Supplementary Material

Tail Index Estimation via Top-k Selection and Rationale for Shape-Only GEV Modeling

Hong-Ji Yang[1] and Chung-I Li*[2]

[1]Department of Applied Mathematics, National Dong Hwa University, Taiwan

[2]Department of Statistics, National Cheng Kung University, Taiwan

1 Exploratory Methodology and Discussion on the Selection of Top-k Order Statistics

The practical application of the tail index estimator $\hat{\xi}_k^T$, where $T = P$ or M represents the Pickands or moment estimator, is often challenged by the selection of k, the number of top-order statistics used. Vermaat et al. (2003) provided a numerical evaluation of the extreme-value theory control chart and found that $k = \max\{5, n/500\}$ is a reasonable choice. However, in theory, choosing k requires balancing bias and variance: increasing k leads to more significant bias due to weaker tail convergence while decreasing k results in higher variance (Kotz and Nadarajah 2000, p. 81). Gomes and Guillou (2015) indicated that the optimal k, which minimizes the mean squared error, depends on the sample size and the unknown parameters ξ and ρ, posing challenges for practical implementation. Notably, ρ is the non-positive second-order parameter that controls the speed of convergence in EVT; however, the estimator $\hat{\rho}$ also depends on k (Alves, de Haan, and Lin 2003).

To confront this problematic and ambiguous practice, as underscored in Section 6.4.2 by Embrechts, Klüppelberg, and Mikosch (2013), it is noted that

> *"The optimal choice of k is a delicate point for which no uniformly best solution exists. It is intuitively clear that one should choose $\hat{\xi}_k$ from a k-region where the plot is roughly horizontal."*

Researchers suggest visualizing $\{(k, \hat{\xi}_k)\}$ to identify a stable regime where the estimator stabilizes, e.g. Drees, de Haan, and Resnick (2000); de Sousa and Michailidis (2004). Theoretical limit theorems dictate that k should increase with n while remaining a decreasing

*CONTACT: Chung-I Li. Email: cili@ncku.edu.tw proportion of the sample size (i.e., $k \equiv k(n) \to \infty$ and $k/n \to 0$, as $n \to \infty$) in order to get asymptotic normality or even consistency for $\hat{\xi}_k$ (Alves, de Haan, and Lin 2003).

In this study, we introduce a feasible statistical justification method to identify stable regimes and determine the smallest k for $\hat{\xi}_k^T$. Inspired by Embrechts, Klüppelberg, and Mikosch (2013), we focus on regions where the plot is approximately horizontal, facilitating stable and consistent estimation. For implementation, we recommend plotting $\{(k, \hat{\xi}_k^T)\}$. For the Pickands estimator ($T = P$), k ranges from 1 to $\lfloor (n+1)/4 \rfloor$, while for the moment estimator ($T = M$), k typically ranges from 1 to $n - 1$. To reduce bias in tail index estimation, we restrict k to $\lfloor (n+1)/4 \rfloor$ rather than $n-1$. The method identifies the earliest stable point $k = k^*$ within extended segments, where $\hat{\xi}_k^T = \hat{\xi}_{k^*}^T$. Our proposed approach is akin to Einmahl, Li, and Liu (2006); a method can be employed to select k, focusing on the first stable part of the graph $\{(k, \hat{\xi}_k^T)\}$ within the visual inspection framework.

To emphasize our methodology, we employ the *recursive segmentation and permutation* (RS/P) approach, initially proposed by Capizzi and Masarotto (2013), which has been demonstrated to be effective in identifying stable regions, assessing stability, and detecting characteristic changes in data points. This approach provides a concise and visually interpretable analysis. The `dfphase1` R package, detailed by Capizzi and Masarotto (2018), facilitates the stability assessment of the location and scale in the i.i.d. data sequence. In our study, we handle dependent data sequences by modeling them as "i.i.d." to approximately assess visual stability horizontally within the data. Specifically, we apply the $\{(k, \hat{\xi}_k^T)\}$ data sequence to assess its ability to detect segments indicative of stability—key for reliable estimations using Pickands and moment methods. The `rsp()` function enriches visual and numerical evaluations, helping to identify stability zones. However, in light of the observations by Embrechts, Klüppelberg, and Mikosch (2013) regarding the selection of k, a definitive and easily implementable methodology has not yet been established. In the absence of superior alternatives to date (to the best of our knowledge), we propose the RS/P application as a viable and straightforward method for exploratory analysis in the selection of k. Using graphing techniques to diagnose stability horizontally within the data, we verify that, within this context, the estimation of the tail index achieves a balance between bias and variance while also demonstrating asymptotic consistency in finite samples (de Sousa and Michailidis 2004).

