Choosing the threshold in extreme value analysis

EVA 2025

Léo Belzile, HEC Montréal (joint work with Sonia Alouini and Anthony Davison)

Friday, Jun 27, 2025

Motivation: peaks over threshold

In the simplest applications:

1. we choose a high threshold u
   - equivalently, choose the number of upper order statistics n_u with threshold $u = X_{(n-n_u)}$.
2. we fit the limiting generalized Pareto distribution with scale σ_u and shape ξ to exceedances $X - u$ over threshold u.
   - if the shape $\xi > 0$, we can use Hill's estimator
3. we use the resulting model for extrapolation beyond u.

Threshold selection

Bias and variance trade-off:

- taking u too high will mean that the number of exceedances n_u is small (increase uncertainty)
- taking u too low will increase n_u, at the risk of a biased extrapolation.

Since the limit holds as u increases to the upper support point of X, we must use a so-called intermediate sequence $n_u/n \to 0$ as $n_u \to \infty$ for consistency.

Limiting distribution and threshold stability

- Limiting distribution of threshold exceedances is generalized Pareto. **If** the limit was exact and $X - u \sim \mathsf{GPD}(\sigma_u, \xi)$
- For higher threshold $v > u$, then

$$X - v \mid X > v \sim \mathsf{GPD}\{\sigma_v = \sigma_u + \xi(v - u), \xi\}.$$

- Provided $\xi > -1$, the expected value of exceedances is

$$\mathsf{E}(X - u) = \sigma_u/(1 + \xi).$$

Literature review

There are earlier reviews of the topic, but the literature keeps increasing...

- The most comprehensive reviews are Scarrott and MacDonald (2012), Caeiro and Gomes (2016) and Langousis et al. (2016).
- Selective numerical comparisons in Gomes and Oliveira (2001), Murphy, Tawn, and Varty (2025) and Schneider, Krajina, and Krivobokova (2021), among others.

Some problems with (most methods)

Consider selecting a threshold among candidates $u_1 < \cdots < u_k$.

- Sample overlap: data above different thresholds are overlapping (so dependence between estimates or test statistics and P-values).
- Non-nested models: we only model exceedances above u_j (so most models are not nested).
- Multiple testing problem.
- Potential for automation.
- Non-stationarity (dependence on covariate, change over time).

Objective

We provide an extensive review of threshold selection mechanisms for peak over threshold analysis, including

- semiparametric methods based on Hill's estimator (not covered in this presentation for the sake of time),
- visual diagnostics,
- goodness-of-fit tests,
- extended generalized Pareto models,
- etc.

How to benchmark methods?

What makes a threshold procedure good? In practice, we care about the extrapolation, often

- a high quantile (return level), or
- a probability of exceedance.

Benchmarking the method based on proximity with the asymptotic shape parameter is **not** a good point of reference.

Scale and shape parameters are negatively correlated.

Penultimate effects

Smith (1987) show that a better approximation is obtained by letting the shape vary with u, with $\xi_t = r'(u)$, where

$$r(t) = \{1 - F(t)\}/f(t)$$

is the reciprocal hazard function.

`mev` package: `smith.penult(family = "norm", method = "pot", qu = c(0.9, 0.95, 0.99))`.

