Application of the Kolmogorov-Smirnov test to estimate the threshold when estimating the extreme value index

J.M. van Zyl

Abstract: The Pareto distribution model assumption in the peaks over threshold method, will be tested by making using of the Kolmogorov-Smirnov goodness of fit method. Pareto distributed variables can be transformed to exponential, and the test will be for exponentiality. It was found that the statistic can be used as an indication of where to choose the threshold and to check the Pareto model assumption.

Keywords: Extreme value index, Pareto model, Kolmogorov-Smirnov test

1. Introduction.

The Kolmogorov-Smirnov test (K-S test) is a nonparametric test which can be modified to test goodness of fit. A Pareto distributed random variable can be transformed to an exponential distributed variable. The specific case of testing for an exponential distribution using the K-S test has been thoroughly investigated also when the parameter is unknown and estimated.

In the peak over threshold (POT) method (Davidson and Smith, 1990), the Pareto model is often used. It is assumed that for x larger than a threshold c,

$$1-F(x)=(x/c)^{-\alpha},\ x>c>0,\alpha\geq 0,$$

where α is called the index of the distribution and the parameter of interest to estimate. The largest m values in a sample of size n, will be transformed to exponential observations. The K-S test will be applied to

check the Pareto model assumption and also to choose a threshold. It was found that the transformation from Pareto to exponential is good when using estimated parameters to perform the transformation, even in small samples.

The K-S test is then used to test if the transformed observations, based on the maximum likelihood estimators are exponential, and thus the original observations Pareto distributed. The choice of the threshold using the K-S statistic will be investigated. The m largest observations which distribution is closest to the Pareto distribution with respect to the K-S test, will be used to estimate the threshold and the index.

Graphical methods are often used to identify the threshold, for example the QQ and the Hill plot (Drees, de Haan, Resnick, 2000). The K-S test has the advantage of not only giving an indication of in which region to choose the threshold, but also to check if the largest observations exhibit Pareto type behaviour.

The Kolmogorov-Smirnov statistic is defined as

$$d_n = \sqrt{n} \sup_x |F_n(x) - F_{\hat{\alpha}}(x)|,$$

the empirical distribution function is denoted by $F_n(x)$, and the estimated distribution function, based on the maximum likelihood estimators, in a sample of size n, denoted by $F_{\hat{\alpha}}(x)$. The large sample properties of testing for an exponential distribution, using the K-S statistic is given by Haywood and Khmaladze (2008). The large sample distribution function of d_n under the null hypothesis of exponentiality, when the parameters are estimated is $F(d_n) = 2\Phi(d_n) - 1$, where Φ denotes the standard normal distribution.

A distribution F is heavy tailed when the distribution has the behaviour that for large x, L(x) a slowly varying function, if

$$1 - F(x) \sim x^{-\alpha} L(x), \ \alpha > 0, x \to \infty .$$

The Pareto model assumes that

$$1-F(x)=(x/c)^{-\alpha},\ x>c>0,\alpha\geq 0,$$

where the extreme value index is denoted by $\gamma=\alpha^{-1}$. The parameter α will be referred to as the index. Applications to financial data can be found, for example in the paper by Rytgaard (1990). The Pareto density function is

$$f(x)=\alpha c^{\alpha}x^{-(\alpha+1)},\ x\geq c.$$

The Pareto distribution has the property that $\log(x/c)$ is exponentially distributed with expected value $1/\alpha$. The maximum likelihood estimators of c and α in a sample of size n+1 Pareto distributed observations, denoted by $x_{(1)}\leq\ldots\leq x_{(n+1)}$, is

$$\hat{\alpha}=n/\sum_{j=2}^{n+1}\log(x_{(j)}/\hat{c}).$$

The maximum likelihood estimator of c is $\hat{c}=x_{(1)}$ and $\hat{\alpha}=n/\sum_{j=2}^{n+1}\log(x_{(j)}/\hat{c})$, $\hat{c}$ and $\hat{\alpha}$ was shown to be consistent estimators of c and α (Johnson, Kotz, Balakrishnan, 1994, p582).

The assumption will be made that if the observations in the tail are Pareto distributed with unknown parameters, that $\log(x_{(j)}/\hat{c}),\ j=2,\ldots,n+1$, is approximately exponentially distributed with expected value $1/\alpha$. Assuming that the transformed observations are exponentially distributed, it follows that the maximum likelihood estimator of the parameter is $\hat{\alpha}=n/\sum_{j=2}^{n}\log(x_{(j)}/\hat{c})$, as in the Pareto model. Using the properties of the Pareto likelihood estimators it follows that asymptotically

$E(\sum_{j=1}^{n}\log(x_{(j)}/\hat{c})/n)=1/\alpha$ and $\mathrm{var}(\sum_{j=1}^{n}\log(x_{(j)}/\hat{c}))=n/\alpha^2$, which shows that the first two moments are similar to that of an exponential distribution. A simulation study shows that even for small samples the distribution of the transformed Pareto observations is very close to exponential.

