Truncated Pareto distribution: parameter estimation and applications

Mark M. Meerschaert
Department of Mathematics & Statistics
University of Otago

NZSA 2005 Meeting
4 July 2005

Abstract

The Pareto distribution is a simple model for nonnegative data with a power law probability tail. In many practical applications, there is a natural upper bound that truncates the probability tail. This talk presents estimators for the truncated Pareto distribution, investigates their properties, and illustrates a way to check for fit. Applications from finance, hydrology and atmospheric science will be included.

Acknowedgments

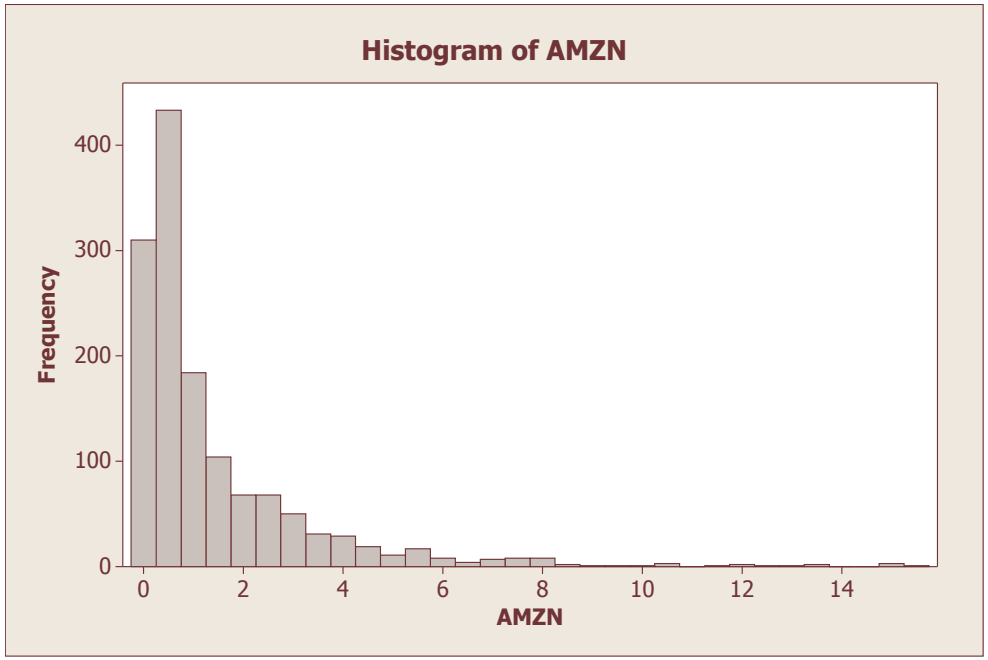
Inmaculada B. Aban, Department of Biostatistics, U. Alabama.

Anna K. Panorska, Department of Maths & Stats, U. Nevada.

This work was partially supported by NSF grants DMS-0139927, DMS-0417869, and ATM-0231781 and by the Marsden Foundation in New Zealand.

