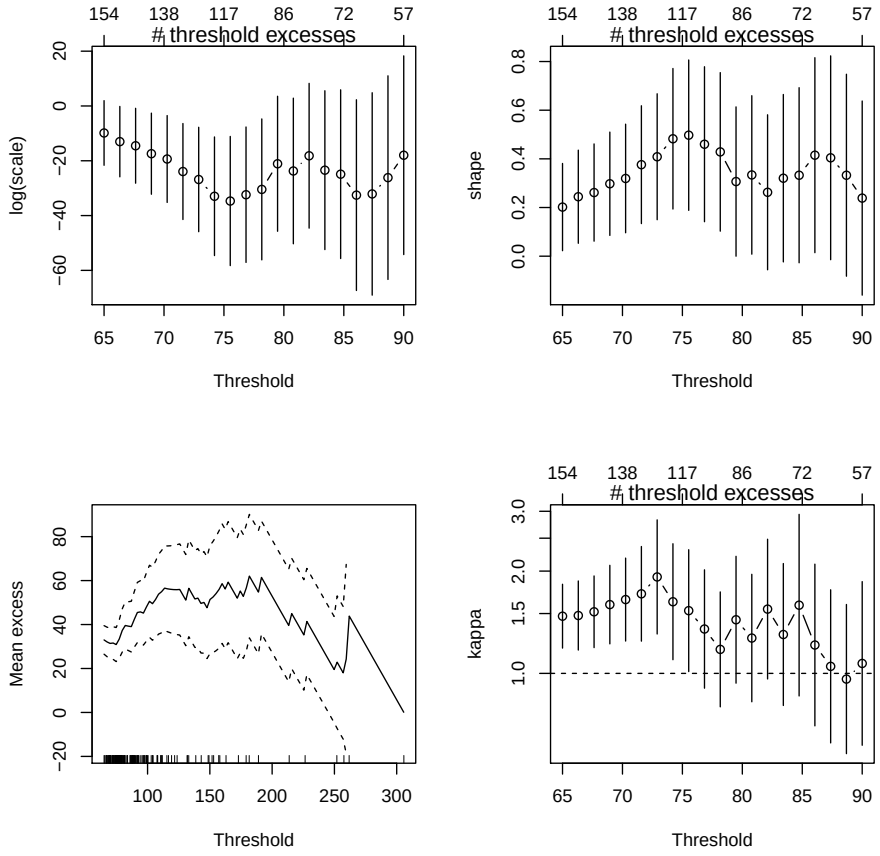

Figure 1: Threshold selection plots for the River Nidd data. The bottom right panel displays the $\hat{\kappa}$ with an approximate 95% confidence interval. The lowest value of $\hat{\kappa}$ for which the confidence interval contains 1 is approximately 75, suggesting this as a threshold above which GPD modelling can be performed.

2.2 Pharmaceutical example

The introduction of κ into the distribution suggests that lower thresholds might be usable for extreme value modelling. We now follow Southworth and Heffernan [5] in analysing some clinical trial safety data. Southworth and Heffernan model all the values of various safety related laboratory variables above the 70^{th} percentile using GPD models allowing $\hat{\xi}$ to vary linearly with dose. In each case, they find a linear relationship with dose, except for bilirubin. Following Papastathopoulos and Tawn, we revisit the issue of threshold selection, hoping to find a lower threshold above which EGP3 models can be fit, thus including more of the available data in the model, increasing the chance of identifying a dose effect if one exists.

