Outlier Identification
in
Spatio-Temporal Processes

by

Shrijita Bhattacharya

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2018

Doctoral Committee:

Professor Stilian Stoev, Co-Chair
Professor George Michailidis, Co-Chair
Professor Veronica Berrocal
Research Scientist Michael Kallitsis
Assistant Professor Gongjun Xu

Shrijita Bhattacharya

shrijita@umich.edu

ORCID iD: 0000-0001-6958-8613

© Shrijita Bhattacharya 2018

All Rights Reserved

This document is dedicated to my parents and my husband

ACKNOWLEDGEMENTS

I would like to express my heartfelt thanks to my adviser Professor Stilian Stoev for his constant guidance, valuable expert suggestions and encouragement. I would like to thank him for giving me the opportunity to work with him at the Department of Statistics, University of Michigan, and providing my research with financial support.

I would also like to thank Professor George Michailidis and Dr. Michalis Kallitsis for their guidance and financial support on two of my projects closely related to the domain of engineering. I would like to thank Merit Network Inc for providing me with data on Internet traffic flows and Electricity consumption on power grids, an indispensable part of the work on anomaly detection in networks.

I would also like to thank all my friends in Ann Arbor and East Lansing, Michigan, especially Raka Mandal, Pramita Bagchi, Sebanti Sengupta, Anwesha Bhattacharya, Diptavo Dutta and Aritra Guha. Without their support and camaraderie, Phd life would have been a lot more difficult to cope up with.

Last, but not the least, I would also like to express my intense gratitude to my parents, Rita and Mrityunjoy Bhattacharya and my husband, Abir Chatterjee for their constant encouragement and motivation all along my Phd Career.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF ALGORITHMS

Algorithm

LIST OF APPENDICES

Appendix

ABSTRACT

This dissertation answers some of the statistical challenges arising in spatio-temporal data from Internet traffic, electricity grids and climate models. It begins with methodological contributions to the problem of anomaly detection in communication networks. Using electricity consumption patterns for University of Michigan campus, the well known spatial prediction method *kriging* has been adapted for identification of false data injections into the system. Events like Distributed Denial of Service (DDoS), Botnet/Malware attacks, Port Scanning etc. call for methods which can identify unusual activity in Internet traffic patterns. Storing information on the entire network though feasible cannot be done at the time scale at which data arrives. In this work, hashing techniques which can produce summary statistics for the network have been used. The hashed data so obtained indeed preserves the heavy tailed nature of traffic payloads, thereby providing a platform for the application of extreme value theory (EVT) to identify heavy hitters in volumetric attacks. These methods based on EVT require the estimation of the tail index of a heavy tailed distribution. The traditional estimators (*Hill* (1975)) for the tail index tend to be biased in the presence of outliers. To circumvent this issue, a trimmed version of the classic Hill estimator has been proposed and studied from a theoretical perspective. For the Pareto domain of attraction, the optimality and asymptotic normality of the estimator has been established. Additionally, a data driven strategy to detect the number of extreme outliers in heavy tailed data has also been presented. The dissertation concludes with the statistical formulation of m-year return levels of extreme climatic events (heat/cold waves). The Generalized Pareto distribution (GPD) serves as good fit for modeling peaks over threshold of a distribution. Allowing the parameters of the GPD to vary as a function of covariates such as time of the year, El-Ni$\tilde{n}$o and location in the US, extremes of the areal impact of heat waves have been well modeled and inferred.

CHAPTER I

Introduction

Research on modeling, analysis and inference of spatio-temporal processes has been gaining pace over the last century owing to their predominance in many application domains such as Internet traffic, communication networks, climatic events, time evolving social networks, the stock market etc. In this dissertation, we analyze a pool of techniques for the detection of anomalous events in multivariate time series, each of which is contingent on the nature of data at hand. For example, electricity consumption patterns are naturally modeled by Gaussian distributions, whereas heavy tailed or power law distributions are prevalent in data arising from the stock markets or Internet traffic. Therefore this dissertation has developed a wide range of methods ranging from kriging, community detection to extreme value theory.

Chapter II. The first part discusses statistical tools for detecting intrusions and failures in the electric smart gridin order to prevent widespread power outages and/or breakdown of the electrical system in an area. A smart grid is composed of a multitude of components, each with its own functionality such as power generation, transmission and distribution. These components exchange information through the so called Advanced Meter Infrastructure (AMI), which facilitates efficient operation of the grid by allowing for near-real time adaptation to changes in demand. The AMI is often vulnerable to attacks from malicious users especially from those with access to the state or topology of the electric grids. An example of this includes tampering of meter readings by introducing attack patterns which evade standard detection algorithms. An added disadvantage is the scarcity of resources or physical access needed for securing all the meters in the network. Kriging is a spatial prediction technique which allows for optimal estimation of unobserved quantities from available closely correlated observations. Therefore if a small subset of nodes in the network is secured, kriging may used to curb false data injections on the remaining untrusted nodes. If the size of the network is large, clustering based techniques may be used to obtain a subset of nodes with similar energy patterns and thus reduce the dimensionality of the problem. In Section 2.1.5, this method has been adapted for application to real-world building data from a large university campus.

The second part of Chapter II involves a rather different spatio-temporal data arising from the monitoring of Internet traffic at a large regional Internet Service Provider (ISP). The predominance of Internet in every day life has made it all the more susceptible to attacks from a multitude of sources. Bank frauds, cyber threats, password hacks etc. have made the reliability of information transferred via Internet questionable. Many companies like Akamai are constantly updating their Content Delivery Network (CDN), to prevent Distributed denial of service (DDoS) attacks on their client networks. Network monitoring is essential to network engineering, capacity planning and prevention / mitigation of threats. Section 2.2 describes an open source architecture AMON (All-packet MONitor) deployed at Merit Network[1] which is currently processing 10Gbps+ live Internet traffic. The main challenge in the analysis of network traffic is the shortage of memory resources for the storage of

[1] https://www.merit.edu/

information exchange for such a large network. Also most of the existing methods for anomaly detection are not scalable to the rate of flow for incoming traffic. To circumvent the issue of memory constraint, AMON partitions traffic into sub-streams by using rapid hashing and keeps track of heavy hitter[2] IPs by employing a Boyer-Moore majority algorithm.

Application of optimally chosen hashing techniques preserve most of the important statistical properties of the underlying network traffic flows. These rapidly (online) computed hash-summaries may be thus used for identification of anomalous events in the network such as heavy hitters, DDoS, scanning or outrages. From statistical perspective, this can be cast into the problem of identifying outliers or change-points in multivariate, heavy-tailed time series. Volumetric attacks are identified as extreme outliers in the data and are best detected by the application of Extreme Value Theory. Specifically, robust and adaptive estimates of the heavy-tail exponent of the data are utilized to calibrate an anomaly detection threshold. On the other hand, the stealthier attacks arising from scanning or low-value DDoS are identified as high-connectivity events from a graphical data structure that quantifies the source-destination communication patterns. This calls for other sophisticated techniques like the ones based on community-detection type statistics. These themes are addressed in Sections 2.2.3, 2.2.4 and 2.2.5.

Chapter III. Most of the detection algorithms of Section 2.2 in Chapter II require the estimation of the tail exponent of a *heavy-tailed* distribution. However, in the presence of outliers, the classic Hill estimator is biased and its variance also compromised. In Chapter III, we thereby introduce and study a trimmed version of the Hill estimator for the index of a heavy-tailed distribution, which is robust

[2] IP contributing to usually large traffic in the network

to perturbations in the extreme order statistics. In the ideal Pareto setting, the estimator is shown to be finite-sample efficient among all unbiased estimators with a given strict upper break-down point. For general heavy-tailed models, the asymptotic normality of the estimator under second order regular variation conditions has been established. The estimator is shown to achieve the minimax optimal rate in the Hall class of distributions. A trimmed Hill plot to visually select the number of top order statistics has been proposed. The main contribution is the development of an automatic, data-driven procedure for the choice of trimming based on exponentially weighted sequential testing. This yields a new type of robust estimator that can adapt itself to the unknown level of contamination in the extremes. As a by-product we also obtain a methodology for identifying extreme outliers in heavy tailed data. The competitive performance of the trimmed Hill and adaptive trimmed Hill estimators is illustrated with simulations against several established robust estimators.

Chapter IV. Extremes of weather conditions, be it high or low, may have a devastating impact on the agricultural and industrial production of a country. As stated by Christopher R. Adams[3]: *"In the US, the 1976 - 1977 winter freeze and drought is estimated to have cost $36.6 billion in 1980. In 1980 the nation saw a devastating heat wave and drought that claimed at least 1700 lives and had estimated economic costs $15 - $19 billion in dollars"*. In Chapter IV, we develop a statistical framework for prediction of areal impact of heat waves. The methodology applies to cold waves as well. The approach adopted is the quantification of the area in US under profound heat wave activity at given time point. The Pickands-Balkema-de Haan theorem as well as extensive model diagnostic plots reveal that the generalized Pareto distribution, GPD serves as an efficient tool in modeling the peaks over threshold for

[3] http://sciencepolicy.colorado.edu/socasp/weather1/adams.html

the so-obtained time series. Several other factors like intensity level of the heat wave, grid network of the US, season of the year and duration of heat wave events have been explored in connection to the analysis of heat wave distribution. As a main contribution, we obtain estimates for the out-of-sample return levels for a variety of heat wave events as a function of the season (time of the year), ENSO index and location in US. These estimates are based on the analysis of daily temperature records for a period of 100 years for 424 stations spread across the continental US.

In summary, the dissertation is organized as follows. Chapter II presents two different methodological contributions to anomaly detection in smart grids and computer networks, respectively. These methods motivate the study of robust estimators of heavy tail index in Chapter III. Chapter IV focuses on spatio-temporal inference in the context of extreme heat wave events over the continental US. Finally the dissertation is concluded with the scope of each chapter.

CHAPTER II

Anomaly Detection In Networks

2.1 Security Of Smard Grid Electrical Units

Contributions and due credit: Much of the material in this section is a collaborative work based on *Kallitsis et al.* (2016c). As the author of the thesis, we have contributed primarily to the statistical aspects in Sections 2.1.2, 2.1.3 and 2.1.4 and the proof of Proposition II.1.

2.1.1 Introduction

Smart grid meters were developed to overcome some of the drawbacks of traditional electric grids. For example, accurate estimation of the state of grid, incorporation of renewable energy sources etc. are some of the features specific only to smart grids. The efficient communication between the various components of the smart grid is facilitated by advanced metering infrastructure (AMI). Engineers are mostly interested in the security of AMI in order to ensure that the network can recover itself in the presence of anomalies/ power outages etc *Farhangi* (2010).

The susceptibility of the smart grids to attacks has increased only very recently. The smart grid infrastructure is often jeopardized by individuals who can manipulate the meter readings or inject unwanted load into the system. Some such activities which have resulted in the breakdown of the grid include the Stuxnet worm and the attacks against Iranian nuclear facilities *Falliere et al.* (2011), the compromise of a steel mill in Germany *Lee et al.* (2014), and the cyber attacks on the Ukrainian power gird *Lee et al.* (2016).

In order to control for these malicious behavior, AMI meters are often accessed remotely. This however does not put an end to network changes caused by spoofed message payloads that carry power demand / supply values. Since the power consumption in an electric grid is usually determined by the state or topology matrix *H Liu et al.* (2009), adversaries with access to it can severely compromise the meter readings. Over the past few years, a lot of research has been done on attack which can evade the security protocols of AMI (see *McDaniel and McLaughlin* (2009); *Metke and Ekl* (2010); *Bed and DOE* (2009); *Yu* (2015)).

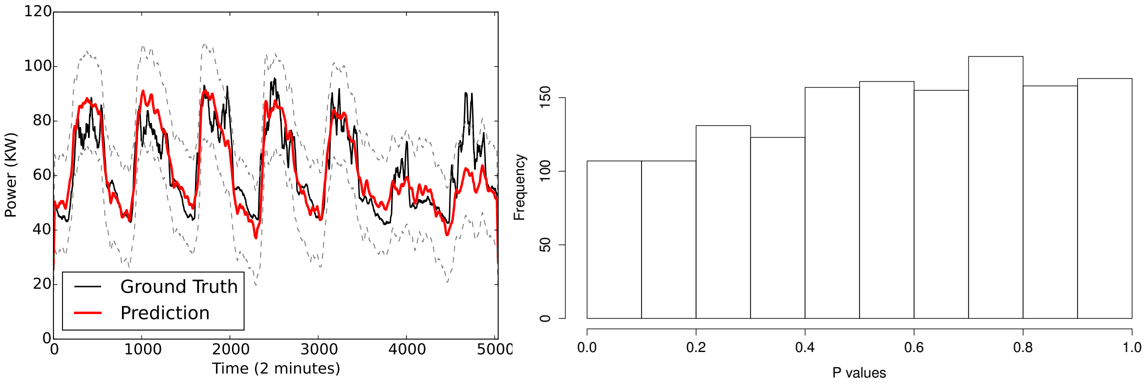


Figure 2.1: *Left*: Power prediction (with 95-percentile bounds). *Right:* Model validation (real-data)
.

In this section, a statistical methodology for the detection of *bad data* injection into wide-area smart grid networks has been proposed. We however make a crucial assumption that the attackers can access only a subset of the total number of available meters. The other meters are however secure and can be used to predict the energy consumption patterns of the remaining ones. Since in an electric network, a large number of meters are closely related in space (e.g., within the same residential neighborhood, university campus, town, etc.), the spatial algorithms which borrow strength from highly correlated observations may be used. Kriging is one such method and is suitable for the data at hand subject to some modifications.

For Section 2.1, we assume that trusted readings involve nodes that transmit encrypted data and whose identity is authenticated *Bi and Zhang* (2014). The set of these trusted nodes shall be referred to as the observed set and the remaining ones are categorized into the unobserved set. The rest of the work is organized as follows: Section 2.1.2 discusses clustering algorithms to group buildings with high correlation index in terms of electricity consumption, Section 2.1.3 details a factor model which illustrates the energy patterns in the network can be explained by just a few factor variables, Section 2.1.4 explains an adaptation of the kriging technique to allow for detection of anomalous observation in the system. Section 2.1.5 finally evaluates the proposed methodology when applied to real world electricity data from University of Michigan campus.

2.1.2 Building grouping

Let $Y(t) = (Y_i(t))_{i \in \mathcal{B}}$ be the time series of electricity usage for a set of $\mathcal{B} = \{1, \cdots, B\}$ buildings recorded over the time $t = 1, 2, \cdots,$. For a monitoring window of size m, we define the $m \times B$ as

$$\mathbf{D}(t_0, m) = [Y_i(t)]_{t_0 - m \leq t < t_0,\ i \in \mathcal{B}}. \tag{2.1}$$

Next the data $Y(t)_{t=1,2,\cdots,}$ is partitioned into windows of size m and the mean consumption for each window is considered as as

$$\mu(t_0, m) = \frac{1}{M} \sum_{w=0}^{M-1} \mathbf{D}(t_0 - wm, m).$$

Treating each vector $\mu(t_0, m)$ as one observation, buildings with similar usage patterns can be identified by applying standard clustering techniques (*K-means Kaufman and Rousseeuw* (1990)) to $\mu(t_0, m)$ for varying values of m.

2.1.3 Modeling power consumption via linear factors

Using *real-world* AMI building data (see Section 2.1.5), we observed that much of the variability in the $Y(t)$'s can be explained by only few principal components of the matrix $\mathbf{Q} = \sum_{n=1}^{N} Y(n)Y(n)^\top$. We therefore consider the eigen value decomposition of $\mathbf{Q}$

$$\mathbf{Q} = V \Lambda V^\top$$

where $V = [v_1, \cdots, v_B]$ is a matrix with B orthonormal columns and Λ is a diagonal matrix with entries $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_B \geq 0$ and propose the following factor model,

$$Y(t) = \mu_Y(t) + Z(t) := \mathbf{F}\beta(t) + Z(t), \tag{2.2}$$

where $\mathbf{F}$ is a matrix $B \times k$ of factors, $\beta(t) \in \mathbb{R}^k$ is a parameter estimable from data and $Z(t)$ is the measurement noise which follows multivariate normal distribution with zero mean and variance-covariance matrix $\mathbf{\Sigma}$. The matrix $\mathbf{F}$ considered is constructed from the first k columns of V which correspond to the eigen vectors for k largest eigen values. The factor model (2.2) serves as an adequate model for capturing the temporal variability of the $Y(t)$'s (see *Vaughan et al.* (2013), Prop. 1). Since the factors $\mathbf{F}$ and

$\mathbf{\Sigma}$ remain constant only over shorter time scales, these are dynamically updated over moving time windows (see Algorithm 1).

2.1.4 Kriging-based prediction and detection

Using the factor model in (2.2), we next propose an anomaly detection methodology. Let $Y(t) = (Y_{i_1}(t), \cdots, Y_{i_b}(t))$, $\{i_1, \cdots, i_b\} \subset \mathcal{B}$ denote the electricity consumption of buildings within the same cluster. We partition meters into observed (trusted) nodes, $\mathcal{O} \subset \{i_1, \cdots, i_b\}$, and unobserved $\mathcal{U} = \{i_1, \cdots, i_b\} \backslash O$. Let $Y_o = (Y_j)_{j \in \mathcal{O}}$ and $Y_u = (Y_j)_{j \in \mathcal{U}}$ denote the partitioned vector Y (for notation simplicity we drop t). Thus, from (2.2),

$$\begin{pmatrix} Y_u \\ Y_o \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{F}_u\beta \\ \mathbf{F}_o\beta \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_{uu} & \mathbf{\Sigma}_{uo} \\ \mathbf{\Sigma}_{ou} & \mathbf{\Sigma}_{oo} \end{pmatrix} \right). \tag{2.3}$$

Given the limited set of observed nodes $\mathcal{O}$, and if the true parameter β is known, the minimum variance unbiased predictor of Y_u is the kriging estimate *Cressie* (1993a); *Vaughan et al.* (2013):

$$\hat{Y}_u(Y_o, \beta) := \mathbf{F}_u\beta + \mathbf{\Sigma}_{uo}\mathbf{\Sigma}_{oo}^{-1}(Y_o - \mathbf{F}_o\beta). \tag{2.4}$$

In practice, the parameter β is unknown, but can be estimated from data on the observed nodes using either a generalized least square regression :

$$\hat{\beta} = (\mathbf{F}_o^\top \mathbf{\Sigma}_{oo}^{-1}\mathbf{F}_o)^{-1}\mathbf{F}_o^\top \mathbf{\Sigma}_{oo}^{-1} Y_o = \mathbf{P} Y_o. \tag{2.5}$$

or ordinary least squares as:

$$\hat{\beta} = (\mathbf{F}_o^\top \mathbf{F}_o)^{-1} \mathbf{F}_o^\top Y_o$$

For predictions for the unobserved nodes $\mathcal{U}$, $\hat{\beta}$ is used as the plug in estimator for β in (2.4). This also simplifies the expression for $\hat{Y}_u$ to $\hat{Y}_u = \mathbf{F}_u \mathbf{P} Y_o + \mathbf{\Sigma}_{uo} \mathbf{\Sigma}_{oo}^{-1} (\mathbf{I} - \mathbf{F}_o \mathbf{P}) Y_o$.

For detecting anomalies in the set of unobserved meters, we need the distribution of the prediction errors (residuals) between the *actual* meter readings and their *predictions*, i.e., $Y_e = Y_u - \hat{Y}_u$.

Proposition II.1. *Under the Null hypothesis of no anomalies and the model of* (2.2)*, the prediction residuals* Y_e *follow a multivariate normal distribution* $Y_e \sim N(0, \mathbf{\Sigma}_{err})$, *with*

$$\mathbf{\Sigma}_{err} = \mathbf{\Sigma}_{uu} - \mathbf{C}\mathbf{\Sigma}_{ou} - \mathbf{\Sigma}_{uo}\mathbf{C}^\top + \mathbf{C}\mathbf{\Sigma}_{oo}\mathbf{C}^\top \tag{2.6}$$

and $\mathbf{C} = \mathbf{F}_u \mathbf{P} + \mathbf{\Sigma}_{uo} \mathbf{\Sigma}_{oo}^{-1} (\mathbf{I} - \mathbf{F}_o \mathbf{P})$.

Proof. Observe that $Y_e = Y_u - \mathbf{C} Y_o$ is a linear transformation of Y, and therefore Y_e has a multivariate normal distribution. The expected error, $\mu_{\text{err}} = \mathbb{E}[Y_u - \mathbf{C} Y_o]$, becomes $\mu_{\text{err}} = \mathbf{F}_u \mathbb{E}[\hat{\beta}] - \mathbf{C}\mathbf{F}_o \mathbb{E}[\hat{\beta}]$ from (2.2). $\mathbf{P}\mathbf{F}_o = \mathbf{I}$, which implies $\mathbf{C}\mathbf{F}_o = \mathbf{F}_u \mathbf{P} \mathbf{F}_o + \mathbf{\Sigma}_{uo} \mathbf{\Sigma}_{oo}^{-1} (\mathbf{I} - \mathbf{F}_o \mathbf{P}) \mathbf{F}_o = \mathbf{F}_u$, and, thus, $\mu_{err} = 0$. For the error variance, $\text{Var}(Y_u - \mathbf{C} Y_o) = \mathbb{E}\big[(Y_u - \mathbf{C} Y_o)(Y_u - \mathbf{C} Y_o)^\top\big]$, and the result follows using (2.3). □

The prediction error Y_e is used for identification of anomalies on the unobserved meters. In this direction, we define the *test statistic* $r^2 = Y_e^\top \mathbf{\Sigma}_{\text{err}}^{-1} Y_e$, which corresponds to the Mahalanobis distance whose p values can be obtained using Proposition (II.1) as

$$p = 1 - F(r^2)$$

```
Algorithm 1 Kriging for detection of data injection attacks.
Input: Training data $\mathcal{D}(t_0) := \{Y_i(t), i \in \{1,\ldots,b\}, t_0 - N \le t \le t_0\ \}$;
Input: Set of “observed” nodes $\mathcal{O}$;
Input: Set of “unobserved” nodes $\mathcal{U} = \{1,\ldots,b\}\backslash\mathcal{O}$;
Output: Sequence of $p$-values for prediction errors.
Obtain $b \times k$ factor matrix $\mathbf{F}$ using PCA on data $\mathcal{D}(t_0)$
Estimate covariance matrix $\mathbf{\Sigma}$ using data $\mathcal{D}(t_0)$
for each new observation $Y = Y(t)$, $t = t_0 + 1, \ldots$ do
    Partition vector $Y$ into $Y_o$ and $Y_u$
    Estimation of $\hat{\beta} = (\mathbf{F}_o^\top \mathbf{\Sigma}_{oo}^{-1}\mathbf{F}_o)^{-1}\mathbf{F}_o^\top\mathbf{\Sigma}_{oo}^{-1}Y_o = \mathbf{P}Y_o$.
    Prediction: $\hat{Y}_u = \mathbf{F}_u\hat{\beta} + \mathbf{\Sigma_{uo}}\mathbf{\Sigma_{oo}}^{-1}(Y_o - \mathbf{F}_o\hat{\beta})$
    Calculate the error covariance matrix $\mathbf{\Sigma}_{\text{err}}$ (see Eq. (2.6))
    With prediction error $Y_e := Y_u - \hat{Y}_u$, get test statistic

        $r^2 = Y_e^\top \mathbf{\Sigma}_{\text{err}}^{-1} Y_e$ (Mahalanobis distance)

    output $p$=1-$F(r^2)$, $F(x)$ is a chi-squared cdf (d.f.= $|\mathcal{U}|$).
end for
```

where $F(x)$ is the chi-squared cumulative distribution function with degrees of freedom equal to rank($\mathbf{\Sigma}_{\text{err}}$). Algorithm 1 explains this entire methodology. To tame the false alarm rate, we apply an *exponential weighted moving average* (EWMA) control chart to the standardized z-scores $z = \Phi^{-1}(1-p)$ (see also *Lambert and Liu* (2006a); *Kallitsis et al.* (2015)), where $\Phi(x)$ is the cumulative distribution for standard normal.

2.1.5 Performance evaluation

This section uses the electricity consumption data from University of Michigan campus for evaluation purposes. The data set comprises of 163 buildings with observations recorded at the time scale of every 2 minutes. The data is available for almost one year with a wide range of buildings such as health services, parking lots, student housing, laboratories etc.

Fig. 2.1 gives a plot for p values are generated by Algorithm 1 for observations from (2.2). As expected, behavior under true model is uniform. Fig 2.1 plots the p values generated from Algorithm 1 but with observations from the data set described above. The uniformity in the p-value consolidates the model assumptions on the real data.

We next describe the simulation setting used for the evaluation of the proposed methodology. In these experiments, the factors and variance covariance matrix are obtained from a two week training period using Algorithm 1. For each building, the next 48 hours (720×48 observations) are predicted using observations from the remaining buildings. Finally EWMA control charts as described in Section 2.1.4 are used for determining out of control values. A simulated attack is injected at a randomly chosen epoch (lasting 1 hour or 30 observations) in the 48 hours span for the building under study. The detection accuracy of Algorithm 1 is determined in terms of the *precision* and *recall* (see *Kallitsis et al.* (2015)). We also evaluate the prediction accuracy for the building under study by the root mean square (rMSE) defined as

$$\text{rMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(Y_i(t) - \hat{Y}_i(t))^2}$$

. For building 1, Table 2.1 reports these values for different pairs (w, L) when averaged over 50 independent realizations. The shift σ denotes the magnitude of the attack injected. As the value of σ grows, the precision and recall improve thereby indicating the successful detection of the false attacks. The false positive rate may be further reduced considering the *two-in-a-row* rule *Lucas and Saccucci* (1990) or additional values for the EWMA pairs (w, L).

To get a unified picture of what happens to all buildings, we first fix a time point to inject attacks. Next a building is chosen and an attack is injected stretching for an hour from that point. The remaining observations are used for predicting the building under study and detecting the occurrence of an attack. This experiment is

w	**L**	**Shift** ($\times\sigma$)	**Precision**	**Recall**	**rMSE (KW)**
1.00	3.719	1	0.07	0.08	12.6
1.00	3.719	2	0.43	0.72	13.5
1.00	3.719	3	0.53	0.99	15.0
0.53	3.714	1	0.05	0.21	12.6
0.53	3.714	2	0.18	0.94	13.6
0.53	3.714	3	0.19	1.00	15.1
0.84	3.719	1	0.07	0.11	12.6
0.84	3.719	2	0.37	0.84	13.6
0.84	3.719	3	0.42	1.00	15.0

Table 2.1: Evaluation of detection performance (meter 1).

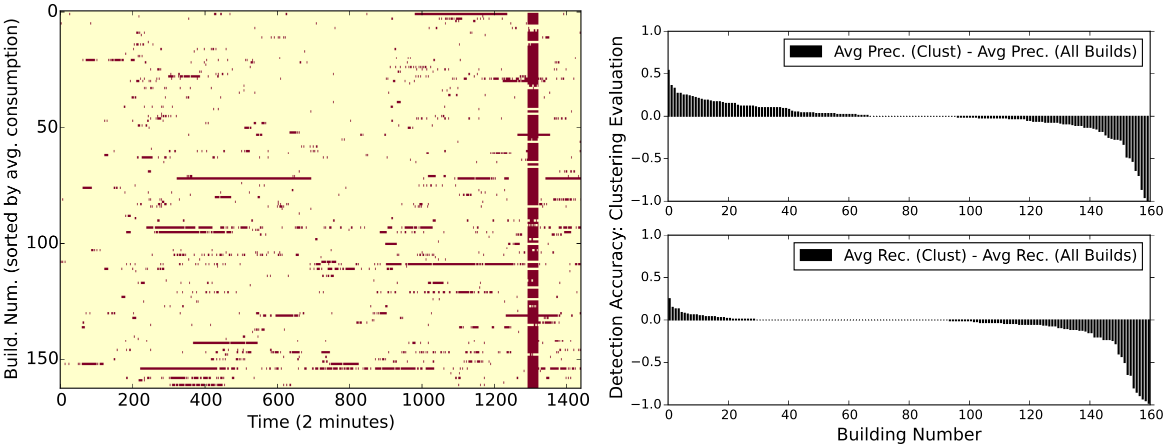


Figure 2.2: Left: Detection alerts (red) over a two-day period. The vertical red stripe denotes an hour-long period of injected anomalies. We study the behavior of each building; the building under study is consider as unobserved (unsecure) and we use observations from the remaining ones. An EWMA control chart was used with $w = 1, L = 3.719$. Right: The effect of clustering in detection performance. (Due to sorting, the building orderings in the top and bottom panels differ.)

repeated for all building and the results are reported in Fig. 2.2 left. In vast majority of cases our methods detect the injected attacks, and that the false positive rate is relatively low in all tests. With increase in the number of observed nodes, the prediction accuracy improves but not substantially (results not reported here).

Lastly, we analyze the effect of clustering when applied with our methods. Fig. 2.2

Cluster number	**2**	3	4	5	6	7	8	9
Silhouette score	**.81**	.59	.60	.56	.52	.47	.43	.45

Table 2.2: Silhouette values for cluster number selection.

right examines the detection accuracy with and without clustering. The number of K-means clusters was selected by looking at the silhouette *Rousseeuw* (1987) scores (see Table 2.2); two main classes were identified. For a large number of buildings, clustering provided better detection accuracy in terms of the precision and recall values.

2.1.6 Conclusions

We have proposed an adaptation of the kriging model wherein the observations from a few trusted nodes are used to predict the energy patterns of the remaining ones. This naturally provides a method for the detection of bad data injection into a smart grid by recording predicted values which go outside the control chart limits. We verified these results in context of data from a wide area network, University of Michigan campus. To handle the issue of dimension when the number of buildings is large, clustering methodology has also been proposed.

Anomaly detection in smart grids can be broadly classified into three main categories signature-, specification- and anomaly-based methods *Berthier et al.* (2010); *Cleveland* (2008); *Wang and Lu* (2013). Each of these methods has its own set of challenges. For example whereas new attacks with signatures agnostic to signature-based intrusion detection system (e.g., *Snort*) evade detection, specification-based systems *Carcano et al.* (2010) are difficult to tune. Our method primarily falls under the third category of anomaly detection. Other works in this direction include *Kallitsis et al.* (2015); *et al.* (2013); *Bi and Zhang* (2014); *Yu* (2015). The idea of using a few trusted node stems from *Bi and Zhang* (2014) which proposes a graph theoretic method for securing an optimal set of meter measurements so that state estimation is not compromised.

The main advantage of our method over the works in *Liu et al.* (2009); *et al.* (2010, 2013); *Bi and Zhang* (2014); *Yu* (2015), is that a prior of the system's topology *DC power flow model Liu et al.* (2009); *et al.* (2010) is not needed and the the power consumption patterns suffice as an input variables. As a part of the future work we wish to incorporate the temporal variations in the data via autoregressive (both univariate and multivariate) processes. In addition, we wish to model the prediction patterns where the set of unobserved and observed nodes is unknown. Predicting the mean consumption patterns by known factors like building type, location, time of the day etc. may be an interesting extension to this work.

Acknowledgements: This work is supported by NSF grant CNS-1422078.

2.2 Online Monitoring Of Internet Traffic: Challenges And Solutions

Contributions and due credit: Much of the material in Sections 2.2.2, 2.2.3, 2.2.4, 2.2.5, 2.2.6, 2.2.7 is based on the collaborative work *Kallitsis et al.* (2016a), done under the leadership of Dr. Michalis Kallitsis. As a author of the thesis, we have contributed primarily to the statistical aspects in 2.2.3, 2.2.4, 2.2.5, which would not have been possible without the real-world Internet data provided by the AMON infrastructure *Kallitsis et al.* (2016a,b) developed by Dr. Kallitsis' team at Merit Network[1].

[1] Merit Network, Inc. operates Michigan's research and education network. It serves a population of more than 1 million users, and its ingress/egress traffic exceeds 40Gbps during peak demand.

2.2.1 Introduction

Over the years, Internet infrastructure has become more accessible but at the same time more vulnerable to security threats. Cyber attacks like Distributed Denial of Service (DDoS), port scanning, outages which are becoming increasingly common; can overwhelm the network rendering it completely dysfunctional. *Czyz et al.* (2014) presents a recent example of a large DDoS attack which was the outcome of mis configured NTP (network time protocol) servers. Other examples include *reflection and amplification* attacks from *Kührer et al.* (2014); *Rossow* (2014) where multiple small requests are sent to several mis-configured NTP servers (or other UDP-based services).

The predominance of Internet in all spheres (banks, universities, industries) calls for the development of sophisticated tools which can help protect the system against malicious users. Some methods such as *Snort* (see, `snort.org`), *Bro* (`bro.org`) and *Suricata*(`suricata-ids.org`) which do exist are easily beguiled by malware existing in varying forms by encryption (*polymorphic*).Thus the problem of anomaly detection in Internet traffic requires a basic understanding of the network features in terms of traffic flows, its composition, capacity, quality of service etc. In the past a lot of statistical work has been done in the direction of analyzing data from Internet, streaming algorithms *Gilbert et al.* (2001); *Muthukrishnan* (2005), tomography *Xi et al.* (2006); *Lawrence et al.* (2006) and analysis of heavy tails and long range dependence *Stoev et al.* (2005, 2006); *Stoev and Michailidis* (2010).

A communication network like Internet involves information exchange across a vast number of IP addresses. Storing information on traffic flows for all nodes is often a challenging task if not impossible. Sketch based algorithms developed in *Krishnamurthy et al.* (2003); *Gilbert et al.* (2007); *Stoev et al.* (2007) provide a solution to this by constructing small summary structures which capture all essential features of the network. However, the amount of time taken in producing these sketches is quite large when compared to the rate of flow of incoming traffic. Another alternative to handling the storage constraint is using information available only through the packet headers like source/destination addresses, application ports, payload size etc. Tools like Netflow, sFlow etc. have been developed to facilitate the compression of packet data by grouping them into flows. However, their compression mechanism does not scale to the order at which packets arrive.

Even if it is possible to bypass the storage issue, there are relatively few algorithms which can detect anomalous events from sketches or Netflows. Such algorithms, if available are mostly sequential in nature and not deployable at shorter time scales. Thus even with the best practices like (e.g., BCP38 recommendation *Senie* (1998)), the identification and mitigation of events like Distributed Denial of Service (DDoS) still seems a distant goal. In this chapter, we thereby develop the statistical framework for AMON; a software which can read packet data from Netflows, compute real time statistical summaries and flag alarms on the onset of any unusual activity in the network (see Figure 2.3). In AMON, data is collected efficiently using a PF_RING ZC module employed at Merit Network (see Section II C in *Kallitsis et al.* (2016b)). The packet data so obtained is then summarized using optimal hashing techniques implementable at the time scale of routing (see Section II A in *Kallitsis et al.* (2016b)).

Our main contribution to this work has been the development of statistical algorithms which when applied to the hashed outputs can detect the onset of both volumetric and low lying DDoS. Some of the statistical approaches for detection of heavy hitters or frequent items in a stream have been developed in *Karp et al.* (2003); *Cormode and Muthukrishnan* (2005b); *Cormode et al.* (2003); *Estan and Varghese*

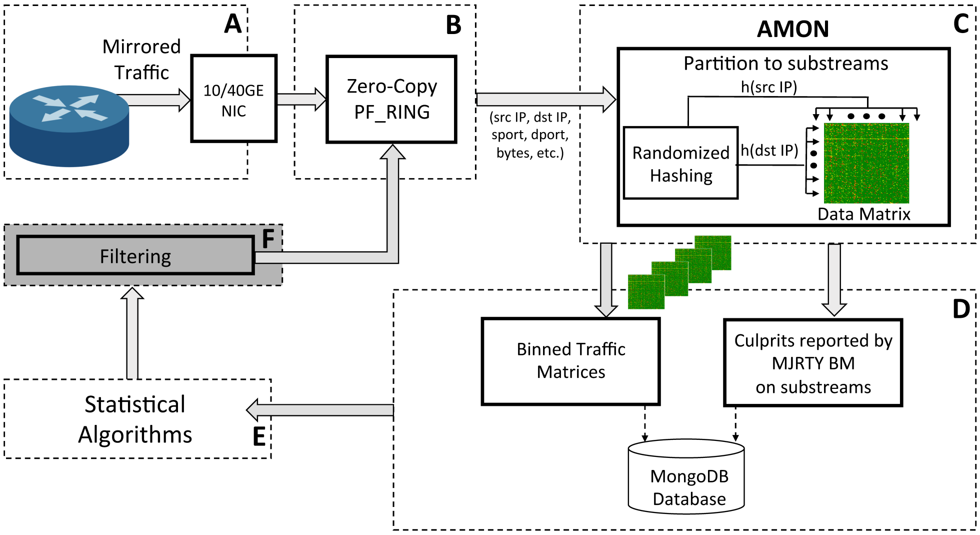


Figure 2.3: High-level architecture of AMON.

(2002); *Gilbert et al.* (2006); *Krishnamurthy et al.* (2003); *Schweller et al.* (2006); *Cormode and Muthukrishnan* (2004, 2005a); *Porat and Strauss* (2012); *Gilbert et al.* (2012). However most of these algorithms provide efficient solutions to the problem data storage rather than modeling the distribution of statistical flows. In this chapter, we show that the summary structures obtained from hashing are indeed heavy tailed and share most of the statistical properties of raw traffic. Thereby techniques from extreme value theory may be used for detection of heavy hitters, i.e. IPs contributing to unusually large traffic in the network. Sometimes smaller magnitude attacks may be launched by a large number of IPs working together in which case a rather different approach needs to be adopted. One such approach relies on recording of sudden appearances communities and cliques, etc.(see *Ranshous et al.* (2015)) in the graph of network flows. With a slight modification, this approach has been extensively used in the paper for identification of low lying attacks.

Once the onset of an anomalous event is recorded, the true IPs associated with it are traced back by using a extended version of the Boyer Moore majority algorithm (see Section II B in *Kallitsis et al.* (2016b)). The rest of the Section is organized as follows: Section 2.2.2 describes how the synopsis data structures are constructed from by application of hashing. It also serves as the link between groups C and D in the architecture of AMON (see Figure 2.3) where the raw data in form of packets is summarized to a form which can be used by the statistical algorithms for anomaly detection. Sections 2.2.3 and 2.2.4 describes two detection algorithms based on the heavy tailed nature of the Internet data; a property which is preserved even after the application of hashing. These algorithm intend at identifying high volumetric attacks. These methods successfully identified the case where a DDoS attack was targeted at a public library of University of Michigan traffic (see Section 2.2.7.1). Section 2.2.5 presents an algorithm for detection of high frequency low magnitude attacks. This method which involves the source destination interaction matrix successfully identified low volume attacks targeted to Michigan servers from an autonomous system in Asia-Pacific (see Section 2.2.7.2). Finally a comparative performance of the methods. The detection accuracy of these proposed methods has been studied under simulation setting of Section 2.2.8.

2.2.2 Hashed binned data matrix and its visualization.

This section describes the construction of source destination flow matrix from the packet header information of payload transfer across the network. Every packet captured at the monitoring station can be viewed as a tuple (w_n, v_n) where w_n is the key and v_n is the payload. In the context of Internet traffic, the key w_n can either denote an individual source/destination address or a source destination pair. The payload v_n can represent file sizes in bytes, packets or network ports. Using the terminology in *Muthukrishnan* (2005), for $w_n \in \Omega$ and $v_n \in V$ we the following function f

$$f : \Omega \to V$$

for every incoming data point where Ω usually denotes the space of IPv4 addresses ($\Omega \in \{0,1\}^{32}$) or their cross product ($\Omega \in \{0,1\}^{64}$ for source destination pair). Storage of the signal f even for a small interval of time is practically impossible owing to the high speed of incoming traffic and large dimension of Ω. Therefore hashing techniques which can the distribute the space of IP addresses to a smaller number of bins have been employed. Precisely, a hash function h is represented as

$$h : [N] \to [m] \tag{2.7}$$

where $[N]$ is the cardinality of Ω and $[m]$ is the cardinality of the space over which the keys are distributed. Extensive details on the choice of the hash function are covered under *Kallitsis et al.* (2014) and *Cormode and Muthukrishnan* (2005b).

When a packet arrives for the key (s, d), one computes $i = h(d)$ and $j = h(s)$ and updates the sketch matrix as

$$X[i, j] = X[i, j] + v \tag{2.8}$$

This sketch matrix is stored at the time scale of every 10s. Direct products of this matrix are the source and destination hashed arrays defined as

$$\begin{aligned} \text{Source} \quad &: \quad X_t(i) = \sum_{j=1}^{m} X_t[j, i] && (2.9) \\ \text{Destination} \quad &: \quad X_t(i) = \sum_{j=1}^{m} X_t[i, j] && (2.10) \end{aligned}$$

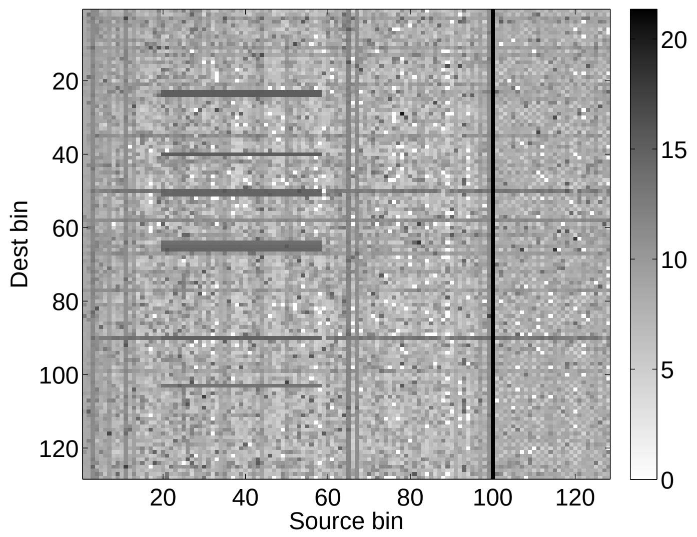


Figure 2.4: Sketch data blocks are used as the basic input structures for our detection algorithms. The databrick matrix; notice the horizontal stripes that signify traffic from multiple destinations to multiple sources. Note also the bold column at source bin 100 that depicts heavy source(s) activity.

The hash binned arrays $(X_t(i))_{i=1,\cdots,m}$ will serve as the input to the statistical detection algorithm of Sections 2.2.3, 2.2.4 and 2.2.5. We next explore their distribution properties for outgoing (Source) traffic at Merit Network for the period 17:30-18:30 EST on July 22, 2015. Figure 2.6 top panel shows a plot of the 360 hashed destination arrays collected for a span of 60 minutes and then stacked one after another. The data reveals a few extreme peaks some of which may be attributed an attack event (see the 'Library' case study, Section 2.2.7.1). A zoomed in version on a short 3-minute period (bottom right panel) show that the extreme peaks, although of lower magnitude, persist. Apparent periodicities in extreme peaks may be attributed to data concatenation.

Data arising from Internet traffic such as file sizes, web page counts etc. are well modeled by heavy tailed or power law distributions *Leland et al.* (1994); *Crovella and Bestavros* (1997); *Faloutsos et al.* (1999). This heavy tailed property is preserved even after the application of hashing (provided the hash function in (2.7) distributes

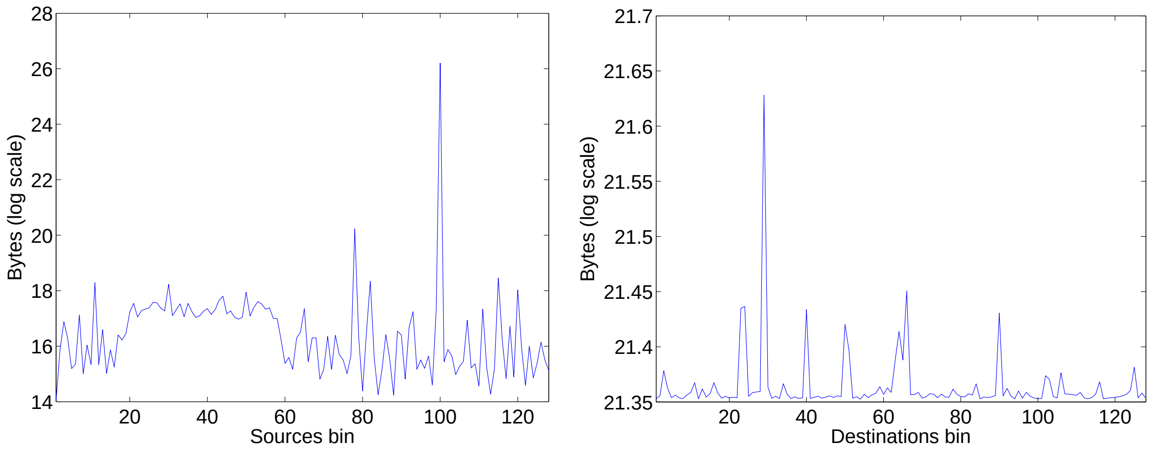


Figure 2.5: *Left:* View of source array, constructed by aggregating over the columns of the matrix. Note the heavy source(s) at bin 100.*Right:* Destinations array; observe how heavy destinations appears as heavy bins. In all cases volume is in bytes in log scale.

the IP space uniformly). We thereby assume that $X = X_t(i)$ satisfies

$$P(X > x) \sim c/x^{\alpha}, \quad \text{as } x \to \infty, \tag{2.11}$$

where $\sim$ indicates asymptotic convergence and c, α are the parameters of power law distribution. Smaller values of α correspond to heavier tails. Indeed for $\alpha < 2$, the mean does not exist and for $\alpha < 1$ both mean and variance are both undefined.

