IBM Research

Frequency Estimation of Rare Events by Adaptive Thresholding

J. R. M. Hosking
IBM Research Division

Graybill VIII / 6th International Conference on Extreme Value Analysis 2009

Motivation

- When managing IT systems, there is a need to identify transactions that are looping or hung, and terminate them.
- An effective solution is to set a threshold, on time or CPU usage, and terminate transactions that exceed the threshold. Threshold should be chosen so that only a small proportion of normal transactions will exceed it.
- How to choose the threshold? Our approach: obtain data for a large number n of normal transaction events, use the data to estimate the event magnitude that has a specified small exceedance probability δ (typically $\delta < 1/n$, so we will need to extrapolate beyond range of data).
- This approach can be applied to many other problems that involve frequency estimation for rare events.

Approach

General approach

- Assume power law tail of distribution, $1-F(x) \propto x^{-1/B}$ for large x.
- Fit a probability distribution (generalized Pareto) to the *m* largest historical events, for a suitably chosen value of *m*.

Choice of *m*

- Deterministic function of sample size (e.g. Weissman, 1978)
- By inspection of graph of tail index vs. *m* (e.g. Embrechts et al., 1997)
- Minimize estimated variance of tail index (e.g. Gomes and Pestana, 2007)

New features

- To reduce bias, choose *m* so that the chosen subsample is consistent with all smaller subsamples; reduce variance by choosing *m* as large as possible subject to the consistency requirement.
- To improve robustness (i.e. maintain accuracy when assumption of Pareto distribution is not valid): test for lack of fit of Pareto distribution, ignore subsamples that fail the test.

Algorithm

1. Define a set of subsample sizes for which generalized Pareto distributions will be fitted.
2. For each subsample size *m*, form the subsample containing the *m* largest data values. Fit a generalized Pareto distribution to this subsample, obtaining a point estimate and a confidence interval for the shape parameter ξ_i.
3. For each subsample, also perform a test of goodness of fit of the generalized Pareto distribution to the subsample. If the goodness of fit is unsatisfactory deem the subsample “unacceptable”.
4. Also deem as “unacceptable” any subsample whose estimated shape parameter lies outside the confidence interval for any smaller sample that was deemed acceptable. Deem as “optimal” the largest acceptable subsample.
5. From the optimal subsample compute quantile estimates for the complete sample.

L-moments

Mean $\lambda_1 = \mathrm{E}X$

Dispersion $\lambda_2 = \frac{1}{2}\mathrm{E}(X_{2:2} - X_{1:2})$

$\lambda_3 = \frac{1}{3}\mathrm{E}(X_{3:3} - 2X_{2:3} + X_{1:3})$

$\lambda_4 = \frac{1}{4}\mathrm{E}(X_{4:4} - 3X_{3:4} + 3X_{2:4} + X_{1:4})$

Skewness $\tau_3 = \lambda_3 / \lambda_2$

Kurtosis $\tau_4 = \lambda_4 / \lambda_2$

Sample estimators

λ_r estimated by ℓ_r, a linear combination of the sample data

τ_r estimated by $t_r = \ell_r / \ell_2$

L-moments have many good properties – Hosking, *J. R. Statist. Soc. B*, 1990

L-moment methods for generalized Pareto distribution

Quantile function

$$Q(u) = \xi + \alpha(1-u)^k / k$$

L-moment estimation of parameters (ξ known)

$$\hat{k} = \ell_1 / \ell_2 - 2, \qquad \hat{\alpha} = \ell_1(\ell_1 - \ell_2) / \ell_2$$

-- very simple to calculate (not iterative)

Goodness-of-fit test based on *L*-moments

$$\boldsymbol{t} \equiv [t_3\ t_4]^{\mathrm{T}} \sim N[\boldsymbol{\mu}(\alpha,k), \boldsymbol{\Sigma}(\alpha,k)]$$ for large samples

