tevt : Threshold methods for Extreme Value Theory

35762970

15 April, 2021

Abstract

`tevt` is an R package which is used for GPD threshold estimation methods. It fixes several issues that other packages in the extremes ecosystem have, as well as the most comprehensive collection of threshold methods in R. Some traditional diagnostic plots are also provided. Exploring the threshold estimation methods when applied to the Danish Fire Insurance dataset, observing thresholds mostly within the interval [2,3]. Further exploration via mixture models with `evmix` give further insight into the distribution of the data, comparing a 2-tailed Kernel GPD with a boundary corrected Kernel GPD, finding no significant difference between the models. Visual inspection gives differing area of lower accuracy, further work could look to integrate these

1 Introduction

The tevt **??** package has a comprehensive suite of threshold estimation models, comparable to current alternatives of eva **??** and ismev **??**. Some

2 Background

This section gives the basic theory behind extreme value theory as well as introducing a newer topic which is not as core to EVT but is required for the understanding of this analysis.

2.1 Extreme Value Theory

Extreme value theory can be used to derive asymptotically justifiable models for the tails of distributions, this is described in more detail in [**?**]. A standard asymptotically motivated model for the exceedances of a threshold u is the generalised Pareto distribution (GPD). Letting the random variable for an exceedance X of u follow a GPD, paramterised by the scale $\sigma_u > 0$ and shape ξ, with a cumulative distribution function (CDF) given by:

$$G(x|u,\sigma_u,\xi) = Pr(X \leq x | X > u) = \begin{cases} 1 - \left[1 + \xi\left(\frac{x-u}{\sigma_u}\right)\right]_+^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp\left[-\left(\frac{x-u}{\sigma_u}\right)_+\right], & \xi = 0, \end{cases}$$

where $x_+ = \max(x,0)$. For $\xi < 0$ the support is $u < x < u - \frac{\sigma_u}{\xi}$, otherwise it is unbounded from above. In some literature the threshold is described as a location parameter.

The GPD is practically applied as an approximation to the upper tail of the population distribution above a sufficiently high threshold. [**?**] again defines some of the conditions required for the GPD to be fitted suitably.

`tevt` implements a custom distribution implementations of the Generalised Pareto Distribution, and also a handler function with which to estimate GPD parameters to data. This improves current packages in multiple areas;
- correctly handling the $\xi = 0$ case - which in other packages causes an error, or is handled incorrectly,
- improves diagnostic plots - allowing individual creation, or automatic creation without need for console input

2.2 Threshold Choice

In the traditional pipeline, the first step in create a GPD model is to allocate a threshold u, which simplistically is a bias against variance trade-off. u must be sufficiently high for the asymptotically motivated GPD to return a reliable approximation of the tail, thus reducing bias, but consequently also increases the estimation variance due to the reduced sample. In contrast, a threshold that is too low may create a GPD model which has a poor approximation of the tail, leading to bias, but providing a larger tail sample and so reducing the estimation variance. This relationship is described as the "threshold exceedance probability" or "tail fraction", which is used when calculating various quantities such as the unconditional survival probability:

$$\Pr(X > x) = \phi_u \left[1 - \Pr(X \leq x | X > u)\right].$$

A deep discussion of threshold diagnostics is covered in the review paper by [**?**]. The notable threshold estimation methods implemented within the `tevt` package are:
- `gpd.hall` - n Implementation of the procedure proposed in [**?**]
- `gpd.danielsson` - An Implementation of the procedure proposed in [**?**]
- `gpd.OSF` - a series of smaller biased Optimal Sample Fraction estimators

These build on previous packages by unifying the parameters and return values, each package that creates a new GPD function has differing parameters causing compatibility issues. In future expansions `tevt` will have conversion handlers to allow model compatibility with other packages seamlessly.

2.3 Kernel Density Estimation

Nonparametric density estimation has an important role in data analysis. In contrast to parametric models, nonparametric estimation can be used to effectively describe the complex data structuure such as bi-modality or $n-$modality without making prior assumptions about their existence. Nonparametric density estimation doesnt assume a specific functional form for the underlying population distribution, only that it "smooth", this provides far more flexibility when estimating the density. Nonparametric distribution was first popularised by [**?**], with the most popular density estimation method being kernel density estimation further discussed in pivotal pieces such as [**?**], [**?**] and the core text book [**?**].