To validate the assumption in (2.11), we consider the max spectrum plot of an hour long time series of hash binned source traffic collected at Merit. The max spectrum is a plot of the mean log block maxima versus the log block sizes of the data. A linear trend indicates the presence of power-law tails with slope giving an estimate for $1/\alpha$. The linearity in the bottom panel of Figure 2.6 shows that power law is a fairly reasonable assumption for the time series of hash binned traffic. Steep slopes in the max spectrum plot correspond to heavier tails (low α). An advantage of the max spectrum plot is its ability to examine various log block sizes thereby allowing

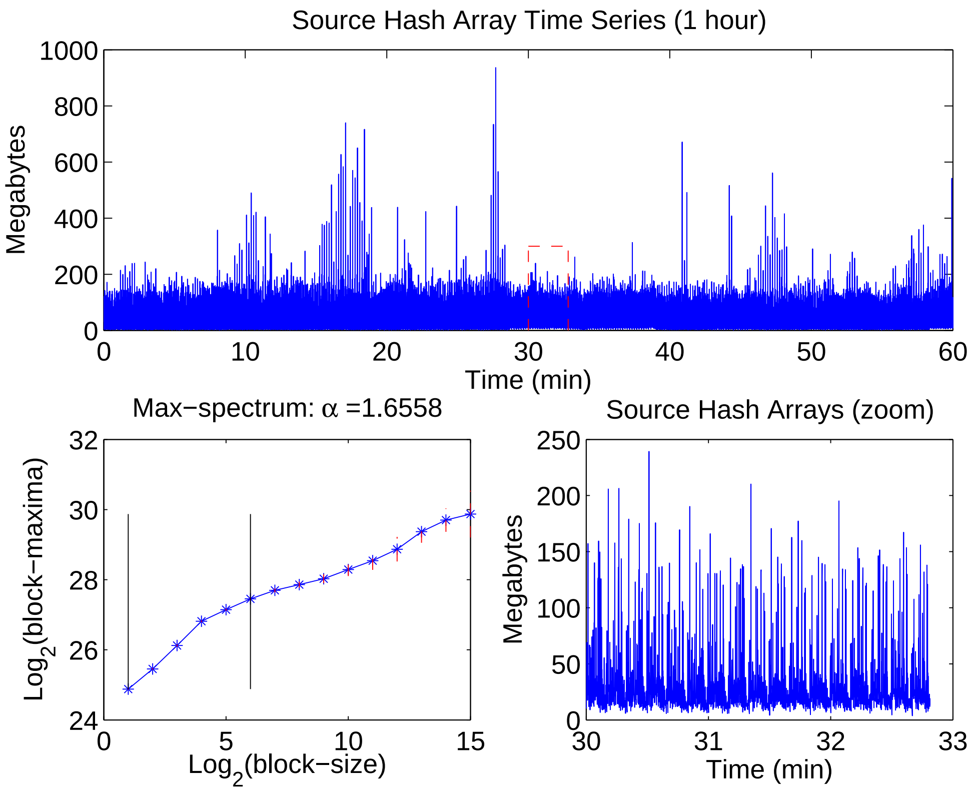


Figure 2.6: Time-series of Source hash-binned arrays (Top) and its zoomed-in version. (Bottom right), computed over 10-second windows. The *max-spectrum* of the entire time series is plotted on the bottom-left. Merit Network: 17:30-18:30 EST, July 22, 2015.

for examination power law behavior in date varying time scales (see *Stoev et al.* (2011)). Extensive experimentation showed that the power-law behavior (linearity in the spectrum) extends over a wide range of time-scales from seconds to minutes with $\alpha \approx 1.6$ and $\alpha \approx 2.5$ for shorter and intermediate time scales respectively (results not reported here). For larger time-scales (hours) complex intermittent non-stationarity and diurnal trends dominate and the heavy tailed characteristic of the data starts to breakdown.

Remark II.2. The hash-array is obtained from the PF_RING-based methodology at the time scale of 10 seconds so that we do not run into the issue of empty bins and algorithms of Sections 2.2.3, 2.2.4 and 2.2.5 can be suitably applied. For higher traffic rates, the methodology can be applied at an even shorter, sub-second time-scale.

2.2.3 Detection of heavy hitters

There has been a lot of work on the estimation of heavy hitters in fast network traffic streams *Karp et al.* (2003); *Cormode and Muthukrishnan* (2005b); *Cormode et al.* (2003); *Estan and Varghese* (2002); *Gilbert et al.* (2006); *Krishnamurthy et al.* (2003); *Schweller et al.* (2006); *Cormode and Muthukrishnan* (2004, 2005a); *Porat and Strauss* (2012); *Gilbert et al.* (2012). The definition of a heavy hitters is not explicit and is often open to interpretations. In this section, we introduce a rather new terminology in connection to the hash array inputs (2.9) where hash bins with abnormally large traffic are identified as heavy hitters. By abnormally large we mean observations which lie well above the quantile threshold for the baseline probability distribution of hashed inputs. Since the nature of traffic changes quite frequently, the baseline model needs to be dynamically updated while accounting for robustness and adaptivity issues (see last paragraph under Section 2.2.3). Lastly the type I error for the proposed test based algorithm is controlled by the choice of the quantile threshold.

In Section 2.2.2, we showed the heavy tailed nature of the hash binned arrays defined in (2.9) (see Figure 2.7). For the source arrays, $X_t(i)$ corresponds to the number of bytes originating from all source IPs ω hashed to bin i, i.e. $h(\omega) = i$ over the time-window t. Optimal hashing techniques *Kallitsis et al.* (2014) and *Cormode and Muthukrishnan* (2005b) randomly distributes the IP addresses which more or less ensures that $X_t(i),\ i = 1, \ldots, m$ are independent and identically distributed (i.i.d.).

Abnormally large values of $X_t(i)$s, for *some* i's correspond anomalous events such as DDoS/outages etc. To this identify them, we consider the sample maximum of the hash-array:

$$D_m(X_t) := \max_{i=1,\dots,m} X_t(i). \tag{2.12}$$

Proposition II.3. *Let $X(i),\ i = 1, \dots, m$ be i.i.d. random variables with heavy tails as in* (2.11)*. Then, as $m \to \infty$, we have that*

$$\frac{1}{m^{1/\alpha}} D_m(X) \equiv \frac{1}{m^{1/\alpha}} \max_{i=1,\dots,m} X(i) \xrightarrow{d} c^{1/\alpha} Z_\alpha, \tag{2.13}$$

where $P(Z_\alpha \le x) = e^{-1/x^\alpha}$ has the standard α-Fréchet distribution and c is the asymptotic parameter in (2.11)*.*

The proof is included in Section A.

A bin $i \in \{1, \dots, m\}$ is a heavy hitter, if its value is large, relative to the *asymptotic approximation* of Proposition II.3. Section 2.2.6 shows that the asymptotic approximation kicks in even at values of m close to 128. We next define a heavy hitter formally as:

$$X_t(i) \ge T_{p_0}(m, \alpha, c) := m^{1/\alpha} c^{1/\alpha} \Phi_\alpha^{-1}(p_0) = \Big(\frac{c}{\log(1/p_0))}\Big)^{1/\alpha}, \tag{2.14}$$

where the sensitivity level p_0 controls the type I error rate and $\Phi_\alpha^{-1}(p) = (\log(1/p))^{-1/\alpha}$, $p \in (0, 1)$ is the inverse of the standard α-Fréchet cumulative distribution function $\Phi_\alpha(x) = e^{-1/x^\alpha},\ x > 0$. By increasing p_0 one can indeed reduce the number of false alarms (see Section 2.2.6). The methodology is summarized in the formal algorithm (Algorithm 2).

Several methods exist for the estimation of the parameters α and c (*Embrechts et al.* (1997),*Hill* (1975)). We however use the max-spectrum method in *Stoev et al.* (2011) since it is computationally efficient and easier to tune. In order to reduced the

Algorithm 2 Detection of heavy-hitter bins in traffic volume hash-arrays.

```
Input: Stream of hash-arrays $X_t = \{X_t(i)\}_{i=1}^m$; probability level $p_0 \in (0,1)$; smoothing coefficient $\lambda \in (0,1)$.
Output: Stream of significant heavy-hitter bins $\mathcal{H}_t \subset \{1,\ldots,m\}$ and their counts $k_t = |\mathcal{H}_t|$.
for each stream item $X_t$ do
    Estimate the tail exponent $\hat{\alpha} := \alpha(X_t)$ and scale coefficient $\hat{c} := c(X_t)$ from the sample $X_t = \{X_t(i)\}_{i=1}^m$ based on the max-spectrum.
    if ($t = 1$) then
        Set $\alpha_t := \hat{\alpha}$ and $c_t := \hat{c}$
    else
        Perform EWMA smoothing: $\alpha_t := \lambda\hat{\alpha} + (1-\lambda)\alpha_{t-1}$ and $c_t := \lambda\hat{c} + (1-\lambda)c_{t-1}$.
    end if
    Compute the significance threshold $T_t := T_{p_0}(m, \alpha_t, c_t)$ using (2.14).
    Estimate the set of heavy hitter bins $\mathcal{H}_t$ at window $t$ as $\mathcal{H}_t := \Big\{i \in \{1,\ldots,m\} : X_t(i) \geq T_t\Big\}$.
    return $\mathcal{H}_t$ and $k_t := |\mathcal{H}_t|$.
end for
```

susceptibility of estimates to outliers, we apply an EWMA smoothing to the values of $\hat{\alpha}$ and $\hat{c}$ (see Step 6 in Algorithm 2). The choice of the smoothing parameter is described in details under Section 2.2.6.

2.2.4 Relative volume

In the previous section we discussed heavy hitters from an absolute threshold standpoint (2.14). In this section, we analyze the scenarios where a small proportion IPs generate abnormally large traffic relative to the remaining ones. In this direction, consider the hash binned arrays in (2.9) and sort them in decreasing order as

$$X_t(i_1) \geq \cdots \geq X_t(i_k) \geq \cdots \geq X_t(i_m) \geq 0.$$

Then for a fixed integer $k \in \{1, \ldots, m\}$, we consider the relative volume of traffic contributed by the top-k bins:

$$V_t(k) := \frac{\sum_{j=1}^{k} X_t(i_j)}{\sum_{j=1}^{m} X_t(i_j)}. \tag{2.15}$$

These top-k bins may easily change from one time-window to another. High volumetric attacks launched by a small subset of k IPs relative to the rest produce large values of $V_t(k)$. In order to obtain significantly large values of $V_t(k)$, we determine its distribution using the heavy tailed property of the $X_t(i)'s$ (see Section 2.2.2). This baseline distribution is updated periodically similar to the previous section.

The following fundamental representation results for the joint distribution of the *order statistics* (see, e.g., p. 189 in *Embrechts et al.* (1997)).

Theorem II.4 (Rényi representation). *Let $U(1), \ldots, U(m)$ be independent and identically distributed Uniform$(0,1)$ random variables. Consider the sorted sample (order statistics) $U(i_1;m) \le \cdots \le U(i_k;m) \le \cdots \le U(i_m;m)$. Then, we have the following stochastic representation:*

$$\Big(U(i_1;m), \cdots, U(i_k;m), \cdots, U(i_m;m)\Big) \overset{d}{=} \Big(\frac{\Gamma_1}{\Gamma_{m+1}}, \cdots, \frac{\Gamma_k}{\Gamma_{m+1}}, \cdots, \frac{\Gamma_m}{\Gamma_{m+1}}\Big),$$

where $\overset{d}{=}$ means equality in distribution and $\Gamma_i = E_1 + \cdots + E_i,\ i = 1, \ldots, m+1$ are Gamma$(i,1)$-distributed random variables, represented as cumulative sums of a fixed set of independent standard Exponential random variables.

In the previous section we argued that $X_t(i),\ i = 1, \ldots, m$ be i.i.d. with distribution as in (2.11). For a continuous distribution function (cdf) F is $U(i) := \overline{F}(X(i)),\ i = 1, \ldots, m$ are i.i.d. Uniform$(0,1)$ where $\overline{F}(x) = 1 - F(x)$ denotes the complementary cdf. Therefore, by Theorem II.4,

$$\Big(\overline{F}(X(i_1)),\cdots,\overline{F}(X(i_k)),\cdots,\overline{F}(X(i_m))\Big) \overset{d}{=} \Big(\frac{\Gamma_1}{\Gamma_{m+1}},\cdots,\frac{\Gamma_k}{\Gamma_{m+1}},\cdots,\frac{\Gamma_m}{\Gamma_{m+1}}\Big).$$

By applying the inverse function $\overline{F}^{-1}$ to all components of the above relation, we obtain

$$\Big(X(i_1),\cdots,X(i_k),\cdots,X(i_m)\Big) \overset{d}{=} \Big(\overline{F}^{-1}\Big(\frac{\Gamma_1}{\Gamma_{m+1}}\Big),\cdots,\overline{F}^{-1}\Big(\frac{\Gamma_k}{\Gamma_{m+1}}\Big),\cdots,\overline{F}^{-1}\Big(\frac{\Gamma_m}{\Gamma_{m+1}}\Big)\Big). \tag{2.16}$$

This yields the following result about the distribution of the relative volume.

Proposition II.5. **(i)** *Under the above assumptions, we have*

$$\{V(k;m),\ k=1,\ldots,m\} \overset{d}{=} \Big\{\frac{\sum_{j=1}^{k}\overline{F}^{-1}(\Gamma_j/\Gamma_{m+1})}{\sum_{j=1}^{m}\overline{F}^{-1}(\Gamma_j/\Gamma_{m+1})},\ k=1,\ldots,m\Big\}. \tag{2.17}$$

(ii) *Under* (2.11)*, for fixed* $1 \le k < \ell$*, we have, as* $m \to \infty$*,*

$$\frac{V(k;m)}{V(\ell;m)} \xrightarrow{d} W_\alpha(k,\ell) := \frac{\sum_{j=1}^{k}\Gamma_j^{-1/\alpha}}{\sum_{j=1}^{\ell}\Gamma_j^{-1/\alpha}}. \tag{2.18}$$

The proof is given in Section A.

Remark II.6. Proposition II.5.(i) is remains valid even for discontinuous and also non-invertible cumulative distribution functions with $\overline{F}^{-1}$ replaced by the left-continuous generalized inverse $\overline{F}^{\leftarrow}(p) := \inf\{x : \overline{F}(x) \le p\}$ (see, e.g. Lemma 4.1.9 on p. 188 of *Embrechts et al.* (1997)).

With the asymptotic distribution of $V(k;\ell)$ available, the occurrence of unusually large values are flagged if

$$V_t(k;\ell) > q_t$$

Algorithm 3 Flagging significant peaks in the relative volume of the top-k hash-bins.

```
Input: Stream of hash-arrays $X_t = \{X_t(i)\}_{i=1}^m$; probability level $p_0 \in (0,1)$; candidate value $k \in \{1, \ldots, m\}$ (preferably $\ll m$); smoothing parameter $\lambda \in (0,1)$.
Output: Binary stream of alarm-flags $f_t \in \{0,1\}$.
 1: for each stream item $X_t$ do
 2:    Estimate the tail exponent $\hat{\alpha} := \alpha(X_t)$ from the sample $X_t = \{X_t(i)\}_{i=1}^m$.
 3:    if $(t = 1)$ then
 4:       Set $\alpha_t := \hat{\alpha}$
 5:    else
 6:       Perform EWMA smoothing: $\alpha_t := \lambda\hat{\alpha} + (1 - \lambda)\alpha_{t-1}$.
 7:    end if
 8:    Compute the relative volume of of the top–$k$ bins $V_t(k)$ as in (2.15).
 9:    Using Monte Carlo simulations, compute numerically the significance threshold
       $q_t = q_t(p_0; k, \alpha_t, m)$, such that
              $P(W_{\alpha_t}(k, m) \le q_t) \approx p_0.$
10:    return $f_t := \mathbb{I}\{V_t(k) > q_t\}$, i.e., flag $V_t(k)$ as significantly large (at level $p_0$) if
       $V_t(k) > q_t$.
11: end for
```

where q_t is the p_0^{th} quantile for the distribution $W_\alpha(k; \ell)$. For application to data from network traffic $\ell = m$ was found to be an adequate choice. This implies the left and right hand side in (2.18) gets replaced by $V(k; m)$ (since $V(m; m) = 1$) and $W_\alpha(k, m)$ respectively.

The methodology is formally described under Algorithm 3.

Remark II.7. When Pareto approximation is not as accurate, lower values of $\ell < m$ need to be used in Algorithm 3. However the choice of $\ell = m$ and $V(k, m) = 1$ worked well enough with the data explored in this chapter.

Similar to Section 2.2.3, the parameter α_t is estimated using the max spectrum of *Stoev et al.* (2011). In order to reduce the susceptibility of the parameters to the presence of outliers, we perform an EWMA smoothing on $\hat{\alpha}$ as in Step 6 of Algorithm 3. High values of $\lambda \approx 1$ imply greater dependence on current estimate whereas small values$\lambda \approx 0$ tend to borrow strength from past observations. Whereas the prior is more adaptive to changing nature of the traffic, the later is more robust to outliers (see Section 2.2.6 for more details).

We next propose a slight modification to the Algorithm 3 so that false alarm rates can be further controlled. This especially useful when anomalies persist over a long period of time. The pvalues are considered based on (2.18) have the form $p_t := P(V_t(k) > W_{\alpha_t}(k,m))$. Following *Lambert and Liu* (2006b), one considers the EWMA on the *z-scores* as

$$z_t := \lambda_p \Phi^{-1}(1 - p_t) + (1 - \lambda_p) z_{t-1},$$

for some $\lambda_p \in (0,1)$. An alarm is raised if $z_t/\sigma_z > L$, for a given level parameter $L > 0$. Section 2.2.6 describes the performance of the Algorithm 3 for choices of the tuning parameter (λ_p, L).

2.2.5 Community detection

In this section we demonstrate a methodology which is more effective in identifying low volume attacks which escape the detection by Algorithm 2.2.3 and 2.2.4 (see Section 2.2.7.2). In order to identify subtle changes in the traffic patterns, we go back to the original hashed matrix $X_t = \{X_t(i,j)\}_{i,j=1}^{m}$ in (2.8) which records the byte payloads, obtained over a certain period of time t. Most of the information on community structure is hidden in the top source-destination flows. We thus construct a binary matrix A_t from $X_t(i,j)$ such that $a_t(i,j) = 1$ if and only if bin (i,j) belongs to the top N entries $X_t[i,j]$.

The binary matrix A_t may be viewed as the adjacency matrix for a graph G_t which essentially represents the connectivity structures between source destination pairs. An event like DDoS would cause the matrix A_t to show multiple 1 entries for the row index equal to the hash value of the targeted IP (see Section 2.2.7.2 and Figure 2.9). Thus we shall use the matrix A_t for detecting structural changes in the graph G_t. A rather simple approach to do this is to consider row/column indices with unusually large in/out degree. Next we only discuss the case where a particular destination is under attack from a multiple sources but each with a comparatively small magnitude.

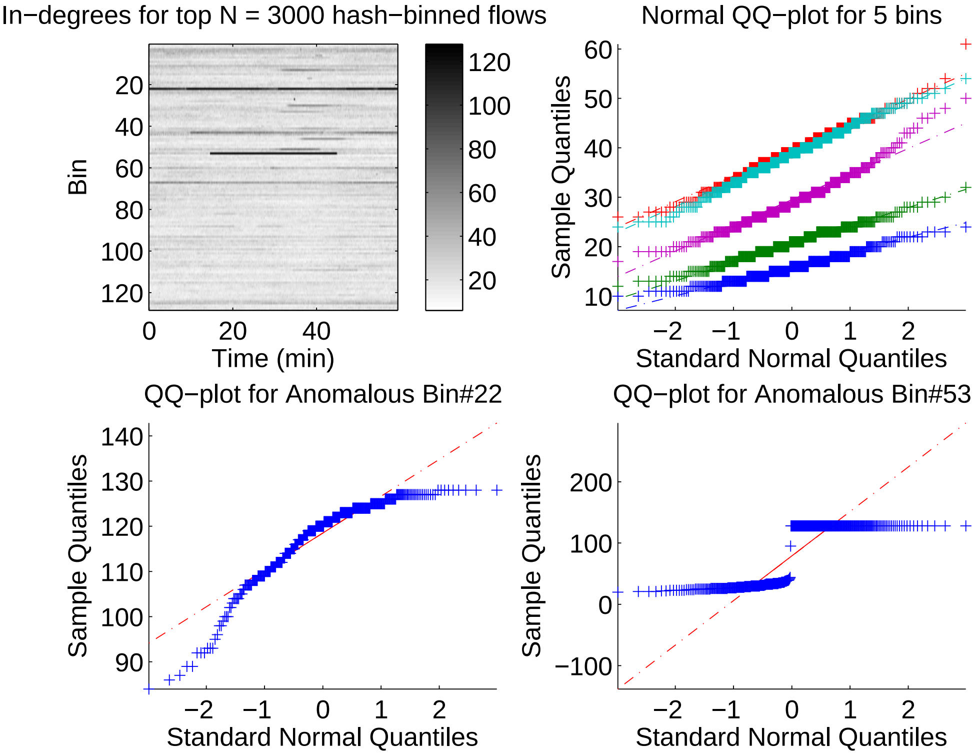


Figure 2.7: Merit Network 16:00-17:00 EST, Aug 1, 2015 – the 'Tor' event in Section refsubsec:tor. (Top left) Ingress connectivity for the top $N = 3000$ hash-binned flows per 10-second windows over 1-hour. (Top-right) QQ-plots demonstrating accuracy of the Normal approximation of typical in-degree distributions. (Bottom plots) QQ-plots for anomalous bins.

Let $I_t(i) := \sum_{j=1}^{m} a_t(i,j), \;\; i = 1, \ldots, m$ be the in-degree of node i for the graph G_t. We wish to find extreme peaks in the values of $I_t(i)$. Since the hash functions from Section 2.2.2 randomly distribute the IPs to m bins, the $I_t(i)'s$ may be assumed to be independent. However in contrast to Sections 2.2.3 and 2.2.4, the counts $I_t(i)$ are no longer heavy-tailed but follow normal distribution. This for fixed i randomization by hashing guarantees the independence of $a_t(i,j)$'s in j. The normal approximation is a direct consequence of CLT when applied to $\sum_{j=1}^{m} a_t(i,j)$ for m sufficiently large. Indeed, Figure 2.6 (top-right plot) shows Normal quantile-quantile plots of $I_t(i),\ t = 1,\dots,T$ for 5 typical (non-anomalous) bins i. The linearity in plot supports the assumption of normality. The heatmap shows the entire array $(I_t(i))_{m\times T}$ of in-degrees computed over 10-second time windows over the duration of 1 hour for $N = 3000$. The bottom plots in this figure show the QQ-plots corresponding to anomalous bins with high in-degree corresponding to the higher intensity lines in the top-left plot. Clearly these bins are quite off from the normal approximation which suggests that extreme peaks in $I_t(i)$ can be used to detect the onset of attacks. Note that N needs to be large to ensure that CLT works but very large values may meddle with the sparse nature of the matrix A_t.

We have thus shown that the in degrees $I_t(i),\ i = 1,\dots,m$ are independent and identically distributed observations from $\mathcal{N}(\mu_t, \sigma_t^2)$. Thus for identifying unusually large in degree values we consider

$$D_t := \max_{i=1,\dots,m} I_t(i),$$

by the independence of the $I_t(i)$'s follows:

$$P(D_t \le x) = \Phi\Big(\frac{x-\mu_t}{\sigma_t}\Big)^m,$$

where Φ is the standard normal CDF. Following the ideas if Algorithm 2, we flag a bin $I_t(i)$ as anomalous if its values exceeds the threshold

$$u_t(p_0) \equiv u(p_0, m, \mu_t, \sigma_t) := \mu_t + \sigma_t \times p_0^{1/m}.$$

where p_0 is the sensitivity parameter controlling for the type I error rate. The empirical means and variances of the data serve as reasonably good estimates for the quantities μ_t and σ_t. However to reduce their susceptibility to outliers, we use the EWMA smoothing on the parameters $\hat{\mu}$ and $\hat{\sigma}$ similar to Step 6 of Algorithm 2.

2.2.6 Detection accuracy

We next describe the detection accuracy of Algorithms 2 and 3 for artificially constructed attacks. Table 2.3 and 2.4 report two metrics, namely *precision* and *recall* (*Kallitsis et al.* (2015)). The data comprises of sketch matrices (databricks) collected at Merit's Detroit monitoring station using *AMON* from an anomaly-free period (one hour during a holiday weekend in July). Attack vectors of varying magnitude (see column three in Tables 2.3 and 2.4) are injected at five randomly chosen points. Three different scenarios are considered: 1) many sources sending traffic to one destination; 2) one source sending traffic to many destinations, and 3) many to many. Each individual experiment is repeated 50 times and we report the average performance in terms of precision and recall. For scenario (3), the input comprises of both destination and source hash binned arrays (see Figure 2.5). The other two scenarios are different where one of destination and source arrays are provided for (1) and (2) respectively. For all situations we allow a grace period of 3 minutes for detection.

Table 2.3 shows the detection accuracy for algorithm 2 for varying values of the pair $(p_0, \lambda) = (p, \lambda_\alpha)$ where the best performance was recorded for $p = 0.95$ and

p	λ_α	Gbps	$P_d^{(1)}$	$R_d^{(1)}$	$P_s^{(2)}$	$R_s^{(2)}$	$P_s^{(3)}$	$R_s^{(3)}$	$P_d^{(3)}$	$R_d^{(3)}$
0.95	0.50	0.50	0.74	1.00	1.00	0.96	1.00	1.00	0.74	1.00
0.95	0.50	1.50	0.73	1.00	1.00	1.00	1.00	1.00	0.74	1.00
0.95	0.50	2.50	0.73	1.00	1.00	0.98	1.00	1.00	0.74	1.00
0.95	0.60	0.50	0.72	0.93	1.00	0.61	1.00	1.00	0.85	1.00
0.95	0.60	1.50	0.74	0.99	1.00	0.88	1.00	0.99	0.85	0.99
0.95	0.60	2.50	0.73	1.00	1.00	0.92	1.00	1.00	0.85	1.00
0.99	0.50	0.50	1.00	0.73	0.74	0.22	1.00	1.00	1.00	0.99
0.99	0.50	1.50	1.00	0.94	0.98	0.50	1.00	1.00	1.00	1.00
0.99	0.50	2.50	1.00	0.97	1.00	0.71	1.00	0.99	1.00	0.99
0.99	0.60	0.50	0.76	0.24	0.36	0.09	1.00	1.00	1.00	1.00
0.99	0.60	1.50	0.92	0.40	0.08	0.02	1.00	1.00	1.00	1.00
0.99	0.60	2.50	1.00	0.55	0.16	0.04	1.00	1.00	1.00	0.99

Table 2.3: Evaluation of detection Algorithm 2 in Section 2.2.3.

$\lambda_\alpha = 0.50$. The parameter p_0 controls for the false alarm rate. The tuning parameter λ is used to robustify the estimation of the tail exponent α (see Algorithm 2). Values of λ close to 1 imply that greater trust is placed on the current estimate of α over the historic ones. This may be disadvantageous when outliers/big spikes are indeed present in the data (large order statistic produce smaller α values). However too much reliability on the past (λ close to 0) makes the algorithm less adaptable to the changes in the traffic.

Table 2.4 illustrates the detection performance of a modification of Algorithm 3, which utilizes EWMA control charts on z-scores, as explained at the end of Section 2.2.4. We employ our methodology for different EWMA pairs $(\lambda_p, L) = \{(2, 0.5), (2, 0.6), (3, 0.5)\}$ and $\lambda = \lambda_\alpha = 0.5$. The simulation settings are the same as that described in the previous paragraph. Best results are obtained for $(\lambda_p, L) = (0.60, 2)$. False alarm rates may be controlled with a higher L and/or decrease further λ_p. These choices are more suitable when the anomalies last for longer durations.

L	λ_p	Gbps	$P_d^{(1)}$	$R_d^{(1)}$	$P_s^{(2)}$	$R_s^{(2)}$	$P_s^{(3)}$	$R_s^{(3)}$	$P_d^{(3)}$	$R_d^{(3)}$
2.00	0.50	0.50	0.45	0.98	0.61	0.97	0.49	0.99	0.35	0.99
2.00	0.50	1.50	0.45	0.99	0.62	0.96	0.41	0.99	0.33	1.00
2.00	0.50	2.50	0.45	0.98	0.64	0.98	0.41	0.99	0.34	1.00
2.00	0.60	0.50	0.61	0.98	0.80	0.96	0.74	0.99	0.45	1.00
2.00	0.60	1.50	0.62	0.97	0.81	0.99	0.50	0.99	0.41	0.99
2.00	0.60	2.50	0.60	0.98	0.81	0.97	0.48	0.99	0.40	0.99
3.00	0.50	0.50	1.00	0.62	0.76	0.25	0.95	0.99	0.78	0.99
3.00	0.50	1.50	0.98	0.74	0.62	0.16	0.73	0.98	0.60	0.98
3.00	0.50	2.50	1.00	0.81	0.90	0.36	0.58	0.98	0.55	0.99
3.00	0.60	0.50	0.96	0.42	0.50	0.13	0.99	0.98	0.92	0.98
3.00	0.60	1.50	1.00	0.59	0.40	0.10	0.79	1.00	0.65	1.00
3.00	0.60	2.50	1.00	0.72	0.56	0.13	0.71	1.00	0.59	1.00

Table 2.4: Evaluation of detection Algorithm 3 in Section 2.2.4.

2.2.7 Case studies

2.2.7.1 Detection and identification of DDoS events

In this section, we apply the detection algorithms of Sections 2.2.3 and Section 2.2.4 to a real world security event recorded recently at Merit. The event which shall be referred to as 'Library' case study' involved a heavy UDP based DNS traffic targeted to a public library in Michigan[2]. Top left panel in Figure 2.8 shows the hashed array for a span of one hour where the attack between 30 to 45 minutes. The top right panel shows a plot of the traffic volume for that hour. The dark strip in the top left panel and the elevated volume in top right panel reveal that our data products captures the attacks at least visually. Alarms raised by the algorithms 2 and 3 have been shown in red in the bottom panel of Figure 2.8. We observe that both algorithms were able successfully identify the period of attack.

[2] This event was also reported from a PeakFlow appliance operated at Merit.

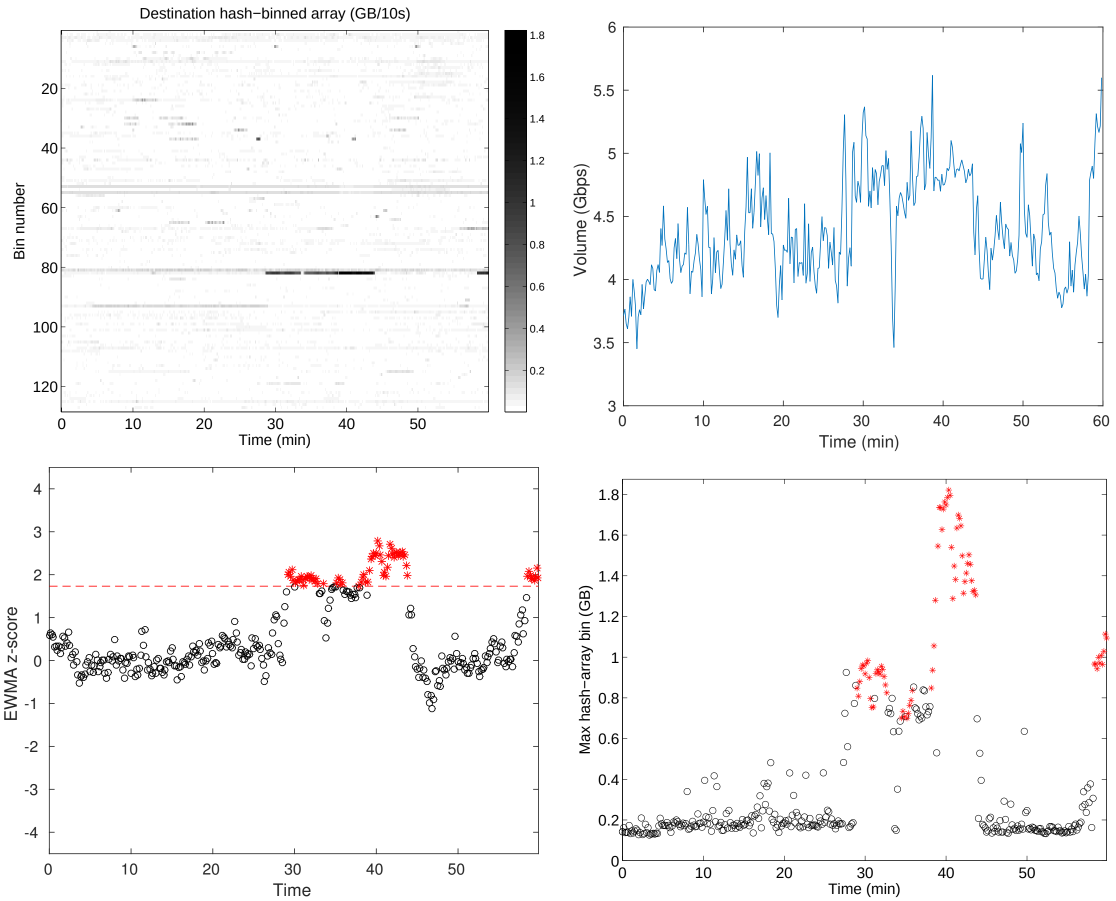


Figure 2.8: Case Study of 'Library' event. Evaluation of our detection algorithm.

2.2.7.2 Case study: detection of structural changes

In this section, we discuss the identification of low volume DDoS attack events. Most of these go unidentified by algorithms (Algorithms 2 and 3) which are better suited for high volumetric attacks. However the community detection algorithm of Section 2.2.5 which aims at identifying structure changes in the traffic comes in more handy. Figure 2.9 presents example of a low volume DDoS where the black lines in the right panel depict malicious activities. The small strip around bin number 52 was due a UDP misuse affecting a Tor exit router within Merit, and the larger strip around bin number 21 represent attempts of SSH-breaking into Michigan-located servers from IPs registered to an autonomous system in the Asia-Pacific region. This case referred to as the 'Tor study' was well identified by the community detection algorithm of Section 2.2.5. However they escaped detection by algorithms 2 and 1 owing to the sparse traffic behavior. Figure 2.10) shows a plot of the highly connected components (exceeding the p-value threshold) for the time duration of 0-60 minutes. The number are connected components rise from 1 to 2 between the time interval 15-45 minutes when both the attacks present themselves.

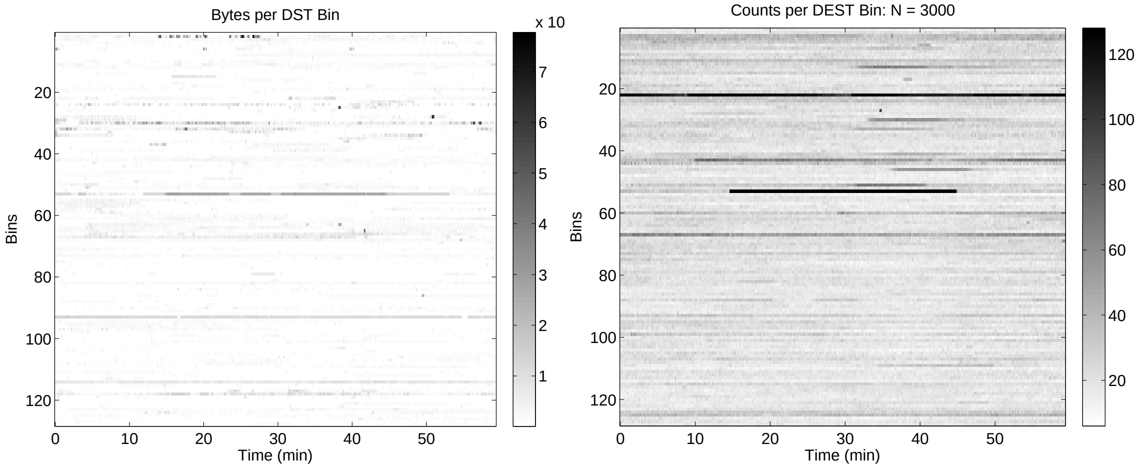


Figure 2.9: 'Tor' case study: Nature of the attacks.

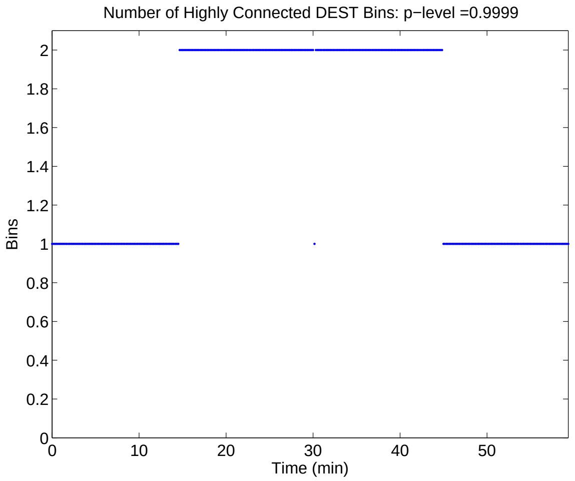


Figure 2.10: 'Tor' case study: Attacks detected by community detection.

To visualize additional structural behavior of the traffic, Figure 2.10 provides a plot of cliques/ clique sizes etc. constructed from the adjacency matrix described in Section 2.2.5. These adjacency matrices provide additional insight into the co-connectivity between sources/ destination via the co-citation/bibliographic matrices. Specifically, let A_t be the matrix constructed as in Section 2.2.5 from the top 90% observations. Then the co-connectivity matrix for sources is given by $S_t = A_t^\top A_t$ and that for destinations is given by $D_t = A_t A_t^\top$. Figure 2.11 shows these co-citation graphs for the period when UDP of misuse of the Tor exit router occurred (The nodes have been relabeled based on the node-degree in decreasing order). The rather large connectivity is clearly evident the src-src graph (top right panel). Indeed these events can be better quantified by considering the clique size for each graph where sudden appearance of large cliques correspond to events of suspicious activity (see bottom panel in Figure 2.11).

2.2.8 Software implementation

We developed the statistical tools which when applied to the hashed array inputs can detect the onset of malicious events like DDoS, power outages, scanning etc. The statistical framework is however incomplete without two main components of AMON viz the data collection software and identification of original IP addresses for heavy hitter bins. The first task is accomplished with a PF_RING ZC module installed at Merit which transmits data on packet headers at a speed of around 10 Gbps+. The explicit details of the software and its online implementation are covered under Section II C of *Kallitsis et al.* (2016b). When an alarm is raised by any one of the three algorithms in Section 2.2.3, 2.2.4 and 2.2.5, one turns to the Boyer Moore algorithm *Boyer and Moore* (1991); *Kallitsis et al.* (2014) for the identification of culprit IPs. For a stream of inputs the Boyer Moore majority algorithm can keep track of the IP contributing to at least 50% of the traffic volume in the stream. Since

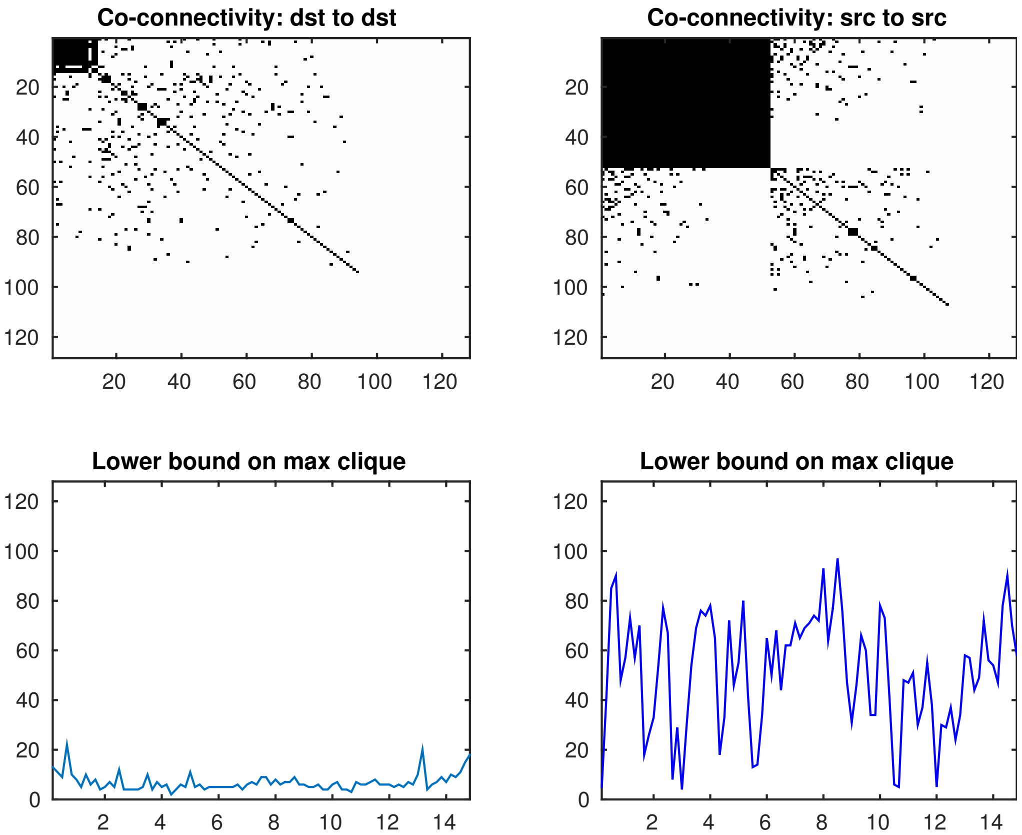


Figure 2.11: Visualizations readily available by our data products. *Top:* Adjacency matrices of co-connectivity graphs (node indices sorted by degree - black corresponds to locations of 1's). *Bottom:* Size of max cliques over time during the 'Tor' case study (Section 2.2.7.2). By observing clique size changes in Dashboards like this, coupled with the detection method of Section 2.2.5, such seemingly innocuous low-volume events are captured.

m is the dimension of hashed IP space, m instances of Boyer Moore are maintained for each of the m substreams. Once an alarm for is flagged, the heavy hitter IP is identified from these sub-streams by a majority vote algorithm. The explicit details of the method and its application to heavy hitter IP identification are covered under Section II B of *Kallitsis et al.* (2016b).

Acknowledgments: Much of this work would have been possible without the sophisticated engineering infrastructure of the AMON developed by Dr. Kallitsis' team at Merit Network. The project was supported by Department of Homeland Security Science and Technology Directorate FA8750-12-2-0314 grant (MK), a National Science Foundation CNS-1422078 grant (MK, GM) and a National Science Foundation grant DMS-1462368 (SS).

Sections 2.2.3 and 2.2.4 heavily depend on the estimation of the tail index α (see 2.14 and 2.18). When outliers are present in the data, the estimates may be severely compromised. Though techniques like exponentially weighted moving average of *Lambert and Liu* (2006a) can reduce the sensitivity to outliers by relying on historical data, they may be ineffective attacks last for a smaller duration of time. This motivated the study of a tail index estimator robust to the presence of outliers which is the topic of our next chapter.

CHAPTER III

Adaptive Trimming of the Hill Estimator

Contributions and due credit: Much of the material in this chapter is based on the work from *Bhattacharya et al.* (2017).

3.1 Introduction

The estimation of the tail index for heavy-tailed distributions is perhaps one of the most studied problems in extreme value theory. Since the seminal works of *Hill* (1975), *Pickands* (1975), *Hall* (1982) and others, numerous aspects of this problem such rate optimality, parameter tuning and applications have been explored (see for example the monographs of *Embrechts et al.* (1997), *Beirlant et al.* (2004b), *de Haan and Ferreira* (2006), *Resnick* (2007) and the references therein).

Given the extensive work on the subject, it may appear naive to hope to say something new. Nevertheless, some curious aspects of this fundamental problem such as sensitivity to outliers, the effective sample size etc. have remained relatively unexplored.

Suppose that $X_1, \cdots, X_n$ is an i.i.d. sample from a heavy tailed distribution F.

Namely,

$$P(X_1 > x) \equiv 1 - F(x) \sim \ell(x)x^{-1/\xi}, \quad \text{as } x \to \infty, \tag{3.1}$$

for some $\xi > 0$ and a slowly varying function $\ell : (0, \infty) \to (0, \infty)$, i.e., $\ell(\lambda x)/\ell(x) \to 1,\ x \to \infty$, for all $\lambda > 0$. The parameter ξ will be referred to as the *tail index* of F. Its estimation is of fundamental importance to the applications of extreme value theory.

The fact that ξ governs the asymptotic tail-behavior of F means that, in practice, one should estimate it by focusing on the most extreme values of the sample. In many applications, one quickly runs out of data since only the largest few order statistics are utilized. In this case, every extreme data-point matters. However the largest order statistics could get *corrupted* in the presence of outliers. Depending on the nature of the outliers, a severe bias may be introduced in the estimation of ξ (see, Tables 3.2 and 3.3 below). In fact, the computed estimate of ξ may be entirely based on these corrupted observations. In such contexts, it is important to have a robust estimator of ξ, which does not necessarily use the most extreme order statistics, perhaps puts less weight on them, or indicates to what extent the most extreme data can be trusted to come from the same distribution.

At first sight, this appears to be an ill-posed problem. Since the tail index ξ is an asymptotic quantity, one *has to focus* on the largest order statistics and if these statistics are corrupted, then there is little or no information left to estimate ξ. Nevertheless, using the joint asymptotic behavior of the extreme order statistics, one can detect statistically significant anomalies in the most extreme order statistics.

The problem of robust estimation of the tail index has already received some attention (see for example *Dutang et al.* (2014), *Goegebeur et al.* (2014), *Knight* (unknown), *Brazauskas and Serfling* (2000), *Peng and Welsh* (2001), *Brzezinski* (2016)).

However, there are still open questions on the optimality and adaptivity of robust estimators to the potentially unknown proportion of extreme outliers. In this chapter, we address these two issues.

Recall the classic *Hill estimator*

$$\widehat{\xi}_k(n) := \frac{1}{k}\sum_{i=1}^{k}\log\left(\frac{X_{(n-i+1,n)}}{X_{(n-k,n)}}\right), \tag{3.2}$$

where $1 \leq k \leq n-1$ and $X_{(n,n)} \geq X_{(n-1,n)} \geq \cdots \geq X_{(1,n)}$ are the order statistics of the sample $X_i,\ i = 1, \cdots, n$. In Section 3.2, we introduce the *trimmed Hill estimator*:

$$\widehat{\xi}^{\rm trim}_{k_0,k}(n) := \sum_{i=k_0+1}^{k} c_{k_0,k}(i)\log\left(\frac{X_{(n-i+1,n)}}{X_{(n-k,n)}}\right), \quad 0 \leq k_0 < k < n. \tag{3.3}$$

Under the Pareto model (3.4), we obtain the *optimal* weights, $c_{k_0,k}(i)$ such that $\widehat{\xi}^{\rm trim}_{k_0,k}$ is the best linear unbiased estimator for ξ (see Proposition III.1, below).

Although the idea of trimming has been considered before by Brazauskas and Serfling *Brazauskas and Serfling* (2000), and most recently by *Zou et al.* (2017), the optimal trimmed Hill estimator has not been derived before. These two works use equal weights in (3.3), thereby producing either suboptimal or biased estimators respectively. Inference for the *truncated* Pareto model has been developed in the seminal work of *Aban et al.* (2006) and recently by *Beirlant et al.* (2016). This should be distinguished from the approach of *trimming* the data in order to achieve robustness, which is the main focus of our work.

Note that the trimmed estimators in (3.3) *do not depend* on the top k_0 order statistics. Therefore, they have a *strong upper break-down point* (see Definition III.3). In the ideal Pareto setting, it turns out that our trimmed Hill estimator is essentially finite-sample optimal among the class of all unbiased estimators of ξ with a fixed *strong upper break-down point* (see Theorem III.4). In Section 3.3.2, we establish the asymptotic normality of the trimmed Hill estimator in the semi parametric regime (3.1), under second order conditions on the regularly varying function ℓ as in *Beirlant et al.* (2006). The rate of convergence of these estimators is the same as that of the classic Hill as long as $k_0 = o(k)$ (see Theorem III.5). The minimax rate–optimality of the trimmed Hill estimators is established in Section 3.3.

These theoretical results though encouraging, are not practically useful unless one has a data-adaptive method for the choice of the trimming parameter k_0. This problem is addressed in Section 3.4. There, we start by introducing *trimmed Hill plot* which can be used to visually determine k_0. Then, by exploiting the elegant joint distribution structure of the optimal trimmed Hill estimators, we devise a weighted sequential testing method for the identification of k_0. The devised sequential testing can be shown to be asymptotically consistent for the general heavy tailed regime. This leads to a new *adaptive trimmed Hill* estimator, which works well even if the degree of contamination in the top order statistics is largely unknown. This novel adaptive robustness property is not present in the existing robust estimators.

