News

The Newsletter of the R Project

Volume 7/1, April 2007

Editorial

by Torsten Hothorn

Welcome to the first issue of R News for 2007, which follows the release of R version 2.5.0. This major revision, in addition to many other features, brings better support of JAVA and Objective C to our desks. Moreover, there is a new recommended package, **codetools**, which includes functions that automagically check R code for possible problems.

Just before the release of R 2.5.0 the fifth developer conference on "Directions in Statistical Computing" was held in Auckland, NZ, the birthplace of R. Hadley Wickham reports on the highlights of this meeting. The R user community is not only active in conferences. Volume 7, like the preceding volumes of R News since 2001, wouldn't be what it is without the outstanding support of our referees. The editorial board would like to say "Thank you!" to all who contributed criticism and encouragement during the last year—the complete list of referees in 2006 is given at the end of this issue.

The scientific part of Volume 7 starts with an article by Paul Murrell, our former editor-in-chief, on handling binary files with tools provided by the **hexView** package. Andrew Robinson teaches how R users can make use of standard Unix tools, for example `mail` for auto-generating large amounts of email (not spam!). Many of us are regularly confronted with data lacking a unique definition of missing values–the **gdata** package can help in this situation, as Gregor Gorjanc explains.

Bettina Grün and Fritz Leisch give an introduction to the **flexmix** package for finite mixture modeling, analyzing a dataset on 21 different whiskey brands. The analysis of field agricultural experiments by means of additive main effect multiplicative interactions is discussed by Andrea Onofri and Egidio Ciriciofolo. Tests and confidence intervals for ratios of means, such as ratios of regression coefficients, implemented in package **mratio** are described by Gemechis Dilba and colleagues. The **npmlreg** package for fitting random effect models is introduced by Jochen Einbeck and his co-workers. Mathieu Ribatet models peaks over a threshold by **POT**, and financial instruments like stocks or options are (back-)tested by Kyle Campbell and colleagues.

Finally, I would like to remind everyone that the next "useR!" conference is taking place in Ames, Iowa, August 8–10. I hope to see you there!

Torsten Hothorn
Ludwig–Maximilians–Universität München
Germany
`Torsten.Hothorn@R-project.org`

Contents of this issue:

POT: Modelling Peaks Over a Threshold

by Mathieu Ribatet

The Generalised Pareto Distribution (**GPD**) is the limiting distribution of normalised excesses over a threshold, as the threshold approaches the endpoint of the variable (Pickands, 1975). The **POT** package contains useful tools to perform statistical analysis for peaks over a threshold using the GPD approximation.

There is many packages devoted to the extreme value theory (**evd**, **ismev**, **evir**, ...); however, the **POT** package is specialised in peaks over threshold analysis. Moreover, this is currently the only one which proposes many estimators for the GPD. A user's guide (as a package vignette) and two demos are also included in the package.

Asymptotic Approximation

Let $X_1, \ldots, X_n$ be a series of *i.id.* random variables with common distribution function F. Let $M_n = \max\{X_1, \ldots, X_n\}$. Suppose there exists constants $a_n > 0$ and b_n such that:

$$\mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq z\right] \longrightarrow G(z), \quad n \to +\infty$$

for $z \in \mathbb{R}$ and where G is a non degenerate distribution function. Then, for $i = 1, \ldots, n$, we have:

$$\mathbb{P}\left[X_i \leq z | X_i > u\right] \longrightarrow H(z), \quad u \to u_{\text{end}} \qquad (1)$$

with

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

where (μ, σ, ξ) are the location, scale and shape parameters respectively, $\sigma > 0$ and $z_+ = \max(z, 0)$ and u_{end} is the right end-point of the variable X_i.

It is usual to fit the GPD to excesses over a (high enough) threshold. Thus we suppose that the asymptotic result given by equation (1) is (approximately) true for the threshold of interest.

Application: Ardières River at Beaujeu

The `ardieres` data frame containing flood discharges (in $m^3 \cdot s^{-1}$) over a period of 33 years of the Ardières river at Beaujeau (FRANCE) is included in the package. There are `NA` values in year 1994 as a flood event damaged record instrumentation. We use this dataset as an example for a typical univariate analysis. First, we have to "extract" independent events from the time series and select a suitable threshold such that asymptotic approximation in equation (1) is good enough.

```
library("POT")
data("ardieres")
tmp <- clust(ardieres, 0.85, tim.cond = 7/365,
             clust.max = TRUE)
par(mfrow=c(2,2))
mrlplot(tmp["obs"], xlim = c(0.85, 17))
diplot(tmp, u.range = c(0.85, 17))
tcplot(tmp["obs"], u.range = c(0.85, 17))
```