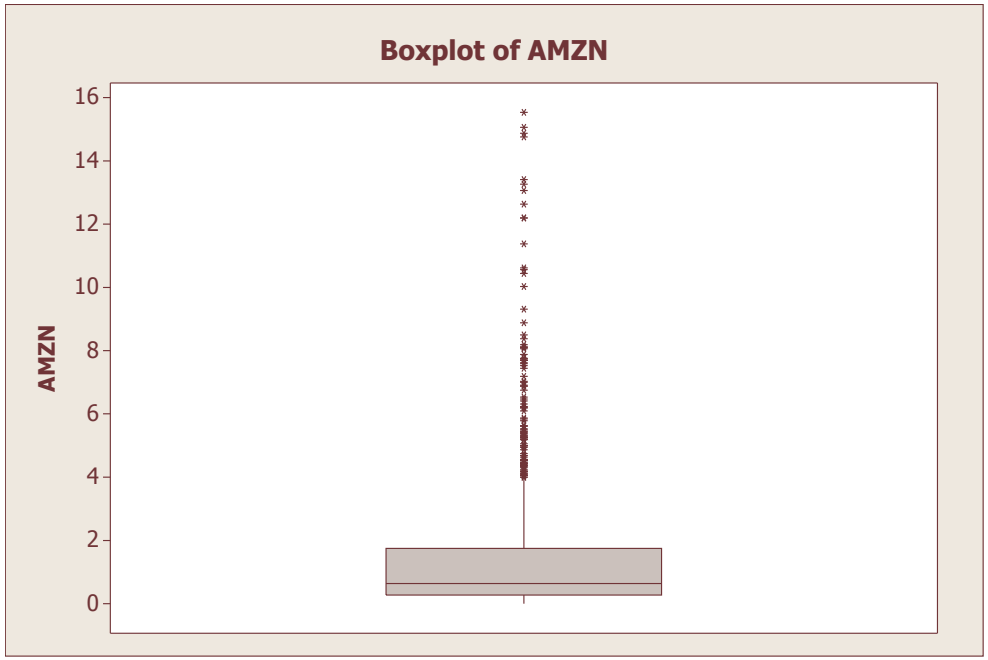
A typical data set from finance

Absolute change in Amazon, Inc. stock prices (US$) from Jan 1, 1998 to June 30, 2003 ($n = 1378$)



Outliers - do not discard!

Absolute change in AMZN stock prices. The outliers are the most important events.



The Pareto distribution

If X is Pareto with $y = P(X > x) = Cx^{-\alpha}$ then

$$\log y = \log C - \alpha \log x.$$

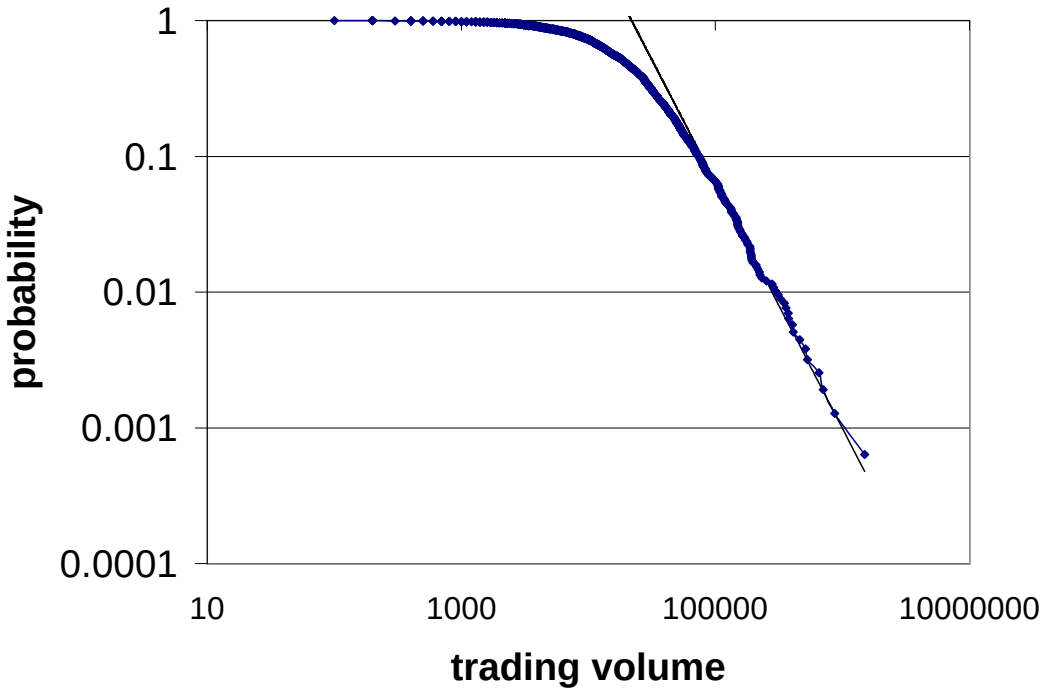
If we order the data $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$ (decreasing order statistics) then we can estimate

$$y = P[X > x]$$

by taking $y = i/n$ and $x = X_{(i)}$. A plot of the points $(x, y) = (\log X_{(i)}, \log i/n)$ should fit a straight line with slope $-\alpha$.