In Section 3.6, we demonstrate the need for adaptive robustness and the advantages of our estimator in comparison with established robust estimators in the literature. The finite–sample performance of the trimmed Hill estimator is studied in the context of various heavy tailed models, tail indices, and contamination scenarios in Section 3.5. We also propose a unified approach which can jointly estimate k_0 along with k so that the method is more suited to practical applications. In Section 3.7, we finally summarize our contributions and outline some future problems and practical challenges.

3.2 The Trimmed Hill Estimator

In this section, we shall focus on the fundamental Pareto(σ, ξ) model and assume that

$$P(X > x) = (x/\sigma)^{-1/\xi},\ x \geq \sigma, \tag{3.4}$$

for some $\sigma > 0$ and a tail index $\xi > 0$.

Motivated by the goal to provide a robust estimate of the tail index ξ and in view of the classical Hill estimator in (3.2), we consider the class of statistics, $\widehat{\xi}^{\text{trim}}_{k_0,k}(n)$ defined in (3.3). Proposition III.1 below finds the weights, $c_{k_0,k}(i)$ for which the estimator in (3.3) is unbiased for ξ and also has the minimum variance. Their optimality and robustness are discussed in Section 3.3.

The following result gives the form of the best linear unbiased trimmed Hill estimator. Its proof is given in Section B.

Proposition III.1. *Suppose $X_1, \cdots, X_n$ are i.i.d. observations from the distribution* Pareto(σ, ξ) *as in* (3.4)*. Then among the general class of estimators given by* (3.3), *the minimum variance linear unbiased estimator of ξ is given by*

$$\widehat{\xi}_{k_0,k}(n) = \frac{k_0+1}{k-k_0}\log\left(\frac{X_{(n-k_0,n)}}{X_{(n-k,n)}}\right) + \frac{1}{k-k_0}\sum_{i=k_0+2}^{k}\log\left(\frac{X_{(n-i+1,n)}}{X_{(n-k,n)}}\right), \quad 0 \leq k_0 < k < n. \tag{3.5}$$

For real data the observations may not be i.i.d. in which case the data may require additional pruning methods like hashing (Section 2.2.2) or declustering (Section 4.3.4). The choice of the trimming parameter k_0 is of key importance in practice. In Section 3.4, we propose an automatic data driven methodology for the selection of k_0, which is motivated by the following result.

Proposition III.2. *The joint distribution of* $\widehat{\xi}_{k_0,k}(n)$ *can be expressed in terms of gamma distributed random variables;*

$$\Big\{\widehat{\xi}_{k_0,k}(n),\ k_0 = 0,\ldots,k-1\Big\} \stackrel{d}{=} \Big\{\xi\frac{\Gamma_{k-k_0}}{k-k_0},\ k_0 = 0,\ldots,k-1\Big\}, \tag{3.6}$$

where the Γ_i*'s are as in* (B.1)*. Consequently, we have that*

$$\mathrm{Cov}(\widehat{\xi}_{i,k}(n), \widehat{\xi}_{j,k}(n)) = \frac{\xi^2}{k - i\wedge j}, \qquad i,j = 0,1,\cdots k-1 \tag{3.7}$$

where $\wedge$ *denotes the min operator. Moreover, as* $k - k_0 \to \infty$*,*

$$\sqrt{k-k_0}(\widehat{\xi}_{k_0,k}(n) - \xi) \overset{d}{\Longrightarrow} N(0,\xi^2) \tag{3.8}$$

The proof is given in Section B.

3.3 Optimality And Asymptotic Properties

3.3.1 Optimality in the ideal Pareto case

The trimmed Hill estimators in (3.3) possess a strict upper breakdown point in the following sense.

Definition III.3. A statistic $\widehat{\theta}$ is said to have a strict upper breakdown point β, $0 \le \beta < 1$, if $\widehat{\theta} = T(X_{(n-[n\beta],n)},\cdots,X_{(1,n)})$ where $X_{(n,n)} \ge \cdots \ge X_{(1,n)}$ are the order statistics of the sample. That is, $\widehat{\theta}$ is unaffected by the values of the top $[n\beta]$ order statistics.

Assuming that all observations are generated from Pareto(σ,ξ), the following theorem describes the optimality properties of the trimmed Hill estimator for both the asymptotic and finite sample regimes for a given value of strict upper break down point.

Theorem III.4. *Consider the class of statistics given by*

$$\mathcal{U}_{k_0} = \Big\{ T = T(X_{(n-k_0,n)}, \cdots, X_{(1,n)}) : \mathbb{E}(T) = \xi,\ X_1, \cdots, X_n \overset{i.i.d.}{\sim} \text{Pareto}(\sigma, \xi) \Big\}$$

which are all unbiased estimators of ξ with strong upper breakdown point $\beta = k_0/n$. Then for $\widehat{\xi}_{k_0,n-1}(n)$ as in (3.5), we have

$$\frac{\xi^2}{n-k_0} \le \inf_{T \in \mathcal{U}_{k_0}} Var(T) \le Var(\widehat{\xi}_{k_0,n-1}) = \frac{\xi^2}{n-k_0-1}. \tag{3.9}$$

In particular, $\widehat{\xi}_{k_0,n-1}$ is asymptotically minimum variance unbiased estimator (MVUE) of ξ among the class of estimators described by $\mathcal{U}_{k_0}$.

The proof is given in Section B.

3.3.2 Asymptotic normality

Here, we shall establish the asymptotic normality of the trimmed Hill estimator, $\widehat{\xi}_{k_0,k}$ under the general semi-parametric regime (3.1). In addition, we also briefly discuss the minimax rate optimality of the estimator.

The key idea used is that the trimmed Hill estimator can be expressed in terms for trimmed Hill for ideal Pareto setting plus a remainder term which goes in probability to 0. Next described are the tools to this end.

Consider the tail quantile function

$$Q(t) = \inf\{x : F(x) \ge 1 - 1/t\} = F^{-1}(1 - 1/t),\ \ t > 1 \tag{3.10}$$

where F^{-1} is the generalized inverse of the distribution function F. As in *Beirlant et al.* (2006), we assume the following equivalent representation of (3.1)

$$Q(t) = t^{\xi} L(t) \tag{3.11}$$

where L is a slowly varying function at ∞ (see, e.g., p. 29 in *Bingham et al.* (1989)).

Let $X_i = Q(Y_i), \quad i = 1, \cdots, n$ for Y_i, $i = 1, \cdots, n$ i.i.d from Pareto$(1, 1)$. Then $X_i, \quad i = 1, \cdots, n$ are i.i.d from the distribution F. Thus expressing (3.5) in terms of the quantile function Q we have

$$\widehat{\xi}_{k_0,k}(n) \;=\; \underbrace{\frac{k_0+1}{k-k_0}\log\left(\frac{Y^{\xi}_{(n-k_0,n)}}{Y^{\xi}_{(n-k,n)}}\right) + \frac{1}{k-k_0}\sum_{i=k_0+2}^{k}\log\left(\frac{Y^{\xi}_{(n-i+1,n)}}{Y^{\xi}_{(n-k,n)}}\right)}_{\widehat{\xi}^{*}_{k_0,k}(n)} + R_{k_0,k}(n)$$

where $Y^{\xi}_{(i,n)}$'s are the order statistics for the Y^{ξ}_i's and the remainder $R_{k_0,k}(n)$ is:

$$R_{k_0,k}(n) = \frac{1}{k-k_0}\Big((k_0+1)\log\frac{L(Y_{(n-k_0,n)})}{L(Y_{(n-k,n)})} + \sum_{i=k_0+2}^{k}\log\frac{L(Y_{(n-i+1,n)})}{L(Y_{(n-k,n)})}\Big). \tag{3.12}$$

Since $X^*_i := Y^{\xi}_i$ follow Pareto$(1, \xi)$, the statistic $\widehat{\xi}^{*}_{k_0,k}(n)$ in (3.12) is nothing but the trimmed Hill estimator for the ideal Pareto data with tail index ξ. Under suitable second order regularity assumptions on the function L, $\sqrt{k-k_0}\; R_{k_0,k}(n)$ converges to a constant in probability which in view of (3.8) leads to the asymptotic normality result for $\widehat{\xi}_{k_0,k}(n)$. These second order conditions also used in *Beirlant et al.* (2006) assume that

$$\forall x > 1 : \frac{L(tx)}{L(t)} = 1 + cg(t)\int_1^x \nu^{-\rho-1} d\nu + o(g(t)), \quad t \to \infty \tag{3.13}$$

where $\rho \geq 0$ and $g : (0,\infty) \to (0,\infty)$ is a $-\rho$ varying function [1]. It can be shown that (3.13) implies

$$\sup_{t \geq t_\varepsilon} \left| \log \frac{L(tx)}{L(t)} - cg(t) \int\limits_1^x \nu^{-\rho-1} d\nu \right| \leq \begin{cases} \varepsilon g(t) & \text{if } \rho > 0 \\ \varepsilon g(t) x^\varepsilon & \text{if } \rho = 0. \end{cases} \tag{3.14}$$

for all $\varepsilon > 0$ and some t_ε dependent on ε and g (see Lemma A.2 in *Beirlant et al.* (2006) for more details). Next stated is the asymptotic normality result for the trimmed Hill estimator assuming that (3.14) holds.

Theorem III.5. *Suppose* (3.13) *holds and let $k \to \infty$, $n \to \infty$ and $k/n \to 0$ be such that for some $\delta > 0$,*

$$k^\delta g(n/k) \to A \tag{3.15}$$

for a constant A. Then,

$$k^\delta \max_{0 \leq k_0 < h(k)} \left| \widehat{\xi}_{k_0,k}(n) - \widehat{\xi}^*_{k_0,k}(n) - \frac{cAk^{-\delta}}{1+\rho} \right| \xrightarrow{P} 0, \tag{3.16}$$

where $h(k) = o(k)$ and $\widehat{\xi}_{k_0,k}(n)$ and $\widehat{\xi}^*_{k_0,k}(n)$ are defined in (3.12)*.*

The proof is given in Section B.

Corollary III.6. *If $k_0 = o(k)$ and $\sqrt{k} g(n/k) \to A$,*

$$\sqrt{k}(\widehat{\xi}_{k_0,k}(n) - \xi) \overset{d}{\Longrightarrow} N\left(\frac{cA}{1+\rho}, \xi^2\right)$$

The proof is a direct consequence of Theorem III.5 for $\delta = 1/2$ and result (3.8).

[1] A ρ varying function g has the form $g(x) = x^\rho h(x)$ where $h(x)$ is a slowly varying function satisfying $\lim_{x\to\infty} \frac{h(\lambda x)}{h(x)} = 1$ (see Theorem 1.4.1 in *Bingham et al.* (1987) for more details).

We end this section with a brief discussion of the rate-optimality of the trimmed Hill estimators in the context of the Hall class.

3.3.3 On the minimax rate–optimality

Consider the class of distributions $\mathcal{D} := \mathcal{D}_\xi(B,\rho)$ with tail index $\xi > 0$, such that (3.11) holds, where

$$L(x) = 1 + r(x), \qquad \text{with} \quad |r(x)| \le Bx^{-\rho},\ (x > 0) \tag{3.17}$$

for some *fixed* constants $B > 0$ and $\rho > 0$ (see also (2.7) in *Boucheron and Thomas* (2015)).

Theorem III.7 (uniform consistency)**.** *Suppose that $k = k(n) \propto n^{2\rho/(2\rho+1)}$ and $h(k) = o(k)$, as $n \to \infty$.*

Then, for every sequence $a(n) \downarrow 0$, such that $a(n)\sqrt{k(n)} \to \infty$, we have

$$\liminf_{n\to\infty} \inf_{F\in\mathcal{D}_\xi(B,\rho)} P_F\left(\max_{0\le k_0<h(k)} |\widehat{\xi}_{k_0,k}(n) - \xi| \le a(n)\right) = 1. \tag{3.18}$$

where by P_F, we understand that $\widehat{\xi}_{k_0,k}(n)$ was built using independent realizations from F.

The proof of this result is given in Section B. Relation (3.18) reads as follows. The estimator $\widehat{\xi}_{k_0,k}(n)$ is *uniformly consistent* (at the rate $a(n)$) in *both* the family of possible distributions $\mathcal{D}$ and in the choice of the trimming parameter k_0, so long as $k_0 = o(k)$. This remarkable property shows that $\widehat{\xi}_{k_0,k}(n)$ are minimax rate-optimal in the sense of *Hall and Welsh* (1984). Indeed, Theorem 1 in Hall and Welsh implies the following.

Theorem III.8 (rate optimality). *Let $\widehat{\xi}_n$ be any estimator of ξ based on an independent sample from a distribution $F \in \mathcal{D}_\xi(B, \rho)$. If we have*

$$\liminf_{n\to\infty} \inf_{F\in\mathcal{D}_\xi(\mathcal{B},\rho)} P_F(|\widehat{\xi}_n - \xi| \le a(n)) = 1 \tag{3.19}$$

then $n^{\rho/(2\rho+1)}a(n) = \infty$.

This result shows that no estimator can be uniformly consistent over the Hall class of distributions $\mathcal{D}$ at a rate better than $n^{\rho/(2\rho+1)}$. This is the *minimax* optimal rate that one could possibly hope to achieve. Observe that this result applies also to the trimmed Hill estimators. As seen in Theorem III.7 above the *trimmed Hill* estimators attain this minimax optimal rates uniformly in $k_0 \in [0, h(k)]$, for any $h(k) = o(n^{2\rho/(2\rho+1)})$.

3.4 Data Driven Parameter Selection

3.4.1 Choice of k_0

Suppose X_i, $i = 1, 2, \cdots, n$ are generated from the distribution F of the form (3.1), then the optimal trimmed Hill statistic, $\widehat{\xi}_{k_0,k}(n)$ is asymptotically an unbiased estimator for the tail index ξ (see Theorem III.5) as long as the parameters k_0 and k satisfy (3.16). However, this result breaks down in the presence of outliers, i.e. $\widehat{\xi}_{k_0,k}(n)$ may be biased estimate of ξ for some $1 \le k_0 \le k-1$. The intuition to this end is illustrated via trimmed Hill plots explained below.

For a fixed value of k, trimmed Hill plot is a plot of the values of $\widehat{\xi}_{k_0,k}(n)$ for varying values of k_0 (see Figure 3.1). The vertical lines correspond to $\widehat{\xi}_{k_0,k}(n)\underline{+}\widehat{\sigma}_{k_0,k}(n)$ where $\widehat{\sigma}_{k_0,k}(n) = \widehat{\xi}_{k_0,k}(n)/\sqrt{k-k_0}$ denotes the plug in estimate of the standard error of $\widehat{\xi}_{k_0,k}(n)$ (see Proposition III.2). In the presence of outliers, a change-point in the form of a knee occurs in the values of $\widehat{\xi}_{k_0,k}(n)$, when k_0 is close to true number of outliers, k_0^* (see the knee at $k_0 = 5$ in Figure 3.1). In order to obtain a robust estimate of the tail index ξ, it is essential to obtain an adaptive estimate of the k_0^*. This can be achieved by estimating the location of the knee, which serves as close approximation to the true number of outliers k_0^*. The plug in statistic, $\widehat{\xi}_{\widehat{k}_0,k}(n)$ based on the so-obtained $\widehat{k}_0$ serves as a robust estimate of the tail index, ξ.

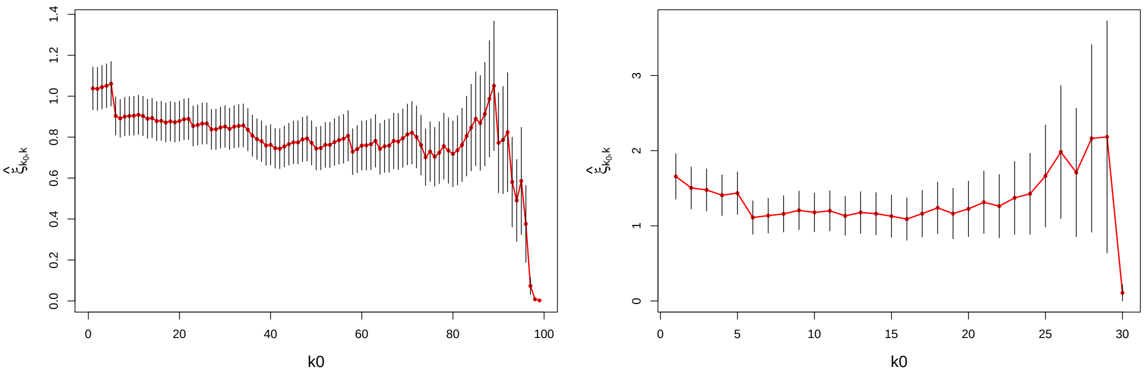


Figure 3.1: Trimmed Hill Plot for 5 outliers and sample size 100. The first knee occurs around $k_0 = 5$. Left: Pareto(1,1) with $k = 99$. Right: Burr(1,0.5,1) with $k = 30$ (see (3.30)).

In order to obtain an accurate estimate for ξ, it is an important task to get an estimate of the parameters k_0 and k. In the first section, we describe the methodology for the estimation of k_0 when k is fixed. Next we describe an iterative algorithm which allows for the estimation of the parameters k_0 and k simultaneously.

Proposition III.9. *Suppose all the X_i's are generated from* Pareto(σ, ξ), *then consider the following class of statistics*

$$T_{k_0,k}(n) := \frac{(k - k_0 - 1)\widehat{\xi}_{k_0+1,k}(n)}{(k - k_0)\widehat{\xi}_{k_0,k}(n)}, \qquad k_0 = 0, 1, \cdots, k - 2. \tag{3.20}$$

The $T_{k_0,k}(n)$*'s are independent and follow* $\mathrm{Beta}(k - k_0 - 1, 1)$ *distribution for* $k_0 = 0, 1, \cdots, k-2$.

Proof. By (3.20) and Proposition III.2, we have

$$\Big(T_{0,k}, \cdots, T_{k-2,k}\Big) \overset{d}{=} \Big(\frac{\Gamma_{k-1}}{\Gamma_k}, \cdots, \frac{\Gamma_1}{\Gamma_2}\Big), \tag{3.21}$$

which implies

$$T_{k_0,k} \overset{d}{=} \frac{\Gamma_{k-k_0-1}}{\Gamma_{k-k_0}} \sim \mathrm{Beta}(k-k_0-1, 1), \quad i = 0, \cdots, k-2.$$

To show the independence of the $T_{k_0,k}$'s, from Relation (B.2) in Lemma B.1 observe that Γ_m and $\{\Gamma_i/\Gamma_m, i = 1, \cdots, m\}$ are independent for all $1 \le m \le k-2$. This in turn implies that

$$\Big(\frac{\Gamma_1}{\Gamma_2}, \frac{\Gamma_2}{\Gamma_3}, \cdots, \frac{\Gamma_{m-1}}{\Gamma_m}\Big) \text{ and } \Gamma_m \text{ are independent.}$$

Since Γ_i, $i = 1, \cdots, m$ and $(E_{m+1}, \cdots, E_k)$ are independent, for all $m = 1, \cdots, k-2$, we have

$$\Big(\frac{\Gamma_1}{\Gamma_2}, \cdots, \frac{\Gamma_{m-1}}{\Gamma_m}\Big) \text{ and } (\Gamma_m, E_{m+1}, \cdots, E_k) \text{ are independent .} \tag{3.22}$$

The independence of the $T_{k_0,k}$'s follows from (3.22) by observing that for all $1 \le m \le k-2$, $\Big(\frac{\Gamma_m}{\Gamma_{m+1}}, \cdots, \frac{\Gamma_{k-1}}{\Gamma_k}\Big)$ is a function of $(\Gamma_m, E_{m+1}, \cdots, E_k)$. □

Remark III.10. We observe that the distribution of $T_{k_0,k}(n)$ depends only on $X_{(n-k_0,n)}$, $\cdots$, $X_{(n-k,n)}$. Therefore the joint distribution of $T_{k_0,k}(n)$'s and hence that of $U_{k_0,k}(n)$'s remains unchanged as long as

$$(X_{(n-k_0,n)}, \cdots, X_{(n-k,n)}) \overset{d}{=} (Y_{(n-k_0,n)}, \cdots, Y_{(n-k,n)})$$

where $Y_{(n,n)} > \cdots > Y_{(1,n)}$ are the order statistics for a sample of n i.i.d. observations from Pareto(σ, ξ). In other words, Proposition III.9 goes through for all $k_0 \geq k_0^*$ provided that the top k_0^* outliers do not perturb the nature of the order statistics $X_{(n-k_0+1,n)}$, $k_0 \geq k_0^*$. This motivates the sequential testing methodology discussed in the next section.

Theorem III.11. *Suppose* (III.5) *in Theorem III.5 holds for some $\delta > 0$. Then,*

$$k^\delta \max_{0 \leq k_0 < h(k)} \left| T_{k_0,k}(n) - T^*_{k_0,k}(n) \right| \xrightarrow{P} 0, \tag{3.23}$$

where $T_{k_0,k}(n)$ and $T^*_{k_0,k}(n)$ are based on $\widehat{\xi}_{k_0,k}(n)$ and $\widehat{\xi}^*_{k_0,k}(n)$ respectively (see (3.12) *and* (3.20) *for explicit expressions).*

The proof of this is described in Section B

Remark III.12. Observe that by Theorem III.11, the asymptotic distribution of $T_{k_0,k}(n)$ and also that of $U_{k_0,k}(n)$ is same as described in Proposition III.9 as long as the number of outliers, $k_0 = o(k)$. This allows us to use the algorithm described below (see Algorithm 4) for the estimation of k_0 in general heavy tailed models (3.1).

3.4.2 Exponentially weighted sequential testing, EWST

Whereas the trimmed Hill plot provides an illustrative estimate of the number of outliers k_0^*, we discuss the weighted sequential testing algorithm for the estimation of k_0^* in a principled manner. One strategy to estimate the true number of outliers,

k_0^*, is to look for the presence of outliers among the set of values, $T_{k_0,k}(n)$. In this context, we define the following statistic

$$U_{k_0,k}(n) := 2|(T_{k_0,k}(n))^{k-k_0-1} - 0.5|, \quad k_0 = 0, 1, \cdots, k-2. \tag{3.24}$$

For i.i.d. observations from Pareto(σ, ξ), $U_{k_0,k}(n)$ are i.i.d. $U(0,1)$ random variables[2]. An estimate of k_0^* is obtained by identifying the largest value of k_0 for which the hypothesis that $U_{k_0,k}(n)$ follows $U(0,1)$ gets rejected.

In this direction, we begin with a large value of $k_0 = f(k)$ and test the hypothesis: $U_{k_0,k}(n) \sim U(0,1)$. If rejected, we stop our search and declare $\widehat{k}_0 = k_0$. Otherwise, we decrease the value of k_0 by 1 and proceed until the hypothesis $U_{k_0,k}(n) \sim U(0,1)$ gets rejected or $k_0 = 0$. The resulting value of k_0 then gives an estimate of k_0^*. The level for these tests increases exponentially with decrease in k_0. This is done in order to guard against large values k_0 close to k.

The methodology is formally described in the following algorithm.

Algorithm 4 Exponentially weighted sequential testing

```
1: Let $q \in (0,1)$ be the significance level.
2: Choose a constant $a > 1$ and set $c = 1/\sum_{i=1}^{k-1} a^i$.
3: Set $k_0 = f(k)$.
4: Compute $T_{k_0,k}(n) = ((k-k_0-1)\widehat{\xi}_{k_0+1,k}(n)/((k-k_0)\widehat{\xi}_{k_0,k}(n))$.
5: Compute $U_{k_0,k}(n) = 2|(T_{k_0,k}(n))^{k-k_0-1} - 0.5|$ as defined in (3.24).
6: If $\log(U_{k_0,k}(n)) < ca^{k-k_0-1}\log(1-q)$, set $k_0 = k_0 - 1$ else goto step 8.
7: If $k_0 \geq 0$, goto step 3 else $k_0 = k_0 + 1$.
8: Return $\widehat{k}_0 = k_0$.
```

Proposition III.13. *For i.i.d. observations from Pareto(σ, ξ) and $q \in (0,1)$, let $\widehat{k}_0(q)$ be the estimate of k_0^* based on Algorithm 4 with $f(k) = k-2$, then under the null hypothesis $H_0 : k_0^* = 0$, we have $P_{H_0}[\widehat{k}_0 > 0] = q$.*

Proof. The type I error for Algorithm 4, $P_{H_0}[\widehat{k}_0 > 0] = 1 - P_{H_0}[\widehat{k}_0 = 0]$ where

[2] If $X \sim \text{Beta}(p, 1)$, $X^p \sim U(0,1)$

$$
\begin{aligned}
P_{H_0}[\widehat{k}_0 = 0] &= P_{H_0}\Big[\log(U_{0,k}) < ca^{k-1}\log(1-q), \cdots, \log(U_{k-2,k}) < ca\log(1-q)\Big] \\
&= \Pi_{i=0}^{k-2} P_{H_0}\Big[U_{i,k} < (1-q)^{ca^{k-i-1}}\Big] \\
&= \Pi_{i=0}^{k-2}(1-q)^{ca^{k-i-1}} \\
&= (1-q)^{c\sum_{i=0}^{k-2} a^{k-i-1}} = 1-q.
\end{aligned}
$$

where the last equality follows since $c = \sum_{i=1}^{k-1} a^i = \sum_{i=0}^{k-2} a^{k-i-1}$. □

Remark III.14. For Pareto case we attain the the exact bound of type I error. The bound is also attained asymptotically for the general heavy tailed distribution in (3.1) but requires additional assumptions. The following theorem sheds light on the reason behind the consistency of EWST for the more general heavy tailed setup.

Theorem III.15. *If* (3.23) *holds for some* $1 < \delta < 2$*, then*

$$k^{(\delta-1)} \max_{0 \le k_0 < h(k)} \left| U_{k_0,k}(n) - U^*_{k_0,k}(n) \right| \xrightarrow{P} 0, \tag{3.25}$$

with $U_{k_0,k}(n)$ *and* $U^*_{k_0,k}(n)$ *are defined in* (3.24)*. Moreover, if* $f(k) = O(k^{\delta-1})$*,*

$$P_{H_0}[\widehat{k}_0 > 0] \xrightarrow{P} q. \tag{3.26}$$

The proof is described in Section B.

3.5 Simulations

In this section, we evaluate the performance of the adaptive trimmed Hill estimator, $\widehat{\xi}_{\hat{k}_0,\hat{k}}(n)$, in terms of the mean squared error, MSE as

$$\mathrm{MSE}(\widehat{\xi}_{\hat{k}_0,\hat{k}}(n)) = \mathbb{E}(\widehat{\xi}_{\hat{k}_0,\hat{k}}(n) - \xi)^2 \tag{3.27}$$

For comparison, we compute the asymptotic relative efficiency, ARE with respect to both the trimmed Hill estimator, $\widehat{\xi}_{k_0,k}(n)$ and the classic Hill, $\widehat{\xi}_k(n)$. The formulas are given by

$$\begin{aligned} \mathrm{ARE}_{\mathrm{TRIM}} &= \mathrm{MSE}(\widehat{\xi}_{k_0,k}(n))/\mathrm{MSE}(\widehat{\xi}_{\hat{k}_0,\hat{k}}(n)) \qquad (3.28) \\ \mathrm{ARE}_{\mathrm{HILL}} &= \mathrm{MSE}(\widehat{\xi}_k(n))/\mathrm{MSE}\widehat{\xi}_{\hat{k}_0,\hat{k}}(n)) \end{aligned}$$

respectively, where k_0 is the true trimming parameter, and k is replaced by its optimal choice as:

$$k^*_{n,k_0} = \underset{k=k_0+1,\cdots,n-1}{\arg\min} \mathrm{MSE}(\widehat{\xi}_{k_0,k}(n)). \tag{3.29}$$

We first explore the performance of exponentially weighted sequential testing algorithm, EWST as described in Section 3.4.2 as an estimator of the trimming parameter k_0. In Sections 3.5.1, 3.5.2 and 3.5.4, we replace $\hat{k}$ in (3.28) by the optimal values k^*_{n,k_0} in (3.29).

In Section 3.5.5, we will address the performance of the adaptive trimmed Hill, $\widehat{\xi}_{\hat{k}_0,\hat{k}}(n)$ where k is unknown and estimated from the data as described in Section 3.5.5.

The efficacy of the proposed algorithms have been explored in the light of the following heavy-tailed distributions.

$$\begin{aligned}
\text{Pareto}(\sigma,\alpha) \quad &: \quad 1-F(x) = \sigma^{\alpha}x^{-\alpha};\ \ x>1, \alpha>0;\ \ \xi = 1/\alpha \qquad (3.30)\\
\text{Frechet}(\alpha) \quad &: \quad 1-F(x) : 1-\exp(-x^{-\alpha});\ \ x>0, \alpha>0;\ \ \xi = 1/\alpha\\
\text{Burr}(\eta,\lambda,\tau) \quad &: \quad 1-F(x) = 1-\left(\frac{\eta}{\eta+x^{-\tau}}\right)^{-\lambda};\ \ x,\eta,\lambda,\tau>0;\ \ \xi = 1/\tau\\
|\text{T}|(t) \quad &: \quad 1-F(x) = \int\limits_{x}^{\infty} \frac{2\Gamma(\frac{t+1}{2})}{\sqrt{n\pi}\Gamma(\frac{t}{2})}\left(1+\frac{w^2}{t}\right)^{-\frac{t+1}{2}} dw;\ \ x>0, t>0;\ \ \xi = 1/t
\end{aligned}$$

In Sections 3.5.2 and 3.5.4, the number of outliers k_0 and the tail index ξ are kept fixed. Varying values of ξ and k_0 are studied in Section 3.5.3.

3.5.1 Performance under H_0 ($k_0 = 0$)

In this section, we let $X_1, \cdots, X_n$ be i.i.d. generated from one of the four distributions in (3.30). The tail index ξ is fixed at 1. We assume that there are no outliers, i.e. $k_0 = 0$ which in turn implies that the trimmed Hill coincides with the classic Hill estimator.

Assuming $k = \hat{k} = k^{*}_{n,0}$, we evaluate the performance of the adaptive trimmed Hill, $\widehat{\xi}_{\hat{k}_0,\hat{k}}(n)$ with respect to the classic Hill, $\widehat{\xi}_{k}(n)$ in terms of ARE using (3.28). The trimming parameter estimate $\hat{k}_0$ is obtained using EWST as in Section 3.4.2. The ARE's are based on 5000 independent Monte Carlo realizations. For EWST , the significance level, q and the exponentiation parameter a are fixed at 0.05 and 1.2 respectively.

n	Pareto(1,1)	Frechet(1)	Burr(1,0.5,1)	T(1)
100	99.17	97.19	86.25	97.22
200	99.53	99.33	96.64	99.83
500	99.85	99.88	98.27	99.85

Table 3.1: ARE of the adaptive trimmed Hill with respect to the classic Hill, $k_0 = 0$ and $\xi = 1$.

As seen in Table 3.1, apart from the Burr distribution, we have fairly large ARE values (almost 100%) even at sample size $n = 100$. This indicates that the EWST algorithm picks up the true $k_0 = 0$ in almost all of the cases. As the sample size grows ($n = 500$), the behavior is more uniform across different distribution and we achieve nearly 100% asymptotic relative efficiency even for the Burr case. This may be explained by the asymptotic Pareto-like behavior of the heavy tailed distributions (see (3.1)).

In the following section, we explore the behavior of adaptive trimmed Hill when there are non zero outliers in the data, i.e. $k_0 > 0$.

3.5.2 Inflated outliers ($k_0 > 0$)

We simulate from one of the distributions in (3.30) with $\xi = 1$. We introduce k_0 outliers by perturbing the top-k_0 order statistics using one of the following two approaches

$$X_{(n-i+1,n)} := X_{(n-k_0,n)} + (X_{(n-i+1,n)} - X_{(n-k_0,n)})^L, \quad i = 1, \cdots, k_0, \quad L > 1 \tag{3.31}$$

$$X_{(n-i+1,n)} := X_{(n-k_0,n)} + C(X_{(n-i+1,n)} - X_{(n-k_0,n)}), \quad i = 1, \cdots, k_0, \quad C > 1 \tag{3.32}$$

For $L, C > 1$, the transformations (3.31) and (3.32) lead to inflation of the top-k_0 order statistics while still preserving their order.

We first fix $k_0 = 10$ and assume that $k = \hat{k} = k^*_{n,10}$. We then obtain the trimming parameter estimate, $\hat{k}_0$ and the corresponding adaptive trimmed Hill estimator, $\widehat{\xi}_{\hat{k}_0,\hat{k}}$ by using the EWST algorithm in Section 3.4.2. The performance is evaluated in terms of the ARE relative to the trimmed Hill and $\widehat{\xi}_{k_0,k}(n)$ and the classic Hill $\widehat{\xi}_k(n)$, as in (3.28). Tables 3.2 and 3.3 show the performance of the adaptive trimmed Hill for varying values of L and C respectively.

n	100				200				500			
L	1.2	1.5	5	20	1.2	1.5	5	20	1.2	1.5	5	20
Pare(1,1)	0.92	0.54	0.94	0.99	0.94	0.77	0.98	1.00	0.95	0.94	1.00	1.00
	1.03	2.95	76.5	1808	1.02	3.58	644.9	1608	1.02	3.21	45.7	1114
Frech(1)	0.82	0.26	0.69	0.96	0.74	0.37	0.89	0.99	0.71	0.56	0.97	1.00
	11.6	3.66	10.4	13.9	21.7	10.8	26.5	28.9	45.4	35.7	59.8	62.2
Bu(1,0.5,1)	1.13	0.33	0.21	0.94	0.87	0.26	0.54	0.96	0.74	0.37	0.88	0.99
	5.19	1.48	0.88	4.24	9.47	2.92	10.3	19.7	5.87	10.3	23.8	25.8
|T|(1)	0.87	0.29	0.56	0.96	0.79	0.36	0.85	0.98	0.75	0.62	0.95	1.00
	15.6	5.11	10.0	17.0	31.7	14.5	33.3	38.0	71.6	58.4	88.6	94.7

Table 3.2: ARE of the adaptive trimmed Hill $k_0 = 10$, $\xi = 1$ and $L > 1$. For each distribution, top row corresponds to ARE_{TRIM} and bottom row indicates ARE_{HILL}.

n	100				200				500			
C	2	10	20	100	2	10	20	100	2	10	20	100
Pareto(1,1)	0.93	0.57	0.64	0.95	0.95	0.73	0.80	0.98	0.98	0.87	0.93	0.99
	1.01	1.85	3.33	12.7	1.01	1.57	2.57	7.22	1.00	1.29	1.79	3.52
Frechet(1)	0.84	0.29	0.37	0.70	0.81	0.39	0.45	0.83	0.81	0.51	0.60	0.92
	11.8	3.94	5.44	10.3	22.7	11.8	13.9	23.7	50.2	31.6	37.7	59.6
Burr(1,0.5,1)	0.98	0.33	0.27	0.37	0.86	0.33	0.33	0.62	0.8	0.44	0.47	0.83
	4.52	1.44	1.20	1.58	9.17	3.58	3.65	6.82	21.3	12.0	12.7	22.3
|T|(1)	0.80	0.28	0.30	0.58	0.77	0.38	0.47	0.86	0.83	0.54	0.65	0.93
	14.5	4.97	5.34	10.6	31.4	15.2	18.9	31.9	80.5	53.0	61.4	87.4

Table 3.3: ARE of the adaptive trimmed Hill $k_0 = 10$, $\xi = 1$ and $C > 1$. For each distribution, top row corresponds to ARE_{TRIM} and bottom row indicates ARE_{HILL}.

We first observe the ARE values compared to the oracle trimmed Hill statistic are relatively stable and improve considerably with the increase in sample size n. For outliers of small magnitude, i.e $L = 1.2$ and $C = 2$, the ARE values are relatively higher as compared to the case of moderate outliers, i.e. $L = 2, 5$ or $C = 10, 20$. This is natural, since small values of L and C are indicative of lower levels of contamination and thus the estimation of ξ is accurate even if k_0 is underestimated. For $L = 2, 5$ and $C = 10, 20$, we have for estimation accuracy for the trimming parameter $\hat{k}_0$ (observed in histograms of $\hat{k}_0$ not reported here). However, the increase in severity of outliers produces a greater error in the estimation of ξ. Outliers of large magnitude, i.e. $L = 20$ and $C = 100$, allow for nearly perfect detection accuracy for the trimming parameter k_0 and hence the ARE values close to 100%.

The estimation of k_0 is best under the Pareto setting followed by Frechet and the T-distribution. Of all cases, the Burr distribution is most challenging. This is explained by the slow rate of convergence of Burr tails to Pareto tails and hence the relatively lower efficiency of the adaptive trimmed Hill. For large sample sizes $n = 500$, sensitivity of the adaptive trimmed Hill to underlying distribution structure decreases and we attain nearly 100% accuracy uniformly across all distributions when $L > 2$ and $C > 20$.

Finally, we observe the unusually large ARE values relative to the classic Hill. It is remarkable that even small perturbations in the top order statistics ($L = 1.2$ and $C = 2$) lead to an unacceptable bias of the classic Hill estimator. The MSE deteriorates by a factor of 14 or 15 in case of the T-distribution for $n = 100$ and it could be as bad as 80 when $n = 500$. This highlights the importance of considering adaptive robust estimators of ξ in real data problems where the observations could be contaminated. For the remaining section, we shall thus consider the ARE values relative to the trimmed Hill only.

3.5.3 Role of ξ and k_0

In this section, we explore the influence of the tail exponent ξ and the extent of contamination k_0 on the EWST algorithm of Section 3.4.2. Figures 3.2 and 3.3

display the ARE values of the adaptive trimmed Hill, $\widehat{\xi}_{\hat{k}_0,\hat{k}}(n)$ for varying values of ξ and k_0 respectively.

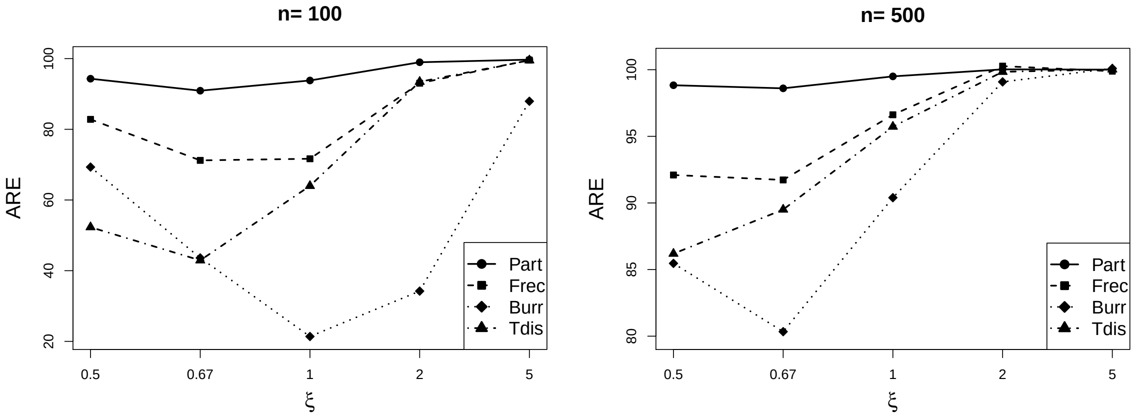


Figure 3.2: Performance for varying values of tail exponent ξ. Left and right panels correspond to $n = 100$ and $n = 500$, respectively.

We first inject $k_0 = 10$ top outlier statistics as in (3.31) with $L = 5$. The underlying distributions from which the data is generated correspond to Pareto$(1, 1/\xi)$, Frechet$(1/\xi)$, Burr$(1, 0.5, 1/\xi)$ and $|T|(1/\xi)$ with ξ in the range $\{0.5, 0.67, 1, 2, 5\}$. With $k = \hat{k} = k^*_{n,10}$, we consider the ARE values of the adaptive trimmed Hill, $\widehat{\xi}_{\hat{k}_0,\hat{k}}(n)$ relative to the trimmed Hill, $\widehat{\xi}_{k_0,k}$ for varying values of ξ. Figure 3.2 shows this behavior for varying sample sizes.

We observe that the efficiency of the proposed adaptive trimmed Hill approaches 100% for $\xi > 0.67$ for all distributions with increase in sample size n. This is expected as the heavy tailed distributions in (3.1) get closer to the Pareto distribution asymptotically. Whereas the Pareto distribution is more or less robust to the change in the tail exponent ξ, the other three heavy tailed distributions suffer from a mild loss in efficiency in the range $\xi < 2$. The superior performance at higher values of ξ indicates easy identification of outliers in heavier tails. The performance of our estimator improves on both sides of $\xi = 0.67$ for all distributions apart from the Burr. The Burr distribution is the most challenging in terms of identification of the trimming parameter k_0 and has relatively low efficiency for $0.67 < \xi < 2$ especially for smaller sample sizes.

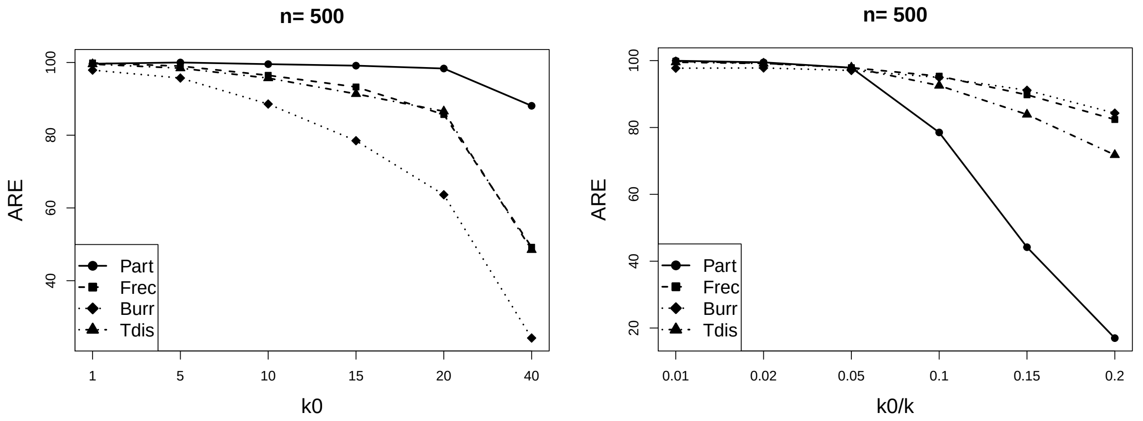


Figure 3.3: Performance for varying amount of outliers. Left and right panels correspond to the number k_0 and the proportion k_0/k of outliers respectively.

We next inject k_0 top outlier statistics as in (3.31) with $L = 5$ and distribution in (3.30) with $\xi = 1$. The underlying distributions from which the data is generated correspond to Pareto$(1, 1)$, Frechet(1), Burr$(1, 0.5, 1)$ and $|T|(1)$. We consider two different scenarios, one where $k_0 \in \{1, 5, 10, 15, 20, 40\}$ and the other where $k_0/k \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$. For scenario 1 (varying k_0), we let $k = \hat{k} = k^*_{n,k_0}$ and for scenario 2 (varying k_0/k), we let $k = \hat{k} = k^*_{n,0}$. We then apply the EWST algorithm for estimation of k_0. Figure 3.3 displays the ARE values of the adaptive trimmed Hill, $\widehat{\xi}_{\hat{k}_0,k}(n)$ relative to the trimmed Hill, $\widehat{\xi}_{\widehat{k_0},k}$.

We observe that with increase in both number (k_0) and proportion of outliers (k_0/k), naturally the efficiency of the proposed adaptive trimmed Hill decreases. This may be attributed to the fact that the detection accuracy of k_0 becomes increasingly difficult with increase in both number and proportion of outliers. From Figure 3.3, we observe that when the number of outliers, k_0 is kept constant, the ARE of Pareto is the greatest while that of Burr is the least. On the other when the proportion k_0/k is kept constant, the performance under Burr is the best while that under Pareto is the worst. This unusual phenomenon can be explained as follows. For same sample size, the optimal $k = k^*_{n,k_0}$ is the largest for Pareto followed by T, Frechet and Burr. Since a large effective sample size allows for better estimation of k_0, therefore the highest ARE values are obtained corresponding to the Pareto distribution in 3.3 left. For k_0/k constant, large $k = k^*n, k_0$ implies large number of outliers, k_0. Since $k^*_{n,0}$ is smallest in case of the Burr distribution, we record largest ARE values for the Burr distribution in Figure 3.3 right.

3.5.4 Deflated outliers, $k_0 > 0$

We simulate from one of the distributions in (3.30) with $\xi = 1$. We introduce k_0 outliers by perturbing the top order statistics as in (3.31) and (3.32) where now $L, C < 1$. For $L, C < 1$, the transformations (3.31) and (3.32) lead to the deflation of the top-k_0 order statistics while still preserving their order.

n	100			200			500		
L	0.005	0.05	0.5	0.005	0.05	0.5	0.005	0.05	0.5
Pareto(1,1)	0.99	0.82	0.90	1.00	0.88	0.90	1.00	0.96	0.93
Frechet(1)	1.35	1.64	2.87	1.20	1.54	1.92	1.12	1.23	1.68
Burr(1,0.5,1)	1.59	3.04	5.14	1.38	1.97	2.91	1.21	1.46	2.06
|T|(1)	1.31	1.59	2.33	1.17	1.22	1.59	1.10	1.13	1.42

Table 3.4: ARE of the adaptive trimmed Hill relative to trimmed Hill for $k_0 = 10$, $\xi =$ and $L < 1$.

With $k_0 = 10$ and $k = \hat{k} = k^*_{n,10}$, we obtain the trimming parameter estimate, $\hat{k}_0$ and the corresponding adaptive trimmed Hill estimator, $\widehat{\xi}_{\hat{k}_0,\hat{k}}$ using the EWST algorithm in Section 3.4.2. Their performance is evaluated in terms of the ARE relative to the trimmed Hill $\widehat{\xi}_{k_0,k}(n)$ in Tables 3.4 and 3.5.

n	100			200			500		
L	0.005	0.05	0.5	0.005	0.05	0.5	0.005	0.05	0.5
Pareto(1,1)	0.97	0.75	1.05	0.98	0.87	1.02	1.00	0.94	1.00
Frechet(1)	1.09	1.68	2.22	1.06	1.97	1.71	1.07	1.75	1.46
Burr(1,0.5,1)	1.11	3.46	3.51	1.05	2.95	2.48	1.08	1.99	1.78
|T|(1)	1.06	1.50	1.94	1.06	1.40	1.52	1.03	1.32	1.29

Table 3.5: ARE of the adaptive trimmed Hill relative to trimmed Hill for $k_0 = 10$, ξ=1 and $C < 1$.

For the Pareto distribution, the ARE is higher in for more severe outliers $L = 0.005$, $C = 0.001$ than the case of moderate outliers $L = 0.05$, $C = 0.1$. This is because more extreme outliers facilitate easier estimation of k_0 and hence the large ARE. However observe that for $L = 0.5, C = 0.5$, we have greater than the case of $L = 0.05$, $C = 0.1$. This is because values of L and C close to 1 under estimation of k_0 does not have a huge impact on the MSE of the adaptive trimmed Hill. For the distributions apart from Pareto, we obtain ARE values which are greater than 100%. The detection accuracy of EWST in determining k_0 has the exact same trend as that for the Pareto case. However, for other heavy tailed distributions, a few downscaled outliers sometimes helps in improving the MSE value of the adaptive trimmed Hill. As a result, the adaptive trimmed Hill outperforms the oracle trimmed Hill benchmark based on the true value of k_0.