To illustrate this procedure, we refer to Figure 1, which shows the Pickands estimates by the top-kth statistics plot of $\{(k, \hat{\xi}_k^P)\}$ for $k = 1, 2, \ldots, \lfloor (n+1)/4 \rfloor = 125$. These estimates were obtained from $n = 500$ i.i.d. samples from the distribution of Exp(1) with $X \sim F_X(x) = 1 - \exp(-x)$ with $x > 0$, known $F_X \in \mathcal{D}(G_{\xi=0})$. After analyzing the changepoints using the RS/P approach in a given location, particularly at $k = 12, 33$, and 65, with a focus on deviation at $k = 26$, the quantity of deviation, represented by $|\hat{\xi}_k^P - \hat{\mu}_i|, k = 1, 2, \ldots, 125$, where $\hat{\mu}_i, i = 1, 2, 3, 4$, indicates the mean of the segment (depicted by the dashed line) at the specified location within the context of Figure 1. Specifically, $\hat{\mu}_1 = \sum_{k=1}^{11} \hat{\xi}_k^P / 11$, $\hat{\mu}_2 = \sum_{k=12}^{32} \hat{\xi}_k^P / 21$, $\hat{\mu}_3 = \sum_{k=33}^{64} \hat{\xi}_k^P / 32$, and $\hat{\mu}_4 = \sum_{k=65}^{125} \hat{\xi}_k^P / 61$. Figure 2 is presented to further examine the extreme value index. Within Figure 2, by consolidating the changepoints in location and deviation, we obtain $k = 12, 26, 33$, and 65. We identify the longest recent segment by computing $\max\{12-1, 26-12, 33-26, 65-33, 125-65\} = 125 - 65$, starting from $k = 65$ and ending at $k = 125$. This analysis concludes that the recent stable point is at $k^* = 65$. Consequently, the corresponding Pickands estimate is $\hat{\xi}^P_{65} = 0.0542$, which is in close agreement with the true value of $\xi = 0$. We can perform a similar process for the moment estimator to identify the recent stable point $k^* = 13$ in the plot of $\{(k, \hat{\xi}^M_k)\}$ for $k = 1, 2, \ldots, 125$. The resulting estimate is $\hat{\xi}^M_{13} = 0.0994$, slightly higher than $\hat{\xi}^P_{14} = 0.0542$, but still close to the true value $\xi = 0$. Refer to Figures 3 and 4 to visualize these results.

However, these results indicate that the recent stable point is at $k^* = 65$ for the Pickands estimator and $k^* = 13$ for the moment estimator. This difference highlights that, in practical applications, the optimal value of k should be determined individually for each method, as the estimators exhibit distinct convergence behaviors and sensitivity to tail characteristics.