Illustration of penultimate effects: normal example

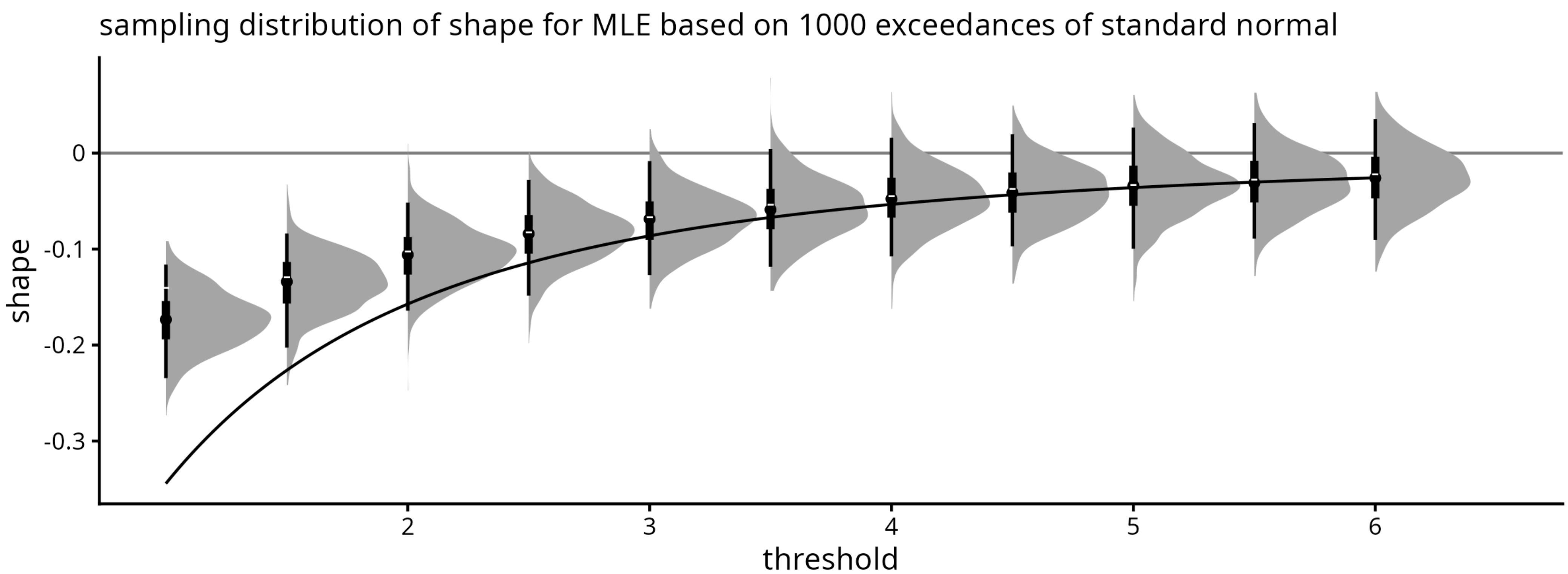
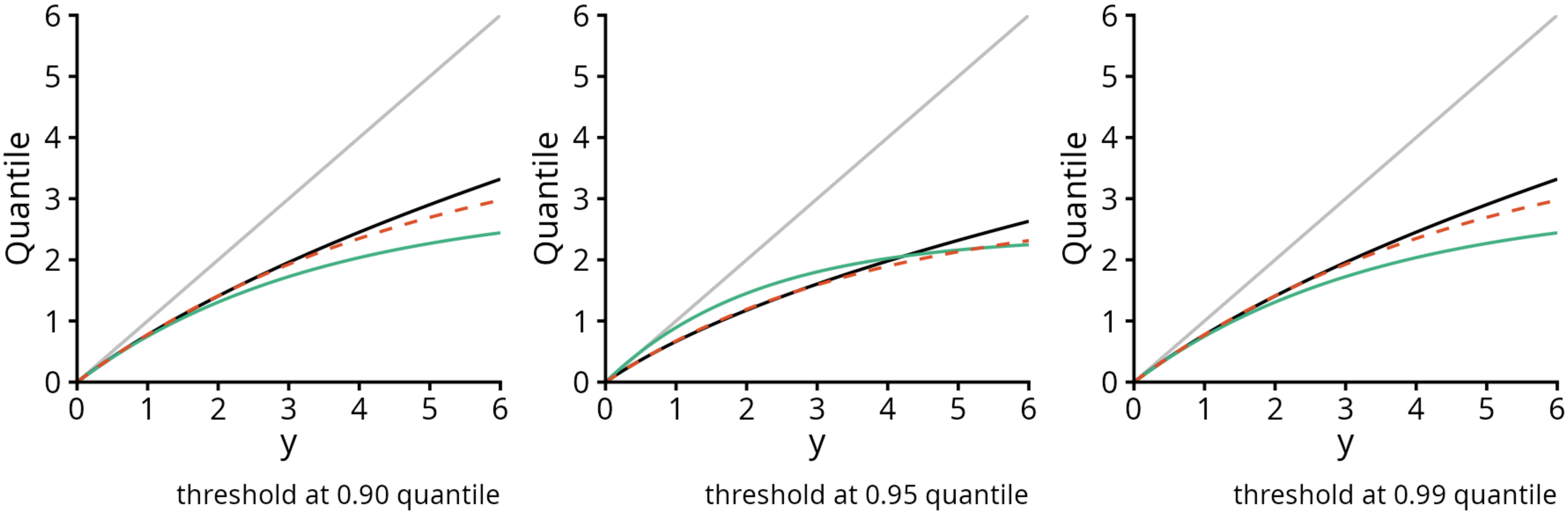


