extremevalues
A package for outlier detection

Version 1.0

M.P.J. van der Loo
mark.vanderloo@gmail.com, www.markvanderloo.eu

December 3, 2009

1 Introduction

This package provides the implementation of the outlier detection method as described in van der Loo (2009). Briefly, the method works as follows: given a one-dimensional dataset $\mathbf{y}$, with values $y_1, y_2, \ldots, y_N$ which is assumed to be drawn from a model distribution with cdf $F(y|\boldsymbol{\theta})$. A value y_i is considered an outlier when it's value exceeds a certain limit ℓ, which is computed as

$$\ell(\rho) = F^{-1}\left(1 - \frac{\rho}{N}\middle|\boldsymbol{\theta}\right). \tag{1}$$

Here, ρ is a parameter which is to be interpreted as the number of expected observations above ℓ, given N independent drawings. If the distribution of the bulk of the dataset is adequately described by the model distribution, a value $\rho < 1$ gives a resonable outlier limit ℓ. The vector $\boldsymbol{\theta}$ consists of the distribution parameters, and they are estimated robustly from a subset of the data between quantiles p_{min} and p_{max}, to be determined by the user.

The main purpose of the extremevalues package is to provide a function which can detect outliers using the method described above. Additionally, a plotfunction is provided for graphical analysis of the result. The package supports four model distributions:

- Lognormal distribution
- Exponential distribution
- Pareto distribution
- Weibull distribution
- Normal distribution

In this document we work through an example to familiarize the reader with the use of the package. I refer to the R help files for a complete description of the functions and to the reference mentioned above for a better explanation of the methodology. In the third section an overview of the package user and internal functions is given and the last section provides an overview of the estimation procedures.

2 A quick example

Generate some lognormally distributed data:

```
> y <- 10^rnorm(100)
```

Let's add an outlier:

```
> y <- c(y,1000)
```

And try to detect it.

```
> L <- getOutliers(y)
```

The number of detected outliers is given by `L$nOut`:

```
> L$nOut
[1] 1
```

The position of outliers in y are stored as an index:

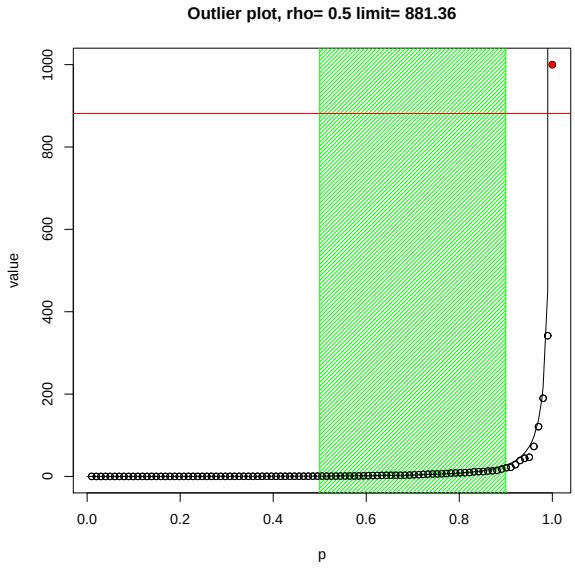
```
> y[L$iOut]
[1] 1000
```

So our added outlier is retrieved. Note that actual results may differ since y is generated randomly and may have other outliers.

To see what we have done, we can plot the results with:

```
outlierPlot(y,L,rho=0.5)
```

The resulting picture looks something like this:



Here, the circles are the sorted y-values, plotted against their estimated quantile values p. Outliers are indicated in red and the red line shows the outlier limit ℓ. The continuous line indcates the estimated cumulative model distribution, in this case the (default) lognormal distribution. The green area indicates which points have been used in the determination of the moder estimate.