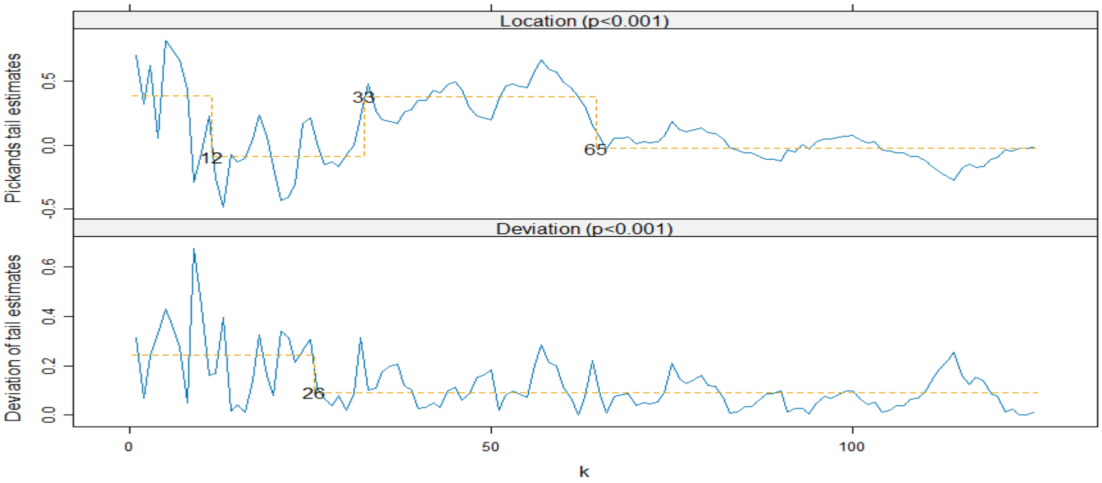

Figure 1: Stability assessment of Pickands tail index estimates in Location and Deviation by top-kth statistics using the R Package `dfphase1`'s `rsp()` function. The significance level is 0.05, which is used to compute the location and deviation estimates; if one of the p-values is greater than 0.05, the corresponding estimate is constant. (Note: the p-values are under the null hypothesis, indicating the absence of any changepoint)

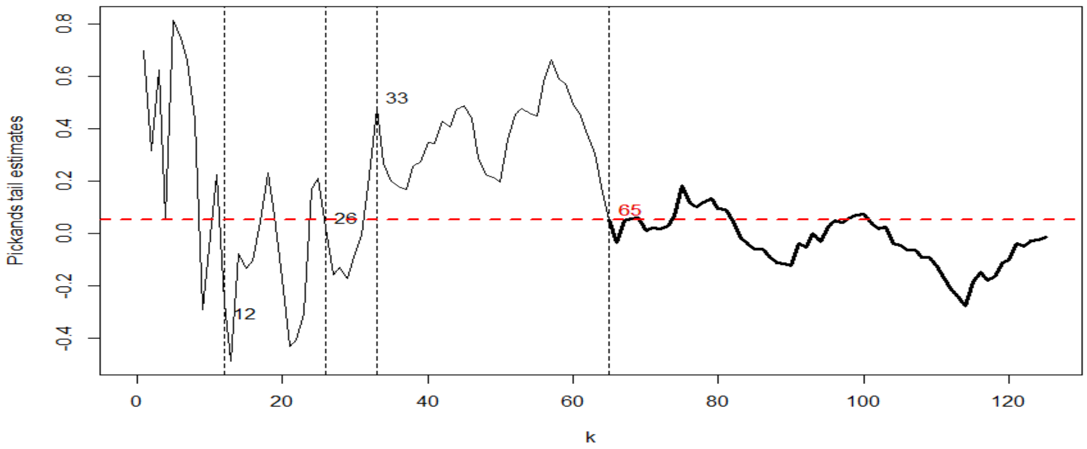

Figure 2: Plot of Pickands tail estimates against top-kth statistics, denoted as $\{(k, \hat{\xi}_k^P)\}$, reveals the recent stable point at $k = 65$, where the estimated tail index is $\hat{\xi}_{65}^P = 0.0542$.