-- Estimate mean and variance, replacing parameters by their estimates

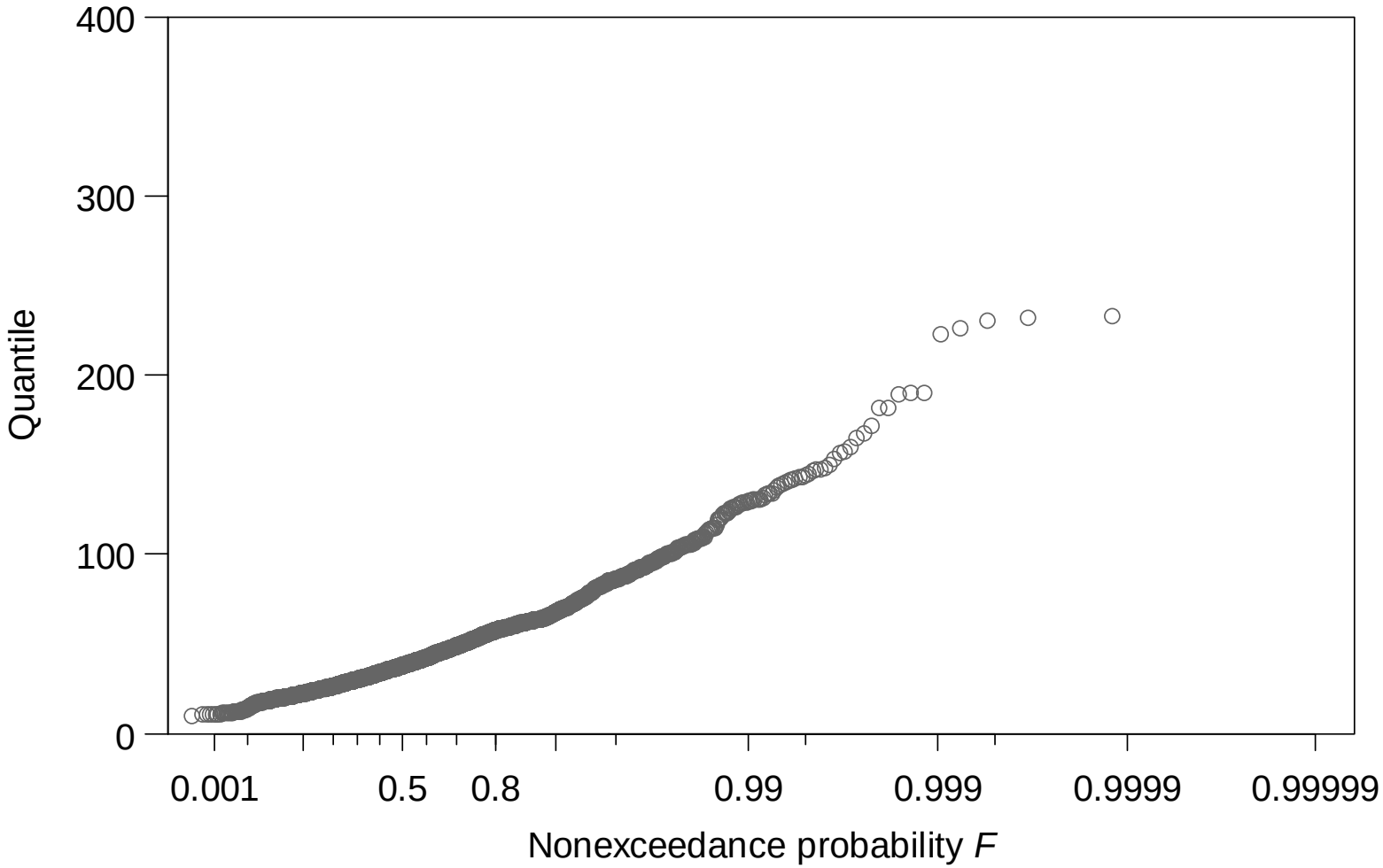
$$(\boldsymbol{t} - \hat{\boldsymbol{\mu}})^{\mathrm{T}}\, \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{t} - \hat{\boldsymbol{\mu}}) \sim \chi^2_2$$