3.5.5 Joint estimation of k and k_0

From Relation (3.18) in Theorem III.7 and Theorem III.8 , we observe that if $k_0 = o(n^{2\rho/(2\rho+1)})$, the asymptotic mean squared error (AMSE) of the trimmed Hill estimator is same as that of the classic Hill. Therefore following *Hall and Welsh*

(1985), the value of k which minimizes the AMSE of the trimmed Hill is

$$k_n^{\text{opt}} \sim \left(\frac{C^2\rho(\rho+1)^2}{2D^2\rho^3}\right)^{1/(2\rho+1)} n^{2\rho/(2\rho+1)}.$$

The finite sample equivalent of k_n^{opt} is given by k_{n,k_0}^* as in (3.29). Drees and Kaufmann in *Drees and Kaufmann* (1998) provide a methodology for the estimation of k_n^{opt} for the classic Hill. Motivated by their approach, we propose a method for the joint estimation of k_0 and k under the following assumptions

$$\begin{aligned} k_0 &= o(n^{2\rho/(2\rho+1)}) \\ 1-F(x) &= Cx^{-1/\xi}(1+Dx^{-\rho/\xi}+o(x^{-\rho/\xi})) \\ F^{-1}(1-t) &= ct^{-\xi}\exp\left(\int_t^1 \frac{\varepsilon(s)}{s}ds\right) \end{aligned} \tag{3.33}$$

The last two assumptions correspond to Eqs (2) and (5) in *Drees and Kaufmann* (1998) respectively.

Suppose the trimming parameter, k_0 is known. We define the modified version of Eq (4) in *Drees and Kaufmann* (1998) as

$$\bar{k}_{n,k_0}(r_n) = \min\left\{k \in \{k_0+1,\cdots,n-1\}\Big| \max_{k_0+1\le i\le k}(i-k_0+1)^{1/2}|\widehat{\xi}_{k_0,i}(n)-\widehat{\xi}_{k_0,k}(n)| > r_n\right\} \tag{3.34}$$

where $\hat{\xi}_{k_0,i}$ is the trimmed Hill based on $i-k_0$ observations. We conjecture a modified version of Theorem 1 in *Drees and Kaufmann* (1998), where the classic Hill estimator gets replaced by its corresponding trimmed version as follows:

Proposition III.16. *Suppose* $r_n = o((n-k_0)^{1/2})$, $\log\log(n-k_0) = o(r_n)$ *and* (3.33) *holds. Then if* $\hat{\rho}_n$ *is any consistent estimator of* ρ *and* $\tilde{\xi}_n$ *is a consistent initial* *estimator of* ξ*, then for* $\epsilon \in (0,1)$ *and* $(\log\log(n-k_0))^{1/2\epsilon} = o(r_n)$*, we have*

$$\hat{k}^{\text{opt}}_{n,k_0} = (2\hat{\rho}_n + 1)^{-1/\hat{\rho}_n}(2\tilde{\xi}_n\hat{\rho}_n)^{1/(2\hat{\rho}_n+1)}\left(\frac{\bar{k}_{n,k_0}(r^{\epsilon}_n)}{(\bar{k}_{n,k_0}(r_n))^{\epsilon}}\right)^{1/(1-\epsilon)} \tag{3.35}$$

is a consistent estimator of k^*_{n,k_0} *in the sense that* $\hat{k}_{n,k_0}/k^*_{n,k_0}$ *converges in probability to 1. In particular,* $\widehat{\xi}_{k_0,\hat{k}^{\text{opt}}_{n,k_0}}(n)$ *has the same asymptotic efficiency as* $\widehat{\xi}_{k_0,k^*_{n,k_0}}(n)$*.*

The trimmed estimator, $\hat{\xi}_{k_0,2\sqrt{n}}$ can be used as an initial consistent estimator of ξ for a wide range distributions from (3.1). As in *Drees and Kaufmann* (1998), it can be shown that for $\lambda \in (0,1)$, a consistent estimator of ρ is given by

$$\hat{\rho}^{(1)}_{n,k_0,\lambda}(r_n) = \log_{\lambda}\frac{\max_{k_0+1\le i\le[\lambda\bar{k}_{n,k_0}(r_n)]}(i-k_0+1)^{1/2}|\widehat{\xi}_{k_0,i}(n) - \widehat{\xi}_{k_0,[\lambda\bar{k}_{n,k_0}(r_n)]}(n)|}{\max_{k_0+1\le i\le[\bar{k}_{n,k_0}(r_n)]}(i-k_0+1)^{1/2}|\widehat{\xi}_{k_0,i}(n) - \widehat{\xi}_{k_0,[\bar{k}_{n,k_0}(r_n)]}(n)|} - \frac{1}{2} \tag{3.36}$$

The detailed presentation of the proof of Proposition III.16 shall be the subject of another work. Here, we shall only demonstrate its application in practice (see Tables 3.6 and 3.7).

We next describe a methodology which allows for the estimation of k when the trimming parameter k_0 is unknown. In this direction, we start with an initial choice of the parameter k. From this initial choice of k, we estimate the trimming parameter, k_0 using EWST Algorithm 4. With this choice of $\hat{k}_0$, we obtain an estimate for k by using Proposition III.16. We iterate between the values of k and k_0, unless convergence is obtained. Next, we describe the methodology more formally:

Algorithm 5 Joint estimation of k_0 and k.

```
1: Set a threshold $\tau$ and $i = 1$.
2: Choose $k$ as a function of $n$. Let $\hat{k}^{(0)}$ be this initial choice.
3: Let $i = i + 1$.
4: With $k = \hat{k}^{(i)}$, obtain $\hat{k}_0^{(i)}$ using Algorithm 4
5: With $k_0 = \hat{k}_0^{(i)}$, obtain $\hat{k}_{(i+1)}$ using (3.35) in Proposition III.16.
6: If $|\hat{k}^{(i+1)} - \hat{k}^{(i)}| > \tau$, goto step 4 else goto step 7
7: Return $\hat{k} = \hat{k}^{(i)}$ and $\hat{k}_0 = \hat{k}_0^{(i)}$.
```

In order to evaluate the performance of Algorithm 5, we first consider the ARE of the adaptive trimmed Hill, $\widehat{\xi}_{\hat{k}_0,\hat{k}}$ relative to the trimmed Hill, $\widehat{\xi}_{k_0,k^*_{n,k_0}}$ where k^*_{n,k_0} is obtained as in (3.29). Table 3.6 shows the ARE values of for Frechet and T distributions with varying tail indices (see (3.30)). The number of outliers k_0 is fixed at 10 and two values of $L = 5, 20$ are chosen. The two columns correspond to the case where ρ is either fixed at constant 1 or estimated using (3.36) for $\lambda = 0.6$.

L	n	Frech(5)		Frech(2)		Frech(1)		|T|(4)		|T|(10)	
		1	$\hat{\rho}^{(1)}$	1	$\hat{\rho}^{(1)}$	1	$\hat{\rho}^{(1)}$	1	$\hat{\rho}^{(1)}$	1	$\hat{\rho}^{(1)}$
5	100	0.74	0.63	0.31	0.18	0.27	0.24	0.86	0.63	0.88	0.64
	200	0.71	0.64	0.37	0.51	0.72	0.49	0.87	0.66	0.86	0.64
	500	0.78	0.60	0.66	0.53	0.71	0.46	0.83	0.62	0.80	0.69
20	100	0.76	0.61	0.87	0.68	0.50	0.55	0.94	0.74	0.90	0.77
	200	0.74	0.71	0.86	0.60	0.74	0.47	0.89	0.66	0.85	0.69
	500	0.79	0.63	0.88	0.51	0.75	0.44	0.82	0.80	0.63	0.68

Table 3.6: ARE of the adaptive trimmed Hill relative to trimmed Hill for $k_0 = 10$.

We observe that the ARE values are nearly 75% for the Frechet and become as large as 90% for the T distribution. The performance is better when $\rho = 1$ rather than estimated from the data using (3.36). Large values of $\alpha = 1/\xi$ lead to greater ARE values for both T and Frechet. This behavior is similar to that observed in the case of known k (see Figure 3.2). Increase in the severity of outliers, L leads to overall improvement in the efficiency, a phenomenon also seen previously in Section 3.5.2.

In order to allow for a comparative baseline to our results in Table 3.6, we replicate the settings of Tables 3 and 6 in *Drees and Kaufmann* (1998). We consider the ratio of root mean squared error of the adaptive trimmed to that of the trimmed as:

$$R = \sqrt{\mathrm{MSE}(\widehat{\xi}_{\hat{k}_0,\hat{k}})}\Big/\sqrt{\mathrm{MSE}(\widehat{\xi}_{k_0,k^*_{n,k_0}})}$$

The results in *Drees and Kaufmann* (1998) correpond to $k_0 = 0$ and $k^*_{n,k_0} = k_n^{\mathrm{opt,sim}}$. As can been from the Table, our results nearly match the ones obtained from *Drees and Kaufmann* (1998) (denoted by ds). This further indicates the efficiency of the proposed Algorithm 5 in the joint estimation of k_0 and k.

n	L	Frech(5)		Frech(2)		Frech(1)		|T|(4)		|T|(10)	
		1	$\hat{\rho}^{(1)}$	1	$\hat{\rho}^{(1)}$	1	$\hat{\rho}^{(1)}$	1	$\hat{\rho}^{(1)}$	1	$\hat{\rho}^{(1)}$
100	5	1.16	1.26	1.80	2.34	1.93	2.04	1.08	1.26	1.06	1.25
	20	1.15	1.28	1.07	1.22	1.41	1.35	1.03	1.16	1.05	1.14
	ds	1.29	1.22	1.08	1.24	1.28	1.12	1.36	1.15	1.24	1.48
200	5	1.18	1.25	1.65	1.40	1.18	1.44	1.07	1.23	1.08	1.25
	20	1.16	1.19	1.08	1.30	1.16	1.46	1.06	1.23	1.09	1.20
	ds	1.19	1.21	1.08	1.23	1.34	1.14	1.28	1.14	1.28	1.46
500	5	1.14	1.29	1.23	1.38	1.18	1.48	1.10	1.27	1.12	1.20
	20	1.12	1.26	1.07	1.40	1.16	1.51	1.11	1.12	1.26	1.21
	ds	1.12	1.18	1.05	1.26	1.30	1.12	1.27	1.14	1.3	1.41

Table 3.7: Ratio of mean squared errors: adaptive trimmed Hill to trimmed Hill for $k_0 = 10$.

3.6 Comparisons With Existing Estimators And Adaptivity

3.6.1 Comparison with other robust estimators

In this section, we present a comparative analysis of the performance of our proposed trimmed Hill estimator, $\hat{\xi}_{k_0,k}$ with respect to the already existing robust tail estimation procedures in the literature. For observations from the Pareto distribution, a robust estimator of α based on the trimmed Hill estimator, $\hat{\xi}_{\mathrm{k_0,n-1}}$ is given by

$$\hat{\alpha}_{\mathrm{TRIM}} = \left(1 - \frac{2}{n}\right)\frac{1}{\hat{\xi}_{k_0,n-1}} \tag{3.37}$$

where $(1 - 2/n)$ is the correction factor for $\hat{\alpha}_{\mathrm{MLE}}$ as in *Brzezinski* (2016).

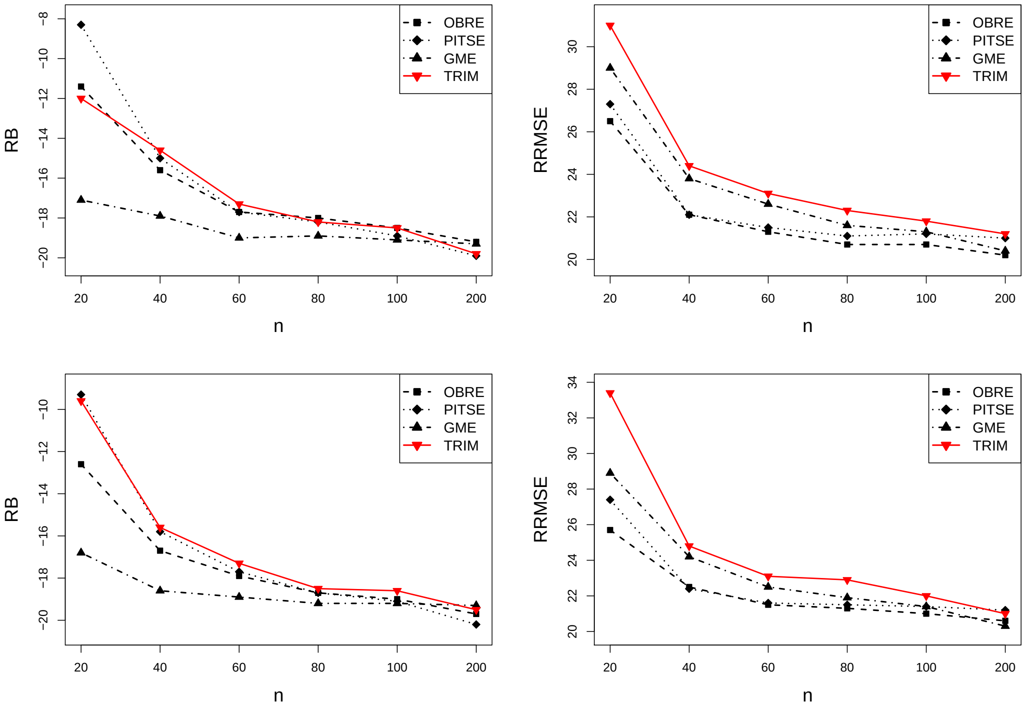


Figure 3.4: Performance of robust estimators for $0.9P(\alpha, 1) + 0.1P(\alpha, 1000)$ at ARE=78%. Top left and right correspond to RB and RRMSE values for $\alpha = 1$. Bottom left and right correspond to RB and RRMSE values for $\alpha = 3$.

The comprehensive comparative analysis in *Brzezinski* (2016) evaluates many robust estimators of the exponent $\alpha = 1/\xi$ with respect to the maximum likelihood estimator $\hat{\alpha}_{\mathrm{MLE}}$ for i.i.d. Pareto observations. The class of estimators used in *Brzezinski* (2016) include the optimal B-robust estimator, (OBRE) proposed in *Victoria-Feser and Ronchetti* (1994), the weighted maximum likelihood estimator (WMLE) introduced in *Dupuis and Victoria-Feser* (2006), the generalized median estimator (GME) of *Brazauskas and Serfling* (2000), the partial density component estimator (PDCE) proposed in *Vandewalle et al.* (2007) and the probability integral transform statistic estimator (PITSE) of *Finkelstein et al.* (2006). Among these estimators of α, the OBRE, PITSE and GME exhibit a superior performance in comparison to the rest and shall be used as the comparative baseline.

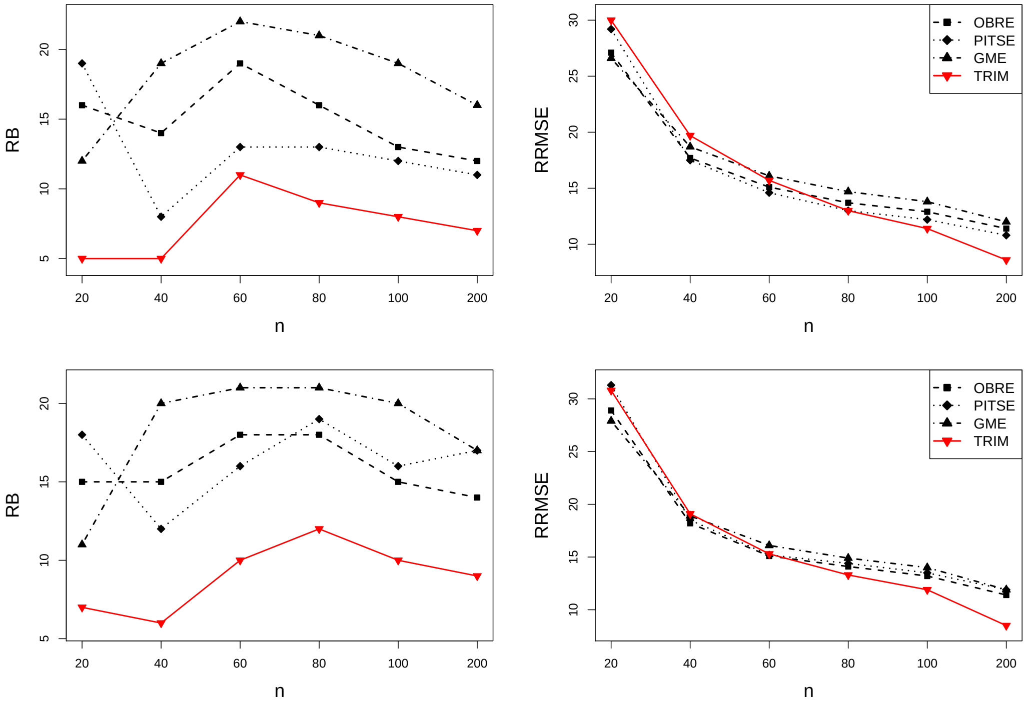


Figure 3.5: Performance of robust estimators where 5% observations of $P(\alpha, 1)$ are inflated by 10 at ARE=78%. Top left and right correspond to RB and RRMSE values for $\alpha = 1$. Bottom left and right correspond to RB and RRMSE values for $\alpha = 3$.

The comparison criterion chosen is the relative bias, RB and relative mean squared error, RRMSE as in *Brzezinski* (2016). The explicit formulas for RB and RRMSE are given by

$$\begin{aligned} \mathrm{RB}(\alpha) &= \frac{1}{\alpha}\Big(\frac{1}{m}\sum_{i=1}^{m}(\widehat{\alpha}_i - \alpha)\Big) \times 100\% \qquad (3.38) \\ \mathrm{RRMSE}(\alpha) &= \frac{1}{\alpha}\Big(\frac{1}{m}\sum_{i=1}^{m}(\widehat{\alpha}_i - \alpha)^2\Big)^{1/2} \times 100\% \end{aligned}$$

where the $\widehat{\alpha}_i$'s are independent realizations of a particular estimator of $\alpha = 1/\xi$.

To be able to compare with *Brzezinski* (2016), we need to determine k_0 in (3.37) so as to match the target ARE (Asymptotic Relative Efficiency) of the estimators considered therein. By relation (3.7) in Proposition III.2 it is easy to see that

$$\text{ARE}(\hat{\alpha}_{\text{TRIM}}) = \frac{\text{Var}(\hat{\alpha}_{\text{MLE}})}{\text{Var}(\hat{\alpha}_{\text{TRIM}})} \approx \frac{1/n}{1/(n-1-k_0)} \tag{3.39}$$

where the last asymptotic equivalence follows by a simple application of delta method to the function form of $\hat{\alpha}_{\text{TRIM}}$ in terms of the statistic $\hat{\xi}_{k_0,n-1}$. Given n, to achieve a target ARE, we use (3.39) to solve for k_0.

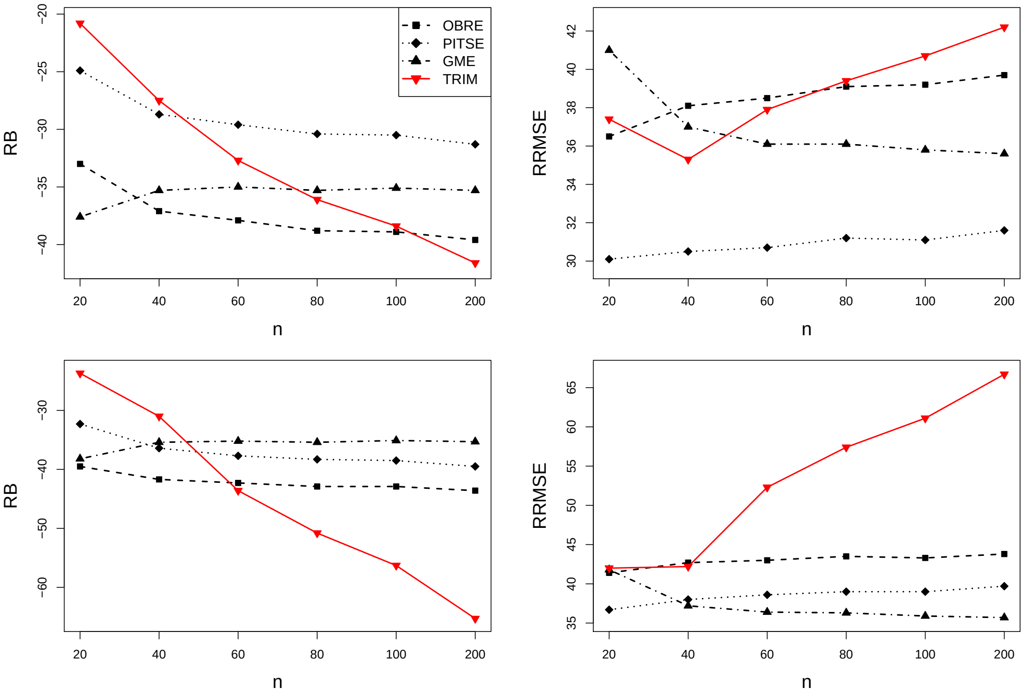


Figure 3.6: Performance of robust estimators for $0.9P(\alpha, 1) + 0.1P(\alpha, 1000)$ at ARE=94%. Top left and right correspond to RB and RRMSE values for $\alpha = 1$. Bottom left and right correspond to RB and RRMSE values for $\alpha = 3$.

As in *Brzezinski* (2016), the data sets are simulated from the Pareto distribution Pareto$(1, 1)$ and contaminated in two ways. In the first method of introducing outliers, we generate observations from the following mixture distribution

$$F = (1-\varepsilon)\ \text{Pareto}(\alpha, 1) + \varepsilon\ \text{Pareto}(\alpha, 1000) \tag{3.40}$$

for $\varepsilon \in (0,1)$ and $\alpha > 0$. In the second method of contamination, s proportion of the observations is randomly selected from Pareto$(\alpha, 1)$ and multiplied by a constant factor of 10.

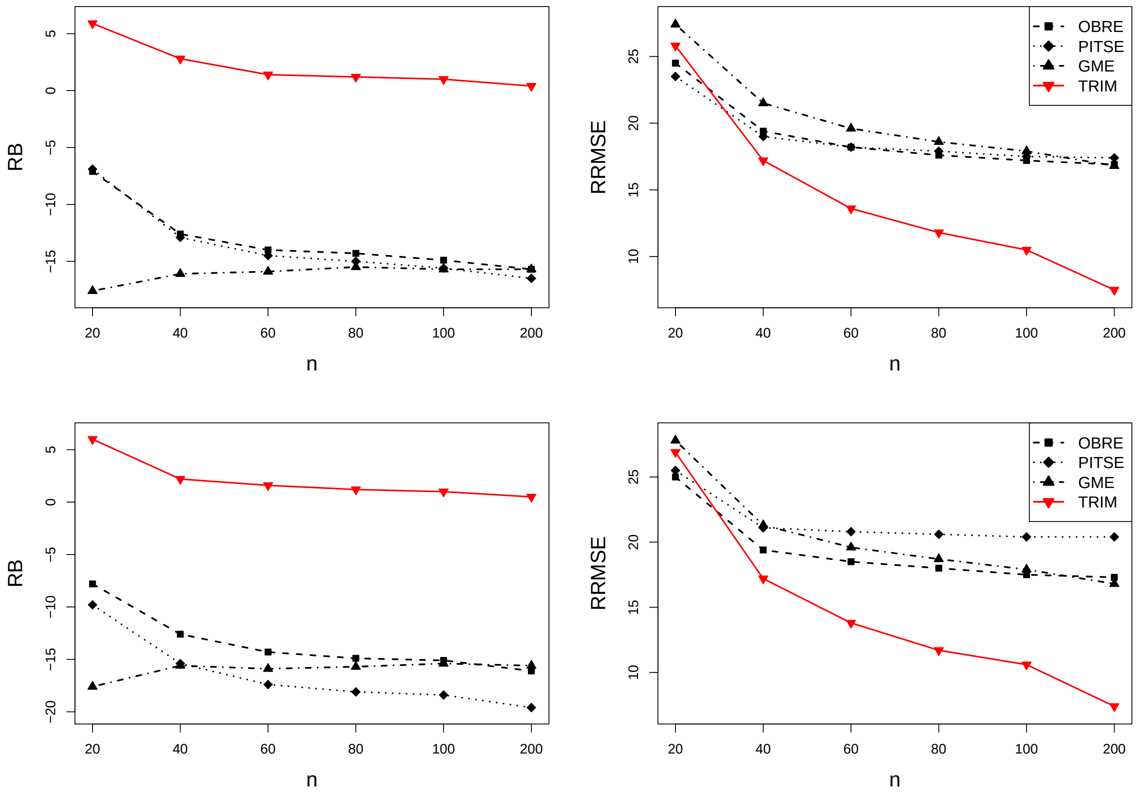


Figure 3.7: Performance of robust estimators for $0.95P(\alpha, 1) + 0.05P(\alpha, 1000)$ at ARE= 94%. Top left and right correspond to RB and RRMSE values for $\alpha = 1$. Bottom left and right correspond to RB and RRMSE values for $\alpha = 3$.

For both methods of data contamination, we analyze performance of the four estimators viz OBRE, PITSE, GME and TRIM. We first fix the asymptotic relative efficiency for these estimators at 78%. Figure 3.4 shows the performance under the first method of data contamination with $\varepsilon = 0.9$ and $\alpha = 1$ and 3. We observe that the performance of $\hat{\alpha}_{\text{TRIM}}$ closely follows that of $\hat{\alpha}_{\text{OBRE}}$, $\hat{\alpha}_{\text{PITSE}}$ and $\hat{\alpha}_{\text{GME}}$. In fact, all the estimators are relatively similar in this case and their difference is relatively small as the sample size n grows. Figure 3.5 on the other hand, shows the performance under the second method of data contamination with $s = 0.05$ and $\alpha = 1$ and 3. For this case, we observe the superior performance of $\hat{\alpha}_{\text{TRIM}}$ in comparison to the estimators. This behavior is more apparent in larger sample sizes ($n = 200$) where the trimmed estimator has more than 50% lower RRMSE values than the rest.

We next fix the asymptotic relative efficiency for these estimators at 94%. Figure 3.6 shows the performance under the first method of data contamination with $\varepsilon = 0.9$ and $\alpha = 1$ and 3. We observe that in this case the performance of $\hat{\alpha}_{\text{TRIM}}$ is relatively poor when compared to that of $\hat{\alpha}_{\text{OBRE}}$, $\hat{\alpha}_{\text{PITSE}}$ and $\hat{\alpha}_{\text{GME}}$ especially for larger sample sizes, n. However this phenomenon gets entirely reversed when $\varepsilon = 0.95$ (see Figure 3.7). The performance of $\hat{\alpha}_{\text{TRIM}}$ improves drastically with increase in sample size n and surpasses the performance of all the other robust estimators. For $n = 200$, the improvement is up to a factor 200% in the RRMSE values. The surprising difference in the performance observed in Figures 3.6 and 3.9 can be explained as follows.

Since the ARE of $\hat{\alpha}_{\text{TRIM}}$ is directly related to the trimming value k_0 (see (3.37)), large ARE or small k_0 values can control against small proportion of contamination ($1-\varepsilon = 0.05$) but not against large proportions ($1-\varepsilon = 0.1$). In scenario of Figure 3.6, setting the ARE as 94% and contaminating 10% of the data, our trimmed estimator is artificially forced to include outliers. This leads to the relatively poor performance of $\hat{\alpha}_{\text{TRIM}}$. For other estimators, the link between ARE and robustness is not as direct which gives them an advantage. At 5% contamination, our trimmed estimator picks up all the outliers at ARE level 94% and hence outperforms the competitors (Figure 3.7).

In the following section, we illustrate an important advantage of our trimmed estimator when k_0 is estimated from the data. This allows us to adapt the degree of robustness to the proportion of outliers.

3.6.2 Adaptive robustness

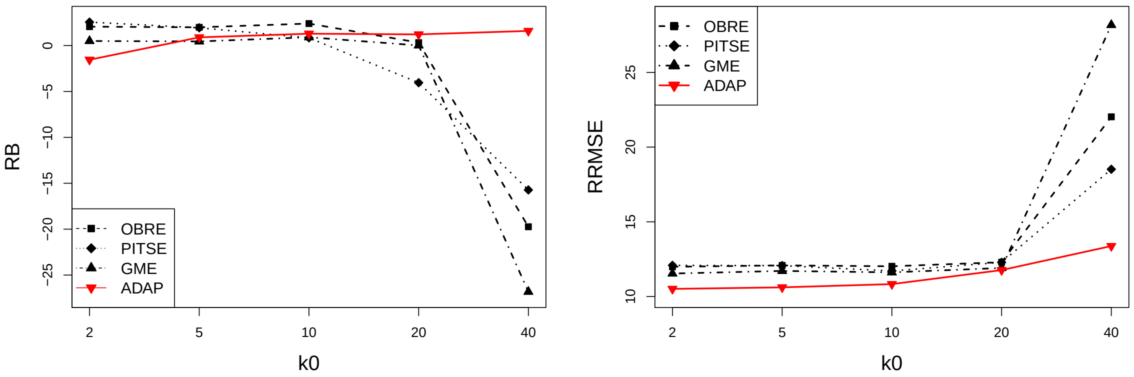


Figure 3.8: Performance of robust estimators at ARE=78%. Top left and right correspond to RB and RRMSE values.

In this section, we describe the superior performance of the adaptive trimmed Hill estimator (ADAP), $\widehat{\xi}_{\hat{k}_0,k}$, relative to several well known existing estimators when the degree of contamination is unknown. The performance of these existing robust estimators depends on the choice of parameters, which is directly related to their asymptotic relative efficiency.

For example, the optimal B-robust estimator (OBRE) requires a suitable choice of the parameter c (see *Victoria-Feser and Ronchetti* (1994)) and the probability integral transform estimator (PITSE) requires a suitable choice of the parameter t (see *Finkelstein et al.* (2006))in order to allow for a given degree of robustness. Unless the degree of contamination is pre specified, it is impossible to accurately determine these parameters, which control the degree of robustness. Our estimator, on the other hand is adaptive in nature and automatically picks the trimming parameter, thereby producing a estimator of the tail index which can adapt to potentially unknown degree of contamination of the top order statistics.

We demonstrate the adaptive property of the proposed estimator, ADAP for the Pareto model where the outliers are injected as in (3.31). For comparative purposes, we use the three best robust estimators, OBRE, PITSE and GME from *Brzezinski* (2016) also described in Section 3.6.1. The comparison is made in terms of RRMSE and RB values as in (3.38). As in Section 3.6.1, we calibrate the parameters of the competing estimators by setting the ARE to be 78% or 94%.

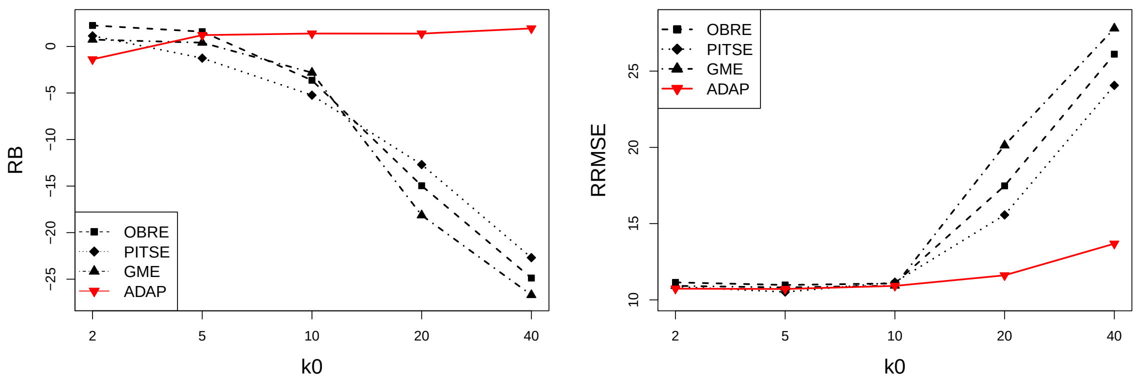


Figure 3.9: Performance of robust estimators at ARE=94%. Top left and right correspond to RB and RRMSE values.

Figure 3.8 demonstrates the performance of ADAP against the three competitors at ARE=78%. Observe that the competitors fail to adapt to the growing degree of contamination and essentially break down at $k_0/n = 40\%$. On the other hand, apart from a mild loss in efficiency, our estimator is resilient to the degree of contamination and adapts itself even to higher values of k_0/n. This feature is even more prominent in Figure 3.9 where the ARE for all estimators is fixed at 94%. Even at contamination proportion as low as 10%, ADAP outperforms all the competitors. This is expected since the performance of the competitors is sensitive to the choice of ARE. Large ARE values (94%) allow for a smaller degree of robustness, hence the poor performance of the OBRE, PITSE and GME even at lower contamination levels. To the best of our knowledge, the remarkable adaptive robustness property inherent to our estimator is not present in any other estimator in the literature.

3.7 Discussion

In this chapter, we introduced the trimmed Hill estimator for the heavy-tail exponent ξ. We established its finite-sample optimality in the ideal Pareto setting and its asymptotic normality under a second order regular variation condition. In Section 3.3.3, we established a uniform consistency result for the trimmed Hill estimator. For the Hall class of distributions, we argued that the trimmed Hill estimator attains the same minimax optimal rate as in the case of no outliers, provided that $k_0 = o(n^{2\rho/(2\rho+1)})$, where $\rho > 0$ is the second order regular variation exponent. One open problem is to establish the minimax optimal rate of the trimmed Hill estimator, in the case when the rate of contamination k_0 exceeds the minimax optimal rate.

In Section 3.5.5, we develop a methodology for the joint selection of the parameters k_0 and k, based on the work of Drees and Kaufman *Drees and Kaufmann* (1998). We formulate an extension of their results when $k_0 = o(n^{2\rho/(2\rho+1)})$. This leads to a practical method for the joint selection of k_0 and k. This method is shown to work as well as the original method of Drees and Kaufman even if the top order statistics are contaminated. As in the case of uncontaminated extremes, however, the main challenge is the accurate estimation of the second order exponent ρ. In the future, perhaps other bootstrap-based methods for the joint estimation of k_0 and k should be explored as in *Danielsson et al.* (2001).

Our key methodological contribution is the data–driven selection of the trimming parameter k_0 using weighted sequential testing. It leads to a robust estimator that adapts to the potentially unknown degree of contamination in the extremes. This unique feature is not available in many other robust estimators, which require the selection of tuning parameters. As demonstrated in Section 3.6.2, the adaptive trimmed Hill estimator has superior performance with practically no tuning. As an added bonus, we obtain a method for the identification of suspect outliers in the extremes of the data, which can be used to perform forensics or detect anomalies *Kallitsis et al.* (2016a).

Finally, we would like to advocate broadly for using robust methods for the estimation of the tail index. Our experience with extensive simulation studies (see e.g., Tables 3.2 and 3.3) convinced us that contamination in small proportion of the extreme order statistics leads to severe bias in the non-robust estimation methods. Trimming and especially data–adaptive trimming provide good alternatives at the expense of little to no loss in efficiency in the case when no contamination is present.

Acknowledgments: This work is supported by National Science Foundation (NSF) grant CNS-1422078 and partially by the NSF grant DMS-1462368.

CHAPTER IV

Extremes Of The Spatial Impact Of Heat waves

4.1 Introduction

Heat waves are becoming more common, especially in the U.S. West, although some of the largest heat waves were recorded during the time period 1930s (caused by the Dust Bowl and other factors). Extreme heat can increase the risk of different types of disasters[1]. *Meehl and Tebaldi* (2004) show based on the severe heat waves in Chicago in 1995 and Paris in 2003 that future heat waves in Europe and North America will become more intense, more frequent, and longer lasting in the second half of the 21st century.

Heat waves are often described by events when the daily maximum temperature remains above a given threshold for a span of Δ consecutive days. This notion of heat waves has been widely used in the papers of *Croitoru et al.* (2016), *Ouzeau et al.* (2016), *Capozzi and Budillon* (2017) and the references therein. Other representatives of heat waves like those based on the Heat Wave Magnitude Index daily (HWMId) has been used in papers like *Ceccherini et al.* (2017). In this chapter, we expand on this definition in *Croitoru et al.* (2016) by adding a spatial context to it. We

[1] `https://www.c2es.org/content/heat-waves-and-climate-change`

study the spatial distribution of the heat waves across the continental US, similar to *Dian-Xiu et al.* (2014) and suggest a unified measure to describe this distribution. This measure aims at computing the proportion of US area under unusually large heat waves at a given point in the year (see (4.9)). The chapter aims at modeling the extremes of the time series thus obtained. Previously, the approach of fitting a Generalized Extreme Value distribution to climatic extremes has been explored in the works of *Jonathan et al.* (2018) and *Huang* (2017). Another approach based on Generalized Pareto Distribution (GPD) for modeling peaks over threshold in climatic events has been explored under *Davison et al.* (2012) and *Davison and Smith* (1990). We shall adopt the later approach in the context of extremes in spatial distribution of heat waves.

Model diagnostic *Coles* (2001) plot reveal that the dependence of only the scale and not the shape of the GPD through covariates well explains the distribution of extremes. The impact of El Niño Southern Oscillation Index[2] (ENSO) on extreme temperature events has been well studied under *McKinnon et al.* (2016) and *Winter et al.* (2016). Inspired by these works, ENSO was added as covariate in the GPD modeling which turned out to be a significant one for the scale parameter. Seasonal and cyclic patterns in the time series of areal extremes is almost inevitable, all of which have been accounted for the covariate described by 365-periodic splines. The overwhelming impact of seasonality motivated us to perform the global analysis on each season individually. The choice of the seasons is based on the work of *Cressie and Kang* (2016). Interestingly enough, depending on the season under study one or more of the covariates may lose their significance in terms of the model fit.

The rest of the chapter is organized as follows. In Section 4.2, we describe the

[2] https://www.ncdc.noaa.gov/teleconnections/enso/indicators/soi/

data that has been used for the evaluation of our proposed methodology. In Section 4.3, we discuss techniques for obtaining a stationary signal for the time series of daily temperatures. Additionally, spatial heat waves and modeling of their block maxima or peak over threshold have been discussed. In Section 4.4, we discuss the GPD model and how its parameters vary as a function of known covariates. Effectiveness of the chosen model has been described in terms of the AIC criterion and model diagnostic plots. In Section 4.5.3, we finally describe the distribution of heat waves with respect to the season and location across US. Section 4.6 concludes all the results by discussing the caveats and roles of varying parametric assumptions in the development of the methodology.

4.2 Data Description

Since 1987, the National Oceanic and Atmospheric Administration's (NOAA) National Centers for Environmental Information (NCEI-NC) has used observations from the U.S. Historical Climatology Network (USHCN) to quantify national- and regional-scale temperature changes in the conterminous United States (CONUS). The USHCN is actually a designated subset of the NOAA Cooperative Observer Program (COOP). The USHCN sites having been selected according to their spatial coverage, record length, data completeness, and historical stability. The first development of USHCN datasets were at NOAA's NCEI in collaboration with the Department of Energy's Carbon Dioxide Information Analysis Center (CDIAC) in a project that dates to the mid-1980s (Quinlan et al. 1987).

Since then, the USHCN dataset has been revised several times (e.g., Karl et al., 1990; Easterling et al., 1996; Menne et al. 2009). The three dataset releases described in Quinlan et al. 1987, Karl et al., 1990 and Easterling et al., 1996 are now referred to as the USHCN version 1 datasets. These version 1 datasets contained adjustments to the monthly mean maximum, minimum, and average temperature data that addressed potential changes in biases (inhomogeneities) in data from USHCN stations documented in NCEIs station history archives. In 2007, USHCN version 2 serial monthly temperature data were released and updates to the version 1 datasets were discontinued. In October 2012, a revision to the version 2.0 dataset was released as version 2.5.

This chapter uses version 2.5 of the data set with daily record maximum and minimum temperature data from 424 U.S. Historical Climatology Network stations meeting stringent data completeness requirements. These 424 stations represent the 2nd phase of defining a subset of the 1218-station U.S. Historical Climatological Network that enables robust assessment of trends in record-setting daily Tmax and Tmin temperatures since 1911 (i.e., a century-scale period of record). Many USHCN stations do not allow for such an assessment due to significant amounts of missing data early in their periods of record.

This 424-station subset includes the initial 200-station subset[3] that was identified using especially stringent missing data requirements[4]. The additional 224 stations have slightly less stringent missing data requirements, but the average percentage of missing data across all 424 stations is still just 2.4%. A graphical user interface for exploring these station data is available online[5]. Also, an inventory file containing metadata for the 424 stations is available through the USHCN website[6].

[3] http://cdiac.ornl.gov/ftp/us_recordtemps/sta200/

[4] /http://cdiac.ornl.gov/climate/temp/us_recordtemps/dayrec.html

[5] http://cdiac.ornl.gov/climate/temp/us_recordtemps/ui.html

[6] http://cdiac.ornl.gov/epubs/ndp/ushcn/daily_doc.html#stations

4.3 Preliminary Analysis

Let $X_t(s)$ be the daily temperature measurement on day t at site s. We have observations at a collection of sites $s_1, \ldots, s_\ell \in \mathcal{D}$ in a region $\mathcal{D}$ over a period of T days. In this chapter, we study the spatial impact of extreme heat-wave events over the entire region $\mathcal{D}$.

4.3.1 Standardization of the data

It is commonly understood that at site s, one experiences a heatwave event if the temperatures are unusually high (relative to the average seasonal temperature) for a prolonged period of time. Depending on the temperature amplitude and temporal duration of the event one encounters different types of heatwaves. We consider heatwave events identified by Δ consecutive days of daily maximum temperatures exceeding a level u_0. To be able to account for trend in temperature patterns (be it seasonal or global), we need to consider a threshold u_0 that varies slowly with the season. Alternatively, one can consider deviations from the mean temperature curve of a station with respect to its standard devation curve.

Specifically, let X_t be the daily temperature maxima over a period of $\Delta = 7$ consecutive days starting on day k. Several sophisticated techniques, both parametric *Ramsay et al.* (2005) and non parametric *Ferraty and Vieu* (2006) techniques exist for modeling of functional curves, we rather use the one motivated from Section 3.5 of *Ramsay et al.* (2005). We assume the following representation for the mean of the time series $X_t(s)$:

$$E(X_t) := \mu(t) = \gamma^{(\mu)} + \underbrace{\sum_{i=1}^{m_1^{(\mu)}} \alpha_i^{(\mu)} \phi_i^{(\mu)}(t)}_{A^{(\mu)}(t)} + \underbrace{\sum_{j=1}^{m_2^{(\mu)}} \beta_j^{(\mu)} \omega_j^{(\mu)}(t)}_{B^{(\mu)}(t)} \tag{4.1}$$

where $\gamma^{(\mu)}$ is a constant term and $\alpha_i^{(\mu)}$ and $\beta_j^{(\mu)}$ are the coefficients corresponding to the covariates of seasonal fluctuations and global trends respectively.

The functions $\phi_i^{(\mu)}, i = 1, \cdots, m_1^{(\mu)}$ denote a 365- periodic cubic[7] spline basis representation[8] of the range $1, 2, \cdots, 365T$ with $T = 100$ with $m = m_1^{(\mu)}$ degrees of freedom. Figure 4.1 left panel displays a plot of 365-periodic cubic splines with 3 degrees of freedom when evaluated on the time span 1986-1990. The quantity $A^{(\mu)}(t)$ thus captures the portion of the mean curve that may be attributed to seasonal variations.

On the other hand, the functions $\omega_i^{(\mu)}, i = 1, \cdots, m_2^{(\mu)}$ denote a cubic spline basis representation[9] of the range $1, \cdots, 365T$. Figure 4.1 middle panel displays a plot of cubic splines with 3 degrees of freedom when evaluated on the time span 1911-2010. $B^{(\mu)}(t)$ thus captures the portion of the mean curve that may be attributed to slowly varying trends over time like global warming, shifting weather conditions etc.

The residuals: $\tilde{X}_t = X_t - \hat{\mu}(t)$ obtained from the linear regression model in (4.1) provides a way for the estimation of the standard error curve. We propose the fol-

[7] *Knott* (2000)

[8] Since seasonal patterns are expected to repeat themselves after a span of 365 days, 365-periodic spline functions with m degrees of freedom are constructed as in *Bojanov et al.* (1993):

$$\phi_i(a) = \phi(a + 365), \quad a = 1, \cdots, 365, i = 1, \cdots, m.$$

The **pbs** function under R package *pbs* (2013) automates this process periodic spline generation.

[9] For m degrees of freedom, the spline basis functions $\omega_j, j = 1, \cdots, m$ are constructed as in *Bojanov et al.* (1993) to describe the yearly trends and phenomenon like global warming, shifting weather conditions etc. The **bs** function under R package *splines* (2000) automates this process the spline generation.

lowing representation of variance of time series $\tilde{X}_t$.

$$E(\log|\widetilde{X}_t|) :\approx \log(c\eta(t)) = \gamma^{(\eta)} + \underbrace{\sum_{i=1}^{m_1^{(\eta)}} \alpha_i^{(\eta)}\phi_i^{(\eta)}(t)}_{A^{(\eta)}(t)} + \underbrace{\sum_{j=1}^{m_2^{(\eta)}} \beta_j\omega_j^{(\eta)}(t)}_{B^{(\eta)}(t)} \tag{4.2}$$

where $\gamma^{(\eta)}$ is a constant term and $\alpha_i^{(\eta)}$ and $\beta_j^{(\eta)}$ are the coefficients corresponding to the covariates of seasonal fluctuations and global trends respectively. A normal approximation to the time series $\tilde{X}_t$ provides a choice for the parameter $c = 0.5298$. Similar to the model in (4.1), the functions $\phi^{(\eta)}(t)$ are 365-periodic cubic splines and $\omega^{(\eta)}(t)$ are ordinary cubic splines for the basis representation of the range $1, 2, \cdots, 365T$ with $m = m_1^{(\eta)}$ and $m = m_2^{(\eta)}$ degrees of freedom respectively. Thus, the quantities $A^{(\eta)}(t)$ and $B^{(\eta)}(t)$ denote the portion of the standard deviation explained by seasonality and global trend respectively.

For known values of $m_1^{(\mu)}$, $m_2^{(\mu)}$, $m_1^{(\eta)}$ and $m_2^{(\eta)}$, the unknown coefficients in models (4.1) and (4.2) may be easily estimated using a simple linear regression approach.