Heavy tails in finance

Price changes and trading volume often have power law tails. Upper tail of trading volume for AMZN fits a Pareto with $\alpha \approx 2.7$.



Hill's estimator

The most popular estimator (Hill 1975; Hall 1982) for α is the Pareto MLE conditional on $X \geq D$.

$$\begin{aligned} \widehat{\alpha}_H &= \left[r^{-1} \sum_{i=1}^{r} \{\ln X_{(i)} - \ln X_{(r+1)}\} \right]^{-1} \\ \widehat{C} &= (r/n)(X_{(r+1)})^{\widehat{\alpha}_H} \end{aligned}$$

In practice we take $D = X_{(r+1)}$ and we select r based on the log-log plot to represent the Pareto tail.

Hill's estimator and regression

The slope of the best fitting line through the points

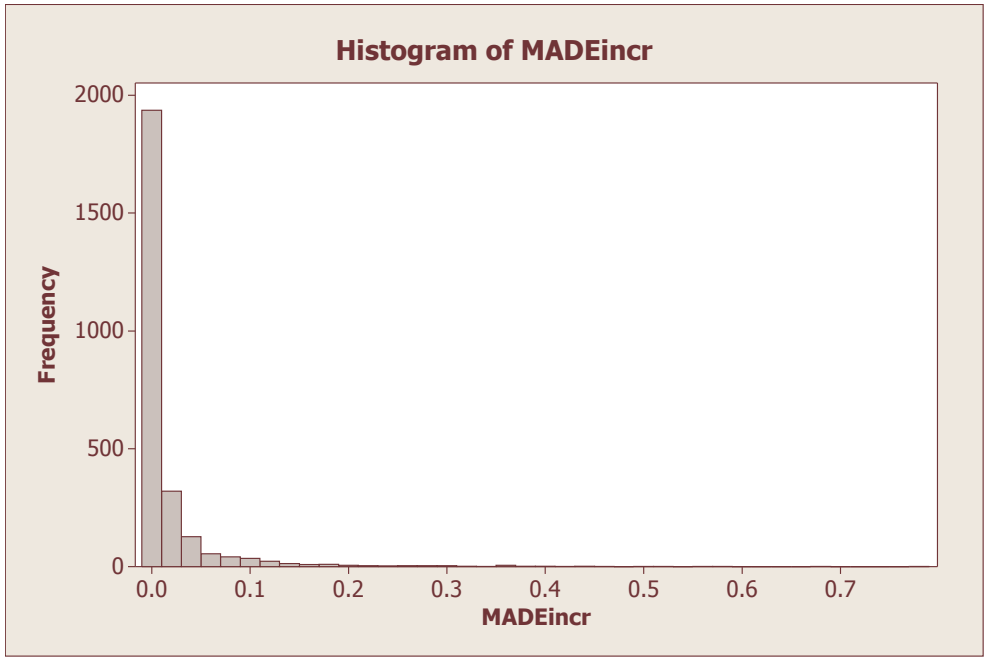
$$(\log(i/n), \log(X_{(i)})) \quad \text{for} \quad 1 \leq i \leq r$$

should be approximately equal to $-1/\alpha$. This is not a classical regression problem because deviations of the y variables $\log(X_{(i)})$ from their respective means are neither independent nor identically distributed.

Generalized linear regression (Aban and Meerschaert, 2004) shows that Hill's estimator is also the BLUE and UMVUE for α.