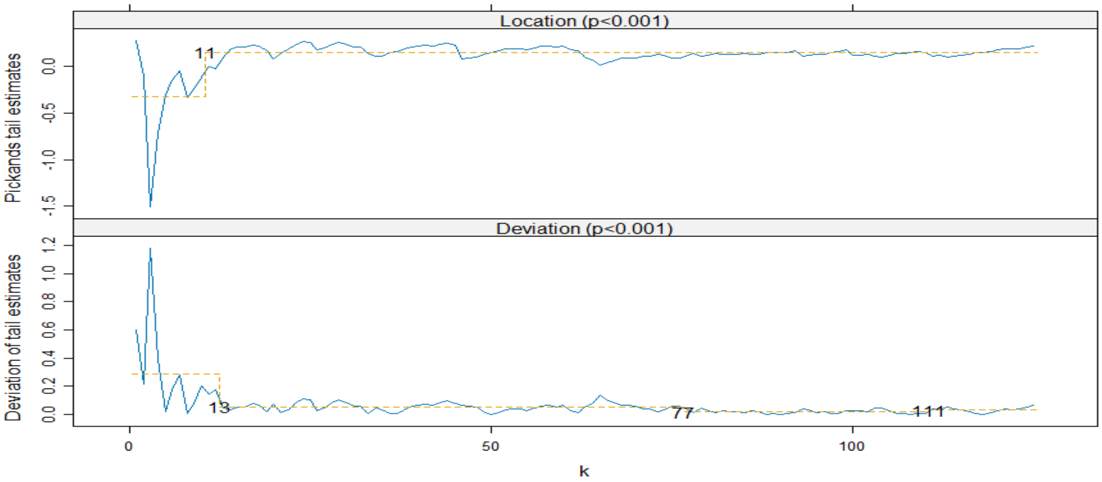

Figure 3: Stability assessment of moment tail index estimates in Location and Deviation by top-kth statistics using the R Package `dfphase1`'s `rsp()` function. The significance level is 0.05, which is used to compute the location and deviation estimates; if one of the p-values is greater than 0.05, the corresponding estimate is constant. (Note: the p-values are under the null hypothesis, indicating the absence of any changepoint)

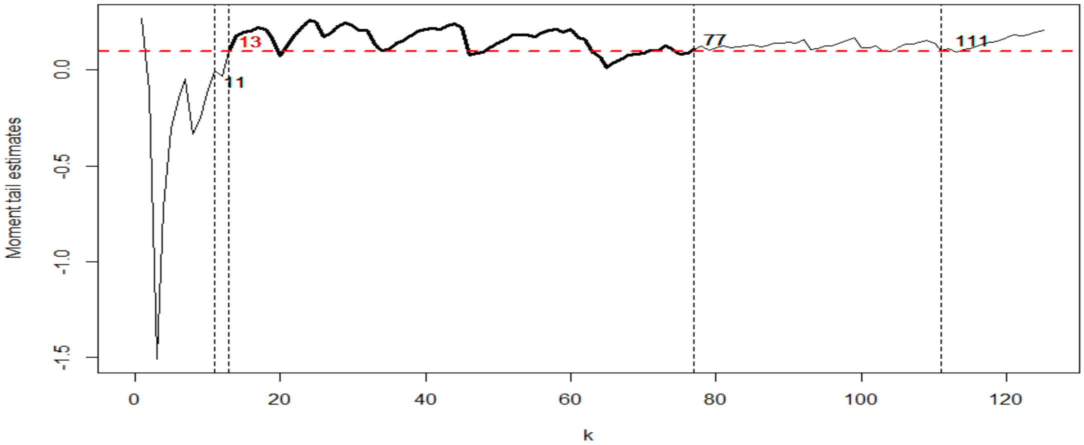

Figure 4: Plot of moment tail estimates against top-kth statistics, denoted as $\{(k, \hat{\xi}_k^M)\}$, reveals the recent stable point at $k = 13$, where the estimated tail index is $\hat{\xi}_{13}^M = 0.0994$.

2 Response to Reviewer Comment: Justification for Omitting Location and Scale Parameters in GEV Modeling

This supplementary section addresses the reviewer's comment regarding why the proposed method estimates only the shape parameter ξ, rather than the full set of location, scale, and shape parameters typically used in the generalized extreme value (GEV) distribution (Coles 2001).

The cumulative distribution function (CDF) of the GEV distribution is defined as:

$$G(x;\mu,\sigma,\xi) = \begin{cases} \exp\left\{-\left[1+\xi\left(\dfrac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, & \text{if } \xi \neq 0, \\ \exp\left\{-\exp\left(-\dfrac{x-\mu}{\sigma}\right)\right\}, & \text{if } \xi = 0, \end{cases}$$

where $\mu \in \mathbb{R}$ is the location parameter, $\sigma > 0$ is the scale parameter, and $\xi \in \mathbb{R}$ is the shape parameter.

This study focuses exclusively on estimating the *shape parameter* ξ of the generalized extreme value (GEV) distribution, as it fundamentally governs the tail behavior that is critical for constructing control limits under the exceedance probability criterion (EPC). In contrast, the *location* μ and *scale* σ parameters primarily affect the central tendency and dispersion of the distribution, and are therefore less relevant when the objective is to extrapolate extreme quantiles.