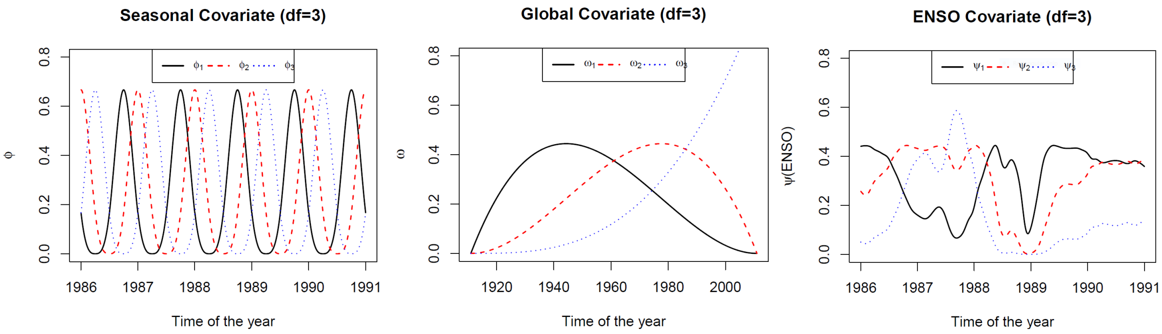
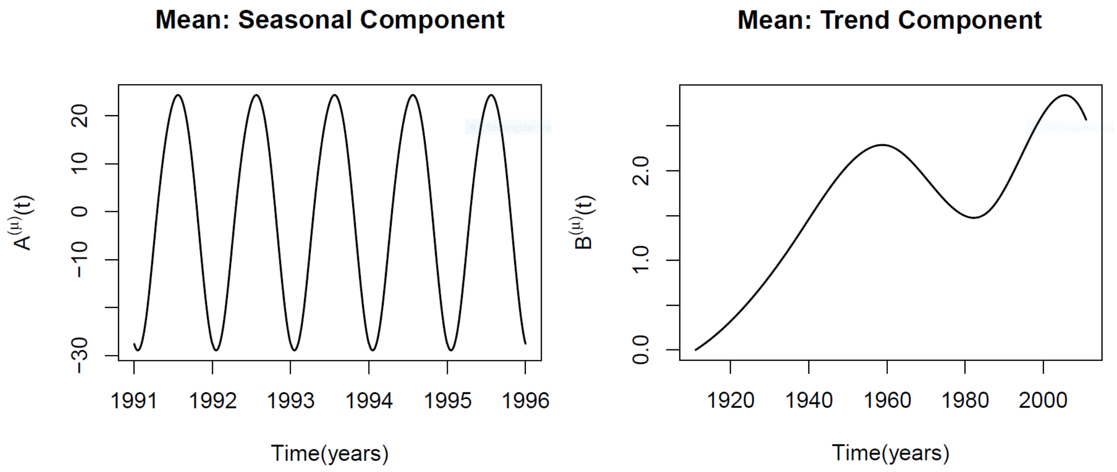


Figure 4.1: Spline basis functions evaluated. Left panel corresponds to the periodic splines for seasonal activity, middle panel corresponds to splines for global activity and right panel corresponds to splines for ENSO activity

Exploratory analysis revealed little sensitivity to the choice of the parameters $m_1^{(\mu)}$, $m_2^{(\mu)}$, $m_1^{(\eta)}$ and $m_2^{(\eta)}$. We however added an additional criterion to guard against large choices of these quantities to en certain that meaningful information does not get lost from the time series in the process of following the mean or standard deviation too closely. We next explain the simultaneous choice of $m_1^{(\mu)}$, $m_2^{(\mu)}$.

$$(m_1^*, m_2^*) = \underset{\{m_1 \le \kappa_1, m_2 \le \kappa_2\}}{\operatorname{argmin}} \quad n \log(RSS/n) + 2(m_1 + m_2) \tag{4.3}$$

where RSS is the residual sum of squares obtained from the linear regression fit in (4.1) with $m_1^{(\mu)} = m_1$ and $m_2^{(\mu)} = m_2$. A choice of thresholds as low as $\kappa_1 = 6$ and $\kappa_2 = 6$ produced standardized z scores which were fairly stationary (see right panel in Figure 4.3) and seemed to provide a promising choice for the remaining analysis. The quantities, $m_1^{(\eta)}$ and $m_2^{(\eta)}$ are also estimated according to the criterion in (4.3) but for the linear model in (4.2). Section 4.6 further sheds some light on the choice of the parameters κ_1 and κ_2.



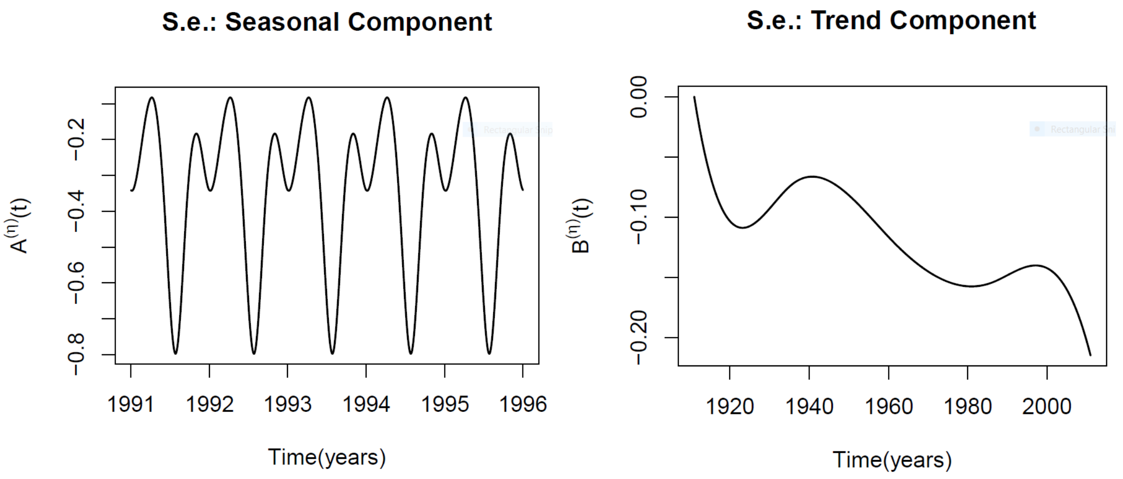


Figure 4.2: Seasonal and trend components for historical temperature mean (top panel) and standard error (bottom panel) curves over the period 1911-2010 for Ann Arbor, MI.

Figure 4.2 top panel illustrates the seasonal and trend components, $A^{(\mu)}(t)$ and $B^{(\mu)}(t)$ of the mean curve $\mu(t)$ respectively. The bottom panel in 4.2 illustrates seasonal and trend components, $A^{(\eta)}(t)$ and $B^{(\eta)}(t)$ of the standard error curve $\eta(t)$. These reported graphs are obtained from the historical daily temperature records over period $1911 - 2010$ at the monitoring station of Ann Arbor in continental US. The degrees of freedom $m_1^{(\mu)}$, $m_2^{(\mu)}$, $m_1^{(\eta)}$, $m_2^{(\eta)}$ are chosen according to the criterion in (4.3) with $\kappa_1 = 6$ and $\kappa_2 = 6$.

4.3.2 Distributional properties of the standardized time series

Our goal is to examine the extreme fluctuations of $X_t(s)$ relative to the estimated mean and standard deviation. To this end, we consider the time series

$$Y_t(s) := \frac{X_t(s) - \hat{\mu}(t)(s)}{\hat{\eta}(t)(s)}, \ t = 1, 2, \ldots, \tag{4.4}$$

where $\hat{\mu}(t)$ and $\hat{\eta}(t)$ are obtained as in Section 4.3.1 for the station s. The time series $\{Y_t(s)\}$ is standardized to have zero mean and unit variance marginal distributions.

Figure 4.3 left shows a plot of the standardized time series and also the auto covariance plot for the station, Ann Arbor. The auto-covariance plots reveal that the standardized time series is fairly stationary and can be suitable for analysis.

We explore next the empirical distribution of the standardized time series $\{Y_t(s)\}$. It turns out that the standard Normal model offers a fairly adequate approximation to the time series for most of the stations s. Indeed, in Figure 4.4, we show normal quantile-quantile plots for the empirical distribution of $\{Y_t(s)\}$ for two stations, Faulkton North West, SD and Pasadena, CA. Qualitatively, the QQ-plots for all other stations look nearly identical. While the standardization does not remove periodic dependencies and non-stationarity, it puts the temperature fluctuations in different seasons across different stations on the same scale.

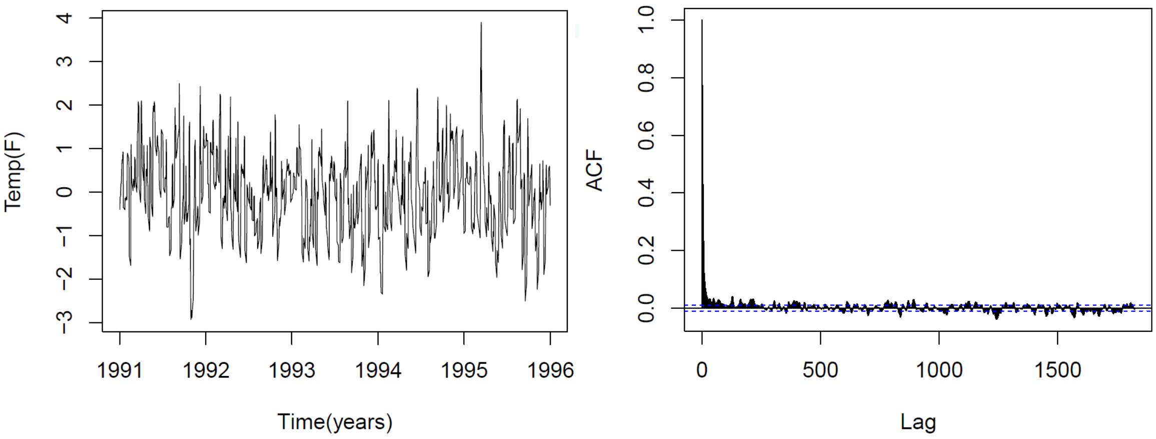


Figure 4.3: Standardized daily time series for Ann Arbor, MI. Left and right panels indicate the standardized time series, $Y_t(s)$ and its corresponding auto-covariance function.

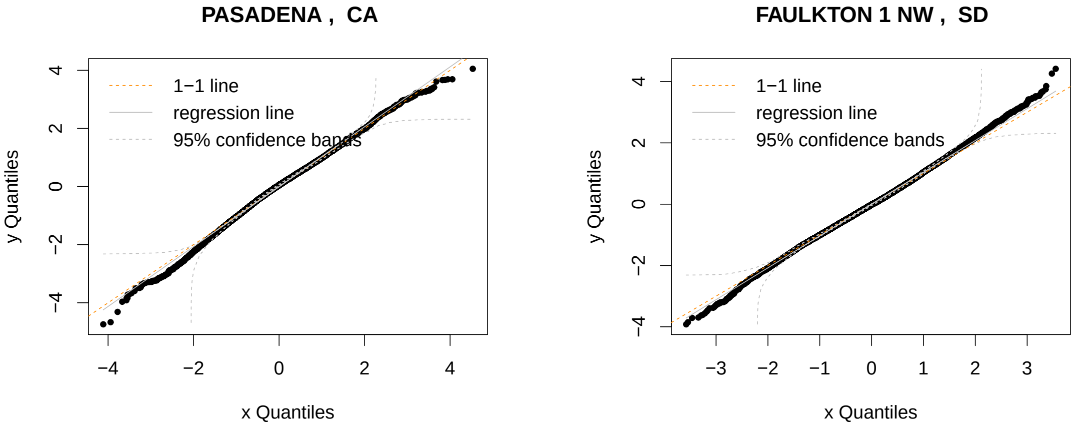


Figure 4.4: Normal quantile-quantile plots for standardized weekly min $Y_t(s)$. Left panel and right panels correspond to stations Faulkton North West, SD and Pasadena, CA respectively.

4.3.3 Defining heat waves

In order to define heat waves, we first define the minimum of the standardized time series $Y_t(s)$ over a span of $\Delta = 7$ days as:

$$Z_{k,\Delta}(s) = \min_{t=k,\ldots,k+\Delta-1} Y_t(s) \tag{4.5}$$

We have a heatwave event at location s starting on day k if

$$Z_{k,\Delta}(s) > U(u_0, s, \Delta) \tag{4.6}$$

where u_0 is the intensity level, s is the station and Δ is size of the window. With Δ fixed the heat wave is defined as: if

$$Z_k(s) > F_s^{-1}(u_0) \tag{4.7}$$

where u_0 is a quantile level and

$$F_s^{-1}(u_0) = \inf\{x \in \mathbb{R} : F_s(x) \geq u_0\} \tag{4.8}$$

with F_s denoting the empirical cumulative distribution function of $Z_k(s)$. Therefore, values of $Z_k(s)$ well above its extreme quantiles correspond to the occurrence of a heat wave event.

To explore the *spatial impact* of the so-defined heat waves, we define

$$A_k(u_0) := \int\limits_D I(Z_k(s) > F_s^{-1}(u_0))ds, \tag{4.9}$$

which is the total area of the sites s in region $\mathcal{D}$ experiencing a heatwave during week k. Since the area $A_k(u_0)$ is bounded above by the total area of the region $\mathcal{D}$, we consider

$$Q_k(u_0) := \frac{A_k(u_0)}{|\mathcal{D}|} \in [0,1],$$

which is the *proportion* of the region $\mathcal{D}$ experiencing a heatwave of intensity level u_0 during week k.

For the stations $s = s_1, s_2, \cdots, s_\ell \in \mathcal{D}$, the series $Z_k(s)$ and $F_s^{-1}(u_0)$ is deterministic. However, for the computation of the integral in (4.9), the series $Z_k(s)$ and $F_s^{-1}(u_0)$ needs to be evaluated at all values of $s \in D$. This is facilitated by using thin plate splines smoothing approach where a surface is fit to the values $Z_k(s_i), i = 1, \cdots, \ell$, with some error allowed at each $s_i, i = 1, \cdots, \ell$. At every iteration, a station is omitted from the estimation of the fitted surface and the mean error is found. This procedure is repeated over a range of values of the smoothing parameter and the value that minimizes the mean error is taken to give the optimum smoothing (also called minimizing the generalized cross validation criterion). Chapter 12 in *Wilson and Mair* (2004) and Section 2.4 in *Tait et al.* (2006) explain the thin spline interpolation when applied to rainfall data. The **tps** function in *fields* (2018) package of R automatically applies this methodology to produce an integral approximation to the quantity (4.9).

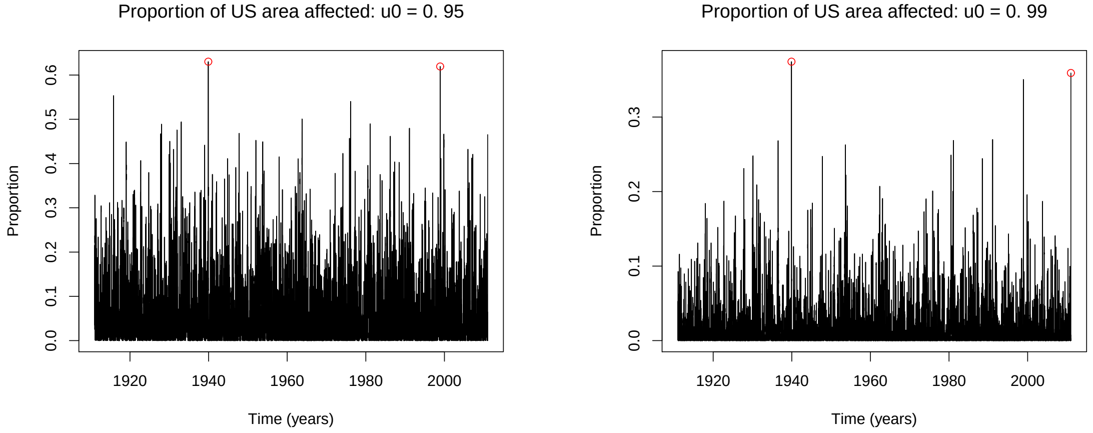


Figure 4.5: Distribution of the time series $Q_k(u_0)$ for varying values of the quantile level u_0 for the period 1911-2010. Extreme events in the series have been marked with red.

Figure 4.5 gives a plot of the time series $Q_k(u_0)$ for two different values of u_0 viz 0.95 and 0.99. As is evident from the plot, depending on value of the quantile u_0, the extremes of the time series Q_k goes on changing. For example, the second largest value for $Q_k(0.95)$ is recorded for December 31, 2010 in contrast to the second largest value for $Q_k(0.95)$ which is recorded on November 28, 1998. The largest value for both $Q_k(0.95)$ and $Q_k(0.99)$ is however on the same date, December 5, 1939. Figure 4.6 gives a plot of the spatially interpolated time series $Z_k(s)$ corresponding to the top record events in the time series $Q_k(u_0)$, $u_0 = 0.95, 0.99$. If one were to interpret these plots, the weeks starting on December 5, 1939 and November 28, 1998 experienced a heat wave of intensity 0.95 over 63% and 61% of the territory in US respectively.

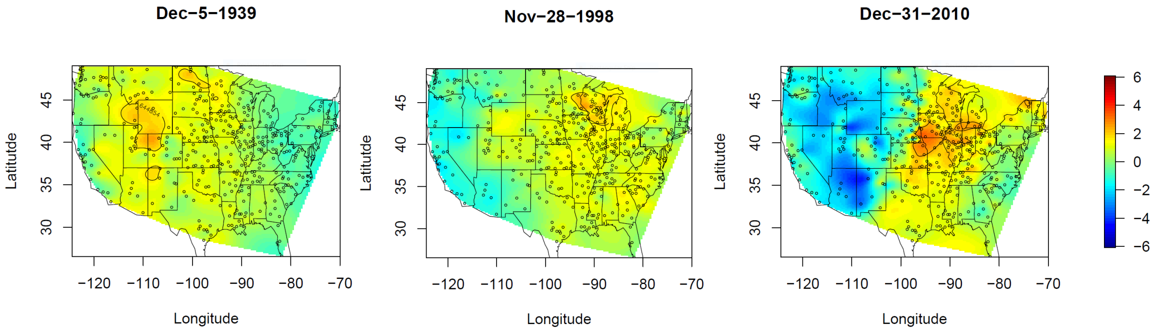


Figure 4.6: Thin spline interpolated time series $Z_k(s)$ for $s \in \mathcal{D}$ on time points corresponding to extreme values of $Q_k(s)$

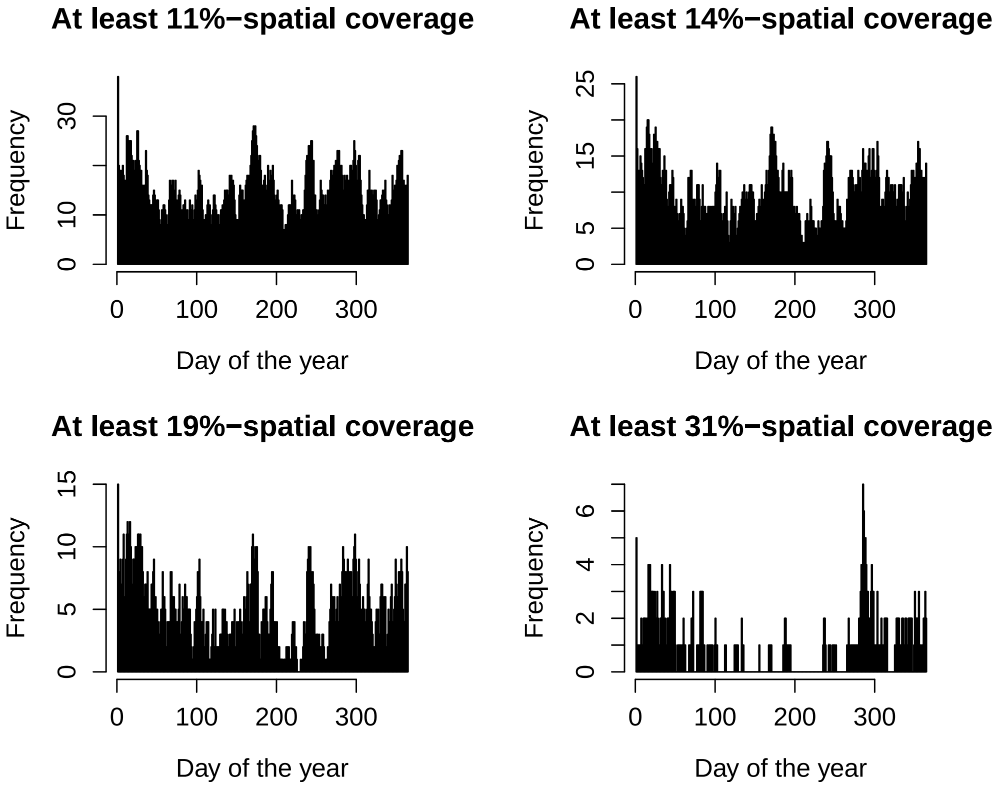


Figure 4.7: Histogram of the seasonal distribution of $Q_k > p$ for varying values of proportion p.

A natural question which arises is the seasonal distribution of the most extreme heat-waves is. In this direction, we consider histogram of

$$\{(k \bmod 365)|\mathrm{Q_k(u_0) > p}\}$$

for $u_0 = 0.95$. The histograms corresponds to nothing but the daily distribution of extreme events with at least $100p\%$ of spatial coverage . Figure 4.7 clearly demonstrates that with increase in proportion p, only the most extreme heat wave events present themselves. For an intensity level of 0.95, it seems spatially extreme heatwave events tend to occur less often in the summer than during the other seasons. This hypothesis is further corroborated under Sections 4.4.4 and 4.5.3.

4.3.4 Declustering

In order to estimate/predict return levels (see Section 4.4.4) of extremes in the time series Q_k, one employs tools from extreme value theory. Specifically, the Generalized Pareto Distribution, GPD (see (4.10)) model serves as a good fit to the peaks over a high threshold and may be used for the extrapolation and computation of out-of-sample tail probabilities (see *Beirlant et al.* (2006)). The accuracy of the approach however hinges to a great extent on the degree to which the excesses may be assumed to be independent.

Since the heatwave events are defined over overlapping windows, a substantial temporal dependence is likely to be present in the time series $\{Q_k(u_0)\}$. Enunciating further, a heat-wave event at site s of duration $d > \Delta = 7$ will trigger relatively large values of $Z_k(s)$ statistics in (4.4) for at least $d - \Delta$ consecutive values of k. This ultimately leads to clustering of extremes for the time series $\{Q_k\}$, which is certainly more difficult to characterize since it involves multiple sites. The latter assumption is well characterized by the extremal index θ of the time series (see Section 10.2.3 in *Beirlant et al.* (2004b)). If the extremal index $\theta \in [0,1]$ is significantly lower than 1, then the excesses tend to cluster with a mean cluster size approximately $1/\theta$.

A well established way to identify clusters of extremes is the method of runs. For a threshold q_0 and a runs parameter $r \geq 1$, a cluster of exceedance begins when Q_k exceeds q_0. The subsequent exceedances of level q_0 belong to the same cluster as long as there are no more than r consecutive time-points where Q_k falls below q_0. After seeing $r+1$ consecutive values of Q_k below q_0, another, separate cluster will commence the first time when Q_k exceeds q_0 again. Thus, a data set $Q_k,\ k = 1, \ldots, T$ is partitioned into clusters $C_i = \{Q_{k(i,1)}, \cdots, Q_{k(i,C_i)}\},\ i = 1, \ldots, n_C$, where C_i is the number of exceedances in the i-th cluster.

Under the assumption of stationarity and mild dependence conditions on the time series $\{Q_k\}$, we have that $E(C_1) \to 1/\theta$, as $q_0 \uparrow x_Q$, where x_Q is the upper end-point of the distribution of the Q_k'. This suggest the following estimator for θ:

$$\hat{\theta}_r := \Big(\frac{1}{n_C}\sum_{i=1}^{n_C} C_i\Big)^{-1}.$$

as in *Beirlant et al.* (2004a), *Smith* (1989). The C_i's and ultimately the estimators $\hat{\theta}_r$ depend on the choice of the threshold v_0 as well as on the runs parameter r. The runs method is somewhat sensitive to the choice of r but when an optimal value of that parameter is available, it has a superb performance. *Ledford and Tawn* (2003) provides some of the diagnostic tools for the choice of the tuning parameter r.

The Ferro-Segers estimator of θ in *Ferro and Segers* (2003) is more robust technique for de-clustering the exceedances in extremes. This estimator relies on the fact that, under appropriate normalization, the time between two consecutive clusters is asymptotically a mixture distribution from degenerate 0 and standard exponential. This de-clustering methodology does not depend on the choice of a tuning parameter. The number of clusters n_C is asymptotically equivalent to $\lfloor K\theta \rfloor$ where K is the total number of exceedances over the threshold q_0 and θ is the extreme value index (see *Beirlant et al.* (2004a)). Thus, Ferro and Segers suggest using inter-exceedance times that exceed the n_C-th order statistic of inter-exceedance times.

Let $T_{i_1}, T_{i_2}, \cdots, T_{i_{n_C}}$ denote these inter-exceedance times and $t_1 = T_{i_1}, t_2 = T_{i_2} - T_{i_1}, \cdots, t_{n_C} = T_{i_{n_C}} - T_{i_{n_C-1}}$ the corresponding differenced series. The Ferro Segers estimator of θ is then defined as:

$$\hat{\theta}_F = \begin{cases} 1 \wedge \frac{2(\sum_{i=1}^{n_C} t_i)^2}{(n_C-1)\sum_{i=1}^{n_C} t_i^2} & \text{if } \max_i t_i \le 2 \\ 1 \wedge \frac{2(\sum_{i=1}^{n_C}(t_i-1))^2}{(n_C-1)\sum_{i=1}^{n_C}(t_i-1)(t_i-2)} & \text{if } \max_i t_i > 2 \end{cases}$$

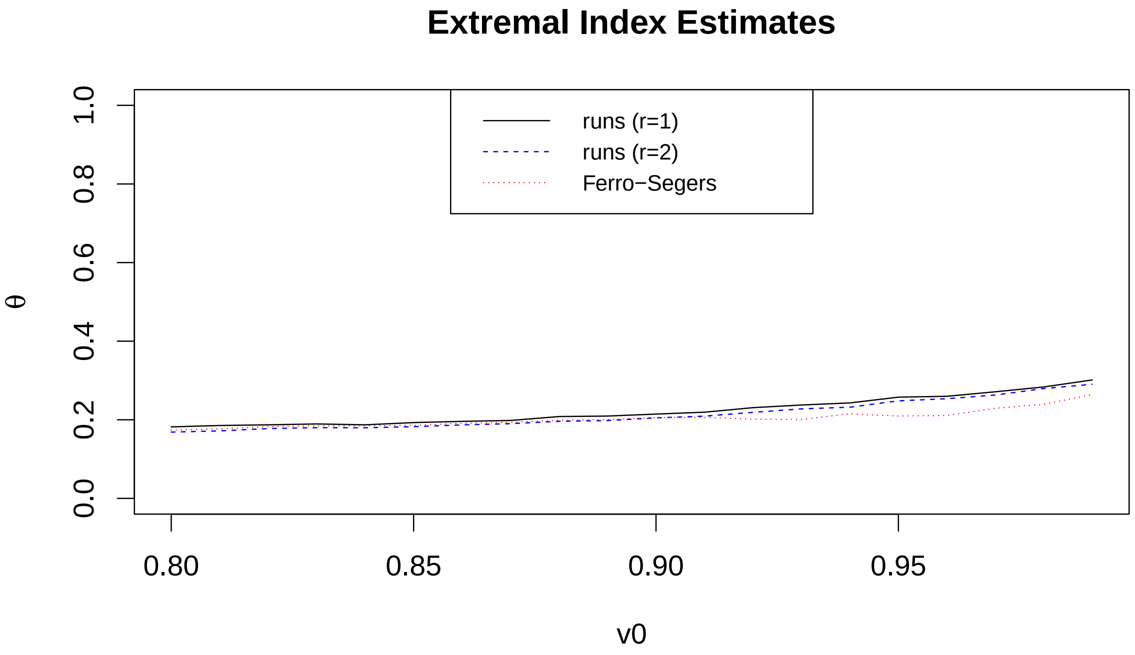


Figure 4.8: Extremal index for the time series Q_k for varying declustering techniques.

The clusters are obtained by considering exceedances separated by the average cluster size, $1/\theta$. The maxima of these clusters are nearly independent and can be modeled as a GPD model as described under Section 4.4.3. The **decluster** function in the R package *extRemes* (2016) automatically applies this declustering procedure and produces independent cluster maximas for GPD model fitting.

Figure 4.8 shows the estimates of the extremal index for $Q_k(0.95)$ for a range of thresholds $q_0 = F^{-1}(v_0)$. Here F is the empirical distribution function the time series Q_k constrained to $Q_k > q_0$ and F^{-1} is given by (4.8). The extremal index estimates for the runs method, r = 1 and r = 2 and the Ferro-Segers closely agree. The average value of θ over all threshold levels for the two runs estimators and the Ferro-Segers estimator are 0.22, 0.21 and 0.20, respectively. These estimates correspond to an average cluster size of about 5. Since the time series Q_k is built by taking minima over overlapping periods of size $\Delta - 1 = 6$, hence a clustering size of around for the extremes.

In the following subsections, we use the method of runs estimate with $r = 2$ for all the analysis and inference on GPD modeling.

4.4 Model

Having obtained a nearly independent sample of excesses – one from each "cluster" – we use maximum likelihood to fit a GPD model with likelihood function of the form

$$P[X > z | X > u] = \left[1 + \xi \left(\frac{z - u}{\sigma}\right)\right]^{-1/\xi} \tag{4.10}$$

By the Pickands-Balkema-de Haan Theorem *Pickands* (1975), *Balkema and de Haan* (1974), the excess over a given threshold are well modeled by the Generalized Pareto

Distribution of the form (4.10).

With a little abuse of notation, we shall henceforth refer to the declustered time series obtained from Q_k also as Q_k. Since the Q_k's are bounded above, theoretically, only models with tail index $\xi \leq 0$ are reasonable. In practice, when considering quantiles that are away from the upper bound, a GPD model with negative tail index may provide a better fit. We shall later show that the resulting estimates based on the data do not reject the common sense hypothesis that $\xi \leq 0$. This is because the pointwise confidence intervals always cover zero for a wide range of thresholds (see Section 4.4.3).

4.4.1 Covariates

For the time series Q_k, we model the scale and shape parameters as a function of k as

$$P[Q_k > z | Q_k > q_0] = \left[1 + \xi(k)\left(\frac{z - q_0}{\sigma(k)}\right)\right]^{-1/\xi(k)} \tag{4.11}$$

for a given threshold value q_0 where the quantities $\sigma(k)$ and $\xi(k)$ are allowed to vary as a function of k. In this direction, we assume the following model for the scale parameter σ:

$$\log(\sigma(k)) = \gamma^{(\sigma)} + \sum_{i=1}^{m_1^{(\sigma)}} \alpha_i^{(\sigma)} \phi_j(k) + \sum_{j=1}^{m_2^{(\sigma)}} \delta_j^{(\sigma)} \psi_j(ENSO(k)) \tag{4.12}$$

where $\gamma^{(\sigma)}$ is a constant term and $\alpha_i^{(\sigma)}$ and $\delta_j^{(\sigma)}$ are the coefficients for diurnal periodic patterns and El-Niño Southern Oscillation (ENSO) activity *Philander et al.* (1989) respectively.

Specifically, ϕ and ψ represent the covariates for daily patterns and ENSO levels respectively. Since seasonal patterns are expected to repeat themselves after a span of

365 days, $\phi_i, i = 1, \cdots, m_1^{(\sigma)}$ represent 365-periodic spline basis representation of the range $1, \cdots, 365T$ as described in Section 4.3.1 with $m = m_1^{(\sigma)}$ degrees of freedom.

With historical data[10] on the ENSO index, $\psi_i, i = 1, \cdots, m_2^{(\sigma)}$ denote a spline basis representation of the range of ENSO values for the period $1, \cdots, 365T$ for $m = m_2^{(\sigma)}$ degrees of freedom. The so-obtained splines are then evaluated at the ENSO index for each time point k. Figure 4.1 right panel displays a plot of cubic splines with 3 degrees of freedom when evaluated at the ENSO values for the time span 1986-1990.

Similar to the scale parameter the regression equation for the shape parameter ξ is given by:

$$\xi(k) = \gamma^{(\xi)} + \sum_{i=1}^{m_1^{(\xi)}} \alpha_i^{(\xi)} \phi_j(k) + \sum_{j=1}^{m_2^{(\xi)}} \delta_j^{(\xi)} \psi_j(ENSO(k)) \tag{4.13}$$

Since $Q_k \sim GPD(\sigma(k), \xi(k))$, the likelihood function for the Q_k's may be expressed as

$$L = \prod \alpha_{k=1}^{N} f(q_k; q_0, \sigma(k), \xi(k))$$

where q_k is the observed q_k and f is the density obtained from the probability distribution function in (4.11). Assuming that the threshold parameter v_0 is fixed, the log likelihood function has the form

$$l(\sigma(k), \xi(k)) = -N \log \sigma(k) - \sum_{k=1}^{N} \left(1 + \frac{1}{\xi(k)}\right) \log\left(1 + \xi(k)\Big(\frac{q_k - qs_0}{\sigma(k)}\Big)\right) \tag{4.14}$$

Thus using (4.12) and (4.13) allows us to write the log likelihood function (4.14) in terms of the parameters $\gamma^{(\xi)}$, $\alpha_1^{(\sigma)}, \cdots, \alpha_{m_1^{(\sigma)}}^{(\sigma)}$, $\delta_1^{(\sigma)}, \cdots, \delta_{m_2^{(\sigma)}}^{(\sigma)}$, $\gamma^{(\xi)}$, $\alpha_1^{(\xi)}, \cdots, \alpha_{m_1^{(\xi)}}^{(\xi)}$, $\delta_1^{(\xi)}, \cdots, \delta_{m_2^{(\xi)}}^{(\xi)}$. The maximum likelihood estimators may be obtained using the **fevd** function of R package *extRemes* (2016) . The explicit details on the method of

[10] https://www.esrl.noaa.gov/psd/enso/data.html

maximum likelihood fitting are covered under Section 4.3.2 in *Coles* (2001).

4.4.2 Model selection and diagnostics

For this section, we keep the value of u_0 fixed at 0.95 and that of $v_0 = 0.9$. We consider the following competing class of models:

$$\begin{aligned} \mathcal{M}_{i^{(\sigma)},j^{(\sigma)},i^{(\xi)},j^{(\xi)}} := m_1^{(\sigma)} &= c_1^{(\sigma)} I_{[i^{(\sigma)}=1]}, \\ m_2^{(\sigma)} &= c_2^{(\sigma)} I_{[j^{(\sigma)}=1]}, \\ m_1^{(\xi)} &= c_1^{(\xi)} I_{[i^{(\xi)}=1]}, \\ m_2^{(\xi)} &= c_2^{(\xi)} I_{[j^{(\xi)}=1]}. \end{aligned} \tag{4.15}$$

where $i^{(\sigma)}, j^{(\sigma)}, i^{(\xi)}, j^{(\xi)} \in \{0, 1\}$ and $c_1^{(\sigma)}$, $c_2^{(\sigma)}$, $c_1^{(\xi)}$, $c_2^{(\xi)}$ are arbitrary constants not equal to 0. Thus the model $\mathcal{M}_{0,1,1,0}$ represents a model where the ENSO covariate is used for explaining the scale parameter and day covariate is used for explaining the shape parameter. The AIC value for models in (4.15) determines which covariates need to be used for explaining the scale and shape parameters of the GPD.

Ideally $c_1^{(\sigma)}$, $c_2^{(\sigma)}$, $c_1^{(\xi)}$, $c_2^{(\xi)}$ should be chosen optimally but for preliminary analysis, we consider each of these to be constant at 3. In order to determine the optimal model, AIC/ BIC criterion is used. The **fevd** function of R package *extRemes* (2016) provides an AIC value corresponding to a maximum likelihood fit for each model. Table 4.1 gives the AIC values for all models in (4.15). The lowest AIC value is recorded for the model $\mathcal{M}_{1,1,1,0}$ which suggests that the scale parameter is explained by both ENSO and day covariate and the shape parameter is explained only through the day covariate. Indeed the model diagnostic plot (see Section 6.2.3 of *Coles* (2001)) in Figure 4.9 left shows the adequacy of fit for model $\mathcal{M}_{1,1,1,0}$.

$(m_1^{(\xi)}, m_2^{(\xi)})\backslash(m_1^{(\sigma)}, m_2^{(\sigma)})$	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	-2172.975	-2178.631	-2209.116	-2210.846
(0,1)	-2174.337	-2173.342	-2208.481	-2205.862
(1,0)	-2193.007	-2196.098	-2209.690	-2212.647
(1,1)	-2194.066	-2192.216	-2211.096	-2208.089

Table 4.1: AIC values for varying models $\mathcal{M}_{i^{(\sigma)},j^{(\sigma)},i^{(\xi)},j^{(\xi)}}$ with $c_1^{(\sigma)} = c_2^{(\sigma)} = c_1^{(\xi)} = c_2^{(\xi)} = 3$.

The analysis so far confirmed that the shape parameter depends on both the covariates $(m_1^{(\sigma)}, m_2^{(\sigma)} \neq 0)$ whereas the scale parameter depends only through day covariate $(m_1^{(\xi)} \neq 0)$. The next step of determining optimal values of $m_1^{(\sigma)}$, $m_2^{(\sigma)}$ and $m_1^{(\xi)}$ is accomplished by using the model with minimum AIC. Performing a grid search over different values $m_1^{(\sigma)}, m_2^{(\sigma)}$ and $m_1^{(\xi)}$, it so turns out that their optimal values are recorded as 5, 3 and 3 respectively.

The model diagnostic plot for the AIC based optimal choice $m_1^{(\sigma)}$, $m_2^{(\sigma)}$ and $m_1^{(\xi)}$ is shown in the right panel of Figure 4.9. It clearly elucidates the adequacy of the chosen parameters in terms of the model fit.

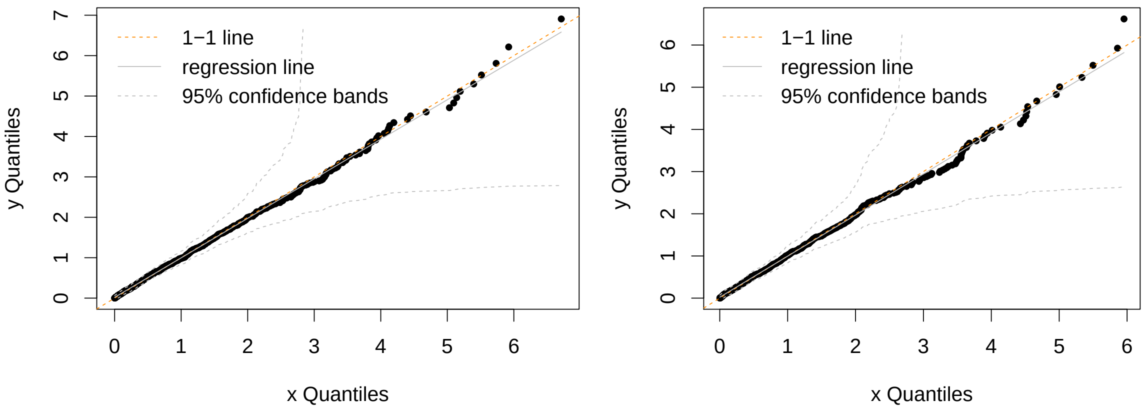


Figure 4.9: Model Diagnostic Plot.

4.4.3 Model estimates

Based on the analysis of previous subsection, we saw that the best model is described by

$$
\begin{aligned}
Q_k &\sim \text{GPD}(\sigma(k), \xi(k)) \qquad (4.16)\\
\log(\sigma(k)) &= \gamma^{(\sigma)} + \sum_{i=1}^{m_1^{(\sigma)}} \alpha_i^{(\sigma)} \phi_j(k) + \sum_{j=1}^{m_2^{(\sigma)}} \delta_j^{(\sigma)} \psi_j(ENSO(k))\\
\xi(k) &= \gamma^{(\xi)} + \sum_{i=1}^{m_1^{(\xi)}} \alpha_i^{(\xi)} \phi_j(k)
\end{aligned}
$$

with $m_1^{(\sigma)} = 5$, $m_2^{(\sigma)} = 3$ and $m_1^{(\xi)} = 3$. We next analyze the estimates for the coefficients and their corresponding standard errors.

Table 4.4.3 gives the estimates of coefficients for day covariate (both scale and shape parameter) and ENSO covariate (only scale parameter) . The ones highlighted in red correspond to significant coefficients ($|\text{Estimate/StandardError}| > \Phi^{-1}(0.95) = 1.645$).

	$\alpha_1^{(\sigma)}$	$\alpha_2^{(\sigma)}$	$\alpha_3^{(\sigma)}$	$\alpha_4^{(\sigma)}$	$\alpha_5^{(\sigma)}$	$\delta_1^{(\sigma)}$	$\delta_2^{(\sigma)}$	$\delta_3^{(\sigma)}$	$\alpha_1^{(\xi)}$	$\alpha_2^{(\xi)}$	$\alpha_3^{(\xi)}$
Est	0.598	-0.341	0.589	0.755	0.892	-2.031	0.406	-1.543	0.762	0.007	0.411
S.e.	0.374	0.240	0.324	0.280	0.307	0.720	0.300	0.510	0.347	0.143	0.338

Table 4.2: Estimates and their standard errors

With estimates $\hat{\alpha}_i^{(\sigma)}$, $i = 1, \cdots, m_1^{(\sigma)}$ and $\hat{\delta}_j^{(\sigma)}$, $j = 1, \cdots, m_2^{(\sigma)}$ at hand, it is easier to visualize the scale parameter σ as a jointly varying function of ENSO level and day of occurrence:

$$
\sigma(k, x) = \exp\left(\gamma^{(\sigma)} + \underbrace{\sum_{j=1}^{m_1^{(\sigma)}} \alpha_j^{(\sigma)} \phi_j(k)}_{A^{(\sigma)}(k)} + \underbrace{\sum_{j=1}^{m_2^{(\sigma)}} \delta_j^{(\sigma)} \psi_j(x)}_{D^{(\sigma)}(x)}\right) = \exp(\gamma^{(\sigma)} + A^{(\sigma)}(k) + D^{(\sigma)}(x)) \qquad (4.17)
$$

In order to study the impact of daily fluctuations on $\sigma(k,x)$, we consider the plot of $A^{(\sigma)}(k)$ as function of k. The effect of ENSO level on $\sigma(k,x)$ is best described through the plot of $D^{(\sigma)}(x)$ versus x. Figure 4.10 gives a plot of the functions $A^{(\sigma)}(.)$ and $D^{(\sigma)}(.)$. The left panel suggests that extreme heat wave events seem to be prevalent during the winter months when compared to summer months. Also during the mid winter and mid summer, the heat waves seem to gain in terms of their intensity. During the fall and spring months, there seem to be little instances of extreme heat wave activity. We study the seasonal behavior in more details under Section 4.5.3. From the middle panel, one may conclude that La Niña phases (ENSO $\in$ [-2,-1.5]) seem to under greater heat wave activity when compared to El Niño phase (ENSO $\in$ [1.5,2]).

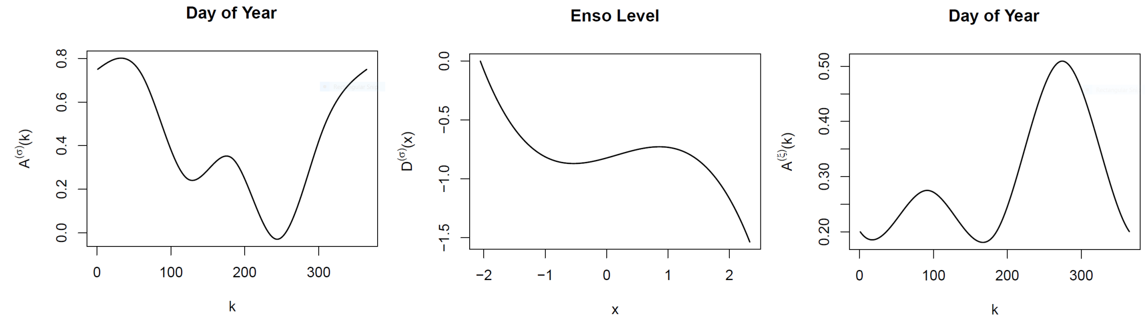


Figure 4.10: Scale parameter and shape parameters as a function of k and x.

With estimates $\hat{\alpha}_i^{(\xi)}$, $i = 1, \cdots, m_1^{(\xi)}$ at hand, it is easier to visualize the shape parameter ξ a function of day of occurrence:

$$\xi(k) = \exp\Big(\gamma^{(\xi)} + \underbrace{\sum_{j=1}^{m_1^{(\xi)}} \alpha_j^{(\xi)} \phi_j(k)}_{A^{(\xi)}(k)}\Big) = \exp(\gamma^{(\xi)} + A^{(\xi)}(k)) \tag{4.18}$$

In order to study the impact of daily fluctuations on $\xi(k)$, we consider the plot of

$A^{(\xi)}(k)$ as function of k. Figure 4.10 right panel gives a plot of function $A^{(\xi)}$. With respect to the shape parameter, extreme heat wave events seem to more prevalent in the fall months. Also during the mid fall and mid spring, the heat waves seem to gain in terms of their intensity. During the summer and winter months, there seem to be little instances of extreme heat wave activity with respect to the shape parameter. The behavior of shape with respect to the seasons is almost contrary to the behavior of scale.

4.4.4 Return levels

If $X \sim \text{GPD}(\sigma, \xi)$, then the unconditional distribution of X using (4.10) may be written as

$$P[X > z] = \tau_u \left[1 + \xi \left(\frac{z-u}{\sigma}\right)\right]^{-1/\xi} \tag{4.19}$$

$\tau_u = P[X > u]$. Thus an m^{th} year return level for the random variable X is given by

$$r_m = \inf\left\{z : P(X > z) \geq \frac{1}{m}\right\} = u + \frac{\sigma}{\xi}[(m\tau_u)^{\xi} - 1] \tag{4.20}$$

where r_m denotes the level that is exceeded on average once every m observations. Since r_m is obtained corresponding to the excess distribution, m should be large enough to guarantee $r_m > u$.

Based on (4.16), the distribution function for time series Q_k is given by:

$$P[Q_k > z] = \tau_{q_0} \left[1 + \xi(k) \left(\frac{z - q_0}{\sigma(k)}\right)\right]^{-1/\xi}$$

Using (4.17) and (4.18), one may equivalently express the return level r_m as a bivariate function of the covariates day in year and ENSO level as:

$$r_m(k,x) = v_0 + \frac{\sigma(k,x)}{\xi(k)}[(m\tau_{v_0})^{\xi} - 1] = v_0 + \frac{\exp(w_1^\top \rho_1)}{w_2^\top \rho_2}[(m\tau_{v_0})^{w_2^\top \rho_2} - 1] \qquad (4.21)$$

with w_1, w_2, ρ_1 and ρ_2 given by

$$\begin{aligned} w_1 &= \begin{bmatrix} 1 & \phi_1(k) & \cdots & \phi_{m_1^{(\sigma)}}(k) & \psi_1(x) & \cdots \psi_{m_2^{(\sigma)}}(x) \end{bmatrix}^\top \\ \rho_1 &= \begin{bmatrix} \gamma^{(\sigma)} & \alpha_1^{(\sigma)} & \cdots & \alpha^{(\sigma)}_{m_1^{(\sigma)}} & \delta_1^{(\sigma)} & \cdots \delta^{(\sigma)}_{m_2^{(\sigma)}} \end{bmatrix}^\top \\ w_2 &= \begin{bmatrix} 1 & \phi_1(k) & \cdots & \phi_{m_1^{(\xi)}}(k) \end{bmatrix}^\top \\ \rho_2 &= \begin{bmatrix} \gamma^{(\xi)} & \alpha_1^{(\sigma)} & \cdots & \alpha^{(\xi)}_{m_1^{(\xi)}} \end{bmatrix}^\top \end{aligned} \qquad (4.22)$$

The estimate for the return level in (4.21) and its standard error are obtained in accordance to Lemma C.1 in Section C.

