On a Notion of Outliers Based on Ratios of Order Statistics

Ahmet Zahid Balcıoğlu* Oğuz Gürerk†

July 27, 2022

Abstract

There are a number of mathematical formalisms of the term "outlier" in statistics, though there is no consensus on what the right notion ought to be. Accordingly, we try to give a consistent and robust definition for a specific type of outliers defined via order statistics. Our approach is based on ratios of partial sums of order statistics to investigate the tail behaviors of hypothetical and empirical distributions. We simulate our statistic on a set of distributions to mark potential outliers and use an algorithm to automatically select a cut-off point without the need of any further a priori assumption. Finally, we show the efficacy of our statistic by a simulation study on distinguishing two Pareto tails outside of the Lévy stable region.

Keywords: order statistics, outliers, anomaly detection, Pareto distribution, exponential distribution

MSC Classification: 62G30, 62E17, 62C05

1 Introduction

The problem of existence of outliers\textsuperscript{1} or outlier detection "[have] been recognized for a very long time, certainly since the middle of the eighteenth century. Daniel Bernoulli, writing in 1777 about the combination of astronomical observations, said:

Is it right to hold that the several observations are of the same weight or moment, or equally prone to any and every error? ... Is there everywhere the same probability? Such an assertion would be quite absurd, which is undoubtedly the reason why astronomers prefer to reject completely observations which they judge to be too wide of the truth, while retaining the rest and, indeed, assigning to them the same reliability. ... I see no way of drawing a dividing line between those that are to be utterly rejected and those that are to be wholly retained; it may even happen that the rejected observation is the one that would have

\textsuperscript{1}[which] are also referred to as abnormalities, discordants, deviants, or anomalies in the data mining and statistics literature [1].
supplied the best correction to the others. Nevertheless, I do not condemn in every case the principle of rejecting one or other of the observations, indeed I approve it, whenever in the course of observation an accident occurs which in itself raises an immediate scruple in the mind of the observer, before he has considered the event and compared it with the other observations. If there is no such reason for dissatisfaction I think each and every observation should be admitted whatever its quality, as long as the observer is conscious that he has taken every care.” [2].

We refer the reader to [1] for a detailed conceptual account as the amount of literature on outliers is vast. However, for an inclusive definition of an outlier, we start with a general one given by Grubbs in 1969, “an outlying observation, or ‘outlier’, may be merely an extreme manifestation of the random variability inherent in the data. ... On the other hand, an outlying observation may be the result of gross deviation from prescribed experimental procedure or an error in calculating or recording the numerical value”[3]. Hawkins in 1980 defined the concept of an outlier as “[a]n outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”[4].

Now, before we start making clear what we imply by an outlier – an observation (or a subset of observations) in a set of data ‘which deviates so much from the remaining data as to arouse suspicions’, we admit that “it is a matter of subjective judgement on the part of the observer whether or not he picks out some observation (or set of observations) for scrutiny” [2]; however, our interest and main lines of inquiry rest in identifying observations which can be characterized as extreme in some way. In this paper, therefore, our purpose is to propose and investigate definitions of types of outliers so that we can minimize the number of data/sample-specific parameters in order for the working definition to have qualities that we expect from a mathematical definition to possess as well as to see if (and how) the new notion of outliers relates to some important results in probability theory and statistics; e.g. to the law of large numbers and the extreme value theory.

We first briefly discuss the work of Klebanov et al in [5] as they recently introduced and analyzed a new formulation for the notion of outliers based on order statistics due to certain drawbacks of the classical (and inherently conceptual) definitions of outliers given in [3], [2], [4] by letting $X_1, X_2, \ldots, X_n$ be i.i.d. non-negative continuous random variables and denoting the corresponding order statistics by $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. Then $X_{(n)}$ is said to be an outlier of order $1/\kappa$ if $X_{(n-1)} \leq \kappa X_{(n)}$, where $\kappa \in (0, 1)$ is some fixed number depending on the choice of the practitioner. Various properties of this new definition were investigated in their paper [5]. This new statistic, although giving a different look at the problem, has certain deficiencies.

