Threshold estimation in marginal modelling of spatially-dependent non-stationary extremes

Philip Jonathan  
Shell Technology Centre Thornton, Chester  
philip.jonathan@shell.com

Paul Northrop  
University College London  
paul@stats.ucl.ac.uk

Environmental Extremes  
Royal Statistical Society  
April 2011
Outline

- Motivation and application.
- Threshold modelling using quantile regression.
- Implications of QR threshold for PP model parameterisation.
- Adjusting for spatial dependence.
- Results for application.
- Initial theoretical & simulation studies.
- Conclusions.
Motivation: Rational design of marine structures

- **Covariate** effects:
  - Location, direction, season, ...
  - Multiple covariates in practice.

- **Cluster** dependence:
  - e.g. storms independent, observed (many times) at many locations.
  - e.g. dependent occurrences in time.

- **Scale** effects:
  - Modelling $H_2^2$ gives different estimates cf. modelling $H_S$.

- **Threshold** estimation; **parameter** estimation.

- **Measurement** issues:
  - Field measurement uncertainty greatest for extreme values.
  - Hindcast data are simulations based on pragmatic physics, calibrated to historical observation.
Motivation: Rational design of marine structures

- **Multivariate** extremes:
  - Waves, winds, currents, ...
  - Componentwise maxima $\Leftrightarrow$ max-stability $\Leftrightarrow$ regular variation:
    - Assumes all components extreme.
    - $\Rightarrow$ Perfect independence or asymptotic dependence **only**.
  - Extremal dependence:
    - Assumes regular variation of joint survivor function.
    - $\Rightarrow$ Asymptotic dependence, asymptotic independence (with $+$ve, $-$ve association).
  - Conditional extremes:
    - Assumes, given one variable being extreme, convergence of distribution of remaining variables.
    - Allows some variables not to be extreme.
  - Inference:
    - ... a huge gap in the theory and practice of multivariate extremes ... (Beirlant et al. 2004)

**Aim:** Useful models with rigourous assessment of model performance, especially in extreme quantiles.
Motivation: Good threshold estimation critical

- Considerable **empirical** evidence from applications that careful estimation of threshold including covariate effects important for satisfactory modelling.
- Often reasonable to assume some (or all) extreme value parameters are **independent** of (some or all) covariates following good thresholding, greatly simplifying model form.
- Quantile thresholds as functions of covariate(s) produce near **constant rates** of threshold exceedence (appealing from design perspective).
Application: Marginal estimation of extreme $H_{SP}^S$

- Data from hindcast of $Y$ storm peak significant wave height (in metres) in the Gulf of Mexico.
  - **Wave height**, $h$: trough to the crest of the wave.
  - **Significant wave height**, $H_S$: the average of the largest 1/3 wave heights $h$ in given period (usually 3 hours).
  - **Storm peak** $H_{SP}^S$: largest value of $H_S$ from a storm (cf. declustering).
- 6 $\times$ 12 grid of 72 sites ($\approx$ 14 km apart).
- Sep 1900 to Sep 2005: 315 storms in total.
- Average of 3 observations (storms) per year, at each site.

**Aim**: Quantify the extremal behaviour of $Y$ at each site, making appropriate adjustment for spatial dependence.
Typical hurricane event in Gulf of Mexico
Spatial dependence

2 distant sites

2 nearby sites

$H_s^\text{sp} / \text{m}, \text{at site 1}$

$H_s^\text{sp} / \text{m}, \text{at site 21}$

$H_s^\text{sp} / \text{m}, \text{at site 31}$
• From single event?
Modelling approach

- Spatial non-stationarity:
  - Model threshold as Legendre polynomial in longitude and latitude using quantile regression.
  - Model spatial variation of PP parameters as Legendre polynomials in longitude and latitude.
  - Lots of other suitable bases: splines, random fields ...

- Spatial dependence:
  - Estimate parameters assuming conditional independence of responses given covariate values.
  - Adjust standard errors etc. for spatial dependence.

- Estimate extreme quantiles.
Extreme value regression model

Conditional on covariates $x_{ij}$ exceedances over a high threshold $u(x_{ij})$ follow a 2-dimensional non-homogeneous Poisson process.

If responses $Y_{ij}, i = 1, \ldots, 72$ (space), $j = 1, \ldots, 315$ (storms) are conditionally independent:

$$L(\theta) = \prod_{j=1}^{315} \prod_{i=1}^{72} \exp \left\{ -\frac{1}{\lambda} \left[ 1 + \xi(x_{ij}) \left( \frac{u(x_{ij}) - \mu(x_{ij})}{\sigma(x_{ij})} \right) \right]^{-1/\xi(x_{ij})} \right\}$$

$$\times \prod_{j=1}^{315} \prod_{i:y_{ij} > u(x_{ij})} \frac{1}{\sigma(x_{ij})} \left[ 1 + \xi(x_{ij}) \left( \frac{y_{ij} - \mu(x_{ij})}{\sigma(x_{ij})} \right) \right]^{-1/\xi(x_{ij})-1}.$$  

$\lambda$: mean number of observations per year.  
$\mu(x_{ij}), \sigma(x_{ij}), \xi(x_{ij})$: PP parameters at $x_{ij}$.  
$\theta$: vector of all model parameters.
Covariate-dependent thresholds

Arguments for:

- Asymptotic justification for EV regression model: the threshold $u(x_{ij})$ needs to be high for each $x_{ij}$.
- Design: spread exceedances across a wide range of covariate values.