While the estimation of μ and σ becomes necessary in applications requiring precise prediction or the transformation of estimated quantiles back to the original data scale, such inversion can be conducted using the GEV quantile function. Specifically, given the cumulative distribution function $G(x_p)$, the corresponding quantile x_p can be recovered as:

$$x_p = \begin{cases} \mu - \dfrac{\sigma}{\xi}\left[1 - (-\log p)^{-\xi}\right], & \text{if } \xi > 0 \text{ and } p \in [0,1) \text{ or } \xi < 0 \text{ and } p \in (0,1], \\ \mu - \sigma \log\left(-\log p\right), & \text{if } \xi = 0 \text{ and } p \in (0,1). \end{cases}$$

However, as emphasized by Tawn (1988), *"In most practical situations it is not the estimation of the parameters of the model [GEV] that is of interest, but the quantiles of the distribution. In particular, in the year n level x_p."* This perspective aligns with our objective: to estimate extreme quantiles that define control limits, rather than fully fitting a parametric model.

By adopting a semiparametric approach grounded in extreme value theory (EVT), the proposed method avoids imposing strict assumptions on the location and scale parameters (μ and σ), which can be particularly sensitive to small sample sizes and susceptible to model misspecification in Phase I settings. Instead, we rely on normalized order statistics and tail index estimation—using the Pickands and moment estimators—to directly characterize tail behavior. This allows for the construction of EPC-compliant control limits without explicit estimation of μ and σ. As demonstrated in Equations (12) and (20) of the manuscript, our method estimates the extreme quantiles x_p corresponding to a desired confidence level by leveraging the properties of the tail index alone. This focus enables robust extrapolation beyond the observed data while maintaining flexibility across diverse distributional shapes—an advantage when full parametric modeling is either impractical or unreliable.

References

Alves, MI Fraga, Laurens de Haan, and Tao Lin. 2003. "Estimation of the parameter controlling the speed of convergence in extreme value theory." *Mathematical Methods of Statistics* 12 (2): 155–176.

Capizzi, Giovanna, and Guido Masarotto. 2013. "Phase I distribution-free analysis of univariate data." *Journal of Quality Technology* 45 (3): 273–284.

Capizzi, Giovanna, and Guido Masarotto. 2018. "Phase I distribution-free analysis with the R package dfphase1." In *Frontiers in Statistical Quality Control 12*, 3–19. Springer.

Coles, Stuart. 2001. *An introduction to statistical modeling of extreme values.* Vol. 208. Springer.

de Sousa, Bruno, and George Michailidis. 2004. "A diagnostic plot for estimating the tail index of a distribution." *Journal of Computational and Graphical Statistics* 13 (4): 974–995.

Drees, Holger, Laurens de Haan, and Sidney Resnick. 2000. "How to make a Hill plot." *The Annals of Statistics* 28 (1): 254–274.

Einmahl, John, J Li, and RY Liu. 2006. *Extreme Value Theory Approach to Simultaneous Monitoring and Thresholding of Multiple Risk Indicators.* Technical Report. Tilburg University, Center for Economic Research.

Embrechts, Paul, Claudia Klüppelberg, and Thomas Mikosch. 2013. *Modelling Extremal Events: for Insurance and Finance.* Vol. 33. Springer Science & Business Media.

Gomes, M Ivette, and Armelle Guillou. 2015. “Extreme value theory and statistics of univariate extremes: a review.” *International statistical review* 83 (2): 263–292.

Kotz, Samuel, and Saralees Nadarajah. 2000. *Extreme value distributions: theory and applications.* world scientific.

Tawn, Jonathan A. 1988. “An extreme-value theory model for dependent observations.” *Journal of Hydrology* 101 (1-4): 227–250.

Vermaat, MB, Roxana A Ion, Ronald JMM Does, and Chris AJ Klaassen. 2003. “A comparison of Shewhart individuals control charts based on normal, non-parametric, and extreme-value theory.” *Quality and Reliability Engineering International* 19 (4): 337–353.