For example, it applies only to cases with a single outlier, it is not robust in the sense that a small change in only a few points in the data may make an outlier a standard outcome, and vice versa. We generalizes this definition by considering the entire sample and looking at the ratios of partial sums of order statistics. Under the assumption that outliers are a subset of sample maxima, we consider the ratios of the form:

$$\frac{\sum_{i=1}^{m} X_{(i)}}{\sum_{i=m+1}^{n} X_{(i)}}$$

for $m \in \{1, \ldots, n\}$. We define $X_{m+1}, \ldots, X_n$ as outliers when the ratio (1) is greater than a (pre-determined) threshold $\kappa$ value.
Our contributions also include a python library focused on tail index estimation and other computational methods related to heavy-tailed distributions [6] as well as the simulations of our statistic. In section 2 we formally introduce our statistic. Section 3 comprises some preliminary lemmas for our calculations. In section 4, we derive the distribution function for our statistic. In section 5 we give a concentration for the error margin in our calculations. In section 6, we discuss outlier generating models and generate comparative simulations of our proposed statistic with well-known distributions. Finally, we propose an algorithmic method for the selection of the \( \kappa \)-threshold value and use it for distinguishing between two Pareto tails.

2 A new notion of outliers

Let \( X_1, X_2, \ldots, X_n \) be i.i.d. random variables with cumulative distribution function \( F \). We may further assume that these are absolutely continuous, and call the common p.d.f. \( f \). Let \( X_{(1)}, \ldots, X_{(n)} \) be the corresponding (increasing) order statistics. We are to investigate the problem of outliers in a way that the number of \( \kappa \)-outliers is defined via moving averages as

\[
\mathcal{O}_n := n - \min \left\{ i : \frac{1}{i} \sum_{j=1}^{i} |X_{(j)}| < \kappa \frac{1}{n-i} \sum_{j=i+1}^{n} |X_{(j)}| \right\}
\] (2)

The definition (2) may also be potentially useful for analyzing time series in a nonparametric way (i.e., without the normality or a similar distributional assumption). Assume that the \( X_i \)'s are nonnegative. We start by defining the following statistics

\[
T_{k,n} \equiv T_k = \sum_{i=n-k+1}^{n} X_{(i)}, \quad 1 \leq k \leq n
\] (3)

denoting the sum of the top \( k \in \{1, \ldots, n\} \) order statistics from a sample of size \( n \) and

\[
S_{m,n} \equiv S_m = \sum_{i=1}^{m} X_{(i)}, \quad 1 \leq m \leq n
\] (4)

denoting the sum of the first \( m \in \{1, \ldots, n\} \) order statistics from the same sample. Subsequently, putting (3) and (4) together, we investigate probabilities of the form \( \mathbb{P} \left( \frac{1}{m} S_m < \kappa \frac{1}{n-m} T_{n-m} \right) \) where \( \kappa \in (0,1) \) is fixed. Furthermore, when \( n \) and \( m \) are fixed, we may redefine \( \kappa \equiv \kappa(n,m) := \frac{m}{n-m} \kappa \), for convenience. In particular, this probability admits a closed form expression which we were able to simplify to an explicit formula for certain special cases. In order to compute the probability

\[
\mathbb{P} \left( \frac{S_m}{T_{n-m}} < \kappa \right),
\]

we exploit the Markov Property (MP) that the order statistics possess in order to compute the following conditional probability (which is slightly different than the probability we are interested in).

\[
\mathbb{P} \left( \frac{S_{m-1}}{T_{n-m}} < \kappa \mid X_{(m)} = u \right)
\] (5)
This is because the conditional distributions of a subset of the order statistics given another subset satisfy some really structured properties including the MP. For reference, see [7], [8], [9], [10]. The following three lemmas we state are instrumental in calculating the probability defined above. They are well-known and follow from direct computations, so we omit the proofs.

3 Preliminaries

**Lemma 3.1** Let \(X_1, X_2, \ldots, X_n\) be independent observations from a continuous cdf \(F\) with density \(f\). Fix \(1 \leq i < j \leq n\). Then, the conditional distribution of \(X_i\) given \(X_j = x\) is the same as the unconditional distribution of the \(i\)-th order statistic in a sample of size \(j - 1\) from a new distribution, namely the original \(F\) truncated at the right at \(x\). In notation,

\[
f_{X_i \mid X_j = x}(u) = \frac{(j - 1)!}{(i - 1)!(j - i)!} \left( \frac{F(u)}{F(x)} \right)^{i-1} \left(1 - \frac{F(u)}{F(x)}\right)^{j-1-i} \frac{f(u)}{F(x)}, u < x.
\]