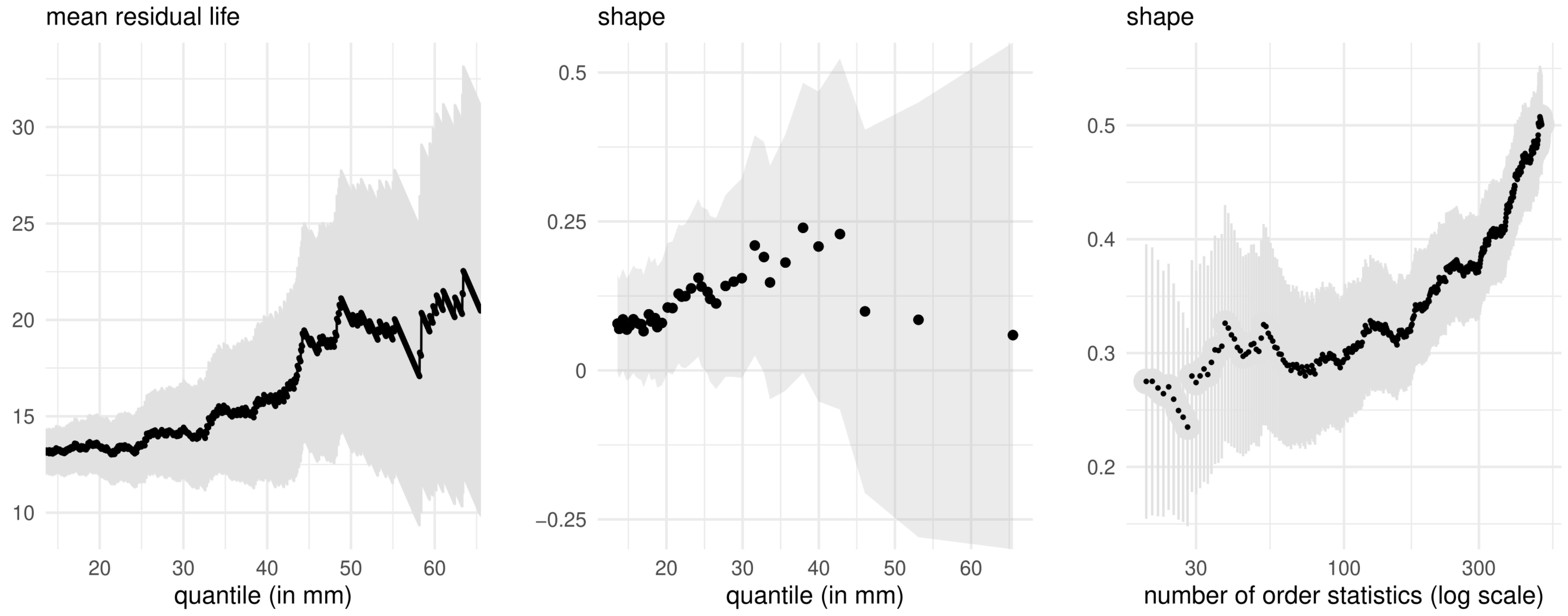
Illustration of penultimate effects: normal quantiles



Conditional quantiles above threshold, rescaled. 45° line shows limiting exponential with true conditional (black), fitted GPD with 1M observations (dashed red) and penultimate (green).

Visual diagnostics on Padova data

Mean residual life (Davison and Smith 1990), parameter stability and Hill plots.