The choice of the threshold is made where the estimated exponential distribution function is closest to the empirical distribution function with respect to the K-S statistic. The distribution function is estimated by making use of the maximum likelihood estimators for a sample size m out of n. Thus when the K-S statistic of the transformed observations $\log(x_{(j)}/\hat{c})$, j=1,…,m is a minimum with respect to the K-S statistic when testing for exponentiality. The K-S statistic can be plotted for various values of m, $d_m=\sqrt{m}\sup_x|F_m(x)-F_{\hat{\alpha}}(x)|$ and the smallest value of d_m corresponding to the threshold $x_{(n-m+1)}$ is used. The p-value for the exponential hypothesis is $F(d_m)=2\Phi(d_m)-1$ and can be used to test if the Pareto model is valid.

The most popular estimator used in peak-over-threshold problems is the Hill estimator which is estimating $\alpha^{-1}=\gamma$. Suppose the largest m in a sample of size n is used, the the Hill estimator is

$$\hat{\gamma}=\frac{1}{m-1}\sum_{j=1}^{m-1}\log(x_{(j)}/\hat{c}),\quad \hat{c}=\min(x_{(1)},...,x_{(m)}),\ (\gamma=1/\alpha).$$

This estimator is consistent but can be biased (Pictet, Dacorogna, Müller, 1998). It can be derived without assuming the strict Pareto model. It is the inverse of the maximum likelihood estimator of α after transforming to exponentially distributed observations and assuming c is unknown.

2. Checking the Pareto assumption and choosing the threshold

Plots were made for various sample sizes n. Between m=50 and m=250 largest observations were chosen. The absolute values of negative observations which are symmetrically distributed around zero were used. Shown are results from t-distributed samples, stable distribution with index 1 (Cauchy) and index 1.5.

In the figure 1 the p-value for the null hypothesis of exponentiality and the K-S statistic in a sample of size n=1000 is shown. It can be seen that for large m, the hypothesis of exponentiality and thus the Pareto model would be rejected.

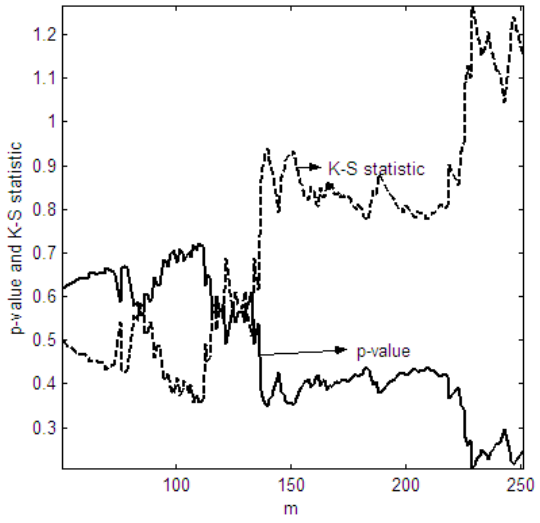

Figure 1. The p-value and estimated index as a function of m, n=1000, t_4 distribution. Best estimate when m=110.

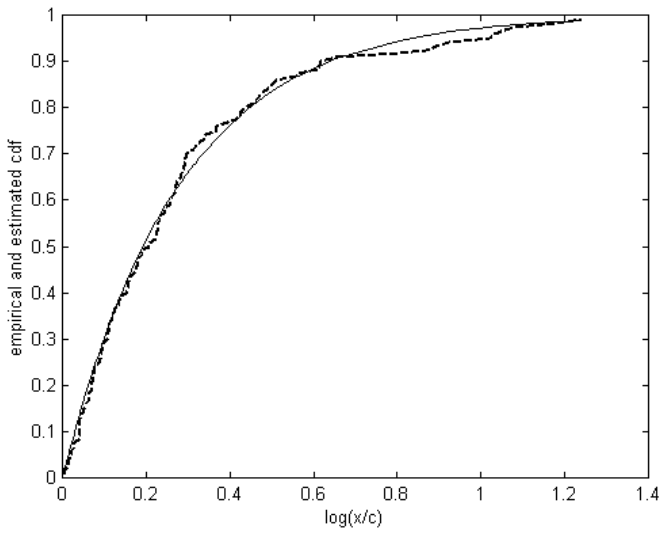

Figure 2. The empirical and estimated cdf of the transformed tail observations, n=1000, t_4 distribution. Best estimate at m=119, estimated 3.5998. Index is 4.