-- Works reasonably well in practice

-- Could be improved, e.g. via parametric bootstrap

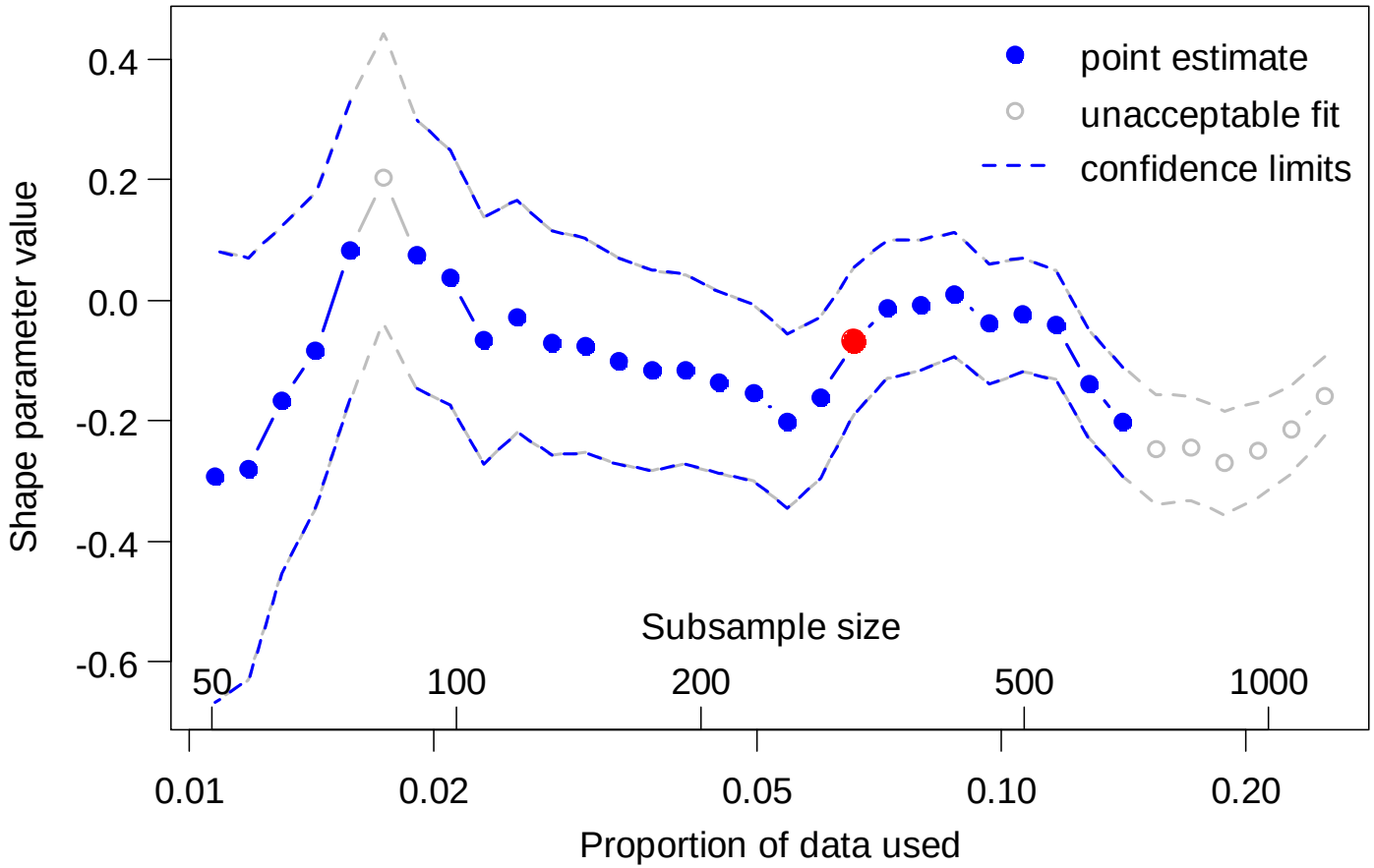
Example -- data

Response times for transactions in an online banking application.
4689 data points, collected during a “typical” period.
Plotted here on an extreme-value (Gumbel) probability scale.