**Lemma 3.2** Let \(X_1, X_2, \ldots, X_n\) be independent observations from a continuous cdf \(F\) with density \(f\). Fix \(1 \leq i < j \leq n\). Then, the conditional distribution of \(X_j\) given \(X_i = x\) is the same as the unconditional distribution of the \((j - i)\)-th order statistic in a sample of size \(n - i\) from a new distribution, namely the original \(F\) truncated at the left at \(x\). In notation,

\[
f_{X_j \mid X_i = x}(u) = \frac{(n - i)!}{(j - i - 1)!(n - j)!} \left( \frac{F(u) - F(x)}{1 - F(x)} \right)^{j-i-1} \left(\frac{1 - F(u)}{1 - F(x)}\right)^{n-j} \frac{f(u)}{1 - F(x)}, u > x.
\]

**Lemma 3.3 (Markov Property)** Let \(X_1, X_2, \ldots, X_n\) be independent observations from a continuous cdf \(F\) with density \(f\). Fix \(1 \leq i < j \leq n\). Then, the conditional distribution of \(X_j\) given \(X_1 = x_1, X_2 = x_2, \ldots, X_i = x_i\) is the same as the conditional distribution of \(X_j\) given \(X_i = x_i\). That is, given \(X_i\), \(X_j\) is independent of \(X_1, X_2, \ldots, X_{i-1}\).

**Corollary 3.1** Let \(X_1, X_2, \ldots, X_n\) be independent observations from a continuous cdf \(F\) with density \(f\). Then, the conditional distribution of \(X_1, X_2, \ldots, X_{n-1}\) given \(X_n = x\) is the same as the unconditional distribution of the order statistics in a sample of size \(n - 1\) from a new distribution, namely the original \(F\) truncated at the right at \(x\). In notation,

\[
f_{X_1, \ldots, X_{n-1} \mid X_n = x}(u_1, \ldots, u_{n-1}) = (n - 1)! \prod_{i=1}^{n-1} \frac{f(u_i)}{F(x)}, \text{ where } u_1 < \cdots < u_{n-1} < x.
\]

A similar result holds for the conditional distribution of \(X_2, X_3, \ldots, X_n\) given \(X_1 = x\)

**Corollary 3.2** If \(F\) is absolutely continuous, the order statistics, \(X_1, \ldots, X_n\), form a (discrete time) Markov chain with transition densities:

\[
f_{i+1|i}(y \mid x) = (n - i) \left( \frac{F(y) - F(x)}{1 - F(x)} \right)^{n-i-1} \frac{f(y)}{1 - F(x)}, \text{ for } y > x; \quad i = 1, \ldots, n - 1.
\]
An important consequence of the two corollaries is that when conditioned on the \( m \)-th order statistic the sums \( S_{m-1} \) and \( T_{n-m} \) are independent. In particular:

**Proposition 3.1** Let \( F \) be absolutely continuous, then for any \( 1 < k < n \), the random vectors
\[
X^{(1)} = (X_{(1)}, \ldots, X_{(k-1)}) \quad \text{and} \quad X^{(2)} = (X_{(k+1)}, \ldots, X_{(n)})
\]
are conditionally independent given that \( X_{(k)} = x_k \), that is to say
\[
\mathbb{P}(X^{(1)} \in B_1, X^{(2)} \in B_2 \mid X_{(k)} = x_k) = \mathbb{P}(X^{(1)} \in B_1 \mid X_{(k)} = x_k) \mathbb{P}(X^{(2)} \in B_2 \mid X_{(k)} = x_k)
\]
for any Borel set \( B_1 \in \mathcal{B}(\mathbb{R}^{k-1}) \) and \( B_2 \in \mathcal{B}(\mathbb{R}^{n-k}) \). Furthermore,
\[
[S_{m-1} \mid X_{(m)} = u] \quad \text{and} \quad [T_{n-m} \mid X_{(m)} = u]
\]
are independent as well since \( S_{m-1} = \sum_{i=1}^{m-1} X_{(i)} \) and \( T_{n-m} = \sum_{i=m+1}^{n} X_{(i)} \) are linear functions of \( X^{(1)} \) and \( X^{(2)} \), respectively.

**4 Finite sample statistics**

As previous, we let \( X_1, X_2, \ldots, X_n \) be i.i.d. non-negative random variables with absolutely continuous distribution function \( F \), and call the common p.d.f. \( f \). Let \( X_{(1)}, \ldots, X_{(n)} \) be the corresponding order statistics.

