TAIL INFERENCE: WHERE DOES THE TAIL BEGIN?

Tilo Nguyen and Gennady Samorodnitsky

October 2014

Estimating the exponent of regular variation

Recall that a univariate distribution (function) F is said to have a regularly varying right tail of index $\alpha > 0$ if the tail function $\bar{F} = 1 - F$ satisfies

$$\lim_{x \to \infty} \frac{\bar{F}(tx)}{\bar{F}(x)} = t^{-\alpha}$$

The index α measures the heaviness of the tail and estimating it is of crucial importance in many applications of stochastic models.

A number of estimators have been designed for that purpose.

The best known estimator of the tail index is the Hill estimator, introduced by Hill (1975).

Let $X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{n,n}$ be the order statistics from a positive sample (or from the positive part of a general sample) $X_1, \ldots, X_n$.

The Hill estimator based on k upper order statistics is defined as

$$H_{k,n} := \frac{1}{k} \sum_{i=0}^{k-1} \log \frac{X_{n-i,n}}{X_{n-k,n}}$$

Suppose that the original observations form an i.i.d. sample from a distribution with a regularly varying right tail with tail index α. If

$$n \to \infty, \ k \to \infty, \ \frac{k}{n} \to 0\,,$$

then

$$H_{k,n} \to \gamma = \frac{1}{\alpha} \text{ in probability}$$

(Mason (1982)).

If, additionally, $k/\log\log n \to \infty$, then we even have

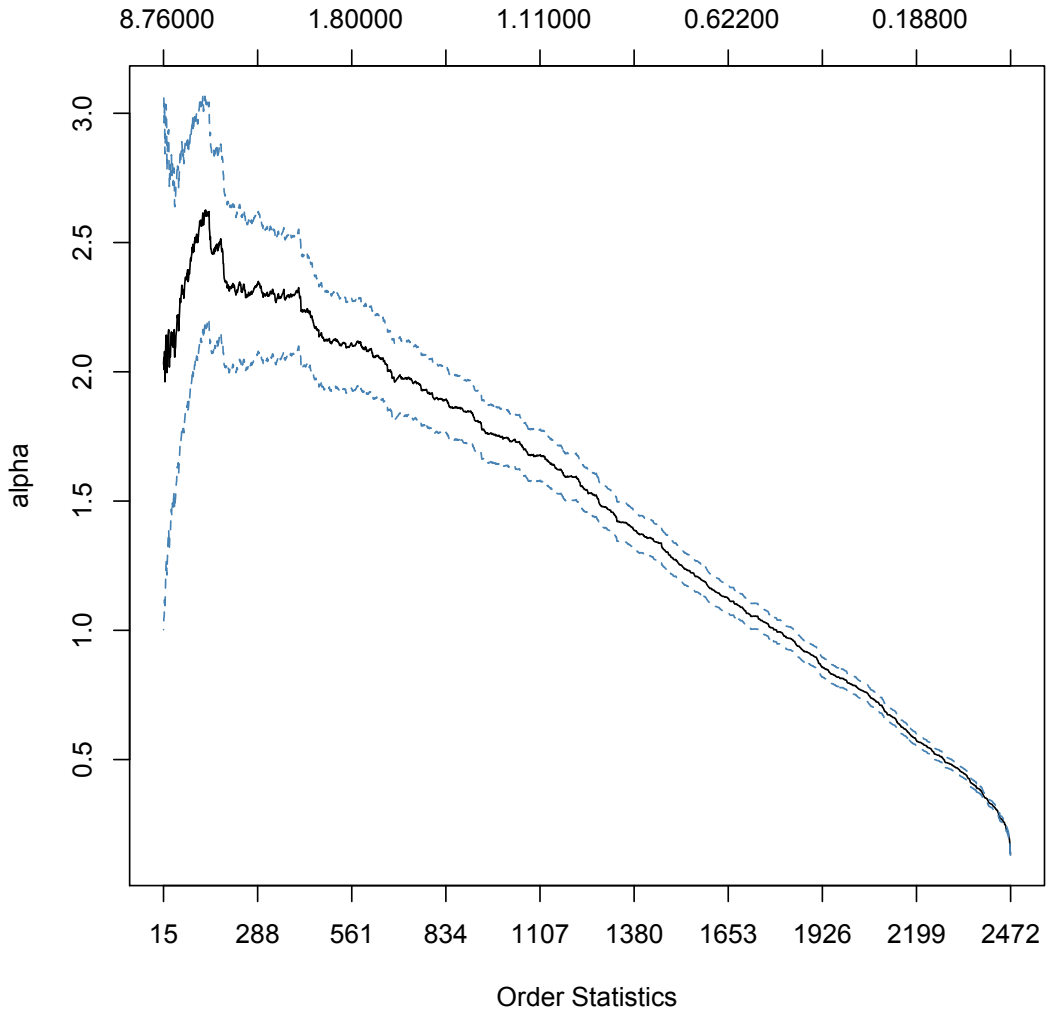
$$H_{k,n} \to \gamma = \frac{1}{\alpha} \text{ a.s.}$$

(Deheuvels, Hausler, Mason (1988)).

Choosing the appropriate number k of upper order statistics when the Hill estimator is applied to a finite sample, is very difficult.