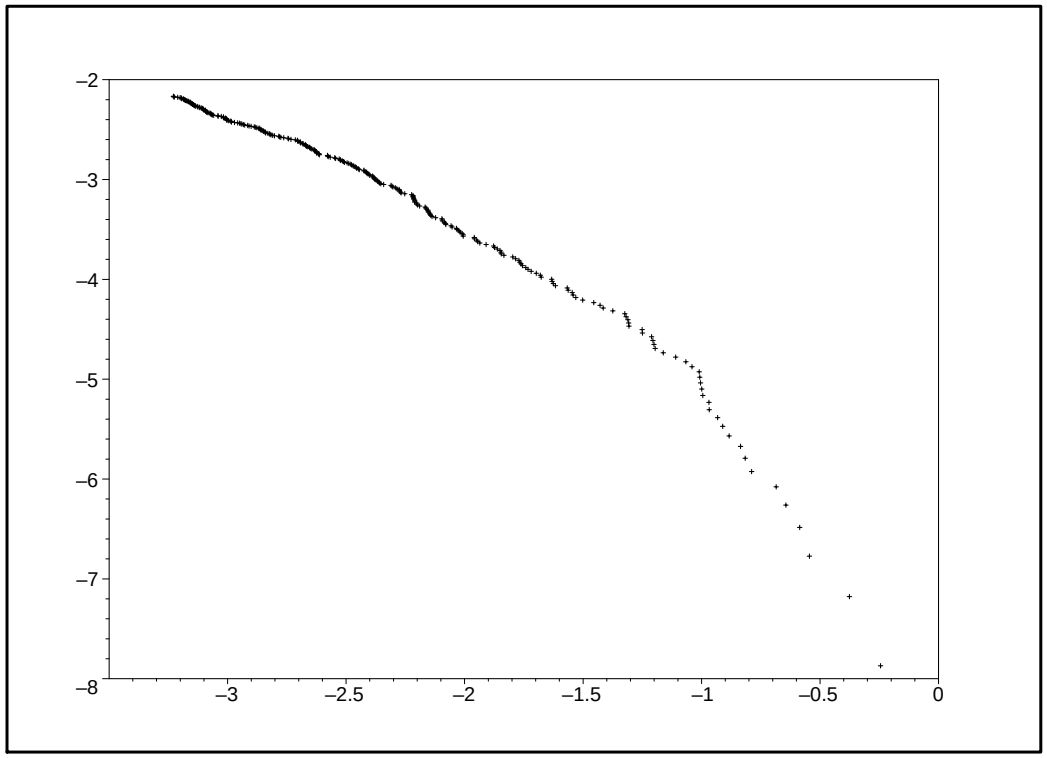
More heavy tails in geophysics

Absolute difference in hydraulic conductivity ($n = 2618$) at the MADE site (Rehfeldt et al. 1992).



Graphical test

A plot of $\log(r)$ versus $\log X_{(r)}$ shows that the upper tail of the MADE data does not fit a simple power law.



Extreme value test

A simple asymptotic level q test ($0 < q < 1$) based on extreme value theory rejects the null hypothesis $H_0 : \nu = \infty$ (Pareto) if and only if

$$X_{(1)} < [(nC)/(-\ln q)]^{1/\alpha}.$$

The corresponding approximate p-value of this test is given by

$$p = \exp\{-n\, C\, X_{(1)}^{-\alpha}\}.$$

In practice, we use Hill's estimator for C and α.

For the MADE data we get $p = 0.012$, strong evidence that an alternative distributional model is needed.