**4.1 Sum of the top order statistics**

Observe, denoting the distribution function of \( X_{(i)} \) by \( F_i \), that
\[
\mathbb{P}(T_k < t) = \mathbb{P}\left( \sum_{i=n-k+1}^{n} X_{(i)} < t \right) = \int \mathbb{P}\left( \sum_{i=n-k+1}^{n} X_{(i)} < t \mid X_{n-k} = u \right) dF_{X_{n-k}}(u)
\]

We know that for each \( n-k < j \leq n \), the conditional distribution of \( X_{(j)} \) given \( X_{(n-k)} = u \) is formed from an i.i.d. sample of size \( k \) having the cdf \( G_u \) given by
\[
G_u(t) = \begin{cases} 0, & t < u \\ \frac{F(t) - F(u)}{1 - F(u)}, & t \geq u. \end{cases}
\]

Hence,
\[
\mathbb{P}(T_k < t) = \int_0^t G_u^{*(k)}(t) dF_{X_{n-k}}(u),
\]
where \( G_u^{*(k)} \) is the \( k \)-fold convolution of \( G_u \).
We note that $T_k$ is also related to the selection differential, which is a familiar term in the genetics literature [11], [12], [13] given by

$$D_k = \frac{1}{\sigma} \left( \frac{1}{k} \sum_{j=n-k+1}^{n} X_{(j)} - \mu \right)$$

where $\mu$ and $\sigma$ are the population mean and standard deviation, respectively, [14]. It also serves as a test statistic for outliers in samples from normal distribution, [2],[12]; additionally, from [15], one can obtain the asymptotic distribution of $D_k$ (when $X_i$’s have finite second moment) under suitable centering and scaling if $n \to \infty$ with $k$ fixed as well as if $k = [np], 0 < p < 1$ and $n \to \infty$.

**Sum of the first $m$ order statistics:** Following a similar argument as before, we get:

$$\mathbb{P} (S_m < t) = \int_{0}^{t} H_u^{* (m)}(t) dF_{X_{m+1}}(u),$$

where $H_u^{* (m)}$ is the $m$-fold convolution of the df $H_u$ given by

$$H_u(t) = \begin{cases} 
F(t), & t \leq u \\
0, & t > u.
\end{cases}$$

4.2 The case of $\mathbb{P} \left( \frac{S_{m-1}}{T_{n-m}} < \kappa \right)$

Now, we are ready to compute the probability (5) in which we were interested in getting a somewhat "nice" expression for

$$\mathbb{P} \left( \frac{S_{m-1}}{T_{n-m}} < \kappa \mid X_{(m)} = u \right)$$

Define for a fixed $m \in \{1, 2, \ldots, n-1\}$,

$$X^{(1)} = (X_{(1)}, \ldots, X_{(m-1)})$$

and

$$X^{(2)} = (X_{(m+1)}, \ldots, X_{(n)}).$$

Then, it follows from Proposition 3.1 that $[X^{(1)} \mid X_{(m)} = u], [X^{(2)} \mid X_{(m)} = u]$ and $[S_{m-1} \mid X_{(m)} = u], [T_{n-m} \mid X_{(m)} = u]$ are independent.

Note that we have

$$f_{X^{(1)} \mid X_{(m)}}(x^{(1)}) \equiv f_{X_{(1)} \ldots, X_{(m-1)} \mid X_{(m)}} (x_1, \ldots, x_{m-1}) = (m-1)! \prod_{i=1}^{m-1} \frac{f(x_i)}{F(u)},$$

for $x_1 < \cdots < x_{m-1} < x_m = u$. Also,

$$f_{X^{(2)} \mid X_{(m)}}(x^{(2)}) \equiv f_{X_{(m+1)} \ldots, X_{(n)} \mid X_{(m)}} (x_{m+1}, \ldots, x_n) = (n-m)! \prod_{i=m+1}^{n} \frac{f(x_i)}{1-F(u)}.$$
for \( u = x_m < x_{m+1} < \cdots < x_n \).