In the following figures the K-S statistic is plotted for various values of m, and the estimated index at those points. The tails of the t_6 distribution is not very heavy and it can be seen that the Pareto model is only valid in the largest values. The K-S statistic lead to a reasonable choice of the threshold.

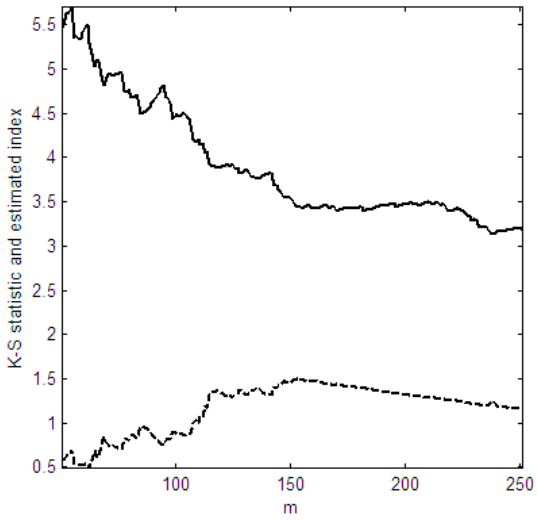

Figure 3. K-S statistic and p-value as a function of m, n=2500 t_6 distribution. Best estimate with m=63, estimated 5.2897. Index is 6. Solid line the estimated index and dashed line K-S statistic.

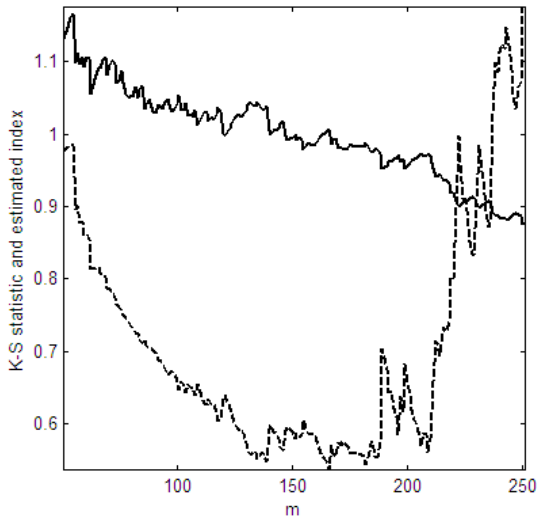

Figure 4. K-S statistic and p-value as a function of m, n=500, stable distribution, index is 1. Best estimate with m=166, estimated index 1.0078. Solid line the estimated index and dashed line K-S statistic.

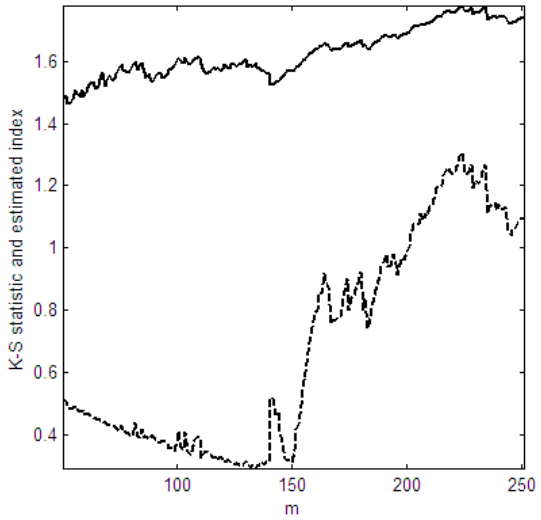

Figure 5. K-S statistic and p-value as a function of m, n=1000, stable distribution, $\alpha = 1.5$. Best estimate at m=134, estimated 1.5937. Solid line the estimated index and dashed line K-S statistic.

It can be seen that in this heavy tailed distribution Cauchy distribution, the Pareto model is valid for large values of m, even in small sample size of

n=500. The estimate of the index is not very sensitive to the choice of the threshold, confirmed by the K-S statistic. A lighter tail, and even for the larger sample size of n=1000, the Pareto model is only valid in the largest few vales, or a small value of m.

