A new approach to threshold selection in extreme value analysis

Paul Northrop and Claire Coleman
University College London
p.northrop@ucl.ac.uk

RMS
22nd January 2014
Outline

• Simple example dataset
• Generalized Pareto (GP) modelling of threshold excesses
• Threshold selection methods
  • parameter stability plots
  • others
• Improved parameter stability plots
• (Automatic?) threshold selection on many datasets

Motivating (classic) example:

• 154 flow rates from the River Nidd (Yorkshire), 1934–1969
• peaks flows: $\approx$ independent, pre-processing has already extracted extreme values
River Nidd 1934-1969: 154 flow rates

Background and motivation
Threshold $u = 90$
Excesses of $u = 90$

![Graph showing flow rate vs. event number]
Threshold $u = 110$

Background and motivation
Excesses of $u = 110$

Background and motivation
Generalized Pareto model

- Set up: $X_1, X_2, \ldots X_n$ are i.i.d.
- $Y = (X - u) \mid X > u$: excess of threshold $u$
- Extreme value theory suggests the GP distribution as a model for $Y$.

GP distribution function:

$$P(Y \leq y) = G(y) = \begin{cases} 1 - (1 + \xi y/\sigma_u)^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-y/\sigma_u), & \xi = 0, \end{cases}$$  \hspace{1cm} (1)

where $y > 0$, $x_+ = \max(x, 0)$.

- For $\xi \geq 0$ we have $y > 0$;
- For $\xi < 0$ we have $0 < y < -\sigma_u/\xi$.

(\ldots how often the threshold is exceeded also matters)
Threshold selection methods

Bias-variance trade-off:

- $u$ too low: GP model inappropriate $\rightarrow$ bias.
- $u$ too high: fewer excesses $\rightarrow$ unnecessary imprecision.

Review paper: Scarrott and MacDonald (2012).

- **Parameter stability plot** : $\hat{\sigma}_u - \hat{\xi}_u$ vs. $u$ and $\hat{\xi}$ vs. $u$.
  Estimates stable above $u^*$?

- **Mean residual life plot** : sample mean excess vs. $u$. Linear above $u^*$?

- **Goodness-of-fit test**: AD or KS $p$-value vs. $u$. For which $u$ don’t we reject GP model?

- **Extend model below** $u$, make $u$ a model parameter, make inferences about $u$, e.g. Wadsworth and Tawn (2012).
Parameter stability: modified scale

Threshold selection
Parameter stability: shape

Threshold selection
Problems with parameter stability plots

• $\hat{\xi}$ and $\hat{\sigma}_u - \hat{\xi}_u$ very strongly negatively associated across $u$: only one plot needed.

• Estimates of $\xi$ based on thresholds $u_1$ and $u_2$ are dependent (one datasets a subset of the other).

• Viewer compares many pairs of thresholds (multiple-testing).

• Viewer invited to ask whether CIs overlap: not the appropriate assessment.

• Threshold choice rather subjective.

Aim: make assessment more formally, by testing (a discrete version of)

$$H_0 : \xi(x) = \xi(u), \text{ for } x > u.$$

Improved parameter stability plots
• Penultimate theory: GP model valid for lowish thresholds, shape parameter $\xi$ varies slowly with threshold.
• Notation: $\xi_i$ is the GP shape local to threshold $u_i$.
• Model $\xi$ as piecewise constant in variable $x$:

$$\xi(x) = \begin{cases} 
\xi_1, & u_1 < x < u_2, \\
\xi_2, & x > u_2.
\end{cases}$$  \hspace{0.5cm} (2)

Test $H_0 : \xi_i = \xi_j$ for all possible pairs $(u_i, u_j)$ from a set of thresholds $(u_1, \ldots, u_m)$.

Drawbacks:

• Simulation required to test $H_0 : \xi_1 = \cdots = \xi_m$ by combining pairwise tests.
• Very computationally-intensive: prohibitively so.
Two-threshold model

$\xi_1$

$\xi_2$

$u_1$

$u_2$

$x$

$f(x)$
Plot of pairwise test results
Multiple-threshold GP model (NC2014)

$$\xi(x) = \begin{cases} 
\xi_i, & u_i < x < u_{i+1}, \quad \text{for } i = 1, \ldots, m - 1, \\
\xi_m, & x > u_m.
\end{cases} \tag{3}$$

$$H_0 : \xi_1 = \cdots = \xi_m \quad \text{vs} \quad H_A : H_0 \text{ not true}.$$

- Scale parameters set to achieve a continuous p.d.f.
- Parameter vector: \( \theta = (\sigma_1, \xi_1, \ldots, \xi_m) \)
- \( \hat{\theta}_0 \): MLE under \( H_0 \).
- \( \hat{\theta} \): MLE under \( H_A \).