Figure 4.11, left explores in details the variations in the 10-year return levels $\hat{r}_m$ as a function of k and x. For $m = 10 \times 365$, the behavior of $\hat{r}_m$ is fairly similar across different values of m and has not been reported in here. The right panel explores in details the variations in the standard error of return level estimates, $\hat{r}_m$. We observe unusually large values of the standard error for ENSO levels close to -2 which may be attributed to the non-availability of observations at such low values of the ENSO. Heat wave activity is most profound during the months of February and December especially with ENSO level close to [-1.7,-1] and [0.3,1.1].

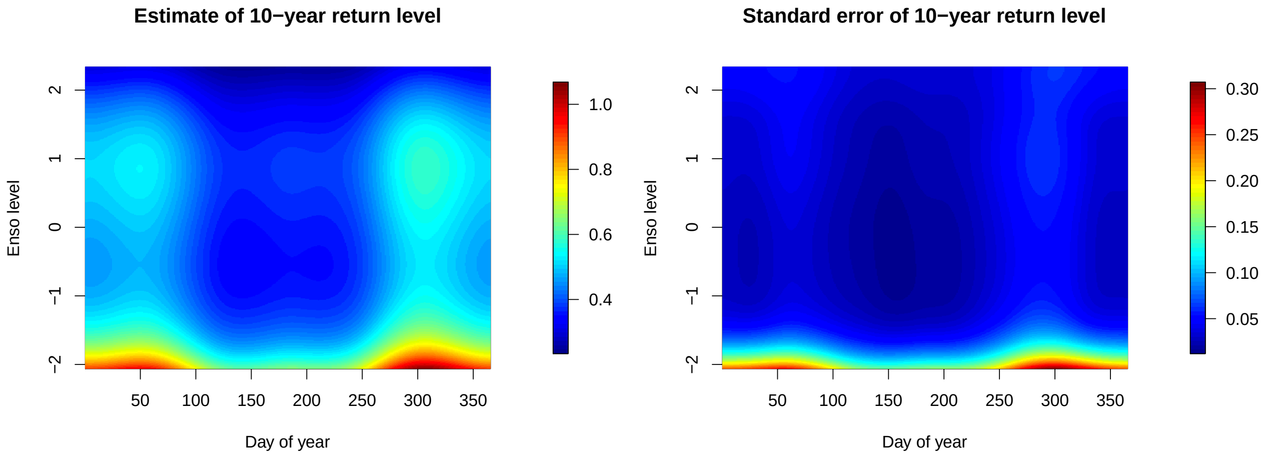


Figure 4.11: Return level $\hat{r}_m$ for $m = 10 \times 365$ as a function of Day and the ENSO.

We next study the behavior of return levels cross sectionally. Figure 4.12 allows us to study the univariate effect of the ENSO factor (x) at at varying levels of day of the year (k). Left panel in Figure 4.12 corresponds to a day in the winter season, middle panel for a day in the summer season and the right panel is for a day in the fall season. The shape of the curve is almost similar to the ENSO variation curve in Figure 4.10 with a change in the location depending on the season under consideration. This is expected since the return level in (4.20) is directly proportional to the scale parameter and the shape parameter is free from the ENSO covariate. As is evident in the graphs, the heat waves seem to be of a larger amplitude for the winter and summer months. They're relatively smaller during the fall season, an observation which is further consolidated under Section 4.5.3.

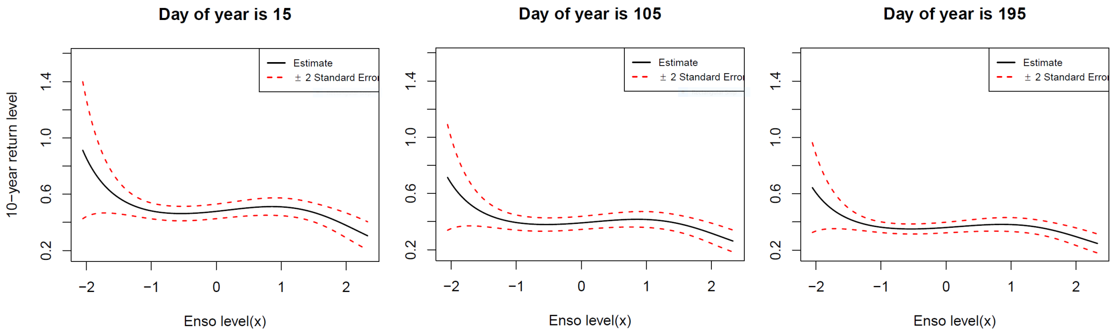


Figure 4.12: Return levels $\hat{r}_m$ for $m = 10 \times 365$ as a function of ENSO on various days of the year.

Figure 4.13 allows us to study the univariate effect of day of the year (k) at varying levels of the ENSO factor (x). Left in Figure 4.13 corresponds to time point from the La Niña phase of ENSO whereas the right panel corresponds to the El Niño phase of ENSO. The shape of the curve is almost similar to the day variation curve in left panel of Figure 4.10 with a change in the location depending on the ENSO value under consideration. This is because the effect of day covariate on the scale predominates the effect on shape parameter (see Table 4.4.3). The figure further supports the hypothesis that heat waves seem to be more severe during the El Niño periods of the climate.

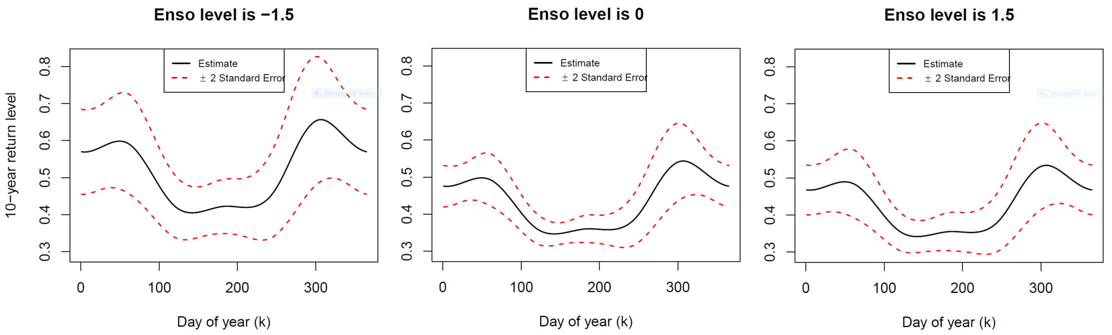


Figure 4.13: Return levels $\hat{r}_m$ for $m = 10 \times 365$ as a function of year for different levels of ENSO.

In the following subsections, we explore other factors which impact the return levels.

4.5 Factors Influencing Return Levels

4.5.1 Role of u_0

In this section, we explore the role of intensity level u_0 in defining the heat wave $Q_k(u_0)$. Two values of u_0 viz 0.92 and 0.97 are chosen and v_0 is fixed at 0.9. Repeating the analysis in Section 4.4.2, we obtain the AIC values for the models described in (4.15) with $c_1^{(\sigma)} = c_2^{(\sigma)} = c_1^{(\xi)} = c_2^{(\xi)} = 3$.

$(m_1^{(\xi)}, m_2^{(\xi)})\backslash(m_1^{(\sigma)}, m_2^{(\sigma)})$	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	-1921.077	-1918.981	-1969.000	-1964.619
(0,1)	-1919.191	-1914.037	-1965.394	-1961.123
(1,0)	-1941.193	-1939.125	-1974.936	-1971.605
(1,1)	-1941.953	-1936.630	-1974.770	-1970.822

Table 4.3: For $u_0 = 0.92$, AIC values for varying models in (4.15) with $c_1^{(\sigma)} = c_2^{(\sigma)} = c_1^{(\xi)} = c_2^{(\xi)} = 3$.

$(m_1^{(\xi)}, m_2^{(\xi)})\backslash(m_1^{(\sigma)}, m_2^{(\sigma)})$	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	-2551.725	-2558.228	-2576.183	-2581.193
(0,1)	-2554.783	-2553.041	-2578.132	-2576.145
(1,0)	-2564.781	-2568.790	-2577.530	-2582.244
(1,1)	-2566.841	-2564.423	-2579.275	-2577.585

Table 4.4: For $u_0 = 0.97$, AIC values for varying models in (4.15) with $c_1^{(\sigma)} = c_2^{(\sigma)} = c_1^{(\xi)} = c_2^{(\xi)} = 3$.

For $u_0 = 0.92$, the optimal model is $\mathcal{M}_{1,0,1,0}$ which implies only the day covariate is significant for both the shape and scale parameter. The ENSO covariate seems to be of little importance be it in terms of explaining the shape or the scale parameter. For $u_0 = 0.97$, the optimal model is $\mathcal{M}_{1,1,1,0}$ which implies only the day covariate is significant for both the shape and scale parameter. The ENSO covariate however only explains variations in the scale parameter.

We end this section by exploring the return levels for all three values of $u_0 = 0.92, 0.0.97$ and comparing it to $u_0 = 0.95$ as in Section 4.4.4 as a function of the season and ENSO covariate as in Section 4.4.4. For $u_0 = 0.92$, since only the day covariate is significant, the return levels and their standard errors are computed using Lemma C.1 of Section C with w_1, ρ_1, w_2 and ρ_2 given by

$$\begin{aligned}
w_1 &= \begin{bmatrix} 1 & \phi_1(k) & \cdots & \phi_{m_1^{(\sigma)}}(k) \end{bmatrix}^\top \\
\rho_1 &= \begin{bmatrix} \gamma^{(\sigma)} & \alpha_1^{(\sigma)} & \cdots & \alpha^{(\sigma)}_{m_1^{(\sigma)}} \end{bmatrix}^\top \\
w_2 &= \begin{bmatrix} 1 & \phi_1(k) & \cdots & \phi_{m_1^{(\xi)}}(k) \end{bmatrix}^\top \\
\rho_2 &= \begin{bmatrix} \gamma^{(\xi)} & \alpha_1^{(\sigma)} & \cdots & \alpha^{(\xi)}_{m_1^{(\xi)}} \end{bmatrix}^\top
\end{aligned} \tag{4.23}$$

where $m_1^{(\sigma)}$ and $m_1^{(\xi)}$ are chosen optimally by the AIC criterion . These values are equal to 3 and 4 respectively. Figure 4.14 top left panel gives the return levels for $u_0 = 0.92$ as a function of day in the year. For $u_0 = 0.97$, the day covariate is significant for both shape and scale parameter and ENSO covariate is significant only for the scale parameter. Therefore the return levels and their standard errors are computed using Lemma C.1 of Section C with w_1, ρ_1, w_2 and ρ_2 replaced by values in (4.26).

Figure 4.14 top right and bottom panel shows a plot of the return levels as a function of day of year for 3 different values of the enso. The return levels curves for $u_0 = 0.97$ match the patterns in Figure 4.13. The effect of season is quite small when we are in the regime of extreme heat waves ($u_0 = 0.97$). However season has a significant impact on heat waves of comparatively lower intensity. Also the return levels decrease monotonically with increase in the intensity quantile u_0.

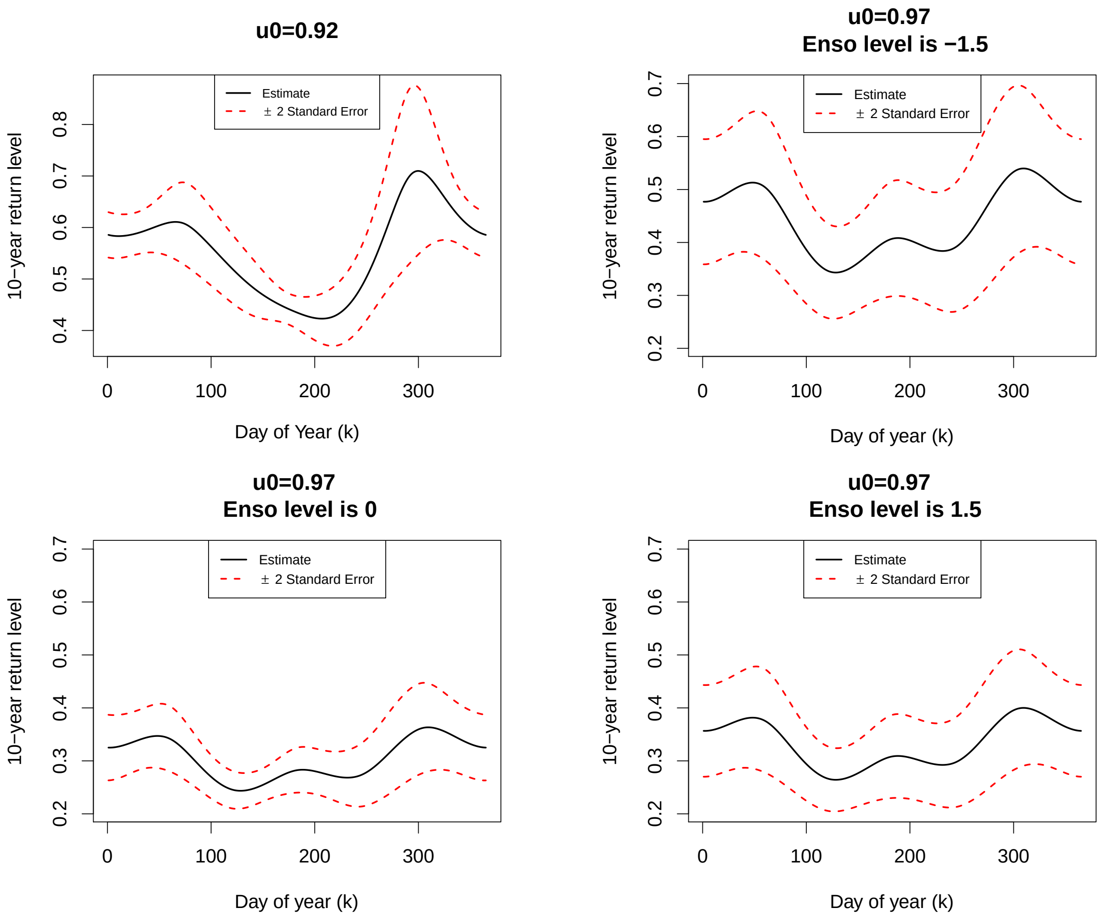


Figure 4.14: Return levels $\hat{r}_m$ for $m = 10 \times 365$ as a function of year for different levels of ENSO for $u_0 = 0.92, 0.97$

4.5.2 Role of Δ

From (4.6), it is clear that the threshold for defining heat waves can vary as a function of s, Δ and u_0. In this section, we explore the role of Δ by two approaches, (1) $U(u_0, s, \Delta) = \Phi^{-1}(u_0)$ where Φ is cumulative distribution function of standard normal(2) $U(u_0, s, \Delta) = F^{-1}_{s,\Delta}(u_0)$ where $F_{s,\Delta}$ is the empirical cumulative distribution function of $Z_{k,\Delta}(s)$ (see (4.5)). Whereas, the first case correspond to the case of an absolute threshold which is free from the parameters s and Δ , the second one denotes a relative threshold. In the following subsections both these cases have been explored for three values of $\Delta = 2, 7, 14$.

4.5.2.1 Absolute threshold

For an absolute threshold $U(u_0, s, \Delta) = \Phi^{-1}(u_0)$, we explore the role of Δ on heat waves. For $u_0 = 0.85$ and $v_0 = 0.85$, the analysis in Section 4.4.2 is repeated, wherein the AIC values for the models in (4.15) are obtained with $c_1^{(\sigma)} = c_2^{(\sigma)} = c_1^{(\xi)} = c_2^{(\xi)} = 3$. Tables 4.5, 4.6 and 4.7 tabulates these AIC values for three different values of $\Delta = 2, 7, 14$.

$(m_1^{(\xi)}, m_2^{(\xi)}) \backslash (m_1^{(\sigma)}, m_2^{(\sigma)})$	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	-3818.555	-3817.642	-3874.286	-3871.342
(0,1)	-3816.362	-3814.471	-3872.872	-3867.557
(1,0)	-3839.510	-3838.591	-3870.374	-3867.406
(1,1)	-3840.595	-3835.849	-3868.910	-3863.821

Table 4.5: For $\Delta = 2$, AIC values for varying models in (4.15).

$(m_1^{(\xi)}, m_2^{(\xi)}) \backslash (m_1^{(\sigma)}, m_2^{(\sigma)})$	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	-6636.493	-6640.659	-6653.301	-6656.352
(0,1)	-6635.730	-6636.347	-6652.806	-6652.012
(1,0)	-6639.743	-6643.132	-6650.894	-6653.515
(1,1)	-6638.546	-6639.437	-6649.615	-6648.591

Table 4.6: For $\Delta = 7$, AIC values for varying models in (4.15).

$(m_1^{(\xi)}, m_2^{(\xi)}) \backslash (m_1^{(\sigma)}, m_2^{(\sigma)})$	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	-9552.274	-9555.518	-9550.450	-9553.115
(0,1)	-9552.516	-9550.439	-9551.234	-9548.361
(1,0)	-9549.035	-9551.949	-9550.942	-9552.849
(1,1)	-9547.751	-9546.568	-9549.003	-9546.918

Table 4.7: For $\Delta = 14$, AIC values for varying models in (4.15).

For $\Delta = 2$, the optimal model is $\mathcal{M}_{1,0,0,0}$ which implies only the day covariate plays a role in terms of determing the scale parameter. The shape parameter is free from the influence of the covariates. For $\Delta = 7$, the optimal model is $\mathcal{M}_{1,1,0,0}$ which implies both day and ENSO covariates influence the shape parameter and the scale parameter is free from the effect of the covariates. Lastly for $\Delta = 14$, the optimal model is $\mathcal{M}_{1,0,0,0}$ which implies that only the shape parameter is influenced by the ENSO covariate. For the case of absolute threshold, a value of $\Phi^{-1}(0.85) = 1.03$ corresponds to very extreme quantiles in the distribution of $Z_{k,14}(s)$ in contrast to $Z_{k,10}(s)$ where it corresponds to moderately large quantiles (see Figure 4.17). These results corroborate the fact that patterns of the unusually large extreme heat waves are determined by the ENSO covariate whereas day covariate influences reasonably large ones.

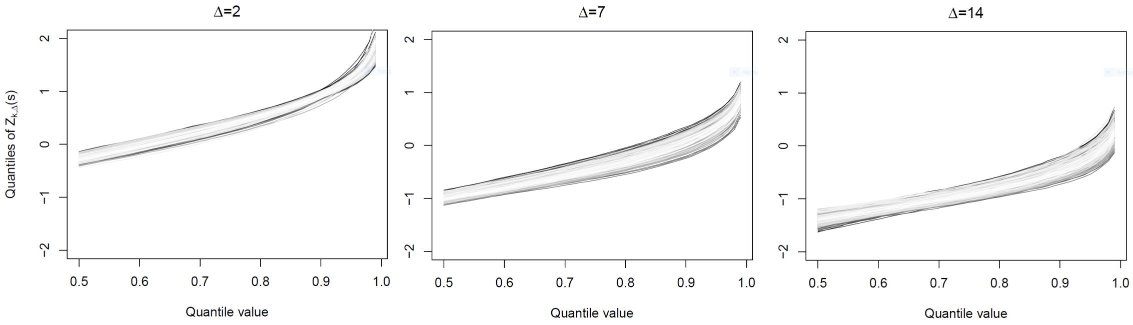


Figure 4.15: Quantiles for the distribution of $Z_{k,\Delta}(s)$ for varying values of Δ. The gray lines represent different stations

For all the three values of $\Delta = 2, 7, 14$, the shape parameter was found to be free from the covariates. For $\Delta = 2$, since the scale parameter is a function of the day covariate only, the return levels and their standard errors are computed using Lemma

C.1 of Section C with w_1, ρ_1, w_2 and ρ_2 given by

$$\begin{aligned} w_1 &= \begin{bmatrix} 1 & \phi_1(k) & \cdots & \phi_{m_1^{(\sigma)}}(k) \end{bmatrix}^\top \\ \rho_1 &= \begin{bmatrix} \gamma^{(\sigma)} & \alpha_1^{(\sigma)} & \cdots & \alpha_{m_1^{(\sigma)}}^{(\sigma)} \end{bmatrix}^\top \\ w_2 &= 1 \\ \rho_2 &= \gamma^{(\xi)} \end{aligned} \tag{4.24}$$

where $m_1^{(\sigma)}$, chosen optimally by the AIC criterion turned out to be 5. Figure 4.16 left panel gives the return levels for $\Delta = 2$ as a function of day in the year. For $\Delta = 14$, since the scale paramter is a function of the ENSO covariate only, the return levels and their standard errors are computed using Lemma C.1 of Section C with w_1, ρ_1, w_2 and ρ_2 given by

$$\begin{aligned} w_1 &= \begin{bmatrix} 1 & \psi_1(x) & \cdots & \psi_{m_2^{(\sigma)}}(x) \end{bmatrix}^\top \\ \rho_1 &= \begin{bmatrix} \gamma^{(\sigma)} & \delta_1^{(\sigma)} & \cdots & \delta_{m_2^{(\sigma)}}^{(\sigma)} \end{bmatrix}^\top \\ w_2 &= 1 \\ \rho_2 &= \gamma^{(\xi)} \end{aligned} \tag{4.25}$$

where $m_2^{(\sigma)}$, chosen optimally by the AIC criterion turned out to be 8. Figure 4.16 right panel displays the return levels for $\Delta = 14$ as a function of the ENSO level. The return levels are much lower for $\Delta = 14$. This is an expected phenomenon because for an absolute threshold $U = \Phi^{-1}(u_0)$, there will be much fewer events of duration $\Delta = 14$ than $\Delta = 2$.

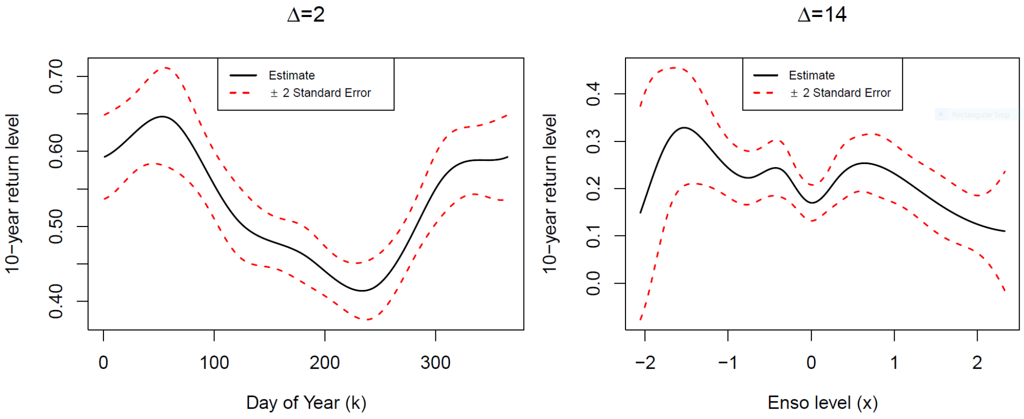


Figure 4.16: Return levels $\hat{r}_m$ for $m = 10 \times 365$. Left panel: for $\Delta = 2$, $\hat{r}_m$ is a function of day in year. Right panel: For $\Delta = 14$, $\hat{r}_m$ is a function of the ENSO value.

For $\Delta = 7$, the scale paramter is a function of both the ENSO covariate and day covariate, the return levels and their standard errors are computed using Lemma C.1 of Section C with w_1, ρ_1, w_2 and ρ_2 given by

$$\begin{aligned} w_1 &= \begin{bmatrix} 1 & \phi_1(k) & \cdots & \phi_{m_1^{(\sigma)}}(k) & \psi_1(x) & \cdots \psi_{m_2^{(\sigma)}}(x) \end{bmatrix}^\top \\ \rho_1 &= \begin{bmatrix} \gamma^{(\sigma)} & \alpha_1^{(\sigma)} & \cdots & \alpha_{m_1^{(\sigma)}}^{(\sigma)} & \delta_1^{(\sigma)} & \cdots \delta_{m_2^{(\sigma)}}^{(\sigma)} \end{bmatrix}^\top \\ w_2 &= 1 \\ \rho_2 &= \gamma^{(\xi)} \end{aligned} \tag{4.26}$$

where $m_1^{(\sigma)}$ and $m_2^{(\sigma)}$, chosen optimally by the AIC criterion turned out to be 5 and 8 respectively. Figure 4.17 shows a plot of the return levels as a function of day in year for varying levels of the ENSO. The return levels are smaller than those for $\Delta = 2$ but larger than $\Delta = 14$. Lower values of ENSO (ENSO=-1.5) produce higher return levels than higher values (ENSO=1.5) similar to the patterns of Figure 4.16 right panel. Winter months lead to greater heat wave activity when compared to summer months which is similar to the phenomenon observed in left panel of Figure 4.16.

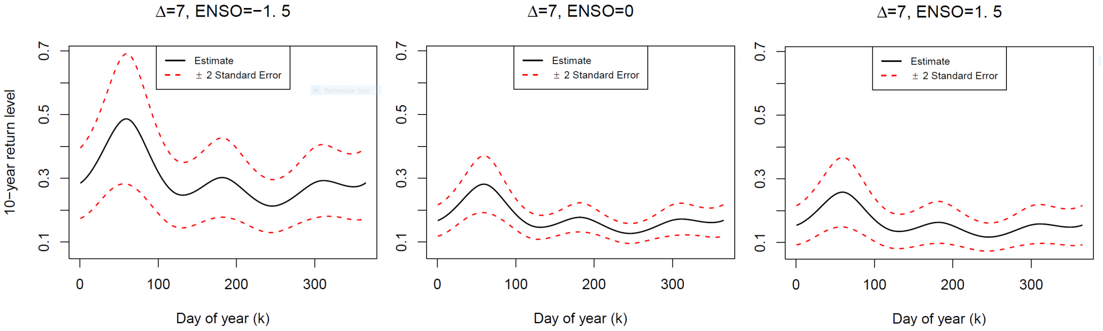


Figure 4.17: Return levels $\hat{r}_m$ for $m = 10 \times 365$ for $\Delta = 7$ at varying values of the ENSO level. Left panel: ENSO=-1.5, Middle panel: ENSO=0, Right panel: ENSO=1.5.

4.5.2.2 Relative threshold

For a relative threshold $U(u_0, s, \Delta) = F^{-1}_{s,\Delta}(u_0)$, we explore the role of Δ on heat waves. For $u_0 = 0.85$ and $v_0 = 0.9$, the analysis in Section 4.4.2 is repeated, wherein the AIC values for the models in (4.15) are obtained with $c_1^{(\sigma)} = c_2^{(\sigma)} = c_1^{(\xi)} = c_2^{(\xi)} = 3$. Tables 4.8, 4.9 tabulates these AIC values for two different values of $\Delta = 2, 14$. The case of $\Delta = 7$ is already covered under Section 4.4.

$(m_1^{(\xi)}, m_2^{(\xi)})\backslash(m_1^{(\sigma)}, m_2^{(\sigma)})$	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	-3479.055	-3477.243	-3517.548	-3512.506
(0,1)	-3475.539	-3473.329	-3513.286	-3507.706
(1,0)	-3488.920	-3485.434	-3513.475	-3508.528
(1,1)	-3485.150	-3481.592	-3509.262	-3503.693

Table 4.8: For $\Delta = 2$, AIC values for varying models in (4.15).

$(m_1^{(\xi)}, m_2^{(\xi)})\backslash(m_1^{(\sigma)}, m_2^{(\sigma)})$	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	-1470.400	-1471.985	-1505.599	-1500.582
(0,1)	-1471.434	-1466.701	-1500.822	-1494.961
(1,0)	-1494.420	-1490.413	-1504.187	-1498.983
(1,1)	-1490.143	-1485.925	-1499.059	-1493.498

Table 4.9: For $\Delta = 14$, AIC values for varying models in (4.15).

The table clearly shows that for $\Delta = 2, 14$, the shape parameter is free from the covariates unlike the case of $\Delta = 7$ where the shape parameter was influenced by day covariate. The day covariate only affects the scale parameter. This is behavior however contradictory to that for $\Delta = 7$ where both day and ENSO covariates were significant for the scale parameter (see Table 4.1).

For $\Delta = 2, 14$, since the scale parameter is a function of the day covariate only, the return levels and their standard errors are computed using Lemma C.1 of Section C with w_1, ρ_1, w_2 and ρ_2 given by

$$\begin{aligned} w_1 &= \begin{bmatrix} 1 & \phi_1(k) & \cdots & \phi_{m_1^{(\sigma)}}(k) \end{bmatrix}^\top \qquad (4.27) \\ \rho_1 &= \begin{bmatrix} \gamma^{(\sigma)} & \alpha_1^{(\sigma)} & \cdots & \alpha_{m_1^{(\sigma)}}^{(\sigma)} \end{bmatrix}^\top \\ w_2 &= 1 \\ \rho_2 &= \gamma^{(\xi)} \end{aligned}$$

where $m_1^{(\sigma)}$, chosen optimally by the AIC criterion turned out to be 3 and 5 for $\Delta = 2$ and $\Delta = 14$ respectively. Figure 4.18 plots the 10-year return levels of heat waves for $\Delta = 2, 14$. The return levels of $\Delta = 14$ seem to greater than $\Delta = 2$ which is in contrast to Figure 4.16. The reason behind the phenomenon is: when relative thresholds are considered the decrease in $Z_{k,\Delta}(s)$ with increase in Δ is accompanied with an increase in the threshold $U(s, u_0, \Delta)$ (see Figure 4.17 which is nothing but a quantile of $Z_{k,\Delta}(s)$. For both values Δ, heat waves are least extreme during the summer months and most extreme for the winter ones.This is exactly similar to the phenomenon as that seen for $\Delta = 7$ (see Figure 4.13) . When compared with Figure 4.13, the return levels for values of ENSO close to -1.5 seem to be higher than all values displayed under Figure 4.18.

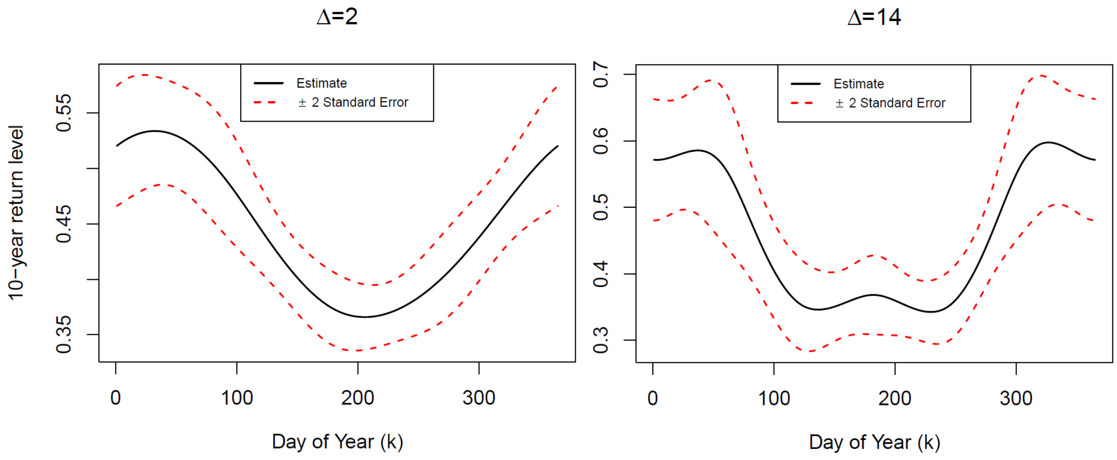


Figure 4.18: Return levels $\hat{r}_m$ for $m = 10 \times 365$ as function of window size Δ.

4.5.3 Role of season

Based on the work of *Cressie* (1993b), we segregate the months of the year into four different seasons as: Spring:{March, April, May}, Summer:{June,July,August}, Fall:{September,October,November}, Winter:{December, January,February}. For a given value of u_0, we consider the time series $Q_k := Q_k(u_0)$ restricted to a season as follows

$$\{Q_k^{\text{seas}}\} = \{Q_k, k \in D^{\text{seas}}\}$$

where D^{seas} corresponds to the day indices in a given season, for example $D^{\text{spring}} = \{(365i + j), i = 1, \cdots, 100, j = 61, \cdots, 150\}$. Moreover, the threshold q_0 for GPD fit is given by$q_0 = (F^{\text{seas}})^{-1}(v_0)$ where F^{seas} is the empirical cdf of the time series Q_k^{seas}. Also, the k varies over the range $1, \cdots, d^{\text{seas}}T$, where d^{seas} denotes the number of days in that season of the year (e.g. $d^{\text{summer}} = 90$).

As in (4.16), Q_k^{seas} is assumed to follow GPD$(\sigma(k), \xi)$ with $\sigma(k)$ as in (4.12). The construction of cubic spline functions, ϕ and ψ differs marginally from that shown in Section 4.4.1. The function ϕ still correspond to periodic patterns and ENSO levels respectively. Since in a given season, patterns are expected to repeat themselves after a span of d^{seas} days, the functions ϕ form d^{seas}-periodic spline basis representation (see Section 4.3.1) of the range $1, \cdots, d^{\text{seas}}T$.

In order to generate the function ψ in (4.12), the range of ENSO values restricted to the particular season are considered and a spline basis representation for that range obtained. The splines are then evaluated at the values of ENSO levels $\{ENSO(k), k = 1, \cdots, d^{\text{seas}}T\}$. Depending on the season under consideration, the effect of either the ENSO or the seasonal covariate may be insignificant.

When restricted to a season, the number of observations in the extremes are very few. Therefore, for seasonal analysis we allow only the scale parameter to depend on the covariates day and ENSO. Table 4.4.3 shows that number of significant coefficients for the scale parameter is comparatively much larger than that for the shape parameter. For most of the cases in Sections 4.5.1 and 4.5.2, the effect of covariates on the shape parameter was negligible. Thus we repeating the analysis in Section 4.4.2 but only for the following four set of models derived from (4.15):

$$\begin{aligned} \mathcal{M}_{i^{(\sigma)},j^{(\sigma)}} := m_1^{(\sigma)} &= c_1^{(\sigma)} I_{[i^{(\sigma)}=1]}, \\ m_2^{(\sigma)} &= c_2^{(\sigma)} I_{[j^{(\sigma)}=1]}, \\ m_1^{(\xi)} &= = 0, \\ m_2^{(\xi)} &= 0. \end{aligned} \tag{4.28}$$

where $i^{(\sigma)}, j^{(\sigma)} \in \{0,1\}$ and $c_1^{(\sigma)}$ snd $c_2^{(\sigma)}$ are arbitrary constants not equal to 0. Thus the model $\mathcal{M}_{0,10}$ represents a model where only the ENSO covariate is used for explaining the scale parameter. The AIC value for models of the for (4.28) determines which covariates need to be used for explaining the scale of the GPD. For $c_1^{(\sigma)} = c_2^{(\sigma)} = 3$, Table 4.10 gives the AIC values for all models in (4.28) for different seasons. The optimal model goes on varying with respect to the season

Season$\backslash(m_1^{(\sigma)}, m_2^{(\sigma)})$	(0,0)	(0,1)	(1,0)	(1,1)
Winter	-623.2626	-623.5936	-619.1384	-618.9644
Spring	-777.1376	-772.8495	-777.6741	-776.1906
Summer	-849.1791	-846.1126	-850.3411	-847.5327
Fall	-776.9355	-772.7973	-791.1087	-786.7266

Table 4.10: AIC values for varying models in (4.28) for $c_1^{(\sigma)} = c_2^{(\sigma)} = 3$ for different seasons.

For the seasons like Spring, Fall and Summer, the optimal model is $\mathcal{M}_{1,0}$ which implies ENSO based covariate is no longer significant. On the contrary, for Winter season the optimal model is $\mathcal{M}_{0,1}$ which implies that only the ENSO and not day covariate is significant.

Since depending on season the significance of a covariate varies, we consider the return levels as a function of day $k^{\text{seas}} = 1, \cdots, d^{\text{seas}}$ in each season. For the spring, fall and summer seasons, the return levels and their standard errors are computed using Lemma C.1 of Section C with w_1, ρ_1, w_2 and ρ_2 given by

$$\begin{aligned} w_1 &= \begin{bmatrix} 1 & \phi_1(k^{\text{seas}}) & \cdots & \phi_{m_1}(k^{\text{seas}}) \end{bmatrix}^\top \\ \rho_1 &= \begin{bmatrix} \gamma^{(\sigma)} & \alpha^{(\sigma)}{}_1 & \cdots & \alpha^{(\sigma)}{}_{m_1^{(\sigma)}} \end{bmatrix}^\top \\ w_2 &= 1 \\ \rho_2 &= \gamma^{(\xi)} \end{aligned} \tag{4.29}$$

which takes into account that the ENSO based covariate is no longer significant. The optimal value for $m_1^{(\sigma)}$ is chosen by the AIC criterion. These values were found to be

3,3 and 8 for spring, summer and fall seasons respectively. For the winter season only the ENSO based covariate is significant, therefore we consider

$$\begin{aligned}
w_1 &= \begin{bmatrix} 1 & \psi_1(x^{\text{seas}}) & \psi_2(x^{\text{seas}}) & \cdots & \psi_{m_2}(x^{\text{seas}}) \end{bmatrix}^\top \\
\rho_1 &= \begin{bmatrix} \gamma^{(\sigma)} & \delta_1^{(\sigma)} & \cdots & \delta^{(\sigma)}{}_{m_2^{(\sigma)}} \end{bmatrix}^\top \\
w_2 &= 1 \\
\rho_2 &= \gamma^{(\xi)}
\end{aligned} \tag{4.30}$$

where x^{seas} belongs to the range of ENSO values restricted to winter season. The optimal value for $m_2^{(\sigma)}$, chosen by the AIC was found to be 3.

For $m = 10 \times d^{\text{seas}}$, Figures 4.19 give a plot of the return levels and the standard errors for the fall, spring, summer and winter seasons. Since the return levels of the fall, spring and summer season are heavily influenced by the value of season based covariate, the return levels are a function of the day in the season only. The return levels for summer season are the largest followed by spring and fall. Towards the end of both season, the return level seems to a bit on the higher end. The return levels for the fall and spring are lower when compared to the winter season. For the winter season, since only the ENSO is significant, we report only the values of the return levels at varying values of the ENSO. At ENSO ≈ -1.2, we observe that the return levels are the highest. This further corroborates that La-Niña phases combined with winter season produce the highest intensity heat waves even greater than those of the summer season. For other values of the ENSO, the return levels in winter are comparable to the spring and fall and smaller than the summer season.

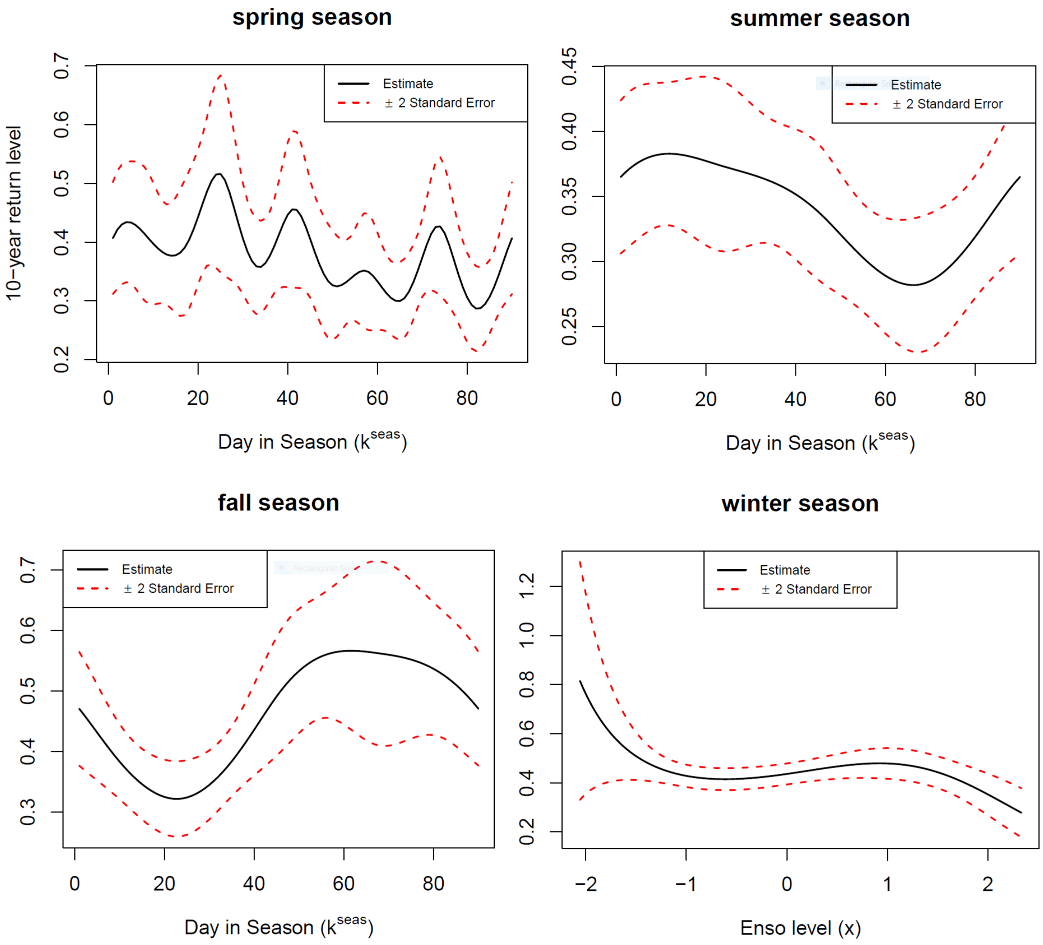


Figure 4.19: Return levels $\hat{r}_m$ for $m = 10 \times 90$.

4.6 Summary And Discussion

In this chapter, we introduce the concept of areal footprint of heat waves and analyzed the behavior of its extremes. The Generalized Pareto Distribution has been used in modeling the peaks over threshold for the time series of areal impact of heat-waves. Covariates like day of the year and ENSO level are shown to be significant in modeling of extreme heat waves events. In Section 4.4.4, we discuss the return levels of the so-defined areal heat waves with respect to variations in the covariates.

Section 4.5 explores the factors which influence the return levels of the areal heat waves. Depending on the value of u_0, either one or both the covariates day and ENSO level may be significant in determining the shape and scale of the GPD. As expected large values of u_0 usually produce smaller return levels which vary with respect to the ENSO levels. In contrast, the return levels are influenced by the day covariate for smaller values of u_0. Section 4.5.2 explores the role of window size Δ in determining the heat waves under two scenarios, one where the threshold of defining heat wave varies as a function of Δ and the other where it is a constant. Large values of Δ correspond to smaller areas of US under extreme heat wave activity. Therefore, under a constant threshold return levels are much smaller for larger Δ. However when relative thresholds are considered, this phenomenon reverses itself. Explaining the return levels as a function of Δ when the threshold varies as a function of Δ is still left for exploration.

An important contribution of the chapter has been the distribution of the heat waves and their return levels with respect to the season. Indeed when restricted to a season, the effect of the week covariate decreases significantly. The ENSO levels seem play a role only during the winter season and do not influence the other seasons. This section also corroborates the observation that heat waves are more severe during winter months. An interesting extension of the chapter is the regional analysis of the extremes in areal distribution of heat waves. A Dirichlet model with time varying coefficients can be employed for modeling the regional distribution over continental US given an extreme event has occurred. These time varying coefficients can be made correlated in space either by a regularization approach or by priors as in *Cooley et al.* (2007), *Besag and Kooperberg* (1995), *Rue and Held* (2005).

CHAPTER V

Conclusions And Future Work

We presented a technique where the security of smart grid networks may be enabled by securing a suitably chosen subset of nodes in the grid. However when the number of these trusted nodes is unknown, the kriging methodology cannot be applied for predictive modeling of energy consumption. Other sophisticated time series based techniques like vector auto regression and dynamic factor models need to be explored. This shall form the platform of our next work of developing a predictive model for which can be used both for anomaly detection and future consumption forecasts.

Cyber activites like Denial of Service, Network Jam etc. all require sophisticated statistical and computational techniques for their detection and prevention. In this work, we presented a few techniques for the identification of heavy hitter IPs for both high and low volume attacks. The community detection algorithms which targeted at detecting structural changes in traffic flow have however not been covered in details. As a part of the future work we wish to develop a principled way for identification of cliques in the co-citation and bibliographic matrices[1] obtained from the adjacency matrix of source-destination flow. Also, the robust estimation of the tail exponents

[1] http://www.pitt.edu/ kpele/Materials15/module1.pdf

may boost the results for our algorithms dependent on the heavy tailed nature of hash binned traffic.

The trimmed Hill estimator presented in this work well identifies the outliers for distribution in the Pareto domain of attraction, i.e. $\xi > 0$. However its behavior with varying nature of the outliers need to be explored. An extension of the work shall be to identify outliers for all domains of attraction, i.e. $\xi <=> 0$. In this direction, a trimmed version of the generalized hill estimator in *Beirlant et al.* (2004b) shall provide a starting point of the research.

We have so far quantified the notion of heat waves and presented their m-year return levels as a function of covariates like Season, Enso and Location. A comparison of the proposed techniques to already existent methods in literature needs to be made. Additionally, the proposed methodology needs to be applied on other climatic databases[2]. Sensitivity of the proposed methodology to other climatic events like cold waves, precipitation and depression shall be explored as a part of the future work.

[2] North American Regional Climate Change Assessment Program (NARCAP) `http://www.narccap.ucar.edu/data/`

APPENDICES

APPENDIX A

AMON

Proof of Proposition II.2. This result is a simple consequence of Theorem 3.3.7, p. 131 in *Embrechts et al.* (1997).

By the independence of the $X(i)'s$, for all fixed $x > 0$, we have

$$\mathbb{P}(m^{-1/\alpha})D_m(X) \leq x) = \mathbb{P}(X \leq m^{1/\alpha}x)^m = (1 - \mathbb{P}(X > m^{1/\alpha}x))^m$$

Since $\mathbb{P}(X > x) \sim c/x^\alpha$, thus with x replaced by $m^{1/\alpha}x$, we have

$$\mathbb{P}(m^{-1/\alpha}D_m(X) \leq x) \sim (1 - c/mx^\alpha)^m \to e^{-c/x^\alpha}, \quad m \to \infty$$

This implies the convergence in (2.13) since for a standard Fréchet distributed random variable Z, $P(Z \leq c^{-1/\alpha}x) = e^{-c/x^\alpha}$.