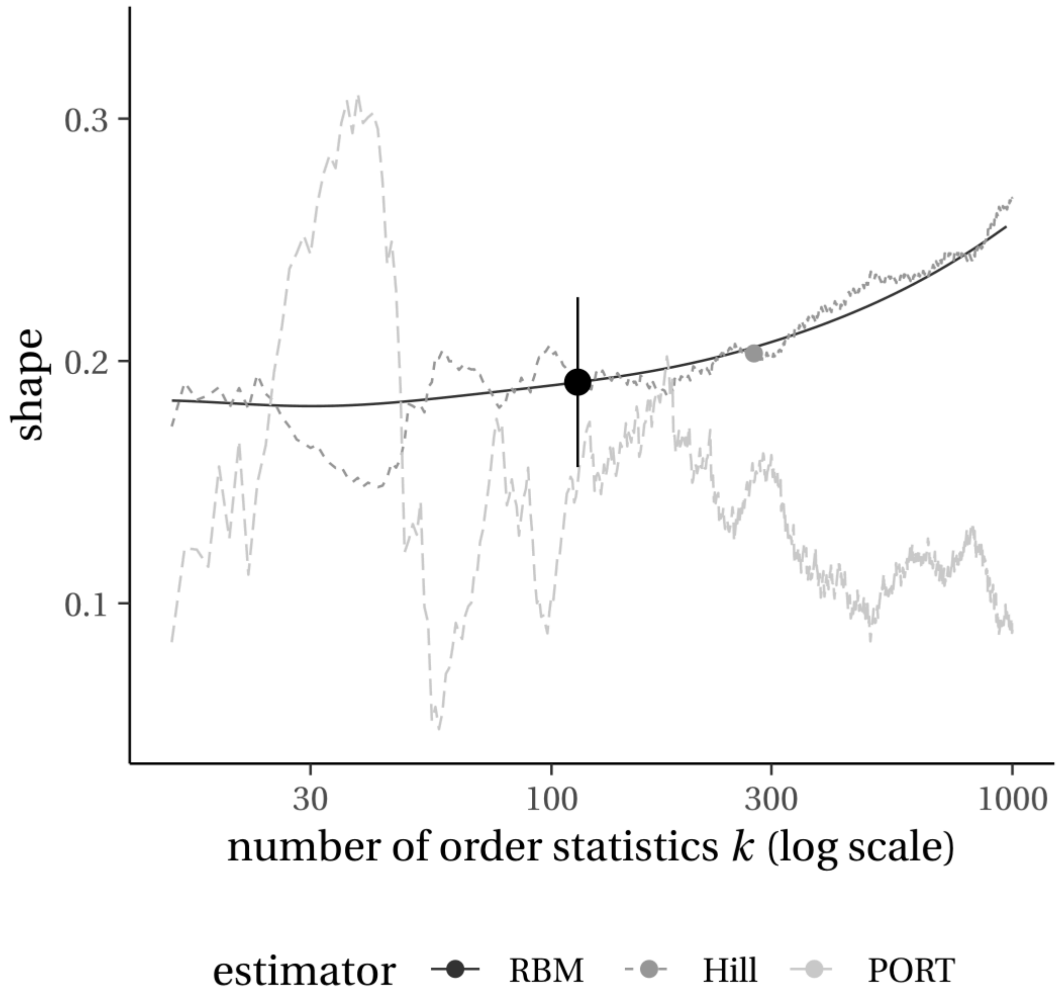


Which threshold would you choose?

Caveats

- Difficulty in automating selection (need visual inspection); proposals in Langousis et al. (2016) or Danielsson et al. (2019), but both fall short.
- Problems: sample overlap leads to dependence between estimates (pointwise confidence intervals).
- Underlying assumptions affected by penultimate effects.
- The plots say nothing about goodness-of-fit!

Stability?



Hill, PORT and random block maxima (Wager 2014) estimates as a function of n_u.

Generalized Pareto model extensions

Build extended models with additional parameters (with continuity constraints) and test for equality of shape.

Some model are inspired by penultimate approximations (GPD-like, but with varying shape).

Northrop and Coleman (2014): piecewise generalized Pareto model coupled with score tests.

- tests are dependent because the data are re-used; no control of overall error rate.
- the power of the test depends on the choice of thresholds and particularly on the largest threshold u_K.

Extended generalized Pareto models

Embed generalized Pareto $F(x;\sigma,\xi)$ in a more flexible model using a continuous distribution function G_κ on $[0,1]$. The EGP(σ,ξ,G_κ) distribution function is then

$$\Pr(X \leq x) = G_\kappa\{F(x;\sigma,\xi)\}.$$

Choose G to keep the tail properties. See the chapter of P. Naveau (2025) for a recent review.

- Papastathopoulos and Tawn (2013): models imply the density at the origin is zero.
- Gamet and Jalbert (2022) propose two models, but one leads to non-regular asymptotics (restriction on boundary of the parameter space).
- Philippe Naveau et al. (2016) additional models with two parameters.

Extended generalized Pareto models

Test for restriction to generalized Pareto sub-model using likelihood ratio tests (profile).

- Could allow for a bit more data to be included, at the expense of additional parameters to estimate (and potentially more variability).
- Same problems as parameter stability plots (sequential tests, overlapping data).

Splicing models

Glue a distribution for the bulk with one for the tail using a mixture of disjoint components below u (bulk) and above u (generalized Pareto). See Scarrott and MacDonald (2012) and Hu and Scarrott (2018).

- Leads to sample contamination, fit may be driven by bulk.
- profile likelihood for threshold u is non-monotone.
- Bayesian version with random threshold (Nascimento, Gamerman, and Lopes 2012) provides threshold selection uncertainty.

Goodness-of-fit measures

1. Fit a generalized Pareto distribution at each candidate threshold.
2. Compute either
   - a suitable statistic which indicates departure from the posulated distribution.
   - a measure of discrepancy between empirical distribution of exceedances above u and generalized Pareto model (via Kolmogorov–Smirnov, Cramér–von Mises, etc.)
3. Perform tests sequentially until rejection, or select the “best” threshold according to the criterion.

Some proposals

- Idea dates back to Pickands (1975).
- Choulakian and Stephens (2001), Bader, Yan, and Zhang (2018) (using ForwardStop).
- Recent proposals using L-moment estimators including Kiran and Srinivas (2021), Solari et al. (2017), Silva Lomba and Fraga Alves (2020).

Note: goodness-of-fit tests null distributions require adjustment for rounded values (estimate null via Monte Carlo).

Testing using maximum likelihood distribution

Thompson et al. (2009) propose using constant values

$$\tau_j = \widehat{\sigma}_j - \hat{\xi}_j u_j$$

and performing Pearson's test of normality for the differences $\tau_{j+1} - \tau_j$ ($j = 1, \ldots, k-1$), stopping whenever the hypothesis is rejected at level $\alpha = 0.2$.

Ad hoc proposal... but works well in simulations.

Sequential analysis

Wadsworth (2016) obtains asymptotic joint distribution of MLE from a superposition of Poisson processes.