General idea: do the data suggest a departure from \( H_0 \) (standard GP model) in the direction of \( m \)-threshold GP model.
$m$-threshold model ($m = 5$)
Likelihood ratio and score tests

LR stat.: compare maximized log-likelihoods under $H_0$ and $H_A$

\[ W = 2 \left\{ l(\hat{\theta}) - l(\hat{\theta}_0) \right\}, \]

Score stat.: how far is the log-likelihood from being flat at $\hat{\theta}_0$?

\[ S = U(\hat{\theta}_0)^T i^{-1}(\hat{\theta}_0) U(\hat{\theta}_0), \]

- $U(\theta)$ is the score function: $U_i = \partial l(\theta)/\partial \theta_i$
- $i(\theta)$ is the expected Fisher information matrix:
  \[ i(\theta)_{ij} = E \left[ -\partial^2 l(\theta)/\partial \theta_i \partial \theta_j \right]. \]
- Derivation of $U(\theta)$ and $i(\theta)$ nasty: it made my head hurt!

If $\xi_m > -1/2$ then $W$ and $S$ are approx. $\chi^2_{m-1}$ under $H_0$.

**LLR test.** Need to fit the full multiple-threshold GP model.

**Score test.** Only fit the null model, i.e. a single $\text{GP}(\sigma, \xi)$ fit.
LR, score and Wald test statistics

Improved parameter stability plots
Effect of number of thresholds

- x - m=7
- - m=12
- o - m=22

Improving parameter stability plots
How to make use of the $p$-values?

Consider $H_0: \xi_i = \cdots = \xi_m$.

- Small $p$-value suggests that $H_0$ isn’t true, i.e. $u_i$ isn’t high enough.
- Large $p$-value: perhaps $u_i$ is high enough.

Possibilities

- Formal: set size of test beforehand, e.g. 5%. Reject $u_i$ as too low if $p$-value is $< 0.05$.
- Informal: view the $p$-values as a measure of the disagreement between the data and the null hypothesis when inspecting plot.

Some multiple-testing remains: we perform tests with lowest thresholds $u_1, u_2, \ldots, u_{m-1}$.
Wave heights from the Gulf of Mexico

- Hindcasts of $H_s$ storm peak significant wave height (in metres) in the Gulf of Mexico.
  - Data from Northrop and Jonathan (2011).
  - wave height: trough to the crest of the wave.
  - significant wave height: the average of the largest 1/3 wave heights. A measure of sea surface roughness.
  - storm peak: largest value from each (hurricane-induced) storm.
- a 6 × 12 grid of 72 sites (≈ 14 km apart).
- Sep 1900 to Sep 2005: 315 storms.
- average of 3 observations (storms) per year, at each site.
Storm damage

- Interested in extremal behaviour of $H_s$ at centre of data grid
- Pool (spatially-dependent) data over space.
- What level (quantile) of threshold is appropriate at site 1, site 2, . . . , site 72?
Gulf of Mexico wave heights

Multiple datasets
Automatic threshold selection

Suppose that we wish to automate the selection of a threshold for each of the datasets, based on tests of size 5%, say.

Two strategies

1. Select the lowest threshold for which $H_0$ is not rejected.
   
   [However, the $p$-values are not constrained to be non-decreasing in the lowest threshold.]

2. Select the lowest threshold with the property that $H_0$ is not rejected at it and at all the higher thresholds considered.
   
   [We need to bear in mind that large variability is expected at the very highest thresholds.]
Automatic threshold selection

Quantile at which threshold chosen

Strategy 1

Strategy 2

Multiple datasets
Concluding remarks

- Makes the (frequentist) fixed threshold selection part of Wadsworth and Tawn (2012) quick.
- Parameter stability is only part of the story (model checking).
- Threshold sensitivity vs. threshold uncertainty:
  - $u$ a tuning parameter vs. $u$ is a model parameter
  - sensitivity/uncertainty shouldn’t be ignored.
- Motivated by slide 28 of Jo’s talk:
  - Chavez-Demoulin et al’s discussion of Northrop and Jonathan (2011);
  - $r$-largest order statistics model for in a multi-site analysis, with (only) latitude and longitude as covariates;
  - ...no need to set explicitly a threshold $u$;
  - ...but what if there is a continuous covariate, such as “distance to nearest gate”.
- Adjustment for serial dependence?


Thank you for your attention.