So, let \( R = S_{m-1}/T_{n-m} \), and observe that

\[
f_{S_{m-1},T_{n-m}|X(m)}(t_1, t_2) = f_{S_{m-1}|X(m)}(t_1) \cdot f_{T_{n-m}|X(m)}(t_2)
\]

where

\[
f_{S_{m-1}|X(m)}(t_1) = (m-1)\text{-fold conv. using } f_X^{(1)}|X(m)
\]

and

\[
f_{T_{n-m}|X(m)}(t_2) = (n-m)\text{-fold conv. using } f_X^{(2)}|X(m)
\]

Hence,

\[
\mathbb{P}(R < \kappa \mid X(m) = u) = \mathbb{P}\left( \frac{S_{m-1}}{T_{n-m}} < \kappa \mid X(m) = u \right)
= \int_0^\infty f_{T_{n-m}|X(m)}(t_2) \left( \int_0^{\kappa t_2} f_{S_{m-1}|X(m)}(t_1)dt_1 \right) dt_2
= \int_0^\infty f_{T_{n-m}|X(m)}(t_2) \cdot H_u^{(m-1)}(\kappa t_2)dt_2
= : F_{R|m}(\kappa)
\]

where \( H_u^{(m-1)} \) is the \((m-1)\text{-fold convolution of the df } H_u \text{ given by}

\[
H_u(t) = \begin{cases} 
\frac{F(t)}{F(u)}, & t \leq u \\
0, & t > u,
\end{cases}
\]

which can be calculated explicitly using \( f_X^{(1)}|X(m) \), but it is generally hard. (However, we will calculate it when we consider specific distributions, e.g. the exponential distribution.)

Differentiating \( F_{R|m}(\kappa) \) with respect to \( \kappa \), we get the pdf \( f_{R|m}(\kappa) \):

\[
f_{R|m}(\kappa) = \frac{d}{d\kappa} \left[ \int_0^\infty f_{T_{n-m}|X(m)}(t_2) \left( \int_0^{\kappa t_2} f_{S_{m-1}|X(m)}(t_1)dt_1 \right) dt_2 \right]
= \int_0^\infty f_{T_{n-m}|X(m)}(t_2) \cdot f_{S_{m-1}|X(m)}(\kappa t_2)dt_2
\]

Therefore,

\[
\mathbb{P}(R < \kappa) = \int_0^\infty \mathbb{P}(R < \kappa \mid X(m) = u) dF_m(u)
= \int_0^\infty F_{R|m}(\kappa)dF_m(u)
= \int_0^\infty \left( \int_0^\infty f_{T_{n-m}|X(m)}(t_2) \left( \int_0^{\kappa t_2} f_{S_{m-1}|X(m)}(t_1)dt_1 \right) dt_2 \right) dF_m(u)
= \int_0^\infty \left( \int_0^\infty f_{T_{n-m}|X(m)}(t_2) \cdot H_u^{(m-1)}(\kappa t_2)dt_2 \right) dF_m(u)
\]

where \( dF_m(u) = f_m(u)du = \frac{n!}{(m-1)!(n-m)!} F^{m-1}(u)[1 - F(u)]^{n-m} f(u)du. \)
4.3 Application: the case of the exponential distribution

Let the parent distribution be the std. exponential; i.e., \( X_i \sim \text{Exp}(1) \) for each \( i = 1, 2, \ldots, n \). Then, it is well-known that (e.g., [9], [16])

\[
X_i \overset{d}{=} \sum_{k=1}^{i} \frac{Z_k}{n-k+1},
\]

where \( Z_k \sim \text{Exp}(1) \) for each \( k \geq 1 \). Now, define \( \hat{Z}_k := \frac{Z_k}{n-k+1} \) so that \( \hat{Z}_k \sim \text{Exp}(\lambda_k^{-1}) \), \( \lambda_k = (n - k + 1) \). Hence,

\[
T_{n-m} = \sum_{i=m+1}^{n} X_i \overset{d}{=} \sum_{i=m+1}^{n} \sum_{k=1}^{i} \hat{Z}_k
\]

\[
= (n - m) \sum_{k=1}^{m+1} \hat{Z}_k + \sum_{k=m+2}^{n} (n - k + 1) \hat{Z}_k
\]

\[
= \sum_{k=1}^{m+1} \beta_k Z_k + W
\]

where \( \beta_k := \frac{n-m}{n-k+1} = (n - m)\lambda_k^{-1} \) and \( W \sim \text{Gamma}(n - m - 1, 1) \).

Now, consider the theorem below by Jasiulewicz, H. and Kordecki, W. to find the pdf of \( \sum_{k=1}^{m+1} \beta_k Z_k \) in the expansion of \( T_{n-m} \).