Notice that the EGP models, like the GPD model, only exist for $x > 0$. Since we will be working with residuals from a linear model, about half of them will be negative. Therefore, we add a positive constant to them to make them all positive. The positive constant is arbitrary, so long as it is greater than the minimum of the residuals. However, in practical applications, the automatic choice of starting values for the optimizer of the likelihood can be poor for some choices of the constant.

```
library(MASS)
rmod <- rlm(log(TBL.M) ~ log(TBL.B) + as.numeric(dose),
            data=liver, method="MM", c=3.44)
liver$r <- resid(rmod) + 1.25

par(mfrow=c(2, 2))
for (dose in LETTERS[1:4]){
  plot(egp3RangeFit(liver$r[liver$dose == dose], umax=1.35))
  title(paste("Dose", dose), line=2)
}
```

The plots of $\hat{\kappa}$ over the range of thresholds in Figure 2 suggests that GPD models ought to fit above a threshold of about 1.3, the 57^{th} percentile, for each dose.

We now pool the residual bilirubin data from all doses and present the standard threshold selection plots as well as the EGP3 plot. The results appear in Figure 3 and also suggest a threshold of 1.3 ought to be ok, gaining us an additional 79 observations compared to the 70^{th} percentile used by Southworth and Heffernan.

```
par(mfrow=c(2, 2))
plot(gpdRangeFit(liver$r, nint=20))

## Fitted values of xi < -0.5
## Fitted values of xi < -0.5
## Fitted values of xi < -0.5

plot(mrl(liver$r))

## Warning:  some values will be clipped

plot(egp3RangeFit(liver$r, nint=20))
```

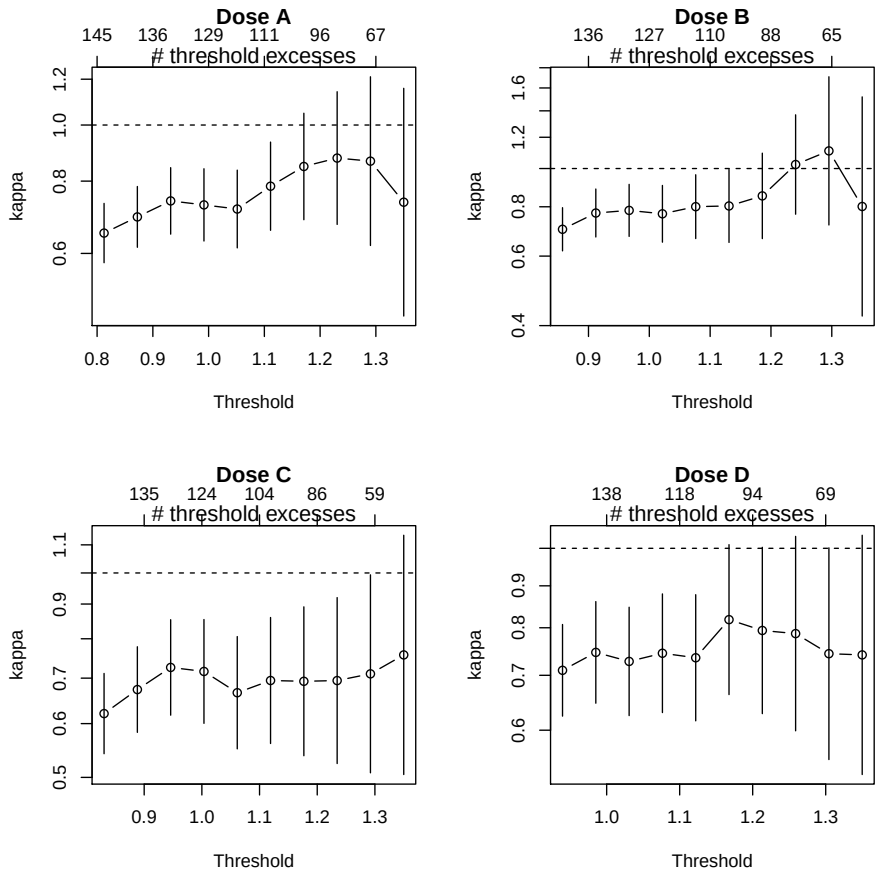

Figure 2: Diagnostic plots for the GPD model with a linear term in dose fitted to the liver data above the 70^{th} percentile.