Truncated Pareto

The truncated Pareto distribution is $P(X > x) = C(x^{-\alpha} - \nu^{-\alpha})$ for $\gamma \leq x \leq \nu$. The MLE conditional on $X \geq D$ is given by $\widehat{\nu} = X_{(1)}$,

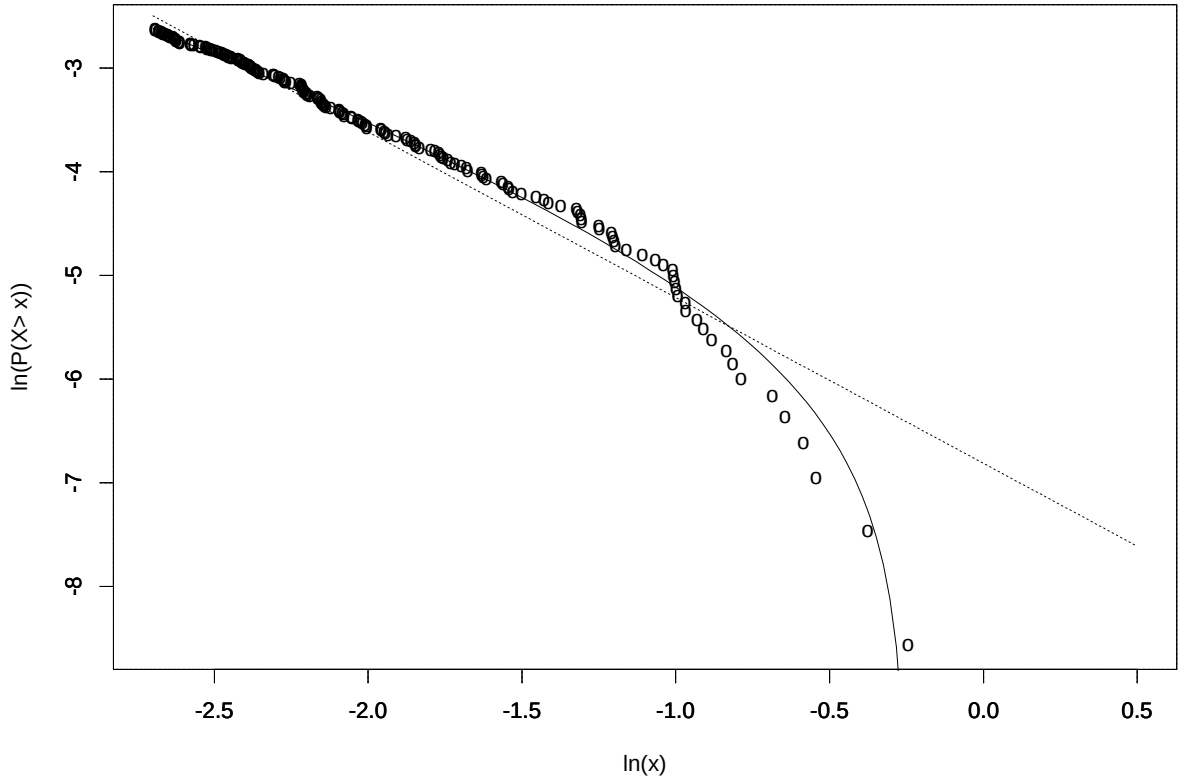
$$\widehat{\gamma} = r^{1/\widehat{\alpha}}(X_{(r+1)}) \left[n - (n-r) \left(X_{(r+1)}/X_{(1)} \right)^{\widehat{\alpha}} \right]^{-1/\widehat{\alpha}}$$

and $\widehat{\alpha}$ solves the equation

$$\frac{r}{\widehat{\alpha}} = \sum_{i=1}^{r} [\ln X_{(i)} - \ln X_{(r+1)}] - \frac{r \left(X_{(r+1)}/X_{(1)} \right)^{\widehat{\alpha}} \ln \left(X_{(r+1)}/X_{(1)} \right)}{1 - \left(X_{(r+1)}/X_{(1)} \right)^{\widehat{\alpha}}}$$