**Theorem 4.1** [17] Let \( X_1, \ldots, X_n \) be \( n \) independent random variables such that every \( X_i \) has a probability density function \( f_{X_i}(t) \) given by

\[
f_{X_i}(t) := \beta_i \exp (-t\beta_i) \mathbb{1}_{(0,\infty)}(t)
\]

for all real number \( t \), where the parameter \( \beta_i \) is positive for all \( i = 1, 2, \ldots, n \). We suppose that the parameters \( \beta_i \) are all distinct. Then the sum \( S_n \) has the following probability density function:

\[
f_{S_n}(t) = \sum_{i=1}^{n} \beta_1 \ldots \beta_n \prod_{j=1}^{n} (\beta_j - \beta_i) \exp (-t\beta_i) \mathbb{1}_{(0,\infty)}(t)
\]

for all \( t \in \mathbb{R} \).

So, by Theorem 4.1 we find the pdf of \( L := \sum_{k=1}^{m+1} \beta_k Z_k \) in the expansion of \( T_{n-m} \) by noting that the \( \beta_i \)'s in the theorem correspond to \( \frac{\lambda_i}{n-m} \) in our notation. Therefore,

\[
f_L(x) = \left( \frac{\lambda_1}{n-m} \right) \left( \frac{\lambda_2}{n-m} \right) \cdots \left( \frac{\lambda_{m+1}}{n-m} \right) \sum_{i=1}^{m+1} \Psi_{i,m+1} \cdot e^{-\left( \frac{\lambda_i}{n-m} \right) x},
\]

where

\[
\Psi_{i,m+1}^{-1} = \prod_{j=1}^{m+1} \frac{(\lambda_j - \lambda_i)}{n-m} = \frac{1}{(n-m)^m} (\lambda_1 - \lambda_i) \cdots (\lambda_{i-1} - \lambda_i) (\lambda_{i+1} - \lambda_i) \cdots (\lambda_{m+1} - \lambda_i)
\]

8
So,
\[ f_L(x) = \frac{\lambda_1 \lambda_2 \cdots \lambda_{m+1}}{(n - m)^{m+1}} \cdot (n - m)^m \sum_{i=1}^{m+1} \psi_{i,m+1} \cdot e^{-\frac{\lambda_i}{n-m}x} , \]
where
\[ \psi_{i,m+1} = (\lambda_1 - \lambda_i) \cdots (\lambda_{i-1} - \lambda_i)(\lambda_{i+1} - \lambda_i) \cdots (\lambda_{m+1} - \lambda_i) \]
and \( \Psi = (n - m)^m \cdot \psi_{i,m+1} \)

Then,
\[ f_L(x) = \frac{\lambda_1 \lambda_2 \cdots \lambda_{m+1}}{n - m} \sum_{i=1}^{m+1} \psi_{i,m+1} \cdot e^{-\frac{\lambda_i}{n-m}x} \]

Now, using convolution formula, we can compute \( f_{T_{n-m}} \) explicitly:
\[ f_{T_{n-m}}(t) = \int_0^\infty f_L(t - x)f_W(x)dx , \]
where \( W \sim \text{Gamma}(n - m - 1, 1) \).

\[ f_{T_{n-m}}(t) = \frac{\lambda_1 \lambda_2 \cdots \lambda_{m+1}}{n - m} \sum_{i=1}^{m+1} \psi_{i,m+1} \int_0^t e^{-\frac{\lambda_i}{n-m}(t-x)} \frac{1}{(n-m-2)!} e^{-x} x^{n-m-2} dx \]
\[ = \frac{1}{(k - 1)!} \frac{\lambda_1 \lambda_2 \cdots \lambda_{m+1}}{k + 1} \sum_{i=1}^{m+1} \psi_{i,m+1} e^{-\frac{\lambda_i}{k+1}t} \int_0^t e^{\left(\frac{\lambda_i}{k+1}t\right)x} x^{k-1} dx , \]
where we put \( n - m - 1 \equiv k \) (for convenience\(^2\)) and \( t > 0 \).