Build independent increments of shape to form a white noise sequence $\xi_i^* = (\hat{\xi}_{u_{i+1}} - \hat{\xi}_{u_i}) / \{(I^{-1}_{u_{i+1}} - I^{-1}_{u_i})^{1/2}_{\xi,\xi}\}$.

- Parameter stability plots with simultaneous confidence intervals!
- White noise test for ξ_i^*, with alternative hypothesis $\mathscr{H}_a : \xi_i^* \sim \mathsf{Normal}(\beta, \sigma) (i = 1, \ldots, j-1)$ and $\xi_i^* \sim \mathsf{Normal}(0,1)$ for $j, \ldots, k-1$ motivated by results on model misspecification (White 1982).

Comments on Wadsworth (2016)

Fails 17% of the time in our simulations with equally spaced quantiles.

- Problems with positive definiteness of information matrices for shape increments.
- Method sensitive to rounding (add noise?)
- Must choose the threshold sequence carefully.

Predictive distribution

Northrop, Attalides, and Jonathan (2017) propose a Bayesian method based on leave-one-out cross validation with a binomial-generalized Pareto (BGP) model and a single validation threshold $v > u_k$ above which we assess the model performance.

The measure of goodness-of-fit proposed is an estimate of the negated Kullback–Leibler divergence,

$$\widehat{T}_v(u_i) = \sum_{i=1}^{n} \log\{\widehat{f}_v(x_r \mid \boldsymbol{x}_{-r}, u_i)\}.$$

The selected threshold is the one among the candidates maximizing this diagnostic.

Further propose using Bayesian model averaging to account for the uncertainty originating from threshold selection.

Metric-based diagnostic

Build quantile-quantile (QQ-) plot, with pointwise confidence intervals can be obtained using a parametric bootstrap.

Each bootstrap sample $b = 1, \ldots, B$, estimate the empirical quantile function F_b, which is evaluated at the plotting position $p_i = i/(n+1)$ for $i = 1, \ldots, n$.

Varty et al. (2021) and Murphy, Tawn, and Varty (2025) propose repeating this with simulated iid data from F_0 (tolerance intervals).

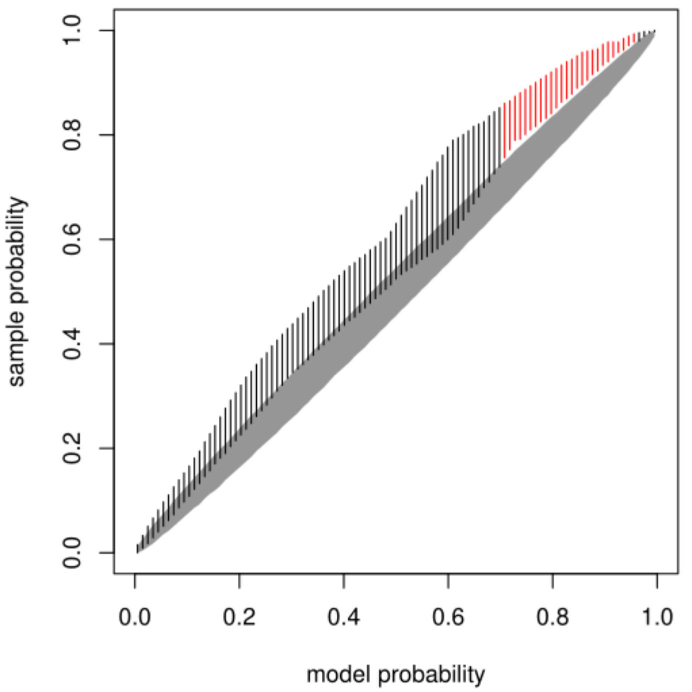


Figure from Varty et al. (2021)

Metric-based adjustment

Build metric based on exponential (or generalized Pareto quantile $F_0^{-1}\{p_i\}$ against $F_b^{-1}(p_i)$ with mean absolute difference or mean squared difference.

Pick the threshold with the smallest average distance is chosen.