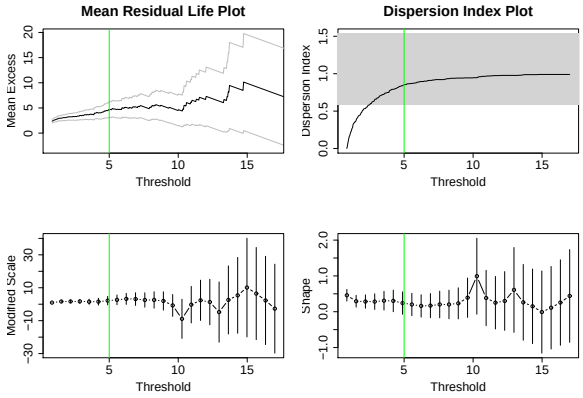

Figure 1: Tools for the threshold selection

The threshold selection stage is a compromise between bias and variance. On one hand, if a too high threshold is selected, the bias decreases as the asymptotic approximation in equation (1) is good enough while the variance increases as there is not enough data above this threshold. On the other hand, by taking a lower threshold, the variance decreases as the number of observations is larger and the bias increases as the asymptotic approximation becomes poorer.

According to Fig. 1, a threshold around five $m^3 \cdot s^{-1}$ should be a "good compromise". Indeed, the mean residual life plot is "linear" on the range $(5, 7)$; for thresholds greater than 5, the dispersion index estimates are "near" the theoretical value 1; and both modified scale and shape estimates seem to be constant on the range $(5, 9)$. Thus, we select only independent values above this threshold by invoking:

```
events <- clust(ardieres, u = 5,
    tim.cond = 7/365, clust.max = TRUE)
```

We can fit the GPD to those excesses according several estimators by setting the `method` option. There is currently 7 estimators: Moments `"moments"`, Unbiased and Biased Probability Weighted Moments `"pwmu"`, `"pwmb"`, Minimum Density Power Divergence `"mdpd"`, medians `"med"`, Pickands `"pickands"`

and Maximum Likelihood (the default) `"mle"` estimators. References for these estimators can be found in (Pickands, 1975; Hosking and Wallis, 1987; Coles, 2001; Peng and Welsh, 2001) and (Juárez and Schucany, 2004). For example, if we want to fit the GPD using the unbiased probability weighted moment estimator:

```
obs <- events[,"obs"]
pwmu <- fitgpd(obs, thresh = 5, "pwmu")
```

Here is the scale and shape parameter estimates of the GPD for the 7 estimators implemented.

```
       scale     shape
mle 2.735991 0.2779359
mom 2.840792 0.2465661
pwmu 2.668368 0.2922964
pwmb 2.704665 0.2826697
mdpd 2.709254 0.2915759
med 2.135882 0.8939585
pick 2.328240 0.6648158
```

By invoking:

```
par(mfrow=c(2,2))
plot(pwmu, npy=2.63)
```

we obtain Fig. 2 which depicts graphic tools for model diagnostic. Profile likelihood confidence intervals can also be computed, even for return levels see Fig. 3, with:

```
gpd.pfrl(mle, 0.995, range = c(20, 120),
         conf = 0.95)
```

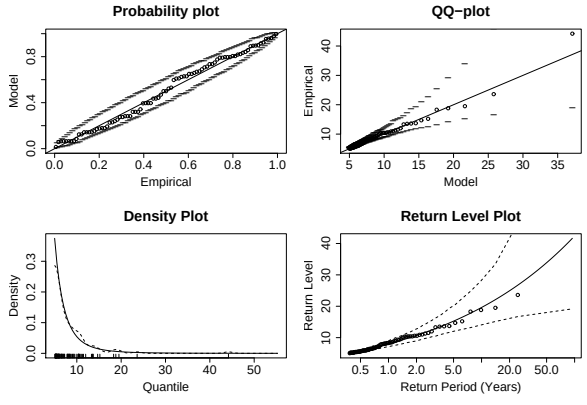

Figure 2: Graphic tools for model diagnostic.