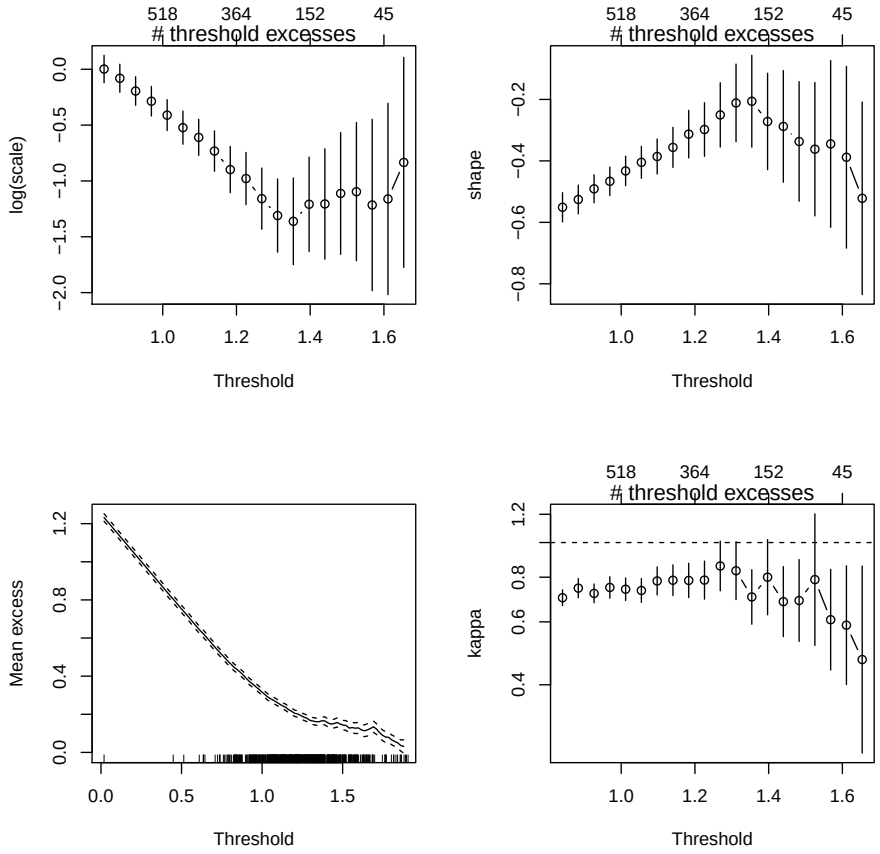

Figure 3: Threshold selection plots for the liver data. The bottom right panel displays our new plot based on the EGP3 distribution.

We now fit the GPD model and produce diagnostic plots and a summary of the model.

```
gmod <- evm(r, data=liver, th=1.3, xi=~as.numeric(dose))
par(mfrow=c(2, 2))
plot(gmod)
summary(gmod)
```

```
## Call: evm(y = r, data = liver, th = 1.3, xi = ~as.numeric(dose))
##
## Family:       GPD
##
## Model fit by maximum likelihood.
##
## Convergence: TRUE
## Threshold: 1.3
## Rate of excess: 0.429
```