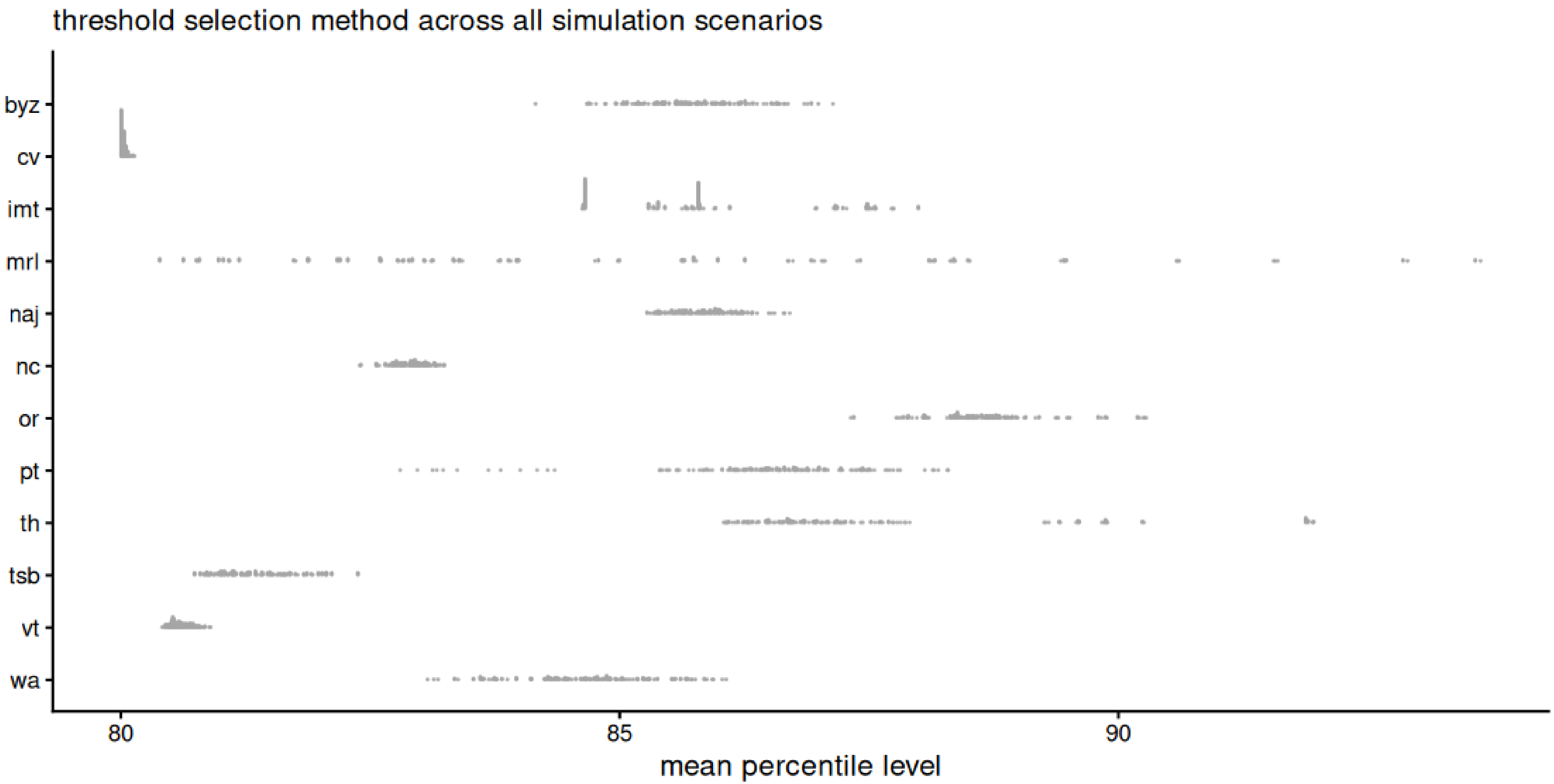
- amenable to different sampling schemes (censoring, non-identically distributed data, time-varying thresholds).
- computationally-intensive by design.

Simulation study: comparison of methods

We considered 13 different distributions from simulation studies in Choulakian and Stephens (2001) and Schneider, Krajina, and Krivobokova (2021).

- Consider 1000 replications of IID data, data with serial correlation in the tail, rounded observations, with different tail behaviour.
- Data of size $n \in \{1000, 2000\}$ with candidate thresholds at the sample $\{0.8, 0.81, \ldots, 0.99\}$ quantiles, keeping a minimum of 20 exceedances.
- Evaluate bias, variance, RMSE of point estimator for 0.999 quantile.
- Compare to oracle (model with the closest quantile estimate among candidates).

Which quantile level on average?



Findings

- No universally better method.
- Many of the automated procedures return the lowest possible threshold.
- Using Forward stop method to account for multiple testing leads to thresholds that are much lower.

More comments

- The oracle method returns average threshold around the 87% and the 90% quantile.
- Mean residual life plots are uniformly scattered.
- Varty et al. (2021)'s metric diagnostic also leads to very small thresholds.
- Northrop, Attalides, and Jonathan (2017) and Thompson et al. (2009) lead to much less agreement and a greater variability of selected quantile levels for the thresholds. For the cross-validation approach, this is in line with findings of Murphy, Tawn, and Varty (2025).
- Wadsworth's sequential testing performs best with heavy tailed distributions, but otherwise is one of the worst in terms of ranking.

Conclusions and future work

Is the problem well-formulated? There is no “correct” threshold, so are we barking up the wrong tree?

Some alternatives:

- Weighting with different threshold choices to account for uncertainty (Stein 2023).
- Should we be fitting sub-asymptotic models to much more data?

Question period

- Thank you for your attention and thanks to the organizers.
- Preprint coming soon.
- Many numerical methods in **R** packages `mev` and `tea`.
- Slides at `lbelzile.github.io/EVA2025-choosing-threshold`

Semiparametric methods

The paper also compares 18 different semiparametric methods using Hill-type estimators for heavy-tailed data.