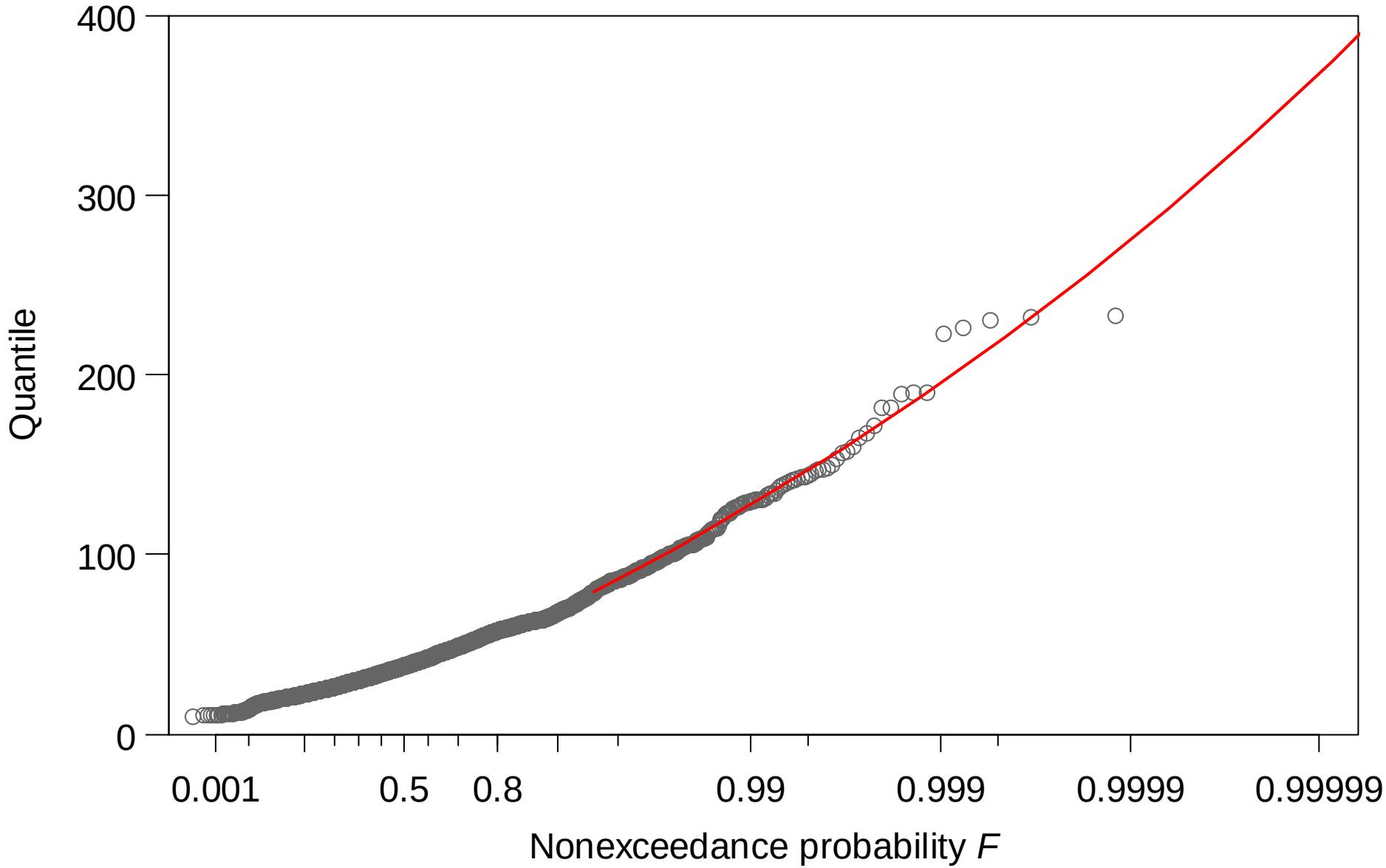


Example -- analysis

1. Fit generalized Pareto distributions (GPDs) to *m* largest values, for various *m*.
2. Plot shape parameter estimate (blue dots) and confidence interval (dashed lines) vs. *m*.
3. Exclude *m* values for which fit of distribution is poor (grey circles).
4. Choose largest *m* for which parameter estimate lies within confidence interval for all smaller *m* (red dot).
5. Using the subsample corresponding to this *m*, estimate GPD and its quantiles (red line).



Example -- results



Summary

- We have developed a method for estimation of extreme quantiles by fitting generalized Pareto distributions to subsets of the data
- New methods for bias-variance tradeoff and robustness
- Convenient calculations using *L*-moments
- Effective in practice

Reference

- S. Heisig and J. R. M. Hosking (2008). Scoring and thresholding for availability. *IBM Systems Journal*, **47**, 653-666.