□

Proof of Proposition II.6. Part (i) is a direct consequence of (2.17). From part (i), we have

$$\{V(k;m),\ k=1,\ldots,m\} \overset{d}{=} \Big\{\frac{\sum_{j=1}^{k}\overline{F}^{-1}(\Gamma_j/\Gamma_{m+1})}{\sum_{j=1}^{m}\overline{F}^{-1}(\Gamma_j/\Gamma_{m+1})},\ k=1,\ldots,m\Big\}. \tag{A.1}$$

From Lemma B.3, we have $\Gamma_j/\Gamma_{m+1} \sim \Gamma_j/m$ as $m \to \infty$ almost surely for all $j = 1,\cdots,\ell$. In view of (2.11), $\overline{F}^{-1}(p) \sim (p/c)^{-1/\alpha}$ as $p \downarrow 0$ and hence for all $j = 1,\cdots,\ell$, with probability one, we have

$$\overline{F}^{-1}\Big(\frac{\Gamma_j}{\Gamma_{m+1}}\Big) \sim \Big(\frac{\Gamma_j}{c\Gamma_{m+1}}\Big)^{-1/\alpha}, \quad m \to \infty.$$

This implies that the right hand side of (A.1) converges almost surely to

$$\frac{\sum_{j=1}^{k}\Gamma_j^{-1/\alpha}(c\Gamma_{m+1})^{1/\alpha}}{\sum_{j=1}^{\ell}\Gamma_j^{-1/\alpha}(c\Gamma_{m+1})^{1/\alpha}} = W_\alpha(k,\ell),$$

which completes the proof of Part (ii). □

APPENDIX B

Robust Hill

Lemma B.1. *Let* $E_j \overset{i.i.d}{\sim} \mathrm{Exp}(1)$*,* $j = 1, 2, \cdots, n+1$ *be standard exponential random variables. Then, the* $\mathrm{Gamma}(i, 1)$ *random variables defined as*

$$\Gamma_i = \sum_{j=1}^{i} E_j \quad i = 1, \cdots, n+1, \tag{B.1}$$

satisfy

$$\Big(\frac{\Gamma_1}{\Gamma_{n+1}}, \cdots, \frac{\Gamma_n}{\Gamma_{n+1}}\Big) \text{ and } \Gamma_{n+1} \text{ are independent.} \tag{B.2}$$

and

$$\Big(\frac{\Gamma_1}{\Gamma_{n+1}}, \cdots, \frac{\Gamma_n}{\Gamma_{n+1}}\Big) \overset{d}{=} (U_{(1,n)}, \cdots, U_{(n,n)}) \tag{B.3}$$

where $U_{(1,n)} < \cdots < U_{(n,n)}$ *are the order statistics of* n *i.i.d. U(0,1) random variables.*

For details on the proof see Example 4.6 on page 44 in *Ahsanullah et al.* (2013). The next result, quoted from page 37 in *de Haan* (2006), shall be used throughout the course of the chapter to switch between order statistics of exponentials and i.i.d. exponential random variables.

Lemma B.2 (Rényi, 1953). *Let $E_1, E_2, \cdots, E_n$ be a sample of n i.i.d. standard exponential random variables and $E_{(1,n)} \leq E_{(2,n)} \leq \cdots \leq E_{(n,n)}$ be the order statistics. By Rényi's (1953) representation, we have for fixed $k \leq n$,*

$$\left(E_{(1,n)}, \cdots, E_{(i,n)}, \cdots, E_{(k,n)}\right) \overset{d}{=} \Big(\frac{E_1^*}{n}, \cdots, \sum_{j=1}^{i} \frac{E_j^*}{n-j+1}, \cdots, \sum_{j=1}^{k} \frac{E_j^*}{n-j+1}\Big) \quad \text{(B.4)}$$

where $E_1^*, \cdots, E_k^*$ are also i.i.d. standard exponentials.

Lemma B.3. *For $\Gamma_m = E_1 + E_2 + \cdots + E_m$ where the $E_i's$ are i.i.d. standard exponential random variables, for any ρ*

$$\sup_{m \geq M} \Big|\Big(\frac{\Gamma_m}{m}\Big)^{-\rho} - 1\Big| \quad \xrightarrow{a.s.} \quad 0, \;\; M \to \infty \quad \text{(B.5)}$$

$$\sup_{m,n \geq M} \Big|\Big(\frac{\Gamma_m/m}{\Gamma_n/n}\Big)^{-\rho} - 1\Big| \quad \xrightarrow{a.s.} \quad 0, \;\; M \to \infty \quad \text{(B.6)}$$

Proof. The proof is a direct consequence of the Strong Law of Large Numbers (SLLN).

□

Lemma B.4. *For all $\rho > 0$, we have*

$$\sup_{m \geq M} \Big|\frac{1}{m} \sum_{i=1}^{m} \Big(\frac{\Gamma_{i+1}}{\Gamma_{m+1}}\Big)^{\rho} - \frac{1}{1-\rho}\Big| \xrightarrow{a.s.} 0, \;\; M \to \infty$$

Proof. It is equivalent to show that, as $m \to \infty$,

$$\Big|\frac{1}{m} \sum_{i=1}^{m} \Big(\frac{\Gamma_{i+1}}{\Gamma_{m+1}}\Big)^{\rho} - \frac{1}{1+\rho}\Big| \xrightarrow{a.s.} 0. \quad \text{(B.7)}$$

For a fixed $\omega \in \Omega$, let us define the following sequence of functions

$$f_m(x) = \sum_{i=1}^{m} (\Gamma_{i+1}/\Gamma_{m+1})^{\rho}(\omega) \mathbf{1}_{(\frac{i-1}{m}, \frac{i}{m}]}(x), \qquad x > 0$$

Suppose $x \in ((i-1)/m, i/m]$, then

$$f_m(x) = (\Gamma_{[mx]+1}/\Gamma_{m+1})^{-\rho}(\omega) = \Big(\frac{[mx]+1}{m}\Big)^{-\rho}\Big(\frac{\Gamma_{[mx]+1}/([mx]+1)}{\Gamma_m/m}\Big)^{\rho}(\omega) \to x^{-\rho} \tag{B.8}$$

where the convergence follows from (B.6). Moreover since $\Gamma_{[mx]+1} < \Gamma_m$ and $\rho < 0$, therefore $|f_m(x)| \le 1$, for all $x > 0$. Thus by dominated convergence theorem,

$$\int_0^1 f_m(x)dx = \frac{1}{m}\sum_{i=1}^{m}(\Gamma_{i+1}/\Gamma_{m+1})^{-\rho}(\omega) \to \int_0^1 x^{-\rho}dx = \frac{1}{1-\rho} \tag{B.9}$$

Since (B.8) hold for all $\omega \in \Omega$ with $P[\Omega] = 1$, so does (B.9). This completes the proof. □

Lemma B.5. *If E_i's $i = 1, \cdots, n$ are i.i.d. observations from* $\mathrm{Exp}(\xi)$*, the best linear unbiased estimator (BLUE) of ξ based on the order statistics, $E_{(1,n)} < \cdots < E_{(r,n)}$ is given by*

$$\hat{\xi} = \frac{1}{r}\sum_{i=1}^{r-1} E_{(i,n)} + \frac{n-r+1}{r}E_{(r,n)}$$

Proof. Let $\widehat{\xi} = \sum_{i=1}^{r}\gamma_i E_{(i,n)}$ denote the BLUE of ξ. By Relation (B.4) in Lemma B.2, the BLUE can then be expressed as

$$\hat{\xi} = \sum_{i=1}^{r}\gamma_i\sum_{j=1}^{i}\frac{E_j^*}{(n-j+1)} = \sum_{j=1}^{r}E_j^*\sum_{i=j}^{r}\frac{\gamma_i}{(n-j+1)} =: \sum_{j=1}^{r}E_j^*\delta_j \tag{B.10}$$

where the E_j^* are i.i.d. from $\mathrm{Exp}(\xi)$.

For i.i.d. observations from $\mathrm{Exp}(\xi)$, the sample mean is the uniformly minimum variance unbiased estimator for ξ (see Lehmann Scheffe Theorem, Theorem 1.11, page 88 in *Lehmann and Casella* (1983)). Thus $\delta_j = 1/r$ yields the required best linear unbiased estimator.

Using the fact that $\sum_{i=j}^{r}\gamma_i = \delta_j(n-j+1) = (n-j+1)/r$, we obtain

$$\gamma_i = \begin{cases} \frac{n-r+1}{r} & i = r \\ \\ \frac{1}{r} & i < r \end{cases}$$

This completes the proof. □

Lemma B.6. *Suppose g is $-\rho$-varying for $\rho \geq 0$ and $Y_{(n-k,n)}$ is the $(k+1)^{th}$ order statistic for n observations from* Pareto$(1,1)$*, then*

$$\frac{g(Y_{(n-k,n)})}{g(n/k)} \xrightarrow{P} 1 \tag{B.11}$$

provided $k \to \infty$, $n \to \infty$ and $k/n \to \infty$.

Proof. Since g is $-\rho$ varying, g may be expressed as $g(t) = t^{-\rho}l(t)$, for some slowly varying function $l(\cdot)$. Thus, we have

$$\frac{g(Y_{(n-k,n)})}{g(n/k)} = \Big(\frac{Y_{(n-k,n)}}{n/k}\Big)^{-\rho}\frac{l(Y_{(n-k,n)})}{l(n/k)}$$

From (B.12), we have $Y_{(n-k,n)} \overset{d}{=} \Gamma_{n+1}/\Gamma_{k+1}$ and therefore, by weak law of large numbers, we have $Y_{(n-k,n)}/(n/k) \xrightarrow{P} 1$.

Thus to prove (B.11), it suffices to show $l(Y_{(n-k,n)})/l(n/k) \xrightarrow{P} 1$. In this direction, observe that for some $\delta > 0$, we have

$$\begin{aligned} P\Bigg[\Big|\frac{l(Y_{(n-k,n)})}{l(n/k)} - 1\Big| > \varepsilon\Bigg] &\leq P\Bigg[\Big|\frac{l(Y_{(n-k,n)})}{l(n/k)} - 1\Big| > \varepsilon, \Big|\frac{Y_{(n-k,n)}}{n/k} - 1\Big| \leq \delta\Bigg] + P\Bigg[\Big|\frac{Y_{(n-k,n)}}{n/k} - 1\Big| > \delta\Bigg] \\ &\leq P\Bigg[\sup_{\lambda\in[1-\delta,1+\delta]}\Big|\frac{l(\lambda n/k)}{l(n/k)} - 1\Big| > \varepsilon\Bigg] + P\Bigg[\Big|\frac{Y_{(n-k,n)}}{n/k} - 1\Big| > \delta\Bigg] \end{aligned}$$

The first term on the right hand side goes to 0 by Theorem 1.5.2 on page 22 in

Bingham et al. (1989). The second term goes to 1 since $Y_{(n-k,n)}/(n/k) \xrightarrow{P} 1$. □

Proof of Proposition III.1. Observe that X_i's can be alternatively written as

$$X_i = \sigma U_i^{-\xi}, \quad i = 1, \cdots, n,$$

where U_i's are i.i.d. $U(0,1)$. Therefore by Relation (B.3) in Lemma B.1, we have

$$(X_{(n,n)}, \cdots, X_{(1,n)}) = \sigma(U_{(1,n)}^{-\xi}, \cdots, U_{(n,n)}^{-\xi}) \overset{d}{=} \sigma\left(\left(\frac{\Gamma_1}{\Gamma_{n+1}}\right)^{-\xi}, \cdots, \left(\frac{\Gamma_n}{\Gamma_{n+1}}\right)^{-\xi}\right) \tag{B.12}$$

where $X_{(n,n)} > \cdots > X_{(1,n)}$ are the order statistics for the X_i's. Hence, for all $1 \leq k \leq n-1$, we have

$$\begin{aligned}\left(\log\left(\frac{X_{(n,n)}}{X_{(n-k,n)}}\right), \cdots, \log\left(\frac{X_{(k)}}{X_{(n-k,n)}}\right)\right) &\overset{d}{=} -\xi\left(\log\left(\frac{\Gamma_1}{\Gamma_{k+1}}\right), \cdots, \log\left(\frac{\Gamma_k}{\Gamma_{k+1}}\right)\right) \\ &\overset{d}{=} -\xi(\log U_{(1,k)}, \cdots, \log U_{(k,k)}),\end{aligned} \tag{B.13}$$

where the $U_{(i,k)}$'s are the order statistics for a sample of k i.i.d. $U(0,1)$ and the last equality in (B.13) follows from Relation (B.3) in Lemma B.1. Since negative log transforms of $U(0,1)$ are standard exponentials, one can define $E_{(i,k)}$, $i = 1, \cdots, k$ as

$$\left(\log\left(\frac{X_{(n,n)}}{X_{(n-k,n)}}\right), \cdots, \log\left(\frac{X_{(k)}}{X_{(n-k,n)}}\right)\right) =: (E_{(k,k)}, \cdots, E_{(1,k)}) \tag{B.14}$$

such that the $E_{(i,k)}$'s are distributed as order statistics of k i.i.d. exponentials with mean ξ, henceforth denoted by $\text{Exp}(\xi)$. One can thereby simplify $\widehat{\xi}_{k_0,k}^{\text{trim}}$ in (3.3) as

$$\widehat{\xi}_{k_0,k}^{\text{trim}} = \sum_{i=k_0+1}^{k} c_{k_0,k}(n-i+1,n) E_{(k-i+1,k)} = \sum_{i=1}^{k-k_0} \delta_i E_{(i,k)} \tag{B.15}$$

where $\delta_i = c_{k_0,k}(k-i+1)$. The optimal choice of weights δ_i's which produce the best linear unbiased estimator (BLUE) for ξ is obtained using Lemma B.5 as:

$$\delta_i^{\mathrm{opt}} = \begin{cases} \frac{1}{k-k_0} & i = 1, \cdots, k-k_0-1 \\ \\ \frac{k_0+1}{k-k_0} & i = k-k_0 \end{cases} \tag{B.16}$$

Rewriting $E_{(i,k)}$'s in terms of $X_{(n-i+1,n)}$'s as in (B.14) completes the proof. □

Proof of Proposition III.2. From (B.15) and (B.16) in Proposition III.1, we have

$$\left\{\widehat{\xi}_{k_0,k},\ k_0 = 0,\ldots,k-1\right\} = \left\{\frac{1}{k-k_0}\sum_{i=1}^{k-k_0-1} E_{(i,k)} + \frac{k_0+1}{k-k_0}E_{(k-k_0,k)},\ k_0 = 0,\ldots,k-1\right\} \tag{B.17}$$

Using Relation (B.4) from Lemma B.2, for all $k_0 = 0, 1, \cdots, k-1$, we have

$$\widehat{\xi}_{k_0,k} = \frac{1}{k-k_0}\sum_{i=1}^{k-k_0-1}\sum_{j=1}^{i}\frac{E_j^*}{(k-j+1)} + \frac{k_0+1}{k-k_0}\sum_{j=1}^{k-k_0}\frac{E_j^*}{(k-j+1)} \tag{B.18}$$

Interchanging the order of summation in the first term in the right hand side of (B.18), for $k_0 = 0, 1, \cdots, k-1$, we obtain

$$\begin{aligned} \widehat{\xi}_{k_0,k} &= \sum_{j=1}^{k-k_0-1}\frac{E_j^*}{k-j+1}\sum_{i=j}^{k-k_0-1}\frac{1}{k-k_0} + \frac{k_0+1}{k-k_0}\sum_{j=1}^{k-k_0}\frac{E_j^*}{(k-j+1)} \\ &= \sum_{j=1}^{k-k_0-1}\frac{E_j^*}{k-j+1}\left(\sum_{i=j}^{k-k_0-1}\frac{1}{k-k_0} + \frac{k_0+1}{k-k_0}\right) + \frac{E_{k-k_0}^*}{k-k_0} \\ &= \sum_{j=1}^{k-k_0-1}\frac{E_j^*}{k-j+1}\frac{(k-j+1)}{k-k_0} + \frac{E_{k-k_0}^*}{k-k_0} \\ &= \frac{1}{k-k_0}\sum_{j=1}^{k-k_0}E_j^*, \end{aligned}$$

Since E_j^*, $j = 1, \cdots, k-k_0$ are rescaled i.i.d. standard exponentials, Relation (3.6) follows.

The covariance structure in (3.7) readily follows from (3.6) and the fact that

$$\mathrm{Cov}\Big(\frac{\Gamma_i}{i}, \frac{\Gamma_j}{j}\Big) = \mathrm{Cov}\Big(\frac{E_1+\cdot\cdot+E_i}{i}, \frac{E_1+\cdot\cdot+E_j}{j}\Big) = \frac{i\wedge j}{ij} = \frac{1}{i\vee j},\ i,j=0,1,\cdots,k$$

where $\vee$ denotes the max operator. This completes the proof. □

Proof of Theorem III.4. Suppose for the moment σ is known and consider the class of statistics:

$$\mathcal{U}_{k_0}^{\sigma} = \Big\{T = T(X_{(n-k_0,n)},\cdots,X_{(1,n)}):\ \mathbb{E}(T)=\xi,\ X_1,\cdots,X_n \overset{i.i.d.}{\sim} \mathrm{Pareto}(\sigma,\xi)\Big\}.$$

Since σ is no longer a parameter, every statistic in $\mathcal{U}_{k_0}^{\sigma}$ can be equivalently written as a function of $\log(X_{(n-i+1,n)}/\sigma)$, $i=k_0+1,\cdots,n$. Therefore, the set of random variables in $\mathcal{U}_{k_0}^{\sigma}$ equals

$$\mathcal{U}_{k_0}^{\sigma} = \left\{S = S\left(\log\Big(\frac{X_{(n-k_0,n)}}{\sigma}\Big),\cdots,\log\Big(\frac{X_{(1,n)}}{\sigma}\Big)\right):\mathbb{E}(S)=\xi, X_1,\cdots,X_n \overset{i.i.d.}{\sim} \mathrm{Pareto}(\sigma,\xi)\right\}.$$

Since X_i's follow Pareto(σ,ξ), we have $\log(X_i/\sigma)\sim \mathrm{Exp}(\xi)$ and therefore

$$\left(\log\Big(\frac{X_{(n-k_0,n)}}{\sigma}\Big),\cdots,\log\Big(\frac{X_{(1,n)}}{\sigma}\Big)\right) \overset{d}{=} \big(E_{(n-k_0,n)},\cdots,E_{(1,n)}\big),$$

where $E_{(1,n)}\le\cdots\le E_{(n,n)}$ are the order statistics of n i.i.d. observations from Exp(ξ). Therefore

$$\mathcal{U}_{k_0}^{\sigma} \overset{d}{=} \Big\{S = S(E_{(n-k_0,n)},\cdots,E_{(1,n)}):\ \mathbb{E}(S)=\xi,\ E_1,\cdots,E_n \overset{i.i.d.}{\sim} \mathrm{Exp}(\xi)\Big\}, \quad \text{(B.19)}$$

where we observe that the distribution of the E_i's does not depend on σ.

Using Relation (B.4) from Lemma B.2, we have

$$\begin{aligned} S(E_{(n-k_0,n)},\cdots,E_{(1,n)}) &= S\Big(\sum_{j=1}^{n-k_0}\frac{E_j^*}{n-j+1},\cdots,\sum_{j=1}^{n-k}\frac{E_j^*}{n-j+1}\Big) \\ &= R(E_1^*,\cdots,E_{n-k_0}^*) \end{aligned}$$

Using this on the right hand side of (B.19), we get

$$\mathcal{U}_{k_0}^{\sigma} \overset{d}{=} \mathcal{V}_{k_0} := \Big\{R = R(E_1^*,\cdots,E_{n-k_0}^*) : \mathbb{E}(R) = \xi,\ E_1^*,\cdots,E_{n-k_0}^* \overset{i.i.d.}{\sim} \text{Exp}(\xi)\Big\}. \tag{B.20}$$

where the first equality is in the sense of finite dimensional distributions.

Therefore, $L = \inf_{T\in\mathcal{U}_{k_0}^{\sigma}} \text{Var}(T) = \inf_{R\in\mathcal{V}_{k_0}} \text{Var}(R)$. The quantity L can be easily obtained as

$$L = \text{Var}(\overline{E}_{n-k_0}^*) = \frac{\xi^2}{n-k_0} \tag{B.21}$$

since the sample mean, $\overline{E}_{n-k_0}^* = \sum_{i=1}^{n-k_0} E_i^*/(n-k_0)$ is uniformly minimum variance estimator (UMVUE), for ξ among the class described by $\mathcal{V}_{k_0}$. This follows from the fact that $\overline{E}_{n-k_0}^*$ is an unbiased and complete sufficient statistic for ξ (see Lehmann Scheffe Theorem, Theorem 1.11, page 88 in *Lehmann and Casella* (1983)).

To complete the proof, observe that every statistic, T in $\mathcal{U}_{k_0}$ is an unbiased estimator of ξ for any arbitrary choice of σ. This implies that $T \in \mathcal{U}_{k_0}^{\sigma}$ and therefore $L \le \inf_{T\in\mathcal{U}_{k_0}} \text{Var}(T)$, which yields the lower bound in (3.9).

For the upper bound in (3.9), we observe that $\widehat{\xi}_{k_0,n-1} \in \mathcal{U}_{k_0}$, which in view of Proposition III.2 implies

$$\inf_{T\in\mathcal{U}_{k_0}} \text{Var}(T) \le \text{Var}(\widehat{\xi}_{k_0,n-1}) = \frac{\xi^2}{n-k_0-1}.$$

This completes the proof. □

We shall next present the proof of Theorem III.5. To begin with we state the following two lemmas which shall be used as a part of the proof.

Lemma B.7.

$$\max_{0\le k_0<k}\left|\frac{k-k_0}{kg(Y_{(n-k,n)})}S_{k_0,k}+\frac{c}{1+\rho}\left(\frac{k_0}{k}\right)^{1+\rho}-\frac{c}{1+\rho}\right|\xrightarrow{P}0. \tag{B.22}$$

where $S_{k_0,k}$ is defined as

$$S_{k_0,k}:=\frac{cg(Y_{(n-k,n)})}{k-k_0}\Big((k_0{+}1)\int\limits_1^{Y_{(n-k_0,n)}/Y_{(n-k,n)}}\nu^{-\rho-1}d\nu{+}\sum_{i=k_0+2}^{k}\int\limits_1^{Y_{(n-i+1,n)}/Y_{(n-k,n)}}\nu^{-\rho-1}d\nu\Big). \tag{B.23}$$

where Y_i's are n i.i.d observations from Pareto(1,1)

Proof. The proof of (B.22) involves two cases: $\rho>0$ and $\rho=0$.

Case $\rho>0$: Using the expression of $S_{k_0,k}$ in (B.23), we get

$$\begin{aligned}\frac{k-k_0}{kg(Y_{(n-k,n)})}S_{k_0,k} &= -\frac{c}{k\rho}\left((k_0+1)\Big(\frac{Y_{(n-k_0,n)}}{Y_{(n-k,n)}}\Big)^{-\rho}+\sum_{i=k_0+2}^{k}\Big(\frac{Y_{(n-i+1,n)}}{Y_{(n-k,n)}}\Big)^{-\rho}-k\right)\\ &= \frac{c}{k\rho}\sum_{i=1}^{k_0}\left\{\Big(\frac{Y_{(n-i+1,n)}}{Y_{(n-k,n)}}\Big)^{-\rho}-\Big(\frac{Y_{(n-k_0,n)}}{Y_{(n-k,n)}}\Big)^{-\rho}\right\}\\ &\quad -\frac{c}{k\rho}\sum_{i=1}^{k}\left\{\Big(\frac{Y_{(n-i+1,n)}}{Y_{(n-k,n)}}\Big)^{-\rho}-1\right\}\end{aligned} \tag{B.24}$$

Expressing the order statistics of Pareto in terms of Gamma random variables as in (B.12), we get

$$\frac{k-k_0}{kg(Y_{(n-k,n)})}S_{k_0,k}+c_{1+\rho}\left(\frac{k_0}{k}\right)^{1+\rho} \overset{d}{=} \underbrace{c_{1+\rho}\left(\frac{k_0}{k}\right)^{1+\rho}+\frac{c_\rho}{k}\sum_{i=1}^{k_0}\left\{\left(\frac{\Gamma_{i+1}}{\Gamma_{k+1}}\right)^\rho-\left(\frac{\Gamma_{k_0+1}}{\Gamma_{k+1}}\right)^\rho\right\}}_{B_{k_0,k}} - \underbrace{\frac{c_\rho}{k}\sum_{i=1}^{k}\left\{\left(\frac{\Gamma_{i+1}}{\Gamma_{k+1}}\right)^\rho-1\right\}}_{A_k}$$

with $c_t = c/t$.

To prove (B.22), we first show that $\max_{0\le k_0<k}|A_k + c/(1+\rho)| \xrightarrow{a.s.} 0$. For this (B.7), we have $|(1/k)\sum_{i=1}^k(\Gamma_{i+1}/\Gamma_{k+1})^\rho - 1/(1+\rho)| \xrightarrow{a.s.} 0$. Therefore for any $\omega \in \Omega$ with $P[\Omega] = 1$,

$$\left|A_k(\omega)+\frac{c}{1+\rho}\right| = \left|\frac{c}{\rho k}\sum_{i=1}^{k}\left(\frac{\Gamma_{i+1}}{\Gamma_{k+1}}\right)^\rho(\omega)-\frac{c}{\rho}+\frac{c}{1+\rho}\right| = \left|\frac{c_\rho}{k}\sum_{i=1}^{k}\left(\frac{\Gamma_{i+1}}{\Gamma_{k+1}}\right)^\rho(\omega)-\frac{c_\rho}{1+\rho}\right|\to 0$$

We next show that $\max_{0\le k_0<k} B_{k_0,k} \xrightarrow{a.s.} 0$. For this observe that for any $\omega \in \Omega$,

$$\begin{aligned}\max_{0\le k_0<M} B_{k_0,k}(\omega) &\le \max_{0\le k_0<M}\left\{c_{1+\rho}\left(\frac{k_0}{k}\right)^{1+\rho}+\frac{c_\rho}{k}\sum_{i=1}^{k_0}\left|\left(\frac{\Gamma_{i+1}}{\Gamma_{k+1}}\right)^\rho(\omega)-\left(\frac{\Gamma_{k_0+1}}{\Gamma_{k+1}}\right)^\rho(\omega)\right|\right\}\\ &\le \max_{0\le k_0<M}\left\{c_{1+\rho}\left(\frac{k_0}{k}\right)^{1+\rho}+c_\rho\frac{2k_0}{k}\right\} \quad (\text{since } (\Gamma_i/\Gamma_{k+1})^\rho \le 1,\ 1\le i\le k,\ \rho>0)\\ &\le \frac{c^{1+\rho}M^{1+\rho}+2c_\rho M}{k} = \frac{B_{0M}}{k} \qquad \text{(B.25)}\end{aligned}$$

Additionally,

$$\begin{aligned}
\max_{M\le k_0<k} \quad B_{k_0,k} \quad (\omega) &\le \max_{M\le k_0<k}\left|c_{1+\rho}\Big(\frac{k_0}{k}\Big)^{1+\rho}+\frac{c_\rho}{k}\sum_{i=1}^{k_0}\Big\{\Big(\frac{\Gamma_{i+1}}{\Gamma_{k+1}}\Big)^{\rho}(\omega)-\Big(\frac{\Gamma_{k_0+1}}{\Gamma_{k+1}}\Big)^{\rho}(\omega)\Big\}\right| \\
&\le \quad \max_{M\le k_0<k}\Big(\frac{k_0}{k}\Big)^{1+\rho}\left|\frac{c_{1+\rho}}{c_\rho}+\Big(\frac{k_0}{k}\Big)^{\rho}\frac{1}{k_0}\sum_{i=1}^{k_0}\Big\{\Big(\frac{\Gamma_{i+1}}{\Gamma_{k+1}}\Big)^{\rho}(\omega)-\Big(\frac{\Gamma_{k_0+1}}{\Gamma_{k+1}}\Big)^{\rho}(\omega)\Big\}\right| \\
&\le \quad \max_{M\le k_0<k}\left|\frac{\rho}{1+\rho}+\Big(\frac{\Gamma_{k_0+1}/k_0}{\Gamma_{k+1}/k}\Big)^{\rho}(\omega)\Big\{\underbrace{\frac{1}{k_0}\sum_{i=1}^{k_0}\Big(\frac{\Gamma_{i+1}}{\Gamma_{k_0+1}}\Big)^{\rho}(\omega)}_{C_{k_0}(\omega)}-1\Big\}\right| \\
&= \quad \max_{M\le k_0<k}\left|\frac{\rho}{1+\rho}+(C_{k_0}(\omega)-1)+(C_{k_0}(\omega)-1)\Big\{\Big(\frac{\Gamma_{k_0+1}/k_0}{\Gamma_{k+1}/k}\Big)^{\rho}-1\Big\}\right| \qquad \text{(B.26)}
\end{aligned}$$

Since $\Gamma_{i+1} < \Gamma_{k_0+1}$ and $\rho > 0$, thereby $|C_{k_0}| < 1$. This allows us to simplify (B.26)

$$\max_{M\le k_0<k} B_{k_0,k}(\omega) \quad \le \quad \underbrace{\sup_{M\le k_0}\left|C_{k_0}(\omega)-\frac{1}{1+\rho}\right|}_{B_{1M}(\omega)} +2\underbrace{\sup_{M\le k_0,k}\left|\Big(\frac{\Gamma_{k_0+1}/k_0}{\Gamma_{k+1}/k}\Big)^{\rho}(\omega)-1\right|}_{B_{2M}(\omega)}$$

Thus

$$\max_{0\le k_0<k} B_{k_0,k}(\omega) \le \frac{B_{0M}}{k} + B_{1M}(\omega) + B_{2M}(\omega).$$

Taking $\limsup$ w.r.t to k on both sides we get

$$\limsup_{k\to\infty}\max_{0\le k_0<k} B_{k_0,k}(\omega) \le B_{1M}(\omega) + B_{2M}(\omega) \tag{B.27}$$

Using Lemmas B and B.3 shows that $B_{1M}(\omega) \to 0$ and $B_{2M}(\omega) \to 0$ for all $\omega \in \Omega$ with $P[\Omega] = 1$.

Thus taking $\limsup$ w.r.t M on both sides of (B.27) completes the proof for $\rho < 0$.

Case $\rho = 0$**:** Using the expression of $S_{k_0,k}$ in (B.23), we get

$$\begin{aligned}
\frac{k-k_0}{kg(Y_{(n-k,n)})}S_{k_0,k} + \frac{ck_0}{k} &= \frac{c}{k}\left((k_0+1)\log\Big(\frac{Y_{(n-k_0,n)}}{Y_{(n-k,n)}}\Big) + \sum_{i=k_0+2}^{k}\log\Big(\frac{Y_{(n-i+1,n)}}{Y_{(n-k,n)}}\Big)\right) + \frac{ck_0}{k} \\
&\overset{d}{=} \frac{c(k-k_0)}{k}\widehat{\xi}^{**}_{k_0,k} + \frac{ck_0}{k} \\
&\overset{d}{=} c\Big(\frac{\Gamma_{k-k_0}}{k} - \frac{k-k_0}{k} + 1\Big) \qquad \text{(B.28)}
\end{aligned}$$

where $\widehat{\xi}^{**}_{k_0,k}$ is the trimmed Hill estimator in (3.5) with X_i's replaced by the i.i.d. Pareto$(1,1)$.

Thus to prove (B.22), it suffices to show $\max_{0\le k_0<k}|\Gamma_{k-k_0} - (k-k_0)|/k \xrightarrow{a.s.} 0$. For every ω in Ω with $P[\Omega] = 1$, we have

$$\begin{aligned}
\max_{0\le k_0<k}\frac{|\Gamma_{k-k_0}(\omega) - (k-k_0)|}{k} &= \max_{0\le k_0<k}\frac{(k-k_0)}{k}\Big|\frac{\Gamma_{k-k_0}}{k-k_0}(\omega) - 1\Big| \qquad \text{(B.29)} \\
&\le \frac{M}{k}\underbrace{\max_{0\le k-k_0<M}\Big|\frac{\Gamma_{k-k_0}}{k-k_0}(\omega)-1\Big|}_{F_{1M}(\omega)} + \underbrace{\sup_{k-k_0\ge M}\Big|\frac{\Gamma_{k-k_0}}{k-k_0}(\omega)-1\Big|}_{F_{2M}(\omega)}
\end{aligned}$$

Observe that by the SLLN, $|\Gamma_n/n - 1| \xrightarrow{a.s.} 0$. Therefore $\sup_n|\Gamma_n(\omega)/n-1|$ is bounded for all $\omega\in\Omega$ with $P[\Omega] = 1$. This implies $F_{1M}(\omega) \le \sup_n|\Gamma_n(\omega)/n-1|$ is bounded.

Thus taking lim sup with respect to k on both sides of (B.29) we get

$$\limsup_{k\to\infty}\max_{0\le k_0<k}\frac{|\Gamma_{k-k_0}(\omega) - (k-k_0)|}{k} \le F_{2M}(\omega)$$

Taking $\lim_{M\to\infty}$ on both sides and using (B.6), the proof follows. □

Lemma B.8. *Assumption* (3.13) *imply*

$$\max_{0\le k_0\le k}\Big(\frac{k-k_0}{kg(Y_{(n-k,n)})}|R_{k_0,k} - S_{k_0,k}|\Big) \xrightarrow{P} 0 \qquad \text{(B.30)}$$

where $R_{k_0,k}$ *and* $S_{k_0,k}$ *are defined in* (3.12) *and* (B.23)*, respectively.*

Proof. The proof of (B.30) involves two cases: $\rho > 0$ and $\rho = 0$.

Case $\rho > 0$: Since $Y_{(n-i+1,n)}/Y_{(n-k,n)} > 1$, $i = 1, \cdots, k$, over the event $\{Y_{(n-k,n)} > t_\varepsilon\}$, by (3.13):

$$\begin{aligned}
(k-k_0) \quad |R_{k_0,k} - S_{k_0,k}| &\leq (k_0+1)\left|\log\frac{L(Y_{(n-k_0,n)})}{L(Y_{(n-k,n)})} - cg(Y_{(n-k,n)}\int_1^{Y_{(n-k_0,n)}/Y_{(n-k,n)}} \nu^{-\rho-1}d\nu\right| \\
+ \quad & \sum_{i=k_0+2}^{k}\left|\log\frac{L(Y_{(n-i+1,n)})}{L(Y_{(n-k,n)})} - cg(Y_{(n-k,n)}\int_1^{Y_{(n-i+1,n)}/Y_{(n-k,n)}} \nu^{-\rho-1}d\nu\right| \\
\leq \quad & (k_0+1)g(Y_{(n-k,n)})\varepsilon + \sum_{i=k_0+2}^{k} g(Y_{(n-k,n)})\varepsilon = g(Y_{(n-k,n)})k\varepsilon.
\end{aligned}$$

Therefore over the event $\{Y_{(n-k,n)} > t_\varepsilon\}$

$$\max_{0\leq k_0\leq k}\Big(\frac{k-k_0}{kg(Y_{(n-k,n)})}|R_{k_0,k} - S_{k_0,k}|\Big) \leq \varepsilon. \tag{B.31}$$

From (B.12) we get $Y_{(n-k,n)} \overset{d}{=} (\Gamma_{k+1}/\Gamma_{n+1})^{-1}$. By Lemma B.3, we have

$$Y_{(n-k,n)} \overset{d}{=} \frac{n}{k}\Big(\frac{\Gamma_{k+1}/k}{\Gamma_{n+1}/n}\Big)^{-1} \xrightarrow{P} \infty$$

which implies $P[Y_{(n-k,n)} > t_\varepsilon] \to 1$ and hence completes the proof.

Case $\rho = 0$: As in the previous case, over the event $\{Y_{(n-k,n)} > t_\varepsilon\}$, by (3.13) we have

$$\begin{aligned}
(k-k_0) \quad |R_{k_0,k} - S_{k_0,k}| &= (k_0+1)\left|\log\frac{L(Y_{(n-k_0,n)})}{L(Y_{(n-k,n)})} - cg(Y_{(n-k,n)}\int_1^{Y_{(n-k_0,n)}/Y_{(n-k,n)}} \frac{d\nu}{\nu}\right| \\
+ \quad & \sum_{i=k_0+2}^{k}\left|\log\frac{L(Y_{(n-i+1,n)})}{L(Y_{(n-k,n)})} - cg(Y_{(n-k,n)}\int_1^{Y_{(n-i+1,n)}/Y_{(n-k,n)}} \frac{d\nu}{\nu}\right| \\
\leq \quad & \varepsilon\Bigg((k_0+1)g(Y_{(n-k,n)})\Big(\frac{Y_{(n-k_0,n)}}{Y_{(n-k,n)}}\Big)^{\varepsilon} + \sum_{i=k_0+2}^{k} g(Y_{(n-k,n)})\Big(\frac{Y_{(n-i+1,n)}}{Y_{(n-k,n)}}\Big)^{\varepsilon}\Bigg)
\end{aligned}$$

Since $Y_{(n-i+1,n)} \geq Y_{(n-k_0,n)}$ for $i = 1, \cdots, k_0 + 1$, we further obtain

$$\max_{0 \leq k_0 \leq k} \Big(\frac{(k-k_0)}{kg(Y_{(n-k,n)})} |R_{k_0,k} - S_{k_0,k}| \Big) \leq \frac{\varepsilon}{k} \sum_{i=1}^{k} \Big(\frac{Y_{(n-i+1,n)}}{Y_{(n-k,n)}} \Big)^{\varepsilon} \leq 2\varepsilon \tag{B.32}$$

over the events $\{Y_{(n-k,n)} > t_\varepsilon\}$ and $\{(1/k) \sum_{i=1}^{k} (Y_{(n-i+1,n)}/Y_{(n-k,n)})^{\varepsilon} < 2\}$.

For $\{(1/k) \sum_{i=1}^{k} (Y_{(n-i+1,n)}/Y_{(n-k,n)})^{\varepsilon} < 2\}$, from (B.12), we observe that

$$\frac{1}{k} \sum_{i=1}^{k} \Big(\frac{Y_{(n-i+1,n)}}{Y_{(n-k,n)}} \Big)^{\varepsilon} \overset{d}{=} \frac{1}{k} \sum_{i=1}^{k} \Big(\frac{\Gamma_{i+1}}{\Gamma_{k+1}} \Big)^{-\varepsilon} = \frac{1}{k} \sum_{i=1}^{k} U_{i,k}^{-\varepsilon} \xrightarrow{P} \frac{1}{1-\varepsilon}$$

where the last convergence follows from weak law of large numbers.

Thus $P[(1/k) \sum_{i=1}^{k} (Y_{(n-i+1,n)}/Y_{(n-k,n)})^{\varepsilon} < 2] \to 1$ as long as $\varepsilon < 0.5$. Since we already proved that $P[Y_{(n-k,n)} > t_\varepsilon] \to 1$, the proof for the case $\rho < 0$ follows. □

Proof of Theorem III.5. Using (3.12), we can rewrite (3.16) as

$$k^{\delta} \max_{0 \leq k_0 < h(k)} \Big| R_{k_0,k} - \frac{k^{-\delta} cA}{(1+\rho)} \Big| \xrightarrow{P} 0 \tag{B.33}$$

In this direction, observe that

$$\Big| R_{k_0,k} - \frac{k^{-\delta} cA}{(1+\rho)} \Big| \leq |R_{k_0,k} - S_{k_0,k}| + \Big| S_{k_0,k} - \frac{k^{-\delta} cA}{(1+\rho)} \Big|$$

where $S_{k_0,k}$ is defined in (B.23).

To prove (B.33), we first show that $k^{\delta} \max_{0 \leq k_0 < h(k)} |R_{k_0,k} - S_{k_0,k}| \xrightarrow{P} 0$. In this direction, we have

$$\begin{aligned} k^{\delta}\max_{0\le k_0<h(k)}|R_{k_0,k}-S_{k_0,k}| &= k^{\delta}\max_{0\le k_0<h(k)}\frac{kg(Y_{n-k,n})}{k-k_0}\Big(\frac{k-k_0}{kg(Y_{(n-k,n)})}|R_{k_0,k}-S_{k_0,k}|\Big)\\ &\le \frac{\Delta_{1k}}{1-h(k)/k}\underbrace{\max_{0\le k_0<h(k)}\Big(\frac{k-k_0}{kg(Y_{(n-k,n)})}|R_{k_0,k}-S_{k_0,k}|\Big)}_{\Delta_{2k}} \end{aligned}$$

where $\Delta_{2k}\xrightarrow{P}0$ by Lemma B.8. Since $h(k)=o(k)$, $1-h(k)/k\to 1$ and $\Delta_{1k}=$ $k^{\delta}g(Y_{(n-k,n)})$

$$\Delta_{1k}=k^{\delta}g(n/k)\frac{g(Y_{(n-k,n)})}{g(n/k)}\xrightarrow{P}A \tag{B.34}$$

where (B.34) follows from assumption (4.9) and Lemma B.6.

Towards the proof of (B.33), we finally show that

$k^{\delta}\max_{0\le k_0<h(k)}|S_{k_0,k}-(k^{-\delta}cA)/(1+\rho)|\xrightarrow{P}0$. In this direction, we have

$$\begin{aligned} k^{\delta}\max_{0\le k_0<h(k)}\Big|S_{k_0,k}-\frac{k^{-\delta}cA}{(1+\rho)}\Big| &= k^{\delta}\max_{0\le k_0<h(k)}\frac{kg(Y_{n-k,n})}{k-k_0}\Big|\frac{k-k_0}{kg(Y_{(n-k,n)})}S_{k_0,k}-\frac{cA(k-k_0)}{k(1+\rho)\Delta_{1k}}\Big|\\ &\le \frac{\Delta_{1k}}{1-h(k)/k}\underbrace{\max_{0\le k_0<h(k)}\Big|\frac{k-k_0}{kg(Y_{(n-k,n)})}S_{k_0,k}-\frac{cA(k-k_0)}{k(1+\rho)\Delta_{1k}}\Big|}_{\Delta_{3k}} \end{aligned}$$

where $\Delta_{1k}\xrightarrow{P}A$ as in (B.34) and $1-h(k)/k\to 1$. Δ_{3k} can be further simplified as

$$\begin{aligned} \Delta_{3k} &\le \underbrace{\max_{0\le k_0<h(k)}\Big|\frac{k-k_0}{kg(Y_{(n-k,n)})}S_{k_0,k}+c\Big(\frac{k_0}{k}\Big)^{1+\rho}-\frac{c}{1+\rho}\Big|}_{\Delta_{4k}}\\ &+ \underbrace{\max_{0\le k_0<h(k)}\Big|\frac{c}{1+\rho}-c\Big(\frac{k_0}{k}\Big)^{1+\rho}-\frac{cA(k-k_0)}{k(1+\rho)\Delta_{1k}}\Big|}_{\Delta_{5k}} \end{aligned}$$

where $\Delta_{4k}\xrightarrow{P}0$ by Lemma B.7. Since $\max_{0\le k_0<k}(k_0/k)^{1+\rho}\le(h(k)/k)^{1+\rho}\to 0$, thus

to prove $\Delta_{5k} \xrightarrow{P} 0$, it suffices to show that

$$\max_{0\leq k_0<h(k)} \Big|\frac{c}{1+\rho} - \frac{cA(k-k_0)}{k(1+\rho)\Delta_{1k}}\Big| \xrightarrow{P} 0$$

In this direction, we observe that

$$\begin{aligned}
\max_{0\leq k_0\leq h(k)} \Big|\frac{c}{1+\rho} - \frac{cA(k-k_0)}{k(1+\rho)\Delta_{1k}}\Big| &\leq \frac{|c|}{1+\rho} \max_{0\leq k_0<h(k)} \left(\Big|1-\frac{A}{\Delta_{1k}}\Big| + \frac{Ak_0}{k\Delta_{1k}}\right) \\
&\leq \frac{|c|}{1+\rho} \left(\Big|1-\frac{A}{\Delta_{1k}}\Big| + \frac{Ah(k)}{\Delta_{1k}k}\right) \xrightarrow{P} 0
\end{aligned}$$

since $h(k)/k \to 0$ and $A/\Delta_{1k} \xrightarrow{P} 1$ as in (B.34). This completes the proof. □

Proof of Theorem III.11. From (3.20) we have

$$\begin{aligned}
k^{\delta} \max_{0\leq k_0<h(k)} |T_{k_0,k} - T^*_{k_0,k}| &= k^{\delta} \max_{0\leq k_0<h(k)} \frac{k-k_0-1}{k-k_0} \Big|\frac{\widehat{\xi}_{k_0+1,k}}{\widehat{\xi}_{k_0,k}} - \frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}^*_{k_0,k}}\Big| \\
&\leq \frac{k^{\delta}}{1-h(k)/k} \max_{0\leq k_0<h(k)} \underbrace{\Big|\frac{\widehat{\xi}_{k_0+1,k}}{\widehat{\xi}_{k_0,k}} - \frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}^*_{k_0,k}}\Big|}_{W_{k_0,k}}
\end{aligned}$$

Since $h(k) = o(k)$, to prove (B.35), we show $k^{\delta} \max_{0\leq k_0<h(k)} W_{k_0,k} \xrightarrow{P} 0$. In this direction, we observe that

$$\begin{aligned}
W_{k_0,k} &\leq \Big|\frac{\widehat{\xi}_{k_0+1,k}}{\widehat{\xi}_{k_0,k}} - \frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}_{k_0,k}} - \frac{cAk^{-\delta}}{(1+\rho)\widehat{\xi}_{k_0,k}}\Big| + \frac{|c|Ak^{-\delta}}{(1+\rho)\widehat{\xi}_{k_0,k}}\Big|1-\frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}^*_{k_0,k}}\Big| \\
&+ \Big|\frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}_{k_0,k}} - \frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}^*_{k_0,k}} + \frac{cAk^{-\delta}}{(1+\rho)\widehat{\xi}_{k_0,k}} \frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}^*_{k_0,k}}\Big| \\
&= \frac{1}{\widehat{\xi}_{k_0,k}} \left(\Big|R_{k_0,k} - \frac{cAk^{-\delta}}{1+\rho}\Big| + \frac{|c|Ak^{-\delta}}{(1+\rho)}\Big|1-\frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}^*_{k_0,k}}\Big| + \frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}^*_{k_0,k}}\Big|\frac{cAk^{-\delta}}{(1+\rho)} - R_{k_0+1,k}\Big|\right)
\end{aligned}$$

where $R_{k_0,k}$ is defined in (3.12). Thus to show $k^{\delta} \max_{0 \le k_0 < h(k)} W_{k_0,k} \xrightarrow{P} 0$, it

$$\begin{aligned}
\max_{0 \le k_0 < h(k)} k^{\delta} W_{k_0,k} &\le \Big(M_{1k} + \frac{|c|A}{(1+\rho)} \max_{0 \le k_0 h(k)} |1 - B_{k_0,k}| + M_{1k} \max_{0 \le k_0 \le h(k)} B_{k_0,k} \Big) \max_{0 \le k_0 < h(k)} \frac{1}{\widehat{\xi}_{k_0,k}} \\
&= \Big(M_{1k} \max_{0 \le k_0 \le h(k)} (1 + B_{k_0,k}) + \frac{|c|A}{(1+\rho)} \max_{0 \le k_0 \le h(k)} |1 - B_{k_0,k}| \Big) \max_{0 \le k_0 < h(k)} \frac{1}{\widehat{\xi}_{k_0,k}}
\end{aligned}$$

where $M_{1k} = k^{\delta} \max_{0 \le k_0 < h(k)} |R_{k_0,k} - (k^{-\delta} cA)/(1+\rho)| \xrightarrow{P} 0$ is a direct consequence of Theorem III.5. Using (B.17), we next observe that

$$\begin{aligned}
\max_{0 \le k_0 \le h(k)} |1 - B_{k_0,k}| &\overset{d}{=} \max_{0 \le k_0 \le h(k)} \left| 1 - \frac{\Gamma_{k-k_0-1}/(k-k_0-1)}{\Gamma_{k-k_0}/(k-k_0)} \right| \\
&\le \frac{1}{1 - h(k)/k} \max_{k-h(k) \le i \le k} \left| \frac{\Gamma_i/i}{\Gamma_{i+1}/(i+1)} - 1 \right| \xrightarrow{a.s.} 0
\end{aligned} \tag{B.35}$$

is a direct consequence of (B.6) in Lemma B.3. (B.35) also proves that $\max_{0 \le k_0 \le h(k)}(1+ B_{k_0,k})$ is bounded in probabibilty.