Now, let's try use the exponential distribution:

```
> M <- getOutliers(y, pval=c(0.6,0.95), method="exponential")
```

The parameter pval controls which data is used in to estimate the model distribution parameter(s). In this case all observed data between the 0.6 and 0.95 quantile are used. We can check the quality of the estimate from the R^2-value of the fit:

```
> M$R2
[1] 0.454511
> L$R2
[1] 0.8366899
```

As can be expected, the exponential distribution describes the observed data less well than the lognormal distribution (since we generated lognormally distributed data). The number of outliers is also larger:

```
> M$nOut
[1] 7
```

3 Function listing

	User functions
getOutliers	Detect outliers
outlierPlot	Plot detection results
rpareto	Draw from pareto distribution
dpareto	Pareto density function
qpareto	Pareto quantile function
invErf	Inverse error function
	Internal functions
fitPareto	Fit data to cumulative pareto distribution
fitLognormal	Fit data to cumulative lognormal distribution
fitExponential	Fit data to cumulative exponential distribution
fitWeibull	Fit data to cumulative weibull distribution
fitNormal	Fit data to cumulative normal distribution
getParetoLimit	Determine outlier limit assuming pareto distribution
getLognormalLimit	Determine outlier limit assuming lognormal distribution
getExponentialLimit	Determine outlier limit assuming exponential distribution
getWeibullLimit	Determine outlier limit assuming weibull distribution
getNormalLimit	Determine outlier limit assuming normal distribution

4 Summary of fit procedures

A set of sorted real observations $\mathbf{y} = y_1, y_2, \ldots y_N$ so that $y_1 \leq y_2 \leq \ldots y_N$ can be interpreted as estimates for an underlying cdf $y_i = F^{-1}(p_i|\boldsymbol{\theta})$ where the p_i may be estimated as

$$p_i = \frac{i - 1/2}{N}. \tag{2}$$

For the outlier detection method we use the vectors $\mathbf{p} = (p_i | i \in \Lambda)$ and $(y_i | i \in \Lambda)$ to estimate $\boldsymbol{\theta}$, where Λ is an index set given by

$$\Lambda = \{i \in \{1, 2, \ldots, N\} | p_{\mathsf{min}} \leq p_i \leq p_{\mathsf{max}}\}. \tag{3}$$

For F is lognormal, pareto, weibull or normal, the parameters $\boldsymbol{\theta}$ can be estimated with regression on the pairs (p_i, y_i), so that

$$\mathbf{b} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{x}. \tag{4}$$

Here, $\mathbf{b}$ is a 2D column vector containing functions of the estimated vector $\hat{\boldsymbol{\theta}}$, $\mathbf{A}$ is a $N \times 2$ matrix containing functions of p_i and $\mathbf{x}$ is a ND vector function of $\mathbf{y}$. Table 1 gives explicit expressions for the distributions mentioned above.

The cdf for the exponential distribution is given by $1 - e^{\lambda y}$, where λ can be estimated as

$$\hat{\lambda} = -\frac{\mathbf{1}' \cdot \ln \mathbf{1} - \mathbf{p}}{\mathbf{1}' \cdot \mathbf{y}'}. \tag{5}$$

Table 1: Explicit expressions for four distributions.

Name	F	$\mathbf{b}$	$\mathbf{A}$	$\mathbf{x}$
Lognormal	$\frac{1}{2} - \frac{1}{2}\mathrm{erf}\{\frac{\ln y - \mu}{\sqrt{2}\sigma}\}$	$(\hat{\mu}, \hat{\sigma})'$	$[\mathbf{1}, \sqrt{2}\mathrm{erf}^{-1}(2\mathbf{p} - 1)]$	$\ln \mathbf{y}$
Weibull	$1 - e^{-(x/\lambda)^k}$	$(\ln \hat{\lambda}, \hat{k}^{-1})'$	$[\mathbf{1}, \ln\ln(\mathbf{1} - \mathbf{p})^{-1}]$	$\ln \mathbf{y}$
Pareto	$1 - (\frac{y_m}{y})^{\alpha}$	$(\ln \hat{y}_m, -\hat{\alpha}^{-1})'$	$[\mathbf{1}, \ln(\mathbf{1} - \mathbf{p})]$	$\ln \mathbf{y}$
Normal	$\frac{1}{2} - \frac{1}{2}\mathrm{erf}\{\frac{y - \mu}{\sqrt{2}\sigma}\}$	$(\hat{\mu}, \hat{\sigma})'$	$[\mathbf{1}, \sqrt{2}\mathrm{erf}^{-1}(2\mathbf{p} - 1)]$	$\mathbf{y}$

References

M. P. J. van der Loo. An outlier detection method for economic data. *Submitted to the Journal of Official Statistics*, 2009.