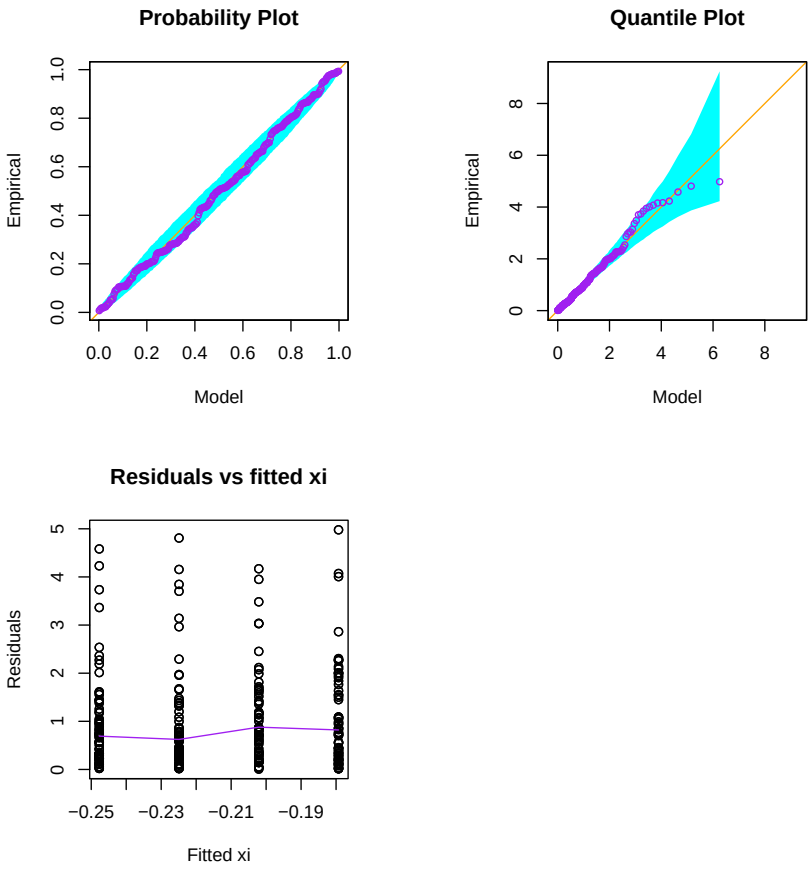

Figure 4: Diagnostic plots for the GPD model with threshold at 1.3.

```
##
## Log-lik. AIC
## 205.6   -405
##
## Coefficients:
##                         Value     SE        t
## phi: (Intercept)       -1.5785    0.0884  -17.8564
## xi: (Intercept)        -0.2705    0.0973   -2.7803
## xi: as.numeric(dose)    0.0228    0.0306    0.7454
##
## 1000 simulated data sets compared against observed data QQ-plot.
## Quantile of the observed MSE:  0.469
## 0 observations (0%) outside the 95% simulated envelope.
```

The GPD model fit to all values above 1.3 seems to fit ok, the diagnostic plots revealing no great cause for concern and all points on the QQ-plot falling inside the simualted envelope. The summary table reveals there to be still no

evidence of a dose effect, at least according to the approximate test implied by the t-value.

We now attempt to model *all* the data using the EGP3 distribution.

```
emod <- evm(r, data=liver, family=egp3, th=0,
            xi=~as.numeric(dose))
par(mfrow=c(2, 2))
plot(emod)
summary(emod)
```

```
## Call: evm(y = r, data = liver, family = egp3, th = 0, xi = ~as.numeric(dose))
##
## Family:       EGP3
##
## Model fit by maximum likelihood.
##
## Convergence: TRUE
##
## Log-lik. AIC
## -51.07   110
##
## Coefficients:
##                         Value      SE         t
## lambda: (Intercept)     2.42164    0.09360   25.87188
## phi: (Intercept)       -0.35378    0.04247   -8.32940
## xi: (Intercept)        -0.36724    0.02007  -18.29892
## xi: as.numeric(dose)    0.00546    0.00193    2.83508
##
## 1000 simulated data sets compared against observed data QQ-plot.
## Quantile of the observed MSE:  0.984
## 301 observations (49.67%) outside the 95% simulated envelope.
```

In Figure 5 there is some noticable structure in the diagnostic plots. The simulation test reveals there to be almost 50% of the observations outside of the simulated envelope, so that the model is a terrible fit to the data. We now attempt a somewhat higher threshold of 1, corresponding to -0.3 on the original scale of the residuals.

```
emod <- evm(r, data=liver, family=egp3, th=1,
            xi=~as.numeric(dose))
par(mfrow=c(2, 2))
plot(emod)
summary(emod)
```

```
## Call: evm(y = r, data = liver, family = egp3, th = 1, xi = ~as.numeric(dose))
##
## Family:       EGP3
##
## Model fit by maximum likelihood.
##
```