Thus, to complete the proof of $k^{\delta} \max_{0 \le k_0 < h(k)} W_{k_0,k} \xrightarrow{P} 0$, we show that $\min_{0 \le k_0 < h(k)} |\widehat{\xi}_{k_0,k}|$ is bounded away from 0 in probability as follows:

$$\min_{0 \le k_0 < h(k)} \widehat{\xi}_{k_0,k} \ge \min_{0 \le k_0 < h(k)} \widehat{\xi}^*_{k_0,k} - \max_{0 \le k_0 < h(k)} |\widehat{\xi}_{k_0,k} - \widehat{\xi}^*_{k_0,k}| \tag{B.36}$$

For $\delta > 0$, Theorem III.5 implies $\max_{0 \le k_0 < h(k)} |\widehat{\xi}_{k_0,k} - \widehat{\xi}^*_{k_0,k}| \xrightarrow{P} 0$. Therefore $\min_{0 \le k_0 < h(k)} \widehat{\xi}_{k_0,k}$ is bounded away from 0 as long as $\min_{0 \le k_0 < h(k)} \widehat{\xi}^*_{k_0,k}$ is bounded away from 0. This can be easily shown because

$$\min_{0 \le k_0 < h(k)} \widehat{\xi}^*_{k_0,k} \overset{d}{=} \min_{0 \le k_0 < h(k)} \frac{\Gamma_{k-k_0}}{k-k_0} \ge 1 - \max_{k-h(k) \le i < k} \left| \frac{\Gamma_i}{i} - 1 \right| \xrightarrow{a.s.} 1$$

where the last convergence is a direct consequence of (B.5) in Lemma B.3. This completes the proof.

□

Lemma B.9. *Assumption* (3.17) *implies there exist* $M > 0$ *such that*

$$\inf_{F \in \mathcal{D}_\xi(B,\rho)} P_F\Big[\max_{0 \le k_0 < k} \sqrt{k}|R_{k_0,k}| \le M\Big] \to 1 \text{ as } h(k) \to \infty \tag{B.37}$$

where $R_{k_0,k}$ *is defined in* (3.12) *and* $k = O(n^{2\rho/(1+2\rho)})$.

Proof. By (3.17), we have $1 - Bx^{-\rho} \le L(x) \le 1 + Bx^{-\rho}$. Therefore

$$\begin{aligned}(k-k_0)R_{k_0,k} &\le (k_0+1)\log\frac{1+BY^{-\rho}_{(n-k_0,n)}}{1-BY^{-\rho}_{(n-k,n)}} + \sum_{i=k_0+2}^{k}\log\frac{1+BY^{-\rho}_{(n-i+1,n)}}{1-BY^{-\rho}_{(n-k,n)}} \quad \text{(B.38)}\\ &\le k\log\frac{1+BY^{-\rho}_{(n-k,n)}}{1-BY^{-\rho}_{(n-k,n)}}\end{aligned}$$

since $Y^{-\rho}_{(n-k,n)} \ge Y^{-\rho}_{(n-i+1,n)}$ for $i = k_0+1, \cdots, k$. Similarly, we also have

$$(k-k_0)R_{k_0,k} \ge k\log\frac{1-BY^{-\rho}_{(n-k,n)}}{1+BY^{-\rho}_{(n-k,n)}} \tag{B.39}$$

and thus, (B.38) and (B.39) together imply

$$\max_{0\le k_0<h(k)} \sqrt{k}|R_{k_0,k}| \le \frac{\sqrt{k}Y^{-\rho}_{(n-k,n)}}{1-h(k)/k}\max_{0\le k_0<h(k)}\frac{1}{Y^{-\rho}_{(n-k,n)}}\log\frac{1+BY^{-\rho}_{(n-k,n)}}{1-BY^{-\rho}_{(n-k,n)}} \tag{B.40}$$

where $h(k) = o(k)$ and

$$\sqrt{k}Y^{-\rho}_{(n-k,n)}\frac{1}{Y^{-\rho}_{(n-k,n)}}\log\frac{1+BY^{-\rho}_{(n-k,n)}}{1-BY^{-\rho}_{(n-k,n)}} \quad \overset{d}{=} \quad \underbrace{\sqrt{k}(\Gamma_{k+1}/\Gamma_{n+1})^\rho}_{\Delta_{1k}}\underbrace{\frac{1}{(\Gamma_{k+1}/\Gamma_{n+1})^\rho}\log\frac{1+B(\Gamma_{k+1}/\Gamma_{n+1})^\rho}{1-B(\Gamma_{k+1}/\Gamma_{n+1})^\rho}}_{\Delta_{2k}}$$

Now, by relation (B.5) in Lemma B.3, we have $((\Gamma_{k+1}/k)/(\Gamma_{n+1}/n))^\rho \xrightarrow{a.s.} 1$. For $k = O(n^{2\rho/(1+2\rho)})$, Δ_{1k} is bounded almost surely and $\Delta_{2k} \xrightarrow{a.s.} 2B$. Therefore there exist M such that

$$\inf_{F\in\mathcal{D}_\xi(B,\rho)} P_F\Big[\max_{0\le k_0<k}\frac{k-k_0}{kY^{-\rho}_{(n-k,n)}}|R_{k_0,k}|\le M\Big]\ge P[\Delta_{1k}\Delta_{2k}\le M]\to 1$$

This completes the proof.

□

Proof of Theorem III.7. Let $P_n=\inf_{F\in\mathcal{D}_\xi(B,\rho)} P_F\Big[\max_{0\le k_0<h(k)}|\widehat{\xi}_{k_0,k}-\xi|\le a(n)\Big]$, then

$$P_n=\inf_{\mathcal{D}_\xi(B,\rho)} P_F\Big[\underbrace{\max_{0\le k_0<h(k)}\sqrt{k}|R_{k_0,k}|\le(\sqrt{k}a(n))/2}_{A_{1n}}\cap\underbrace{\max_{0\le k_0<h(k)}\sqrt{k}|\widehat{\xi}^*_{k_0,k}-\xi|\le(\sqrt{k}a(n))/2}_{A_{2n}}\Big]$$

Since $\sqrt{k}a(n)\to\infty$, by Lemma B.9, $\inf_{F\in\mathcal{D}_\xi(B,\rho)}P_F[A_{1n}]\to 1$. We also have that,

$$\inf_{F\in\mathcal{D}_\xi(B,\rho)} P_F[A_{2n}]=P\Big[\max_{0\le k_0<h(k)}\sqrt{k}|\widehat{\xi}^*_{k_0,k}(n)-\xi|\le(\sqrt{k}a(n))/2\Big]$$

since $\widehat{\xi}^*_{k_0,k}$ does not depend on $F\in\mathcal{D}_\xi(B,\rho)$.

By using Donsker's principle, we will show that

$$\max_{0\le k_0<h(k)}|\widehat{\xi}^*_{k_0,k}(n)-\xi|=o_P(a(n)),$$

which will imply $P_F(A_{2n})\to 1$. Indeed, without loss of generality, suppose $\xi=1$ and let $E_i,\ i=1,2,\dots$ be independent standard exponential random variables. For every $\epsilon\in(0,1)$, we have that

$$W_k=\{W_k(t),\ t\in[\epsilon,1]\}:=\left\{\frac{\sqrt{k}}{[kt]}\sum_{i=1}^{[kt]}(E_i-1),\ t\in[0,1]\right\}\xrightarrow{d}\{B(t)/t,\ t\in[\epsilon,1]\},\tag{B.41}$$

as $k \to \infty$, where $B = \{B(t),\ t \in [0,1]\}$ is the standard Brownian motion, and where the last convergence is in the space of cadlag functions $\mathbb{D}[\epsilon, 1]$ equipped with the Skorokhod J_1-topology. (In fact, since the limit has continuous paths, the convergence is also valid in the uniform norm.)

Recall that by (3.6), we have

$$\{\widehat{\xi}^*_{k_0,k}(n),\ 0 \le k_0 < k\} \stackrel{d}{=} \left\{ \sum_{i=1}^{k-k_0} E_i/(k-k_0),\ 0 \le k_0 < k \right\}.$$

Thus,

$$\sqrt{k} \max_{0 \le k_0 < h(k)} |\widehat{\xi}^*_{k_0,k}(n) - \xi| \stackrel{d}{=} \sup_{t \in [1-h(k)/k, 1]} |W_k(t)| \le \sup_{t \in [\epsilon,1]} |W_k(t)|, \tag{B.42}$$

where the last inequality holds for all sufficiently large k, since $1 - h(k)/k \to 1$, as $k \to \infty$. Since the supremum is a continuous functional in J_1, the convergence in (B.41) implies that the right–hand side of (B.42) converges in distribution to $\sup_{t\in[\epsilon,1]} |B(t)/t| = O_P(1)$, which is finite with probability one. This completes the proof since $a(n)\sqrt{k(n)} \to \infty$. □

Proof of Theorem III.15. We first begin with the proof of (3.25). For this from (3.24) we have

$$\begin{aligned}
\max_{0 \le k_0 < h(k)} |U_{k_0,k} - U^*_{k_0,k}| &= \max_{0 \le k_0 < h(k)} 2 \Big| |(T_{k_0,k})^{k-k_0-1} - 0.5| - |(T^*_{k_0,k})^{k-k_0-1} - 0.5| \Big| \\
&\le 2 \max_{0 \le k_0 < h(k)} \Big| |(T_{k_0,k})^{k-k_0-1} - (T^*_{k_0,k})^{k-k_0-1}| \Big| \\
&\le 2 \max_{0 \le k_0 < h(k)} \left| \left(\frac{T_{k_0,k}}{T^*_{k_0,k}} \right)^{k-k_0-1} - 1 \right|
\end{aligned}$$

where the last inequality holds since $T^*_{k_0,k} \le 1$(see III.9). Thus to prove (3.25), it suffices to show

$$k^{\delta-1}\max_{0\le k_0<h(k)}\Big|\Big(\frac{T_{k_0,k}}{T^*_{k_0,k}}\Big)^{k-k_0-1}-1\Big|\xrightarrow{P}0 \tag{B.43}$$

To prove (B.43), we begin by showing

$$k^{\delta}\max_{0\le k_0<h(k)}|\frac{T_{k_0,k}}{T^*_{k_0,k}}-1|\xrightarrow{P}0. \tag{B.44}$$

In this direction, from (3.23), we observe that

$$k^{\delta}\max_{0\le k_0<h(k)}\Big|\frac{T_{k_0,k}}{T^*_{k_0,k}}-1\Big|\le\frac{1}{\min_{0\le h(k)}T^*_{k_0,k}}\max_{0\le k_0<h(k)}\underbrace{k^{\delta}|T_{k_0,k}-T^*_{k_0,k}|}_{\Delta_k}$$

From (3.23), we have $\Delta_k\xrightarrow{P}0$. Thus (B.44) holds as long as $\min_{0\le k_0<h(k)}T^*_{k_0,k}$ is bounded away from 0 in probability. This can be easily seen as follows

$$\begin{aligned}\min_{0\le k_0<h(k)}T^*_{k_0,k}&\overset{d}{=}\min_{0\le k_0<h(k)}\frac{\Gamma_{k-k_0-1}/(k-k_0-1)}{\Gamma_{k-k_0}/(k-k_0)}\\&\ge1-\max_{k-h(k)\le i<k}\Big|\frac{\Gamma_i/i}{\Gamma_{i+1}/(i+1)}-1\Big|\xrightarrow{a.s.}1\end{aligned}$$

where the last convergence is a direct consequence of (B.6) in Lemma B.3.

In view of (B.35), for a subsequence, $\{k_l\}$ there exists a further subsequence $\widetilde{k}$ such that

$$\widetilde{k}^{\delta}\max_{0\le k_0<h(\widetilde{k})}\Big|\frac{T_{k_0,\widetilde{k}}}{T^*_{k_0,\widetilde{k}}}-1\Big|\xrightarrow{a.s.}0$$

This implies there exists M such that for every $\widetilde{k}\ge M$ and $0\le k_0<h(\widetilde{k})$,

$$1-\frac{\epsilon}{\widetilde{k}^{\delta}}\le\Big(\frac{T_{k_0,\widetilde{k}}}{T^*_{k_0,\widetilde{k}}}\Big)(\omega)\le1+\frac{\epsilon}{\widetilde{k}^{\delta}} \tag{B.45}$$

for all $\omega \in \Omega$ with $P[\Omega] = 1$. (B.45) further implies

$$\underbrace{\widetilde{k}^{\delta-1}\Big(\Big(1-\frac{\epsilon}{\widetilde{k}^{\delta}}\Big)^{\widetilde{k}-h(\widetilde{k})-1}-1\Big)}_{-a_{\widetilde{k}}} \le \quad \widetilde{k}^{\delta-1}\Big(\Big(\frac{T_{k_0,\widetilde{k}}}{T^*_{k_0,\widetilde{k}}}\Big)^{\widetilde{k}-k_0-1}(\omega)-1\Big) \quad \le \underbrace{\widetilde{k}^{\delta-1}\Big(\Big(1+\frac{\epsilon}{\widetilde{k}^{\delta}}\Big)^{\widetilde{k}-1}-1\Big)}_{b_{\widetilde{k}}}$$

Therefore,

$$\widetilde{k}^{\delta-1}\max_{0\le k_0<h(\widetilde{k})}\Big|\Big(\frac{T_{k_0,\widetilde{k}}}{T^*_{k_0,\widetilde{k}}}\Big)^{\widetilde{k}-k_0-1}(w)-1\Big| \le a_{\widetilde{k}} \vee b_{\widetilde{k}} \tag{B.46}$$

First observe that both the sequences $a_{\widetilde{k}}$ and $b_{\widetilde{k}}$ converge to ϵ as $\widetilde{k} \to \infty$. Thereby, taking limsup w.r.t $\widetilde{k}$ on both sides of (B.46), we get

$$\limsup_{\widetilde{k}\to\infty}\widetilde{k}^{\delta-1}\max_{0\le k_0<h(\widetilde{k})}\Big|\Big(\frac{T_{k_0,\widetilde{k}}}{T^*_{k_0,\widetilde{k}}}\Big)^{\widetilde{k}-k_0-1}(w)-1\Big| \le \epsilon \tag{B.47}$$

Since (B.47) holds for all $\epsilon > 0$, we have

$$\widetilde{k}^{\delta-1}\max_{0\le k_0<h(\widetilde{k})}\Big|\Big(\frac{T_{k_0,\widetilde{k}}}{T^*_{k_0,\widetilde{k}}}\Big)^{\widetilde{k}-k_0-1}(w)-1\Big| \to 0$$

This entails the proof of convergence in probability of (B.43).

We next begin with the proof (3.26). To this end, we have

$$1 - P_{H_0}[\widehat{k}_0 = 0] \quad = \quad 1 - P_{H_0}\Big[\underbrace{\bigcap_{i=0}^{f(k)}\{U_{i,k} < (1-q)^{ca^{k-i-1}}\}}_{A_k}\Big]$$

where we shall show $P[A_k] \to 1-q$ as follows.

$$\begin{aligned} P_{H_0}\Big[A_k\Big] &\leq P_{H_0}\Big[A_k \cap \underbrace{\{k^{\delta-1} \max_{0\leq i<f(k)} (U_{i,k} - U^*_{i,k}) < \epsilon\}}_{B_{1k}}\Big] + P[B^c_{1k}] \\ &\leq P_{H_0}\Big[\underbrace{\bigcap_{i=0}^{f(k)} \{U^*_{i,k} < (1-q)^{ca^{k-i-1}} + \epsilon k^{1-\delta}\}}_{A^*_{1k}}\Big] + P[B^c_{1k}] \end{aligned}$$

since $A_k \cap B_{1k} \implies A^*_{1k}$ and $P[B^c_{1k}] \to 0$ by (3.25). It remain to show $P_{H_0}[A^*_{1k}] \to 1-q$. In this direction, we observe that

$$\begin{aligned} P_{H_0}[A^*_{1k}] &= \prod_{i=0}^{f(k)} (1-q)^{ca^{k-i-1}} \prod_{i=0}^{f(k)} \Big(1 + \frac{\epsilon k^{1-\delta}}{(1-q)^{ca^{k-i-1}}}\Big) \\ &\leq \underbrace{\Big(1-q\Big)^{\frac{a^{(k-1)}-a^{(k-f(k)-2})}{a^{(k-1)}-1}}}_{c_{0k}} \underbrace{\Big(1 + \frac{\epsilon}{(1-q)k^{\delta-1}}\Big)^{f(k)}}_{c_{1k}} \end{aligned}$$

since $(1-q)^{ca^{k-i-1}} \geq (1-q)$. Observe that for $f(k) \to \infty$, we have

$$c_{0k} \to 1-q. \tag{B.48}$$

For $f(k) = O(k^{\delta-1})$, $\limsup_{k\to\infty} c_{1k} \leq (1 + M\epsilon/(1-q))$ for some $M > 0$. Thus

$$\limsup_{k\to\infty} P_{H_0}[A^*_{1k}] \leq (1-q) + M\epsilon$$

holds for every $\epsilon > 0$ which implies $\limsup_{k\to\infty} P_{H_0}[A^*_{1k}] \leq (1-q)$. Additionally,

$$\begin{aligned} P_{H_0}\Big[A_k\Big] &\geq P_{H_0}\Big[A_k \cap \underbrace{\{k^{\delta-1} \max_{0\leq i<f(k)} (U_{i,k} - U^*_{i,k}) > -\epsilon\}}_{B_{2k}}\Big] \\ &\geq P_{H_0}\Big[\underbrace{\bigcap_{i=0}^{f(k)} \{U^*_{i,k} < (1-q)^{ca^{k-i-1}} - \epsilon k^{1-\delta}\}}_{A^*_{2k}}\Big] - P_{H_0}[B^c_{2k}] \end{aligned}$$

since $A^*_{2k} \cap B_{2k} \implies A_k \cap B_{2k}$ and $P[B^c_{2k}] \to 0$ by (3.25). It remain to show $P_{H_0}[A^*_{2k}] \to 1-q$. In this direction, we observe that

$$\begin{aligned} P_{H_0}[A^*_{2k}] &= \prod_{i=0}^{f(k)} (1-q)^{ca^{k-i-1}} \prod_{i=0}^{f(k)} \Big(1 - \frac{\epsilon k^{1-\delta}}{(1-q)^{ca^{k-i-1}}}\Big) \\ &\geq \underbrace{\Big(1-q\Big)^{\frac{a^{(k-1)}-a^{(k-f(k)-2)}}{a^{(k-1)}-1}}}_{c_{0k}} \underbrace{\Big(1 - \frac{\epsilon}{(1-q)k^{\delta-1}}\Big)^{f(k)}}_{c_{2k}} \end{aligned}$$

since $(1-q)^{ca^{k-i-1}} \geq (1-q)$. As before, for $f(k) \to \infty$, we have that $c_{0k} \to 1-q$. For $f(k) = O(k^{\delta-1})$, $\limsup_{k\to\infty} c_{1k} \geq (1 - M\epsilon/(1-q))$ for some $M > 0$. Thus

$$\limsup_{k\to\infty} P_{H_0}[A^*_{2k}] \geq (1-q) - M\epsilon$$

holds for every $\epsilon > 0$ which implies $\limsup_{k\to\infty} P_{H_0}[A^*_{2k}] \geq (1-q)$.

Thus $\lim_{k\to\infty} P_{H_0}[A^*_{2k}] = 1-q$ which completes the proof. □

APPENDIX C

Spatial Extremes

Lemma C.1. *Suppose $X \sim GPD(\sigma, \xi)$ for $X > u$ and σ and ξ are a linear function of known covariates w_1 and w_2 as:*

$$\begin{aligned} \log(\sigma) &= {w_1}^\top \rho_1 \\ \xi &= {w_2}^\top \rho_2. \end{aligned}$$

Then the return level in (4.21) *is expressed as*

$$r_m = v_0 + \frac{\exp({w_1}^\top \rho_1)}{{w_2}^\top \rho_2}[(m\tau_{v_0})^{{w_2}^\top \rho_2} - 1] \tag{C.1}$$

with standard error given by

$$\mathrm{Var}(\hat{r}_m) = \nabla^\top r_m \, \Sigma \, \nabla r_m \tag{C.2}$$

where the variance covariance matrix, Σ *of*$[\hat{\tau}_{v_0}, \hat{\rho}^{(\sigma)}, \hat{\rho}^{(\xi)}]$ *has form*

$$\Sigma = \begin{bmatrix} \tau_{v_0}(1-\tau_{v_0})/n & 0 \\ 0 & H^{-1}_{\rho_1,\rho_2} \end{bmatrix}$$

where H*, the hessian matrix of* (ρ_1, ρ_2) *is a direct product of the maximum likelihood fit to the time series* Q_k *and the expressions for* ∇r_m *are given by:*

$$\nabla^\top r_m = \left[\frac{\partial r_m}{\partial \tau_{v_0}}, \nabla^\top_{\rho_1} r_m, \nabla^\top_{\rho_2} r_m\right]$$

where

$$\frac{\partial r_m(k,x)}{\partial \tau_{v_0}} = \exp({w_1}^\top \rho_1) m^{{w_2}^\top \rho_2} \tau_{v_0}^{{w_2}^\top \rho_2 - 1}$$

$$\nabla^\top_{\rho_1} r_m(k,x) = \exp({w_1}^\top \rho_1) \Big(\frac{(m\tau_u)^{{w_2}^\top \rho_2} - 1}{{w_1}^\top \rho_1}\Big) w_1$$

$$\nabla^\top_{\rho_2} = -\frac{\exp({w_1}^\top {\rho_1}^\top)}{{w_2}^\top \rho_2} \left(\frac{(m\tau_{v_0})^{{w_2}^\top \rho_2} - 1}{{w_2}^\top \rho_2} - (m\tau_{v_0})^{{w_2}^\top \rho_2} \log(m\tau_{v_0})\right) w_2$$

.

Proof. The proof (C.1) is straightforward. The proof of (C.2) is a direct application of Delta method and Section 4.3.3 in *Coles* (2001). □

BIBLIOGRAPHY

Aban, I. B., M. M. Meerschaert, and A. K. Panorska (2006), Parameter estimation for the truncated pareto distribution, *Journal of the American Statistical Association*, *101*(473), 270–277.

Ahsanullah, M., V. B. Nevzorov, and M. Shakil (2013), *An introduction to order statistics*, *Atlantis Studies in Probability and Statistics*, vol. 3, x+244 pp., Atlantis Press, Paris, doi:10.2991/978-94-91216-83-1.

Balkema, A. A., and L. de Haan (1974), Residual life time at great age, *Ann. Probab.*, *2*(5), 792–804, doi:10.1214/aop/1176996548.

Bed, N. S. T., and U. DOE (2009), Study of Security Attributes of Smart Grid Systems - Current Cyber Security Issues.

Beirlant, J., Y. Goegebeur, J. Segers, J. Teugels, D. De Waal, and C. Ferro (2004a), *Statistics of Extremes: Theory and Applications*, Wiley Series in Probability and Statistics, John Wiley & Sons.

Beirlant, J., Y. Goegebeur, J. Teugels, and J. Segers (2004b), *Statistics of extremes*, Wiley Series in Probability and Statistics, xiv+490 pp., John Wiley & Sons, Ltd., Chichester, doi:10.1002/0470012382, theory and applications, With contributions from Daniel De Waal and Chris Ferro.

Beirlant, J., C. Bouquiaux, and B. J. Werker (2006), Semiparametric lower bounds for tail index estimation, *Journal of Statistical Planning and Inference*, *136*(3), 705 – 729, doi:http://dx.doi.org/10.1016/j.jspi.2004.08.018.

Beirlant, J., I. F. Alves, and I. Gomes (2016), Tail fitting for truncated and non-truncated pareto-type distributions, *Extremes*, *19*(3), 429–462, doi:10.1007/s10687-016-0247-3.

Berthier, R., W. Sanders, and H. Khurana (2010), Intrusion detection for advanced metering infrastructures: Requirements and architectural directions, in *IEEE SmartGridComm*, pp. 350–355, doi:10.1109/SMARTGRID.2010.5622068.

Besag, J., and C. Kooperberg (1995), On conditional and intrinsic autoregression, *Biometrika*, *82*(4), 733–746.

Bhattacharya, S., M. Kallitsis, and S. Stoev (2017), Trimming the Hill estimator: robustness, optimality and adaptivity, *ArXiv e-prints*.

Bi, S., and Y. J. Zhang (2014), Graphical methods for defense against false-data injection attacks on power system state estimation, *Smart Grid, IEEE Transactions on*, *5*(3), 1216–1227, doi:10.1109/TSG.2013.2294966.

Bingham, N., C. Goldie, and J. Teugels (1989), *Regular Variation*, no. no. 1 in Encyclopedia of Mathematics and its Applications, Cambridge University Press.

Bingham, N. H., C. M. Goldie, and J. L. Teugels (1987), *Regular Variation*, Cambridge University Press.

Bojanov, B., H. Hakopian, and B. Sahakian (1993), *Spline Functions and Multivariate Interpolations*, Mathematics and Its Applications, Springer.

Boucheron, S., and M. Thomas (2015), Tail index estimation, concentration and adaptivity, *Electron. J. Statist.*, *9*(2), 2751–2792, doi:10.1214/15-EJS1088.

Boyer, R., and J. Moore (1991), MJRTY - a fast majority vote algorithm, in *Automated Reasoning*, *Automated Reasoning Series*, vol. 1, edited by R. Boyer, pp. 105–117, doi:10.1007/978-94-011-3488-0_5.

Brazauskas, V., and R. Serfling (2000), Robust estimation of tail parameters for two-parameter Pareto and exponential models via generalized quantile statistics, *Extremes*, *3*(3), 231–249 (2001), doi:10.1023/A:1011455027066.

Brzezinski, M. (2016), Robust estimation of the pareto tail index: a monte carlo analysis, *Empirical Economics*, *51*(1), 1–30, doi:10.1007/s00181-015-0989-9.

Capozzi, V., and G. Budillon (2017), Detection of heat and cold waves in montevergine time series (1884–2015), *Advances in Geosciences*, *44*, 35–51, doi: 10.5194/adgeo-44-35-2017.

Carcano, A., I. N. Fovino, M. Masera, and A. Trombetta (2010), State-based network intrusion detection systems for SCADA protocols: A proof of concept, in *Proceedings of CRITIS'09*, pp. 138–150.

Ceccherini, G., S. Russo, I. Ameztoy, A. F. Marchese, and C. Carmona-Moreno (2017), Heat waves in africa 1981–2015, observations and reanalysis, *Natural Hazards and Earth System Sciences*, *17*(1), 115–125, doi:10.5194/nhess-17-115-2017.

Cleveland, F. (2008), Cyber security issues for advanced metering infrastructure, in *Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century*, pp. 1–5, doi:10.1109/PES.2008.4596535.

Coles, S. (2001), *An introduction to statistical modeling of extreme values*, Springer Series in Statistics, Springer-Verlag, London.

Cooley, D., D. Nychka, and P. Naveau (2007), Bayesian spatial modeling of extreme precipitation return levels, *J. Amer. Statist. Assoc.*, *102*(479), 824–840.

Cormode, G., and S. Muthukrishnan (2004), What's new: finding significant differences in network data streams, in *INFOCOM 2004*, vol. 3, pp. 1534–1545 vol.3, doi:10.1109/INFCOM.2004.1354567.

Cormode, G., and S. Muthukrishnan (2005a), What's hot and what's not: Tracking most frequent items dynamically, *ACM Trans. Database Syst.*, *30*(1), 249–278, doi:10.1145/1061318.1061325.

Cormode, G., and S. Muthukrishnan (2005b), An improved data stream summary: the count-min sketch and its applications, *J. Algorithms*, *55*(1), 58–75, doi:10.1016/j.jalgor.2003.12.001.

Cormode, G., F. Korn, S. Muthukrishnan, and D. Srivastava (2003), Finding hierarchical heavy hitters in data streams, in *VLDB 03*, pp. 464–475.

Cressie, N. (1993a), *Statistics for Spatial Data*, John Wiley, New York, USA.

Cressie, N. (1993b), *Statistics for Spatial Data: revised ed*, John Wiley, New York.

Cressie, N., and E. L. Kang (2016), Hot enough for you? a spatial exploratory and inferential analysis of north american climate-change projections, *Mathematical Geosciences*, *48*(2), 107–121, doi:10.1007/s11004-015-9607-9.

Croitoru, A.-E., A. Piticar, A.-F. Ciupertea, and C. F. Roca (2016), Changes in heat waves indices in romania over the period 19612015, *Global and Planetary Change*, *146*, 109 – 121, doi:https://doi.org/10.1016/j.gloplacha.2016.08.016.

Crovella, M., and A. Bestavros (1997), Self-similarity in world wide web traffic: evidence and possible causes, *Networking, IEEE/ACM Transactions on*, *5*(6), 835–846, doi:10.1109/90.650143.

Czyz, J., M. Kallitsis, M. Gharaibeh, C. Papadopoulos, M. Bailey, and M. Karir (2014), Taming the 800 pound gorilla: The rise and decline of ntp ddos attacks, in *Proceedings of the 2014 Conference on Internet Measurement Conference*, IMC '14, pp. 435–448, ACM, New York, NY, USA, doi:10.1145/2663716.2663717.

Danielsson, J., L. de Haan, L. Peng, and C. de Vries (2001), Using a bootstrap method to choose the sample fraction in tail index estimation, *Journal of Multivariate Analysis*, *76*(2), 226 – 248, doi:https://doi.org/10.1006/jmva.2000.1903.

Davison, A. C., and R. L. Smith (1990), Models for exceedances over high thresholds, *Journal of the Royal Statistical Society. B*, *52*(3), 393–442.

Davison, A. C., S. A. Padoan, and M. Ribatet (2012), Statistical modeling of spatial extremes, *Statist. Sci.*, *27*(2), 161–186, doi:10.1214/11-STS376.

de Haan, L. (2006), *Extreme Value Theory, An Introduction*, Springer.

de Haan, L., and A. Ferreira (2006), *Extreme value theory*, Springer Series in Operations Research and Financial Engineering, xviii+417 pp., Springer, New York, an introduction.

Dian-Xiu, Y., Y. Ji-Fu, C. Zheng-Hong, Z. You-Fei, and W. Rong-Jun (2014), Spatial and temporal variations of heat waves in china from 1961 to 2010, *Advances in Climate Change Research*, *5*(2), 66 – 73, doi:https://doi.org/10.3724/SP.J.1248.2014.066.

Drees, H., and E. Kaufmann (1998), Selecting the optimal sample fraction in univariate extreme value estimation, *Stochastic Processes and their Applications*, *75*(2), 149 – 172, doi:https://doi.org/10.1016/S0304-4149(98)00017-9.

Dupuis, D., and M.-P. Victoria-Feser (2006), A robust prediction error criterion for pareto modeling of upper tails, *Canadian Journal of Statistics*, *34*(4), 639–358, iD: unige:6462.

Dutang, C., Y. Goegebeur, and A. Guillou (2014), Robust and bias-corrected estimation of the coefficient of tail dependence, *Insurance Math. Econom.*, *57*, 46–57, doi:10.1016/j.insmatheco.2014.05.003.

Embrechts, P., C. Klüppelberg, and T. Mikosch (1997), *Modelling Extreme Events*, Springer-Verlag, New York.

Estan, C., and G. Varghese (2002), New directions in traffic measurement and accounting, *SIGCOMM Comput. Commun. Rev.*, *32*(4), 323–336, doi:10.1145/964725.633056.

et al., R. B. B. (2010), Detecting false data injection attacks on DC state estimation, in *SCS Workshop*.

et al., Y. H. (2013), Bad data injection in smart grid: attack and defense mechanisms, *Communications Magazine, IEEE*, doi:10.1109/MCOM.2013.6400435.

extRemes (2016), Package 'extremes', `https://cran.r-project.org/web/packages/extRemes/extRemes.pdf`.

Falliere, N., L. Murch, and E. Chien (2011), W32.stuxnet dossier.

Faloutsos, M., P. Faloutsos, and C. Faloutsos (1999), On power-law relationships of the internet topology, *SIGCOMM Comput. Commun. Rev.*, *29*(4), 251–262, doi:10.1145/316194.316229.

Farhangi, H. (2010), The path of the smart grid, *IEEE Power and Energy Magazine*, *8*(1), doi:10.1109/MPE.2009.934876.

Ferraty, F., and P. Vieu (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, Springer Series in Statistics, Springer New York.

Ferro, C. A. T., and J. Segers (2003), Inference for clusters of extreme values, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(2), 545–556, doi:10.1111/1467-9868.00401.

fields (2018), Package 'fields', `https://cran.r-project.org/web/packages/fields/fields.pdf`.

Finkelstein, M., H. G. Tucker, and J. A. Veeh (2006), Pareto tail index estimation revisited, *North American Actuarial Journal*, *10*(1), 1–10, doi:10.1080/10920277.2006.10596236.

Gilbert, A., Y. Kotidis, S. Muthukrishnan, and M. Strauss (2001), Surfing wavelets on streams: one-pass summaries for approximate aggregate queries, in *Procedings of VLDB, Rome, Italy.*

Gilbert, A., Y. Li, E. Porat, and M. Strauss (2012), Approximate sparse recovery: Optimizing time and measurements, *SIAM Journal on Computing*, *41*(2), 436–453, doi:10.1137/100816705.

Gilbert, A. C., M. J. Strauss, J. A. Tropp, and R. Vershynin (2006), Algorithmic linear dimension reduction in the ℓ-1 norm for sparse vectors, in *Allerton 2006.*

Gilbert, A. C., M. J. Strauss, J. A. Tropp, and R. Vershynin (2007), One sketch for all: Fast algorithms for compressed sensing, in *STOC '07*, pp. 237–246, NY, USA, doi:10.1145/1250790.1250824.

Goegebeur, Y., A. Guillou, and A. Verster (2014), Robust and asymptotically unbiased estimation of extreme quantiles for heavy tailed distributions, *Statist. Probab. Lett.*, *87*, 108–114, doi:10.1016/j.spl.2014.01.010.

Hall, P. (1982), On some simple estimates of an exponent of regular variation, *J. Roy. Stat. Assoc.*, *44*, 37–42, series B.

Hall, P., and A. H. Welsh (1984), Best attainable rates of convergence for estimates of parameters of regular variation, *The Annals of Statistics*, *12*(3), 1079–1084.

Hall, P., and A. H. Welsh (1985), Adaptive estimates of parameters of regular variation, *Ann. Statist.*, *13*(1), 331–341, doi:10.1214/aos/1176346596.

Hill, B. M. (1975), A simple general approach to inference about the tail of a distribution, *Annals of statistics*, *3*(5), 1163–1174.

Huang, W. K. (2017), Statistics of extremes with applications in climate, Ph.D. thesis, copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2017-12-05.

Jonathan, J., F. AnneCatherine, B. Claude, and A. JeanFranois (2018), A spatiotemporal model for extreme precipitation simulated by a climate model, with an application to assessing changes in return levels over north america, *Journal* *of the Royal Statistical Society: Series C (Applied Statistics)*, *66*(5), 941–962, doi: 10.1111/rssc.12212.

Kallitsis, M., S. Stoev, and G. Michailidis (2014), Hashing Pursuit for Online Identification of Heavy-Hitters in High-Speed Network Streams, `http://arxiv.org/abs/1412.6148`.

Kallitsis, M., S. A. Stoev, S. Bhattacharya, and G. Michailidis (2016a), Amon: An open source architecture for online monitoring, statistical analysis, and forensics of multi-gigabit streams, *IEEE Journal on Selected Areas in Communications*, *34*(6), 1834–1848, doi:10.1109/JSAC.2016.2558958.

Kallitsis, M., S. A. Stoev, S. Bhattacharya, and G. Michailidis (2016b), Amon: An open source architecture for online monitoring, statistical analysis, and forensics of multi-gigabit streams, doi:10.1109/JSAC.2016.2558958.

Kallitsis, M. G., G. Michailidis, and S. Tout (2015), Correlative monitoring for detection of false data injection attacks in smart grids, in *SmartGridComm*.

Kallitsis, M. G., S. Bhattacharya, S. Stoev, and G. Michailidis (2016c), Adaptive statistical detection of false data injection attacks in smart grids, in *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 826–830, doi:10.1109/GlobalSIP.2016.7905958.

Karp, R. M., S. Shenker, and C. H. Papadimitriou (2003), A simple algorithm for finding frequent elements in streams and bags, *ACM Trans. Database Syst.*, *28*(1), 51–55, doi:10.1145/762471.762473.

Kaufman, L., and P. J. Rousseeuw (1990), *Finding Groups in Data: An Introduction to Cluster Analysis,*, John Wiley and Sons, Inc., Hoboken, NJ, USA.

Knight, K. (unknown), A simple modification of the hill estimator with applications to robustness and bias reduction.

Knott, G. (2000), *Interpolating Cubic Splines*, Interpolating Cubic Splines, Birkhauser.

Krishnamurthy, B., S. Sen, Y. Zhang, and Y. Chen (2003), Sketch-based change detection: methods, evaluation, and applications, in *3rd ACM SIGCOMM IMC*, pp. 234–247, NY, USA, doi:10.1145/948205.948236.

Kührer, M., T. Hupperich, C. Rossow, and T. Holz (2014), Exit from hell? reducing the impact of amplification ddos attacks, in *Proceedings of the 23rd USENIX Conference on Security Symposium*, SEC'14, pp. 111–125, USENIX Association, Berkeley, CA, USA.

Lambert, D., and C. Liu (2006a), Adaptive thresholds: Monitoring streams of network counts, *online, J. Am. Stat. Assoc*, pp. 78–89.

Lambert, D., and C. Liu (2006b), Adaptive thresholds: monitoring streams of network counts, *J. Amer. Statist. Assoc.*, *101*(473), 78–88, doi:10.1198/016214505000000943.

Lawrence, E., G. Michailidis, V. N. Nair, and B. Xi (2006), Network tomography: a review and recent developments, in *Frontiers in statistics*, pp. 345–366, Imp. Coll. Press, London.

Ledford, A. W., and J. A. Tawn (2003), Diagnostics for dependence within time series extremes, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(2), 521–543, doi:10.1111/1467-9868.00400.

Lee, R., M. Assante, and T. Conway (2014), German steel mill cyber attack.

Lee, R., M. Assante, and T. Conway (2016), Analysis of the cyber attack on the ukrainian power grid.

Lehmann, E., and G. Casella (1983), *Theory of Point Estimation*, 88 pp., Springer.

Leland, W., M. Taqqu, W. Willinger, and D. Wilson (1994), On the self-similar nature of ethernet traffic (extended version), *Networking, IEEE/ACM Transactions on*, *2*(1), 1–15, doi:10.1109/90.282603.

Liu, Y., P. Ning, and M. K. Reiter (2009), False data injection attacks against state estimation in electric power grids, in *Proceedings of CCS '09*, CCS '09, pp. 21–32, ACM, New York, NY, USA, doi:10.1145/1653662.1653666.

Lucas, J. M., and M. S. Saccucci (1990), Exponentially weighted moving average control schemes: Properties and enhancements, *Technometrics*, *32*(1), 1–29, doi:10.2307/1269835.

McDaniel, P., and S. McLaughlin (2009), Security and privacy challenges in the smart grid, *Security Privacy, IEEE*, *7*(3), 75–77, doi:10.1109/MSP.2009.76.

McKinnon, K., A. Rhines, M. P. Tingley, and P. Huybers (2016), Long-lead predictions of eastern united states hot days from pacific sea surface temperatures, *9*.

Meehl, G. A., and C. Tebaldi (2004), More intense, more frequent, and longer lasting heat waves in the 21st century, *Science*, *305*(5686), 994–997, doi:10.1126/science.1098704.

Metke, A., and R. Ekl (2010), Security technology for smart grid networks, *Smart Grid, IEEE Transactions on*, *1*(1), 99–107, doi:10.1109/TSG.2010.2046347.

Muthukrishnan, S. (2005), Data streams: Algorithms and applications, *Found. Trends Theor. CS*, *1*(2), doi:10.1561/0400000002.

Ouzeau, G., J.-M. Soubeyroux, M. Schneider, R. Vautard, and S. Planton (2016), Heat waves analysis over france in present and future climate: Application of a new method on the euro-cordex ensemble, *Climate Services*, *4*, 1 – 12, doi:https://doi.org/10.1016/j.cliser.2016.09.002.

pbs (2013), Package 'pbs', `https://cran.r-project.org/web/packages/pbs/pbs.pdf`.

Peng, L., and A. Welsh (2001), Robust estimation of the generalized pareto distribution, *Extremes*, *4*(1), 53–65, doi:10.1023/A:1012233423407.

Philander, S., J. Holton, and R. Dmowska (1989), *El Nino, La Nina, and the Southern Oscillation*, International Geophysics, Elsevier Science.

Pickands, J. (1975), Statistical inference using extreme order statistics, *Ann. Statist.*, *3*, 119–131.

Porat, E., and M. J. Strauss (2012), Sublinear time, measurement-optimal, sparse recovery for all, in *ACM-SIAM SODA*.

Ramsay, J., J. Ramsay, and B. Silverman (2005), *Functional Data Analysis*, Springer Series in Statistics, Springer.

Ranshous, S., S. Shen, D. Koutra, S. Harenberg, C. Faloutsos, and N. F. Samatova (2015), Anomaly detection in dynamic networks: a survey, *Wiley Interdisciplinary Reviews: Computational Statistics*, *7*(3), 223–247, doi:10.1002/wics.1347.

Resnick, S. I. (2007), in *Heavy-Tail Phenomena Probabilistic and Statistical Modeling*, edited by S. M. R. Thomas V. Mikosch, Sidney I. Resnick, Springer Series in Operations Research and Financial Engineering, pp. 114–115.

Rossow, C. (2014), Amplification Hell: Revisiting Network Protocols for DDoS Abuse, in *Proceedings of the 2014 Network and Distributed System Security (NDSS) Symposium*.

Rousseeuw, P. (1987), Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.*, *20*(1), 53–65, doi:10.1016/0377-0427(87)90125-7.

Rue, H., and L. Held (2005), *Gaussian Markov Random Fields: Theory And Applications (Monographs on Statistics and Applied Probability)*, Chapman & Hall/CRC.

Schweller, R., Z. Li, Y. Chen, Y. Gao, A. Gupta, Y. Zhang, P. Dinda, M.-Y. Kao, and G. Memik (2006), Reverse hashing for high-speed network monitoring: Algorithms, evaluation, and applications, in *INFOCOM 2006*, pp. 1–12, doi:10.1109/INFOCOM.2006.203.

Senie, D. (1998), Network ingress filtering: Defeating denial of service attacks which employ ip source address spoofing.

Smith, R. L. (1989), Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone, *Statist. Sci.*, *4*(4), 367–377, doi:10.1214/ss/1177012400.

splines (2000), splines package function - r documentation, https://www.rdocumentation.org/packages/splines/versions/3.4.3/.

Stoev, S., and G. Michailidis (2010), On the estimation of the heavy–tail exponent in time series using the max–spectrum, *Applied Stochastic Models in Business and Industry*, *26*(3), 224–253.

Stoev, S., M. S. Taqqu, C. Park, and J. S. Marron (2005), On the wavelet spectrum diagnostic for Hurst parameter estimation in the analysis of Internet traffic, *Computer Networks*, *48*, 423–445.

Stoev, S., M. Taqqu, C. Park, G. Michailidis, and J. S. Marron (2006), LASS: a tool for the local analysis of self-similarity, *Computational Statistics and Data Analysis*, *50*, 2447–2471.

Stoev, S., M. Hadjieleftheriou, G. Kollios, and M. Taqqu (2007), Norm, point, and distance estimation over multiple signals using max-stable distributions, in *IEEE 23rd International Conference on Data Engineering*, pp. 1006–1015, doi:10.1109/ICDE.2007.368959.

Stoev, S., G. Michailidis, and M. Taqqu (2011), Estimating heavy–tail exponent through max self–similarity, *IEEE Transactions on Information Theory*, *57*(3), 1615–1636.

Tait, A., R. Henderson, R. Turner, and X. Zheng (2006), Thin plate smoothing spline interpolation of daily rainfall for new zealand using a climatological rainfall surface, *International Journal of Climatology*.

Vandewalle, B., J. Beirlant, A. Christmann, and M. Hubert (2007), A robust estimator for the tail index of pareto-type distributions, *Comput. Stat. Data Anal.*, *51*(12), 6252–6268, doi:10.1016/j.csda.2007.01.003.

Vaughan, J., S. Stoev, and G. Michailidis (2013), Network-wide statistical modeling, prediction, and monitoring of computer traffic, *Technometrics*, *55*(1), 79–93, doi:10.1080/00401706.2012.723959.

Victoria-Feser, M.-P., and E. Ronchetti (1994), Robust methods for personal-income distribution models, *Canadian Journal of Statistics*, *22*(2), 247–258, doi:10.2307/3315587.

Wang, W., and Z. Lu (2013), Survey cyber security in the smart grid: Survey and challenges, *Comput. Netw.*, doi:10.1016/j.comnet.2012.12.017.

Wilson, D. C., and B. A. Mair (2004), *Thin-Plate Spline Interpolation*, Birkhäuser Boston, Boston, MA, doi:10.1007/978-0-8176-8212-5_12.

Winter, H. C., J. A. Tawn, and S. J. Brown (2016), Modelling the effect of the el nio-southern oscillation on extreme spatial temperature events over australia, *Ann. Appl. Stat.*, *10*(4), 2075–2101, doi:10.1214/16-AOAS965.

Xi, B., G. Michailidis, and V. N. Nair (2006), Estimating network loss rates using active tomography, *J. Amer. Statist. Assoc.*, *101*(476), 1430–1448.

Yu, W. e. a. (2015), An integrated detection system against false data injection attacks in the smart grid, *Security and Communication Networks*, *8*(2), doi:10.1002/sec.957.

Zou, J., R. A. Davis, and G. Samorodnitsky (2017), Extreme Value Analysis Without the Largest Values: What Can Be Done?, *ArXiv e-prints*.