2. Simulation study results

The exponential approximation using estimated parameters to transform Pareto observations will be considered first. For various sample sizes n, Pareto observations simulated and transformed to exponential $\log(x_{(j)}/\hat{c})$, j=2,…,n, where $\hat{c}$ is the smallest value in the Pareto distributed sample. The transformed sample was tested for exponentiality, using the K-S test and $\hat{\alpha} = n/\sum_{j=1}^{n}\log(x_{(j)}/\hat{c})$. The expected value of the transformed variables should be $1/\alpha$ and with variance $1/\alpha^2$. The results of 5000 repetitions, is given in table 1. The RMSE is calculated with respect to the true parameter, $\alpha = 2$. The results give an indication that the transformation works well, even for small sample and could be used when n>25.

n	Sample mean ($1/\alpha$=0.5)	Sample variance ($1/\alpha^2 = 0.25$)	Mean K-S Statistic	Mean p-value	RMSE
15	0.4991	0.2431	0.00006	0.5943	1.7647
25	0.4978	0.2466	0.0001	0.5940	1.7685
50	0.4986	0.2491	0.0001	0.5356	1.5763
100	0.5009	0.2512	0.0001	0.5164	1.5315
250	0.5000	0.2499	0.0001	0.5000	1.5133

Table 1. Estimated parameters and average values of the K-S test in 5000 repetitions. Data transformed using estimated parameters.

The estimate when using the K-S threshold was compared with the estimates when using a constant m=50 and m=100 highest observations. The RMSE is calculated with respect to the index or $\hat{\alpha} = 1/\hat{\gamma}$. The minimum number of largest observations included was 30 with a maximum

of m=200. The study was conducted for t_4, t_6, t_{10} , Cauchy and from a Generalized Pareto distribution with $\alpha = 1/0.3333, \alpha = 1/0.25$. The index and threshold was calculated from samples of sizes n=500 and the results is the RMSE of 500 repetitions.

Distribution	**Mean m K-S method**	**RMSE K-S method**	**RSMSE fixed m=50**	**RMSE fixed m=100**
t_4	59.3100	1.0814	1.4193	1.8073
t_6	54.7240	2.4046	3.0318	3.5586
t_8	53	4.0311	4.8484	5.4241
Cauchy	80.4440	1.1573	1.1081	0.0854
GPD ($\alpha = 3$)	65.6160	0.6985	0.8332	0.9846
GPD ($\alpha = 4$)	68.3580	1.3235	1.5414	1.7600

Table 2. Comparison of the various the RMSE's with the K-S method. Sample size n=500, 500 repetitions.

Distribution	**Mean m K-S method**	**RMSE K-S method**	**RSMSE fixed m=50**	**RMSE Fixed m=100**
t_4	58.9080	1.0869	1.4153	1.8047
t_6	54.2370	2.4108	3.0337	3.5633
t_8	53.5430	4.0162	4.8196	5.4368
Cauchy	77.8500	1.1747	1.1174	0.0924
GPD ($\alpha = 3$)	69.9050	0.7163	0.8273	0.9778
GPD ($\alpha = 4$)	66.9350	1.3268	1.5517	1.7596

Table 3. Comparison of the various the RMSE's with the K-S method. Sample size n=1000, 500 repetitions.

The K-S method gives a good choice of a threshold, but performs weaker in the Cauchy data sample. For the t-distribution with 6 and 8 degrees of freedom, this method performs well. Overall it seems the method performed better with smaller sample sizes and where the tails are heavy,

but say with $\alpha > 4$. In such problems the K-S statistic is a good method to choose an threshold.

3. Conclusions

The Kolmogorov-Smirnov statistic can be used as an indication of whether the Pareto model is valid in the largest values of a sample, when estimating the index. It also gives a good indication where the Hill estimator is best, and can thus be used to get an indication of where to choose the threshold. The rate of convergence to a Pareto distribution of the complement of the distribution function in the tails is dependent on the distribution involved, and in the case of unknown distribution it would be important to check the Pareto principle.

An interesting aspect that in many cases an excellent fit to the tails was found using the Pareto assumption, but the estimated index was very biased, which can show that for a given distribution, the estimation method may be good, but the index is a function of n, m and x, and only reaches the true value of the index in the limit.

References;

Davidson, A.C. and Smith, R.L. (1990). Models of Exceedances Over High Thresholds, *Journal of the Royal Statistical Society* 52 (Ser. B), 393-442

Drees, H., de Haan, L., Resnick, S. (2000). How to make a Hill Plot, *Annals of Statistics*, 28(1), 254 –274.

Haywood, J., Khmaladze, E. (2008). On distribution-free goodness-of-fit testing of exponentiality, *Journal of Econometrics*, 143, 5 – 18.

Johnson, L. , Kotz, S., Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Volume 1, Wiley-Interscience, New York.

Pictet, O.V., Dacorogna, M.M., Müller,U.A. (1998). *Estimators for heavy tails*, in A Prcatical Guide to Heavy Tails (Adler, R., Feldman, R, Raqqu, M., eds.),Birkhäuser, Boston.

Rytgaard, M. (1990). Estimation in the Pareto Distribution, *Astin Bulletin*, 20(2), 201 –216.