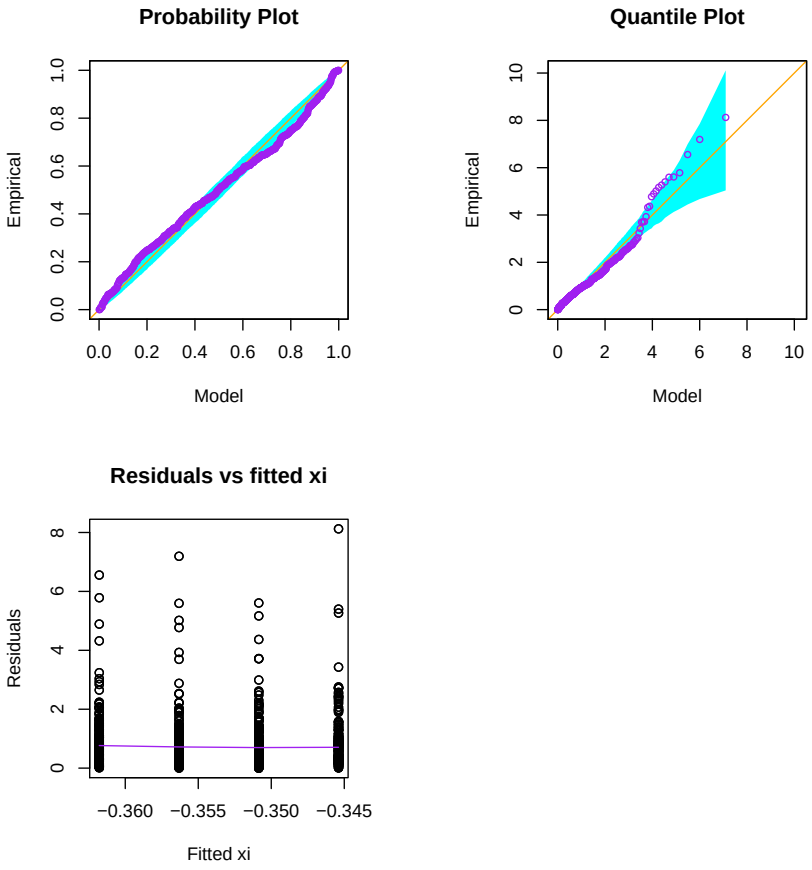

Figure 5: Diagnostic plots from fitting the EGP3 distribution to all of the residual bilirubin data.

```
## Convergence: TRUE
## Threshold: 1
## Rate of excess: 0.855
##
## Log-lik. AIC
## 155.4   -303
##
## Coefficients:
##                         Value     SE        t
## lambda: (Intercept)     0.4307    0.0715    6.0267
## phi: (Intercept)       -1.1227    0.0730  -15.3794
## xi: (Intercept)        -0.3407    0.0500   -6.8166
## xi: as.numeric(dose)    0.0145    0.0123    1.1833
##
## 1000 simulated data sets compared against observed data QQ-plot.
## Quantile of the observed MSE:  0.609
## 11 observations (2.124%) outside the 95% simulated envelope.
```

We see from the output that the model appears to fit the data, and still there is no evidence of a dose effect.

2.3 Discussion

In the pharmaceutical example, we were able to claw back some additional data into the model by using the EGP3 distribution, at the expense of an additional parameter. However, when all of the data were used, the fit was awful. It appears that the EPG3 distribution's most useful function might be to provide an extra threshold selection plot, or even a test to decide on a lower threshold. For modelling extreme values it will, at least in some examples, allow inclusion of a larger proportion of observations, but some care will need to be taken in selection of the threshold, and no obvious threshold selection methods for EPG3 are currently available.

3 The EGP3 distribution: some technical details

The EGP3 family introduces a new parameter, κ, to the familiar parameters σ and ξ in the GPD family. The threshold u remains. The EPG3 distribution function is then obtained by raising the GPD distribution function to $\kappa > 0$.