Set $u(x_{ij})$ so that $P(Y > u(x_{ij}))$, is approx. constant for all $x_{ij}$.

- Set $u(x_{ij})$ by trial-and-error or by discretising $x_{ij}$, e.g. different threshold for different locations, months etc.
- **Quantile regression (QR)**: model quantiles of a response $Y$ as a function of covariates.
Constant threshold

\[ Y \]

\[ X \]

- estimate of 90\% quantile
Quantile regression

\[ \text{estimate of 90\% quantile} \]
Simple quantile regression in outline

- Data \( \{x_i, y_i\}_{i=1}^n \)
- \( \tau^{th} \) conditional quantile function \( Q_y(\tau|x) = x\phi(\tau) \) estimated by solving:
  \[
  \min_{\phi} \sum_{i=1}^{n} \rho_\tau(y_i - x_i\phi) 
  \]
  where \( \rho_\tau(r) = \tau r - r I(r < 0) \), or (with \( r_i = r_i(\phi) = y_i - x_i\phi \)):
  \[
  \min_{\phi} \left\{ \tau \sum_{r_i \geq 0} |r_i| + (1 - \tau) \sum_{r_i < 0} |r_i| \right\}
  \]
- As a linear program:
  \[
  \min_{\phi, u, v} \left\{ \tau 1_n^T u + (1 - \tau) 1_n^T v \mid x\phi + u - v = y \right\}
  \]
  where \( \{u_i\} \) and \( \{v_i\} \) are **slack** variables corresponding to (absolute values of) positive and negative residuals.
Model parameterisation

Let $p(x_{ij}) = P(Y_{ij} > u(x_{ij}))$. Then, if $\xi(x_{ij}) = \xi$ is constant,

$$p(x_{ij}) \approx \frac{1}{\lambda} \left[ 1 + \xi \left( \frac{u(x_{ij}) - \mu(x_{ij})}{\sigma(x_{ij})} \right) \right]^{-1/\xi}.$$ 

If $p(x_{ij}) = p$ is constant then:

$$u(x_{ij}) = \mu(x_{ij}) + c \sigma(x_{ij}),$$

for some constant $c$.

The form of $u(x_{ij})$ is determined by the extreme value model:

- if $\mu(x_{ij})$ and/or $\sigma(x_{ij})$ are linear in $x_{ij}$: linear QR.
- if $\log(\mu(x_{ij}))$ and/or $\log(\sigma(x_{ij}))$ is linear in $x_{ij}$: non-linear QR.
Adjustment for spatial dependence

- **Independence** log-likelihood:

\[
\ell_{IND}(\theta) = \sum_{j=1}^{k} \sum_{i=1}^{72} \log f_{ij}(y_{ij}; \theta) = \sum_{j=1}^{k} \ell_j(\theta)
\]

(storms) (space)

- If **correct** model specification:

\[
\hat{\theta} \rightarrow N(\theta_0, I^{-1})
\]

- If **model mis-specified**, in regular problems, as \( k \rightarrow \infty \):

\[
\hat{\theta} \rightarrow N(\theta_0, I^{-1} V I^{-1})
\]

- \( I = \) Expected information: \(-E \left( \frac{\partial^2}{\partial \theta^2} \ell_{IND}(\theta_0) \right)\).
- \( V = \) var \( \left( \frac{\partial}{\partial \theta} \ell_{IND}(\theta) \right) \).
Adjustment of $\ell_{IND}(\theta)$

- **Idea:** Adjust $\ell_{IND}(\theta)$ to have correct curvature near $\hat{\theta}$ using sandwich estimate.

\[
\ell_{ADJ}(\theta) = \ell_{IND}(\hat{\theta}) + (\theta - \hat{\theta})' \left( -\hat{I}^{-1} \hat{V} \hat{I}^{-1} \right)^{-1} (\theta - \hat{\theta}) \frac{1}{(\theta - \hat{\theta})'(-\hat{I})(\theta - \hat{\theta})} \left( \ell_{IND}(\theta) - \ell_{IND}(\hat{\theta}) \right)
\]

- Estimate $I$ by observed information at $\hat{\theta}$.

- Estimate $V$ by $\sum_{j=1}^{k} U_j^2(\hat{\theta})$, $U_j(\theta) = \frac{\partial \ell_j(\theta)}{\partial \theta}$.

- **Vertical** adjustment preserves asymptotic distribution of likelihood ratio statistic.