Recall that we wanted to get an expression for \( \mathbb{P}(R < \kappa) \), where \( R = S_{m-1}/T_{n-m} \):
\[ \mathbb{P} (R < \kappa) = \int_0^\infty \mathbb{P} \left( R < \kappa \mid X_{(m)} = u \right) dF_m(u) \]
\[ = \int_0^\infty \left( \int_0^\infty f_{T_{n-m}}(t_2) \cdot H_u^{(m-1)}(kt_2) dt_2 \right) dF_m(u) \]
where \( dF_m(u) = f_m(u)du = \frac{n!}{(m-1)!(n-m)!} (1 - e^{-u})^{m-1} e^{-(n-m)} e^{-u} du \), and \( H_u^{(m-1)} \) is the \((m - 1)\)-fold convolution of the df \( H_u \) given by
\[ H_u(t) = \begin{cases} 
\frac{1-e^{-t}}{1-e^{-u}}, & t \leq u \\
0, & t > u.
\end{cases} \]

Hence, an explicit expression for our target probability is available as all the ingredients are ready (up to scaling) to be employed.

---

\(^2\)For a given \( k \), one can evaluate the integral:
\[ \int_0^t e^{\left(\frac{\lambda_i}{k+1}t\right)x} x^{k-1} dx = t^k \left( t - \frac{\lambda_i t}{k+1} \right)^{-k} \left( \Gamma(k) - \Gamma \left( k, t - \frac{\lambda_i t}{k+1} \right) \right) \]
5 Approximating $\mathbb{P}(S_m/T_{n-m} \leq \kappa)$

As usual, for $X_1, X_2, \ldots, X_n$ non-negative i.i.d. random variables with absolutely continuous distribution function $F$, letting

$$T_{k,n} \equiv T_k = \sum_{i=n-k+1}^{n} X(i), \quad 1 \leq k \leq n$$

and

$$S_{m,n} \equiv S_m = \sum_{i=1}^{m} X(i), \quad 1 \leq k \leq n$$

observe, for $m = 1, 2, \ldots, n - 1$, that

$$\frac{S_m}{T_{n-m}} = \frac{S_{m-1} + X(m)}{T_{n-m}} = \frac{S_{m-1}}{T_{n-m}} + \frac{X(m)}{T_{n-m}}$$

where $T_{n-m} = \sum_{i=m+1}^{n} X(i)$.

Note that

$$\frac{X(m)}{T_{n-m}} = \frac{X(m)}{\sum_{i=m+1}^{n} X(i)} \leq \frac{1}{n-m} \quad \text{almost surely.}$$

Thus, we have

$$\mathbb{P}\left(\frac{S_{m-1}}{T_{n-m}} \leq \kappa\right) \leq \mathbb{P}\left(\frac{S_m}{T_{n-m}} \leq \kappa\right) \leq \mathbb{P}\left(\frac{S_{m-1}}{T_{n-m}} \leq \kappa + \frac{1}{n-m}\right)$$

Recall that we have, for $R = S_{m-1}/T_{n-m}$, that

$$f_{S_{m-1},T_{n-m}|X(m)}(t_1, t_2) = f_{S_{m-1}|X(m)}(t_1) \cdot f_{T_{n-m}|X(m)}(t_2)$$

So,

$$\mathbb{P}\left(R < \kappa + \frac{1}{n-m} \mid X(m) = u\right) = \mathbb{P}\left(\frac{S_{m-1}}{T_{n-m}} < \kappa + \frac{1}{n-m} \mid X(m) = u\right)$$

$$= \int_0^{\infty} f_{T_{n-m}|X(m)}(t_2) \left(\int_0^{(\kappa+\frac{1}{n-m})t_2} f_{S_{m-1}|X(m)}(t_1)dt_1\right)dt_2$$

Now note that

$$\int_{0}^{(\kappa+\frac{1}{n-m})t_2} f_{S_{m-1}|X(m)}(t_1)dt_1 = \int_{0}^{\kappa t_2} f_{S_{m-1}|X(m)}(t_1)dt_1 + \int_{\kappa t_2}^{(\kappa+\frac{1}{n-m})t_2} f_{S_{m-1}|X(m)}(t_1)dt_1$$

$$\leq \int_{0}^{\kappa t_2} f_{S_{m-1}|X(m)}(t_1)dt_1 + \frac{t_2}{n-m}$$

(6)
Also,
\[ \int_0^\infty f_{T_{n-m}\mid X(m)}(t_2) \frac{t_2}{n-m} \, dt_2 = \frac{\mathbb{E}[T_{n-m} \mid X(m)]}{n-m} \]