3.1 Distribution function, probability density function and random number generation

The cumulative distribution function for the EGP3 family is

$$F(x) = \begin{cases} \left[1 - \left[1 + \xi \frac{x-u}{\sigma}\right]^{-1/\xi}\right]^{\kappa} & \xi \neq 0 \\ \left[1 - \exp\left(-\frac{(x-u)}{\sigma}\right)\right]^{\kappa} & \xi = 0 \end{cases} \tag{1}$$

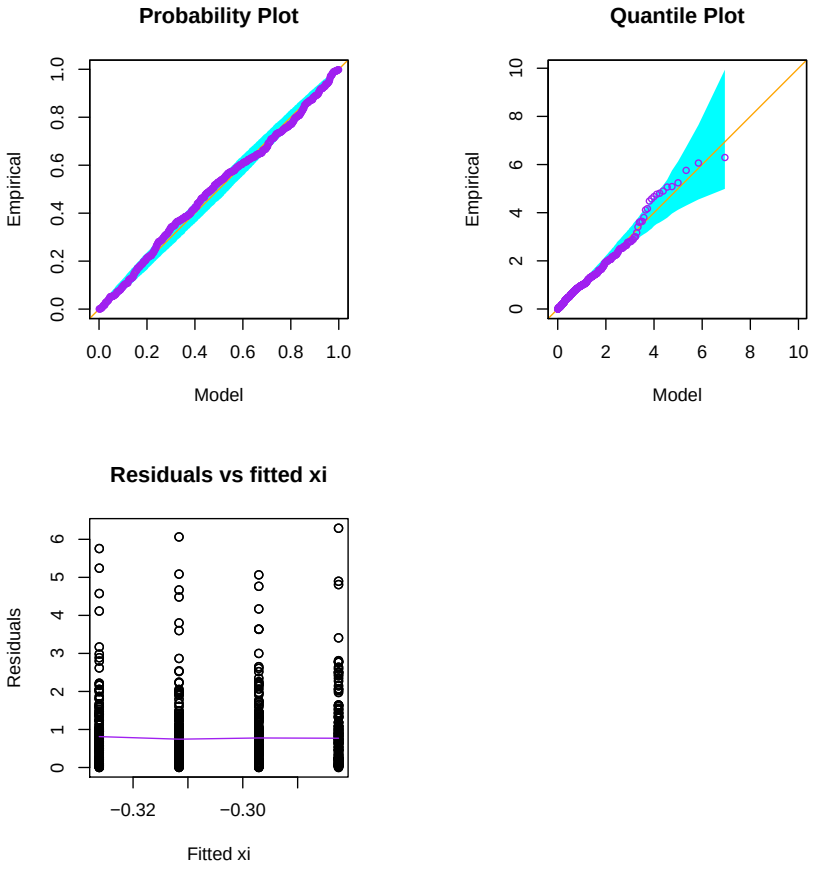

Figure 6: Diagnostic plots for the EGP3 model when using a threshold of 1.

which yields probability density function

$$f(x) = \begin{cases} \frac{\kappa}{\sigma}\left\{1-(1+\xi(x-u)/\sigma)^{-1/\xi}\right\}^{\kappa-1}(1+\xi(x-u)/\sigma)^{-1/\xi-1} & \xi \neq 0 \\ \frac{\kappa}{\sigma}e^{-(x-u)/\sigma}\left(1-e^{-(x-u)/\sigma}\right)^{\kappa-1} & \xi = 0. \end{cases} \tag{2}$$

Equations (1) and (2) are implemented in texmex in the functions `pegp3` and `degp3`.

Inversion of (1) yields

$$z = \begin{cases} u + \frac{\sigma}{\xi}\left[\left(1-x^{1/\kappa}\right)^{-\xi}-1\right] & \xi \neq 0 \\ u - \sigma\log\left(1-x^{1/\kappa}\right) & \xi = 0 \end{cases} \tag{3}$$

enabling random number generation as implemented in `regp3`.