The traditional univariate kernel density estimator is formally defined as:

$$f(\hat{x;}\mathbf{x},h)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-\mathbf{x}_i}{h}\right),$$

where $K()$ is the kernel function often defined as a symmetric and unimodal pdf and h is the bandwith parameter. Additionally the kernel function is commonly given a series of constraints; $K(x)\geq 0$ and $\int K(x)\mathrm{d}x=1$. [**?**] states that the value of the kernel estimate at point x is simply the average of the n kernel ordinates at that point. Various kernel functions have been proposed and compared in literature including uniform, normal, biweight and others, [**?**] however described that the coice of the kernel function is less critical than the choice of bandwidth size. As with traditional threshold choice, the bias-variance trade-off is present.

3 Literature Review

3.1 Extremal Mixture Models

3.2 Danish Fire Example

4 Danish Fire Example

This section will explore the applications of `tevt` to the danish insurance dataset, then transition to explore mixture models with `evmix`. The focus of applying mixture models is to examine the use of a boundary corrected model or 2-tailed model.

4.1 Introduction

The Danish dataset contains 2167 fire insurance claims in Denmark from January 1960 to December 1990. It is available within the `tevt` library in `R`. Figure 1 shows the density histogram, it has been extensively studied in various literature including [**?**] and [**?**] which apply the Peaks Over Threshold (POT) method to estimate the quantile of claim size. [**?**] uses several diagnostic plots for the independence assumption and it concludes that there is insufficient evidence to reject the independence hypothesis. He acknolwedges the work from McNeil in its applications for the insurance claim use case. In the field of Extreme Value Mixture Models, [**?**], [**?**] and [**?**] have applied their mixture model to this heavy tailed data.

Figure 1: The histogram of the Danish fire insurance data in log scale

4.2 Threshold Estimation

The focus of `tevt` is **T**hreshold methods for **E**xtreme **V**alue **T**heory, currently there are 11 threshold methods currently implemented which are summarised in Table 1, for the danish dataset, most of these give threshold estimations within [2,3.3], the Drees Kaufmann method was unable to converge. The diagnostic plot of the danielsson threshold estimation is shown in Figure 2 in both the standard form and logarithmic transform of the data.

Figure 2: Diagnostic Plot for Danielsson Threshold the Danish fire insurance data. Danish left, Log(Danish) Right

Figure 3 illustrates the mean residual life plot and Figure 4 gives the parameter stability plots. [**?**] tried some subjective thresholds using POT method. Based on these figures, 3 threshold are considered [4.5,9,10] as the initial value of the threshold for the likelihood inference. These three thresholds are corresponding to 90%, 95% and 99% quantile of the data. According to Figure 3, the threshold is chosen at the location where the mean excess above the threshold is linear. The parameter stability plot fits GPD over a certain values of thresholds against the shape and scale parameter. The threshold u should be chosen where the shape and modified scale parameter remain constant after taking the sampling variability into account. Any of the 3 stated thresholds are valid, however $u=4.5$ generates more excesses compared to 9 or 10, and will be used further.

Figure 3: The mean residual life plot and Pickand plot for the Danish fire insurance data. The red line is where $u=4.5$, green $u=9$ and red $u=10$. The dashed line shows the 95% CI. The black line represents the key metric

4.3 Fitting a Mixture Model

Comparing various mixed model methods, including Normal with GPD (NG) - with both specifications of the tail fraction, 2-tailed Normal with GPD (GNG), 2-tailed Kernel Density Estimated with GPD (GKG) in addition to Boundary Corrected Kernel Density