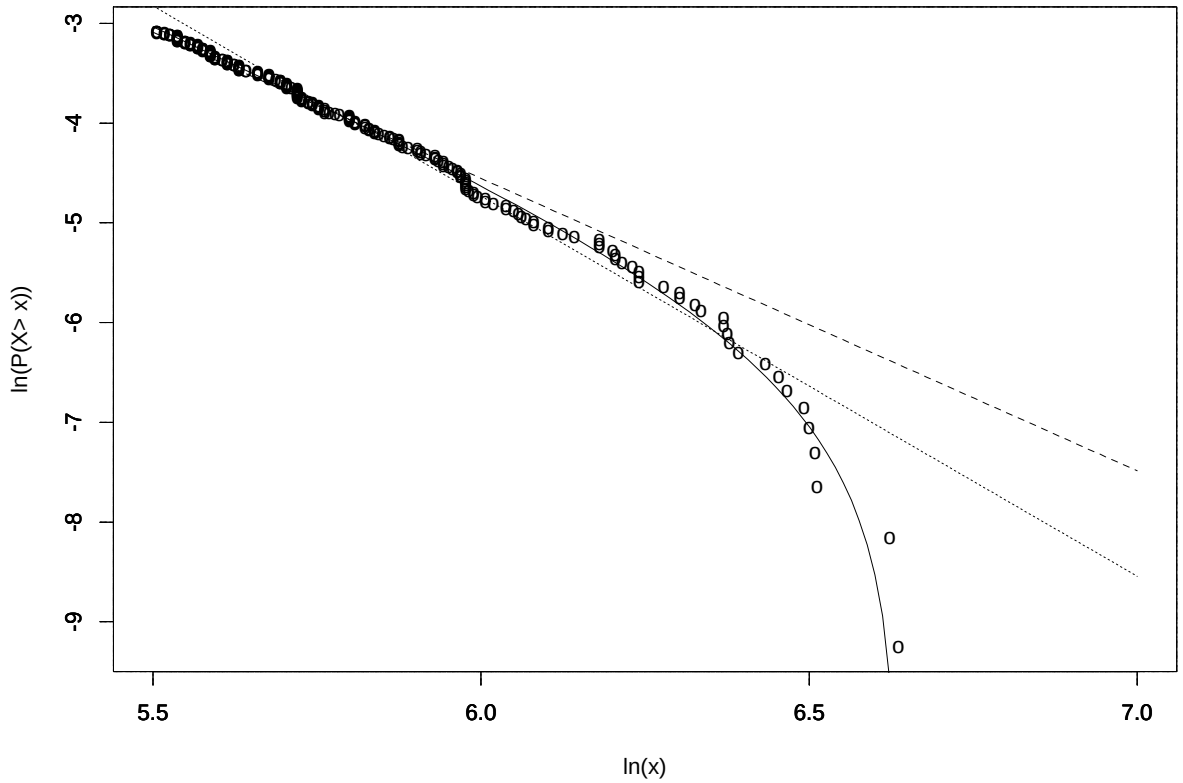
Truncated Pareto model for MADE data

Pareto ($\widehat{\alpha}_H = 1.6$) and TP ($\widehat{\alpha}_{TP} = 1.2$) fit to the MADE data.
Possible source of truncation: volume averaging.