3.2 Return levels

Following Coles ([1] Section 4.3.3) computation of return levels proceeds as follows. We note that

$$P(X > x | X > u) = 1 - F(X)_{X>u}$$

so that

$$P(X > x) = \theta_u\{1 - F(X)\}$$

where $\theta_u = P(X > u)$ for threshold u. Therefore, the level x_M that is exceeded on average once every M observations is the solution to

$$\frac{1}{M} = \theta_u\{1 - F(X)\}. \tag{4}$$

For the EGP3 distribution, we solve (4) to yield

$$x_M = \begin{cases} u + \frac{\sigma}{\xi}\left[\left\{1-\left(1-\frac{1}{M\theta_u}\right)^{1/\kappa}\right\}^{-\xi}-1\right] & \xi \neq 0 \\ u - \sigma\log\left[1-\left(1-\frac{1}{M\theta_u}\right)^{1/\kappa}\right] & \xi = 0 \end{cases} \tag{5}$$

correcting Papastathopoulos and Tawn.

3.2.1 Deriviatives

In order to compute approximate standard errors for return levels, we need derivatives of (5) with respect to each of κ, σ and ξ. These are found to be

$$\frac{dx_M}{d\kappa} = -\frac{\left(1-\frac{1}{M\theta_u}\right)^{1/\kappa}\sigma\left(1-\left(1-\frac{1}{M\theta_u}\right)^{1/\kappa}\right)^{-\xi-1}\log\left(1-\frac{1}{M\theta_u}\right)}{\kappa^2}$$

$$\frac{dx_M}{d\sigma} = \frac{\left(1-\left(1-\frac{1}{M\theta_u}\right)^{1/\kappa}\right)^{-\xi}-1}{\xi}$$

$$\frac{dx_M}{d\xi} = -\frac{\sigma\left(1-\left(1-\frac{1}{M\theta_u}\right)^{1/\kappa}\right)^{-\xi}\log\left(1-\left(1-\frac{1}{M\theta_u}\right)^{1/\kappa}\right)}{\xi} - \frac{\sigma\left(\left(1-\left(1-\frac{1}{M\theta_u}\right)^{1/\kappa}\right)^{-\xi}-1\right)}{\xi^2}$$

3.3 Upper endpoint

When $\xi < 0$, the GPD has upper endpoint $u - \frac{\sigma}{\xi}$. This value is obtained by setting the distribution function to 1 and solving. Working with (1), setting it to 1 and solving reveals the EGP3 distribution to have the same upper endpoint as the GPD.

4 Appendix

4.1 Information on the R session

Information on the R session, in the interests of reproducibility.

```
## R version 3.1.1 (2014-07-10)
## Platform: i386-w64-mingw32/i386 (32-bit)
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.1252
## [2] LC_CTYPE=English_United Kingdom.1252
## [3] LC_MONETARY=English_United Kingdom.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.1252
##
## attached base packages:
## [1] parallel  stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
## [1] MASS_7.3-33   texmex_2.3    mvtnorm_1.0-0 knitr_1.6
##
## loaded via a namespace (and not attached):
## [1] evaluate_0.5.5 formatR_1.0    highr_0.3      stringr_0.6.2
## [5] tools_3.1.1
```

References

[1] S. Coles. *An Introduction to Statistical Modelling of Extreme Values*. Springer, 2001.

[2] I. Papastathopoulos and J. A. Tawn. Extended generalised pareto models for tail estimation. *Journal of Statistical Planning and Inference*, 143:131 – 143, 2013.

[3] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[4] H. Southworth and J. E. Heffernan. *texmex: Threshold exceedences and multivariate extremes*, 2015. R package version 2.3.

[5] H. Southworth and J. E. Heffernan. Multivariate extreme value modelling of laboratory safety data from clinical studies. *Pharmaceutical Statistics*, to appear.