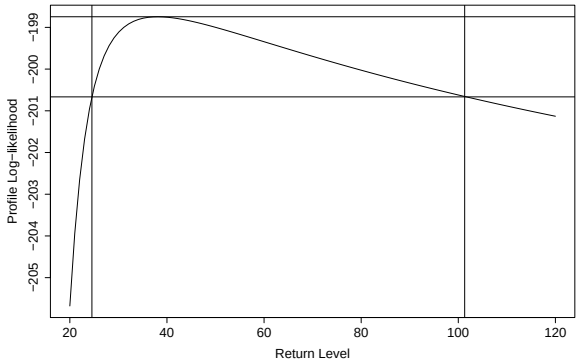

Figure 3: 95% profile confidence interval for the return level associated to non exceedance probability 0.995.

Miscellaneous Features

The **POT** package can also:

- simulate and compute density, quantile and distribution functions for the GPD;
- fit the GPD with a varying threshold using MLE;
- fit the GPD with held fixed parameters using MLE;
- perform analysis of variance for two nested models;
- estimate the extremal index using two estimators;
- display a L-Moment plot (Hosking and Wallis, 1997);
- compute sample L-moments;
- convert non exceedance probabilities to return periods and vice-versa;
- compute "averaged" time series using an average mobile window.

Currently, most of the package developments are devoted to bivariate peaks over threshold. For this purpose, the **POT** package can also:

- fit a bivariate GPD using 6 parametric dependence functions;
- fit a first order Markov chain with a fixed extreme value dependence structure to all threshold exceedances;
- simulate first order Markov chains with a fixed extreme value dependence structure;
- plot the Pickands' dependence and the spectral density functions.

Bibliography

S. Coles. *An Introduction to Statistical Modelling of Extreme Values*. Springer Series in Statistics. London, 2001. 35

J. Hosking and J. Wallis. Parameter and quantile estimation for the generalised Pareto distribution. *Technometrics*, 29(3):339–349, 1987. 35

J. Hosking and J. Wallis. *Regional Frequency Analysis*. Cambridge University Press, 1997. 35

S. Juárez and W. Schucany. Robust and efficient estimation for the generalised Pareto distribution. *Extremes*, 7(3):237–251, 2004. ISSN 13861999. 35

L. Peng and A. Welsh. Robust estimation of the generalised Pareto distribution. *Extremes*, 4(1):53–65, 2001. 35

J. Pickands. Statistical inference using extreme order statistics. *Annals of Statistics*, 3:119–131, 1975. 34, 35

Mathieu Ribatet
Cemagref Unité de Recherche HH Lyon, France
ribatet@lyon.cemagref.fr

Backtests

by Kyle Campbell, Jeff Enos, Daniel Gerlanc and David Kane

Introduction

The **backtest** package provides facilities for exploring portfolio-based conjectures about financial instruments (stocks, bonds, swaps, options, et cetera). For example, consider a claim that stocks for which analysts are raising their earnings estimates perform better than stocks for which analysts are lowering estimates. We want to examine if, on average, stocks with raised estimates have higher future returns than stocks with lowered estimates and whether this is true over various time horizons and across different categories of stocks. Colloquially, "backtest" is the term used in finance for such tests.

Background

To demonstrate the capabilities of the **backtest** package we will consider a series of examples based on a single real-world data set. StarMine[1] is a San Fransisco research company which creates quantitative equity models for stock selection. According to the company:

> StarMine Indicator is a 1-100 percentile ranking of stocks that is predictive of future analyst revisions. StarMine Indicator improves upon basic earnings revisions models by:
>
> - Explicitly considering management guidance.
> - Incorporating SmartEstimates, StarMine's superior estimates constructed by putting more weight on the most accurate analysts.
> - Using a longer-term (forward 12-month) forecast horizon (in addition to the current quarter).
>
> StarMine Indicator is positively correlated to future stock price movements. Top-decile stocks have annually outperformed bottom-decile stocks by 27 percentage points over the past ten years across all global regions.

These ranks and other attributes of stocks are in the `starmine` data frame, available as part of the **backtest** package.

```
> data("starmine")
> names(starmine)

 [1] "date"       "id"         "name"
 [4] "country"    "sector"     "cap.usd"
 [7] "size"       "smi"        "fwd.ret.1m"
[10] "fwd.ret.6m"
```

`starmine` contains selected attributes such as sector, market capitalisation, country, and various measures of return for a universe of approximately 6,000 securities. The data is on a monthly frequency from January, 1995 through November, 1995. The number of observations varies over time from a low of 4,528 in February to a high of 5,194 in November.

```
      date count
1995-01-31  4593
1995-02-28  4528
1995-03-31  4569
1995-04-30  4708
1995-05-31  4724
1995-06-30  4748
1995-07-31  4878
1995-08-31  5092
1995-09-30  5185
1995-10-31  5109
1995-11-30  5194
```

[1] See www.starmine.com for details.