- Visual techniques are used: the estimator is plotted for a range of k, and then one looks for a part of the plot that looks stable.
- Several smoothing techniques have been introduced to assist in this visual analysis (Resnick and Starica (1997)).

Hill estimator applied to i.i.d. SαS sample of 5000, $\alpha = 1.7$.



Systematic ways of selecting the number of upper order statistics

- Hall (1990) suggested a procedure minimizing the asymptotic MSE of the estimator based on the assumption of 2nd order regular variation.
- Danielsson et. al (2001) improved the above approach via a two-step bootstrap procedure that uses minimal *a priori* information.
- Drees and Kaufmann (1998) introduced a thresholding approach that works under certain additional assumptions.

Most of existing approaches to selecting the number k in a tail estimator is via optimizing asymptotic efficiency.

We view it as the problem of deciding which part of a given sample contains reliable information on the tail of the distribution F.

Where does the tail begin?

Importance is even higher in a highly dimensional multivariate context, where we need to test repeatedly for tail independence. This is highly sensitive to the contamination of the tail by the center of the distribution.

Our approach is based on **a simple idea**. Under the assumption of regular variation, vague convergence of point processes holds:

$$N_n = \sum_{i=1}^{n} \delta_{X_i/a_n} \xrightarrow{v} N_* \,,$$

where:

- δ_x is a point mass at x;
- (a_n) a positive sequence satisfying $\bar{F}(a_n) \sim 1/n$ as $n \to \infty$;
- N_* is a Poisson random measure on $(0, \infty]$ with mean measure $\mu_*(x, \infty] = x^{-\alpha}, \ x > 0$.

We interpret this result as follows:

any upper order statistics in the sample that fall in the tail region behave like points of a Poisson random measure with a power intensity.

This property can be tested statistically, and sequentially.

One can perform appropriate statistical tests on the subsamples $X_{n-k+1,n}, X_{n-k+2,n}, \ldots, X_{n,n}$ with increasing k.

Terminate the procedure once the k upper order statistics stop resembling points of a Poisson random measure with a power intensity.

In order to avoid taking into account too many order statistics, it is desirable to make it easier to reject the null hypothesis for larger k.

We achieved this by selecting an increasing sequence $\theta_n \uparrow \infty$ and set

$$N_n := \inf \left\{ k : 1 \leq k \leq n, \, |Q_{k,n}| \geq \omega \sqrt{\frac{\theta_n}{k}} \right\} .$$

Under a suitable growth condition on θ_n, this definition of N_n makes it, roughly, proportional to θ_n.

Theorem Let $\theta_n = o\left(n^{\frac{2|\rho|}{1+2|\rho|}}\right)$ as $n \to \infty$. Then

- The Hill estimator based on N_n upper order statistics is consistent:

$$H_{N_n,n} = \frac{1}{N_n} \sum_{i=0}^{N_n-1} \log \frac{X_{n-i,n}}{X_{n-N_n,n}} \xrightarrow{P} \gamma, \ \ n \to \infty.$$

- the asymptotic behaviour of the estimator is given by

$$\sqrt{\theta_n} \left(\frac{H_{N_n,n}}{\gamma} - 1 \right) \Rightarrow \frac{G}{(\tau_\omega)^{1/2}},$$

where G is a standard normal random variable independent of the first hitting time τ_ω of the set $\pm\omega$ by a standard Brownian motion.

Simulated data, $n = 5000$

Method		Hill		N_n, $\theta_n = (\log n)^2$		Bootstrap		$\hat{k}_{opt}$	
Dist.	α	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE
Student(4)	4	2.7794	1.7098	3.4568	.6510	3.6135	.6859	3.4270	.6629
Student(3)	3	2.1719	.9843	2.7726	.3657	2.8490	.4383	2.7669	.3358
Student(1)	1	.9326	.3937	1.0109	.0890	.9881	.0502	.9965	.0391
Stable(1.7)	1.7	2.0347	.6951	2.0013	.3887	2.2515	.5654	2.2138	.5283
Stable(1)	1	.8965	.1683	1.0099	.0855	.9945	.0689	.9912	.0404
MA(1)	3	2.8335	1.8239	3.1434	.5232	3.1365	.5647	3.0955	.3708

Simulated data, $n = 50000$

Method		Hill		N_n, $\theta_n = (\log n)^2$		Bootstrap		$\hat{k}_{opt}$	
Dist.	α	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE
Student(4)	4	3.2954	1.5064	3.7958	.4743	3.7690	.5282	3.6080	.4217
Student(3)	3	2.5280	.6734	2.9391	.2245	2.8900	.3013	2.8490	.1839
Student(1)	1	.9499	.2182	1.0103	.0697	.9959	.0215	.9970	.0159
Stable(1.7)	1.7	2.0894	.5303	1.7733	.1670	2.2276	.5288	2.1057	.4079
Stable(1)	1	.9608	.1357	1.0079	.0764	.9918	.0204	.9965	.0165
MA(1)	3	3.6480	5.4884	3.1893	.4743	3.1014	.2113	3.0898	.1775

Conclusions

- The suggested choice of the sample fraction works well with various distributions;
- it works well even with modest sample sizes;
- it is very efficient computationally;
- we still need to understand its behaviour under tail dependence.