- See Davison (2003), Chandler and Bate (2007).
Summary of modelling of wave height data

- **Threshold selection:**
  - Based on $\mu$ (and $u$) quadratic in longitude and latitude, $\sigma$ and $\xi$ constant . . .
- **Spatial model:**

\[
\mu = \sum_{i=0}^{q_x} \sum_{j=0}^{q_y} \mu_{i+jq_y} \phi_{xi}(l_x) \phi_{yj}(l_y)
\]

where:

- $\phi_0(\cdot) = 1$.
- $\phi_{x1}(l_x) = \frac{1}{5.5} (l_x - 6.5)$, $\phi_{y1}(l_y) = \frac{1}{2.5} (l_y - 3.5)$.
- $\phi_2(\cdot) = \frac{1}{2} (3 \phi_1^2(\cdot) - 1)$, for $l_x, l_y \in [-1, 1]$. 
Threshold selection: $\mu$ intercept

\[ \hat{\mu}_0 \]

\[ \text{probability of exceedance} \]

\[ 0.5 \quad 0.4 \quad 0.3 \quad 0.2 \quad 0.1 \]

\[ 2.5 \quad 3.0 \quad 3.5 \quad 4.0 \quad 4.5 \quad 5.0 \]

\[ \mu^0 \]
Threshold selection: $\mu$ coefficient of latitude

$\hat{\mu}_2$

Probability of exceedance

0.5 0.4 0.3 0.2 0.1

$-0.35 -0.30 -0.25 -0.20 -0.15 -0.10 -0.05$

$\hat{\mu}_2$

Probability of exceedance

0.5 0.4 0.3 0.2 0.1

$-0.35 -0.30 -0.25 -0.20 -0.15 -0.10 -0.05$
Threshold selection: $\xi$
Summary of modelling of wave height data

- Choice of $p$: look for stability in parameter estimates. Use $p = 0.4$.
- $\hat{\xi} = 0.07$, with 95% confidence interval ($-0.05, 0.22$).
- Estimated 200 year return level at (long=7, lat=1) is 15.8m with 95% confidence interval (12.9, 22.3)m.
- Close agreement between parameter estimates for threshold $u$ and point process mean $\mu$. 
Marginal 200 year return levels

![3D surface plot showing marginal 200 year return levels with latitude and longitude axes]
Data-generating process: for covariate values $x_1, \ldots, x_n$:

$$Y_i \mid X = x_i \overset{\text{indep}}{\sim} \text{GEV}(\mu_0 + \mu_1 x_i, \sigma, \xi).$$

Set threshold:

$$u(x) = u_0 + u_1 x.$$

For each $u_1$, set $u_0$ such that the expected proportion of exceedances is kept constant at $p$.

- Calculate Fisher expected information for $(\mu_0, \mu_1, \sigma, \xi)$.
- Invert to find asymptotic V-C of MLEs $\hat{\mu}_0, \hat{\mu}_1, \hat{\sigma}, \hat{\xi}$ and hence $\text{var}(\hat{\mu}_1)$.
- Find the value of $u_1$ that minimises $\text{var}(\hat{\mu}_1)$. 
Findings of *Toy* study 1

Let $\tilde{u}_1$ be the value of $u_1$ that minimises $\text{var}(\hat{\mu}_1)$.

- If covariate values $x_1, \ldots, x_n$ are symmetrically distributed then: $\tilde{u}_1 = \mu_1$ (quantile regression).
- If $x_1, \ldots, x_n$ are positive (negative) skew then $\tilde{u}_1 < \mu_1$ ($\tilde{u}_1 > \mu_1$).

...but the loss in efficiency from using $\tilde{u}_1 = \mu_1$ appears to be small.
Simulation study 2

• 30 years of daily data on a spatial grid.
• Spatial dependence: mimics that of wave height data.
• **Temporal** dependence: moving maxima: extremal index $1/2$ (no declustering)
• Spatial variation: location $\mu$ linear in longitude and latitude.

• $\xi$: $-0.2, 0.1, 0.4, 0.7$.
• Thresholds: 90th, 95th, 99th percentiles.
• SE adjustment: data from distinct years are independent.
• Simulations with no covariate effects and/or no spatial dependence for comparison.
Findings of simulation study 2

- Estimates of regression effects from QR and PP models are very close: both estimate extreme quantiles from the same data.
- Uncertainties in covariate effects of threshold are negligible compared to the uncertainty in the choice of threshold level.
- To a large extent fitting the PP model accounts for uncertainty in the covariate effects at the level of the threshold.
- Slight underestimation of standard errors: uncertainty in threshold ignored.
Conclusions

Quantile regression:

- An intuitive and effective strategy to set thresholds for non-stationary EV models.
- Works well in initial applications.
- Supported by initial theoretical and simulation studies.

Ideas:

- Simultaneous threshold and PP model would avoid iteration (mixed-integer optimisation; see Beirlant et al. 2004).


Northop, P. J. and Jonathan, P. Threshold modelling of spatially-dependent non-stationary extremes with application to hurricane-induced wave heights. Accepted for *Environmetrics*.
Thank you for your attention.