- Careful: many numerical implementations don't specify a minimum sample size!
- Primer: best methods include Caeiro and Gomes (2016), Gomes, Figueiredo, and Neves (2012), Wager (2014).

Don't use the following

- Methods based on minimization of the asymptotic mean squared error can break down catastrophically for particular data sets.
- Methods by Gomes et al. (2013) and Hall and Welsh (1985) behave erratically with small shape parameters: these procedures lead to strongly biased shape parameter estimates. This is due to them keeping more than 15% of the data for inference.
- Drees and Kaufmann (1998) bias-reduction method leads to large width of confidence intervals (unwanted variability), an
- Many other methods are extremely variable.

References

Bader, Brian, Jun Yan, and Xuebin Zhang. 2018. "Automated Threshold Selection for Extreme Value Analysis via Ordered Goodness-of-Fit Tests with Adjustment for False Discovery Rate." *The Annals of Applied Statistics* 12 (1): 310–29. https://doi.org/10.1214/17-aoas1092.

Caeiro, Frederico, and M. Ivette Gomes. 2016. "Threshold Selection in Extreme Value Analysis." In *Extreme Value Modeling and Risk Analysis: Methods and Applications*, edited by Dipak K. Dey and Jun Yan, 69–86. Boca Raton, FL: CRC Press. https://doi.org/10.1201/b19721.

Choulakian, V, and M. A Stephens. 2001. "Goodness-of-Fit Tests for the Generalized Pareto Distribution." *Technometrics* 43 (4): 478–84. https://doi.org/10.1198/00401700152672573.

Danielsson, Jon, Lerby Ergun, Casper G. de Vries, and Laurens de Haan. 2019. "Tail Index Estimation: Quantile-Driven Threshold Selection." https://doi.org/10.34989/swp-2019-28.

Davison, A. C., and R. L. Smith. 1990. "Models for Exceedances over High Thresholds (with Discussion)." *Journal of the Royal Statistical Society. Series B. (Methodological)* 52 (3): 393–442. https://doi.org/10.1111/j.2517-6161.1990.tb01796.x.

Drees, Holger, and Edgar Kaufmann. 1998. "Selecting the Optimal Sample Fraction in Univariate Extreme Value Estimation." *Stochastic Processes and Their Applications* 75 (2): 149–72. https://doi.org/10.1016/s0304-4149(98)00017-9.

Gamet, Philémon, and Jonathan Jalbert. 2022. "A Flexible Extended Generalized Pareto Distribution for Tail Estimation." *Environmetrics* 33 (6). https://doi.org/10.1002/env.2744.

Gomes, M. Ivette, Fernanda Figueiredo, and M. Manuela Neves. 2012. "Adaptive Estimation of Heavy Right Tails: Resampling-Based Methods in Action." *Extremes* 15 (4): 463–89. https://doi.org/10.1007/s10687-011-0146-6.

Gomes, M. Ivette, Lígia Henriques-Rodrigues, M. Isabel Fraga Alves, and B. G. Manjunath. 2013. "Adaptive PORT–MVRB Estimation: An Empirical Comparison of Two Heuristic Algorithms." *Journal of Statistical Computation and Simulation* 83 (6): 1129–44. https://doi.org/10.1080/00949655.2011.652113.

Gomes, M. Ivette, and Orlando Oliveira. 2001. "The Bootstrap Methodology in Statistics of Extremes—Choice of the Optimal Sample Fraction." *Extremes* 4 (4): 331–58. https://doi.org/10.1023/a:1016592028871.

Hall, Peter, and A. H. Welsh. 1985. "Adaptive Estimates of Parameters of Regular Variation." *The Annals of Statistics* 13 (1): 331–41. https://doi.org/10.1214/aos/1176346596.

Hu, Yang, and Carl Scarrott. 2018. "`evmix`: An R Package for Extreme Value Mixture Modeling, Threshold Estimation and Boundary Corrected Kernel Density

Estimation." *Journal of Statistical Software* 84 (5): 1–27. https://doi.org/10.18637/jss.v084.i05.

Kiran, K. G., and V. V. Srinivas. 2021. "A Mahalanobis Distance-Based Automatic Threshold Selection Method for Peaks over Threshold Model." *Water Resources Research* 57 (e2020WR027534). https://doi.org/10.1029/2020wr027534.

Langousis, Andreas, Antonios Mamalakis, Michelangelo Puliga, and Roberto Deidda. 2016. "Threshold Detection for the Generalized Pareto Distribution: Review of Representative Methods and Application to the NOAA NCDC Daily Rainfall Database." *Water Resources Research* 52 (4): 2659–81. https://doi.org/https://doi.org/10.1002/2015WR018502.