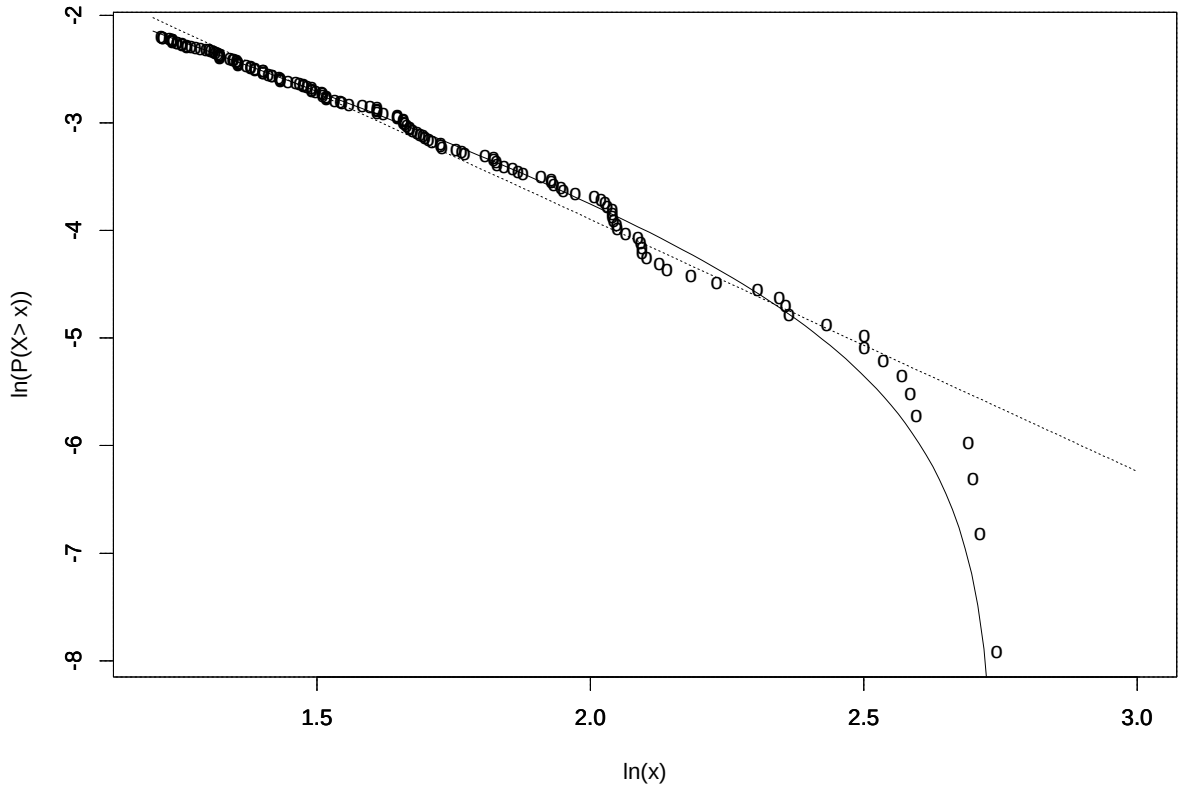
Pareto and TP model for precipitation data

Log-log plot for the 100 largest observations of daily total precipitation in Tombstone AZ between July 1893 to December 2001. An upper bound called the "probable maximum precipitation" is typically computed. Here $\widehat{\alpha}_H = 3.8$, $\widehat{\alpha}_{TP} = 3.0$, and $p = 0.017$.



Pareto and TP model for AMZN data

Pareto ($\widehat{\alpha}_H = 2.3$) and TP ($\widehat{\alpha}_{TP} = 1.7$) fit to the AMZN price change data. Here $p = 0.007$. One advantage of the truncated Pareto model is that all moments exist.



References

1. I.B. Aban and M.M. Meerschaert (2004) Generalized least squares estimators for the thickness of heavy tails. *Journal of Statistical Planning and Inference*, **119**(2), 341-352.

2. I.B. Aban, M.M. Meerschaert and A.K. Panorska (2005) Parameter estimation for the truncated Pareto distribution. *Journal of the American Statistical Association: Theory and Methods*, to appear.

3. P. Hall (1982) On some simple estimates of an exponent of regular variation. *J. Royal Statist. Soc. B*, **44**, 37–42.

4. B. Hill (1975) A simple general approach to inference about the tail of a distribution. *Annals of Statistics* **3** 1163–1173.

5. K.R. Rehfeldt, J. M. Boggs and L. W. Gelhar (1992) Field study of dispersion in a heterogeneous aquifer. 3: Geostatistical analysis of hydraulic conductivity. *Water Resources Research*, **28**(12), 3309-3324.