Table 1: tevt Threshold Methods

models	parameters	k0	threshold	tail
Gomes	1.0237, -0.7255	21	3.3055	6.5716
Danielsson	-0.371	9	3.8395	6.6791
Eye	n/a	34	3.0362	6.2964
Hall	n/a	80	2.5276	5.1348
Min Dist	n/a	230	1.6869	3.1813
Path Stability	n/a	1145	0.5431	1.4105
Drees Kaufmann	-0.2937	0	0	Inf
Guillou Hall	n/a	58	2.7652	6.1008
Reiss Thomas	n/a	57, 57	2.768, 2.768	6.0558, 6.0558
dAMSE	1.0237, -0.7255	124	2.1461	3.7708
Himp	1.0237, -0.7255	20	3.3083	6.791

Figure 4: The parameter stability plot for the Danish fire insurance data. The red line is where $u=4.5$, green $u=9$ and red $u=10$. The dashed line shows the 95% CI. The black line represents the key metric

Esimated (BCK). The modelling leverages functions from [**?**]. Each of these functions are not assigned a threshold, but each individually operate a grid search over a sequence of threshold to locate the value that maximises the profile likelihood, which is then used to calculate other non-threshold parameters. An important distinction is the comparison of boundary-corrected

Figure 5 compares the fitted density of a normal distribution and an extreme value mixture model with a normal for the bulk and a GPD for the tail. Variants of this with both specifications of the tail fraction have been fitted. The usual unbiased estimates of the normal parameters are $\mu = 0.787$ and $\sigma = 0.7167$. The poor fit of this model is clear as the red line is far from the sample histogram density estimator - this is due to both empirical tails being heavier than the asymptotically exponential form of the normal distribution. This does not occur with the bulk model, shown in blue, and parametrised tail fraction, shown in dashed green, which both provide almost identical density fits. The bulk model based tail fraction does not modify the density below the threshold, so density estimations are similar when below the threshold. The estimated normal bulk parameters are $\hat{\mu} = 0.238$ and $\hat{\sigma} = 0.1041$ which are somewhat similar to the unbiased estimators stated previously.

Figure 5: Density histogram of Danish fire insurance data. A fitted normal distribution is shown by the red solid line. The fitted density function of the NG with bulk model based (blue solid line) and parameterized tail fraction (green dashed line).

Figure 6: Return level plots for NG model fits with bulk model based tail fraction (left) and parameterized tail fraction (right).

The parameterised tail fraction provides an extra degree of freedom used to re-scale both the tail and bulk components. The tail fit has improved, as the tail fraction is estimated from the sample fraction of exceedances so is not biased by the poor bulk model which undermines the bulk model based tail fraction estimate. The re-scaling of the bulk density allows the mixture to provide a much better fit below the estimated threshold. Both the bulk model and parameterised tail fraction approaches produce similar fits for the highest return periods above 1000, shown in Figure 6. However, when nearing the threshold, represented by the horizontal line, the fit is better for the parameterised tail fraction approach, shown by the distance between the data points and red line.

The bulk model not only affects the tail fraction, but also the threshold which is biased downwards. It is lower than indicated by the traditional graphical diagnostics, as dicussed previously and shown in Table 1. The estimated threshold are the lowest values considered in the profile grid search. The low thresholds allow the GPD to overcome as much of the poor tail fit as it is capable - this result is an indictor of a poor bulk model. A potential problem that could also occur is local maxima/minima in which some optimisation methods may get stuck, this is an ongoing area of research and has not been well explored. A future compatability update with `tevt` will explore optimisation algorithms with relation to methods from `evmix`.

A marginally better model is obtained by using a GPD for both tails, this attempts to capture the empirical heavy tail. Shown in Figure 7, the density is a better match at the start and end, but fails to capture the central section correctly. This model has 3 extra parameters, lower tail threshold and GPD parameters, compared to the previous model leading to a deviance increase of 20.5579 which provides some level of validation as it above the $\chi^2_3(0.95) = 7.815$ critical value.

The semi-parametric and non-parametric bulk models are more computationally complex than the parametric alternatives, the

mixture of gammas bulk model utilises the EM algorithm, which significantly increases computation time. The non-parametric KDE bulk model also requires significant computing time for a cross-validation likelihood calculation.

Figure 7: Density histogram of Danish fire insurance data. The fitted density function of the GNG with bulk model based tail fraction.