Murphy, Conor, Jonathan A. Tawn, and Zak Varty. 2025. "Automated Threshold Selection and Associated Inference Uncertainty for Univariate Extremes." *Technometrics* 67 (2): 215–24. https://doi.org/10.1080/00401706.2024.2421744.

Nascimento, Fernando Ferraz do, Dani Gamerman, and Hedibert Freitas Lopes. 2012. "A Semiparametric Bayesian Approach to Extreme Value Estimation." *Statistics and Computing* 22 (2): 661–75. https://doi.org/10.1007/s11222-011-9270-z.

Naveau, P. 2025. "Jointly Modeling Bulk and Tails." In *Handbook of Statistics of Extremes*, edited by M. de Carvalho, R. Huser, P. Naveau, and B. J. and Reich, to appear. Boca Raton: Chapman & Hall/CRC.

Naveau, Philippe, Raphael Huser, Pierre Ribereau, and Alexis Hannart. 2016. “Modeling Jointly Low, Moderate, and Heavy Rainfall Intensities Without a Threshold Selection.” *Water Resources Research* 52 (4): 2753–69. https://doi.org/https://doi.org/10.1002/2015WR018552.

Northrop, Paul J., Nicolas Attalides, and Philip Jonathan. 2017. “Cross-Validatory Extreme Value Threshold Selection and Uncertainty with Application to Ocean Storm Severity.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 66 (1): 93–120. https://doi.org/https://doi.org/10.1111/rssc.12159.

Northrop, Paul J., and Claire L. Coleman. 2014. “Improved Threshold Diagnostic Plots for Extreme Value Analyses.” *Extremes* 17 (2): 289–303. https://doi.org/10.1007/s10687-014-0183-z.

Papastathopoulos, Ioannis, and Jonathan A. Tawn. 2013. “Extended Generalised Pareto Models for Tail Estimation.” *Journal of Statistical Planning and Inference* 143 (1): 131–43. https://doi.org/10.1016/j.jspi.2012.07.001.

Pickands, James. 1975. “Statistical Inference Using Extreme Order Statistics.” *The Annals of Statistics* 3: 119–31. https://doi.org/10.1214/aos/1176343003.

Scarrott, Carl, and Anna MacDonald. 2012. “A Review of Extreme-Value Threshold Estimation and Uncertainty Quantification.” *REVSTAT – Statistical Journal* 10 (1): 33–60. https://doi.org/10.57805/revstat.v10i1.110.

Schneider, Laura Fee, Andrea Krajina, and Tatyana Krivobokova. 2021. “Threshold Selection in Univariate Extreme Value Analysis.” *Extremes* 24 (4): 881–913.

https://doi.org/10.1007/s10687-021-00405-7.

Silva Lomba, Jessica, and Maria Isabel Fraga Alves. 2020. "L-Moments for Automatic Threshold Selection in Extreme Value Analysis." *Stochastic Environmental Research and Risk Assessment* 34 (3): 465–91. https://doi.org/10.1007/s00477-020-01789-x.

Smith, Richard L. 1987. "Approximations in Extreme Value Theory." Department of Statistics; Operations Research, University of North Carolina. https://lbelzile.bitbucket.io/papers/Smith-1987-Approximations_in_extreme_value_theory.pdf.

Solari, Sebastián, Marta Egüen, María José Polo, and Miguel A. Losada. 2017. "Peaks over Threshold (POT): A Methodology for Automatic Threshold Estimation Using Goodness of Fit p-Value." *Water Resources Research* 53 (4): 2833–49. https://doi.org/10.1002/2016wr019426.

Stein, Michael L. 2023. "A Weighted Composite Log-Likelihood Approach to Parametric Estimation of the Extreme Quantiles of a Distribution." *Extremes* 26: 469–507. https://doi.org/10.1007/s10687-023-00466-w.

Thompson, Paul, Yuzhi Cai, Dominic Reeve, and Julian Stander. 2009. "Automated Threshold Selection Methods for Extreme Wave Analysis." *Coastal Engineering* 56 (10): 1013–21. https://doi.org/https://doi.org/10.1016/j.coastaleng.2009.06.003.

Varty, Zak, Jonathan A. Tawn, Peter M. Atkinson, and Stijn Bierman. 2021. "Inference for Extreme Earthquake Magnitudes Accounting for a Time-Varying Measurement Process." https://doi.org/10.48550/arXiv.2102.00884.

Wadsworth, J. L. 2016. "Exploiting Structure of Maximum Likelihood Estimators for Extreme Value Threshold Selection." *Technometrics* 58 (1): 116–26. https://doi.org/10.1080/00401706.2014.998345.

Wager, Stefan. 2014. "Subsampling Extremes: From Block Maxima to Smooth Tail Estimation." *Journal of Multivariate Analysis* 130: 335–53. https://doi.org/10.1016/j.jmva.2014.06.010.

White, Halbert. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50 (1): 1–25. https://doi.org/10.2307/1912526.