Figure 8: Density histogram of Danish fire insurance data overlaid with the fitted GKG with bulk model based tail fraction density (left) and the return level plot (right).

In an attempt to reduce computation times, using a standard KDE between the upper and lower threshold (using the thresholds obtained from the previous model). The fitted density and return level plot are shown in Figure 8, the density has a superior fit to that of the GNG model, with the return level plot also showing the Monte Carlo uncertainty intervals. The shape parameter estimate for the upper tail is -0.0501133 whilst the upper tail fit shown in the return level plot is also strong.

The final model considers a boundary corrected KDE, Figure 9 gives the Danish data when estimating the bandwidth using maximum cross-validated likelihood estimation with the reflection method. The reflection method is known to be the most computationally efficient of the boundary correction methods in `evmix` (as stated in [**?**]). The standard KDE, shown in green, is very biased near the boundary, the reflection method, in red, still has minor bias with the simple method returning the best results at the cost of computation. The return plot for this method is similar to that of previous models with the exception of the distinctive flick around $m = 900$, which is not seen in any other model. The confidence intervals display this action too.

Figure 9: Density histogram of Danish fire insurance data overlaid with the fitted Boundary Corrected Kernel Density Estimation with density (left) and the return level plot (right).

The diagnostic plots of the GKG model and BCK model are shown in Figure **??**, the confidence intervals on the probability plot are far tighter on the BCK model than the GKG model, which sees some small deviations from the expected red line. Additionally the density plot of the GKG model fits very well to the underlying histogram represented by the green and black line fitting very closely. The BCK model is different in that the black line matches the underlying distribution greater, but does not match the green line. The quantities plot show that the BCK sits far closer to the red line, with less deviations than the GKG model.

The question of which model is best fit for the Danish dataset is better stated as which type of model is preferred to be used, as the log-liklihood values are 1646.857193 and 1629.8831103 respectively - which is a statistically insignificant difference. The initial histogram of data, Figure 1, shows that when the data is presented on a log scale, the peak density is not at 0, it is unknown if the data before the peak follows a transformed normal distribution, this is what motivates the use of 2-tailed models such as the GNG and GKG. However these methods are suited to distributions which can become negative, such as financial markets, so naturally a boundary corrected KDE model is far more suited to the data, as there is an explicit boundary at 0, as a claim cannot be below 0. The question of the matter, is whether an extreme model is applicable to the lower end of data, whereby the claimsize is very small or whether this data can naturally fit into a normal distribution, the conclusion of that requires further investigation.

The results indicate that the kernel density estimator based models are superior or close to the best performer, at the expense of a huge computational burden, with the return plot for BCK requiring 4.5 hours of computation time. If the bulk model is correctly specified, a bulk model is preferred, but otherwise using bulk models are a massive disadvantage. A KDE based approach does not always fall into these problems as it makes no assumptions and is cross-validated.

5 Conclusion

This report thoroughly evaluates the application of both threshold estimation and mixture models to the Danish dataset, including implementing GPD parameter estimation and various threshold estimation methods in an `R` package. The purpose of the `R` package is to extend functionality within current packages, and allowing absent function calls (such as those within RMarkdown or RSweave). The `R` package is available on github to increase usability of such models and allow the package to fit into the EVT ecosystem.

Initially the EVT basics were reviewed briefly including a basic GPD definition, how threshold estimation operates in principle and also exploring kernel density estimation, which is required for the understanding of mixture models, each section briefly mentions where `tevt` fits into the ecosystem. Some previous important literature is briefly discussed in regards to the application to the Danish dataset

The application of a 2-tailed normal GPG by [**?**], 2-tailed kernel GPD by [**?**] and the BCK GPD by [**?**] provide extensive coverage of the current mixed method literature.

Further work could include exploring further threshold methods, extending the capabilities of `tevt`, and as previously mentioned extending some compatibility layers for package interconnectivity. Exploring more advanced mixture models applications to danish dataset could see increased computation performance which is desirable given how long it takes. Further more, exploring the nature of the lower tail of the Danish dataset is required to establish which model would be the best fit for the underlying distribution.