(7)

Therefore,
\[ \mathbb{P} \left( R < \kappa + \frac{1}{n-m} \right) = \int_0^\infty \mathbb{P} \left( R < \kappa + \frac{1}{n-m} \mid X(m) = u \right) \, dF_m(u) \]
\[ \leq \int_0^\infty \left( \mathbb{P} \left( R < \kappa \mid X(m) = u \right) + \frac{\mathbb{E}[T_{n-m} \mid X(m)]}{n-m} \right) \, dF_m(u) \text{ by (6) and (7)}, \]
\[ = \mathbb{P} (R < \kappa) + \frac{1}{n-m} \int_0^\infty \mathbb{E}[T_{n-m} \mid X(m) = u] \, dF_m(u) \]
\[ = \mathbb{P} (R < \kappa) + \frac{1}{n-m} \mathbb{E}[T_{n-m}] \]
\[ = \mathbb{P} (R < \kappa) + o(1), \quad n \to \infty \]

since if $X_i$’s have finite moment, then so is $T_{n-m}$, so that $\frac{\mathbb{E}[T_{n-m}]}{n-m} = o(1), \quad n \to \infty$. So, the rough approximation above gives an error bound vanishing in the order of $1/n$.

6 Simulations and Experiments

Empirically it is important to choose a good value of $\kappa$ which can distinguish between normal and anomalous observations. There are a few points which are of concern. First for a given distribution, a good $\kappa$ value should be able to distinguish between the centre and the tail of the distribution. Informally, $\kappa$ value should be natural to choose. Furthermore, it is important that, for a given distribution, the concentration of potential $\kappa$ values should be tight, potentially depending on the family of the given distribution.

Consequently, it is of importance to know how the statistic $R = \frac{S_m}{T_{n-m}}$ is distributed. We expect to see similar values of $\kappa$ on lower quantiles for most distributions. The differences between the $\kappa$ values should increase gradually towards the tail. Finally, at the tail we expect the $\kappa$ values to differ the most, with concentrations dependent on the tail index. From another perspective, when considering the moving averages of our statistic for the sample, we expect to have a sharp increase towards the end of the tail.

We will show simulations of $R$ for a set of distributions and anomaly generating models. As most outlier generator models depend upon the exponential distribution, we will use the exponential distribution for comparison. For anomaly generating model we will use the identified outliers model, in which observations $X_1, \ldots, X_n$ are not i.i.d. but some $k \in \{1, \ldots, n\}$ come from a separate distribution. We will consider the simplistic exponential case as described below [18].

**Identified Outliers Model**  
$X = \{X_1, \ldots, X_{n-k} : 1 - e^{-x/\theta}; \theta > 0\}, k$ is known and fixed, say for simplicity let’s suppose $k = 1$, $X_1, \ldots, X_{n-k}$ are independent and the index of the contaminant is also known. If we assume further that the distribution function of the contaminant is

\[ G(x) = F \left( b^{-1} x \right) = 1 - e^{-(b\theta)^{-1} x}, \quad x \in \mathbb{R} \]
for some \( b \geq 1 \), then, without loss of any generality, the joint distribution function of \( X_1, \ldots, X_n \) is given by

\[
F_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = \left[ \prod_{i=1}^{n-1} \left( 1 - e^{-x_i/\theta} \right) \right] \left( 1 - e^{-x_n/b\theta} \right)
\]

for \( x > 0, \theta > 0, b \geq 1 \).

The simulations for the identified outliers model. In these sets of simulations we took
\( k = 100, \theta = 1 \) for the original sample, and \( b = 3 \) for outliers.

**Exponential Distribution**  The simulations for exponential distribution, \( f(x; \theta) = \theta e^{-x\theta} \), with \( \theta = 1 \).

**Half-Normal Distribution**  We will also use the absolute value of the normal distribution, the half-normal distribution, as a way to quantify the outliers in a normally distributed sample. The behaviour of moving averages of \( R \)-statistic for the normal distribution is specifically relevant for any empirical study. As characterizing the tail and cut-off of the normal distribution has many applications. For the simulations of Half-Normal distribution, we will use the standard normal \( Z \sim N(0,1) \) and simulate \( Y = |Z| \).

Simulations are generated using the following procedure, first we generate and sort an i.i.d. sample of size \( n = 1000 \). Then we compute the \( R \)-statistic for some values of \( m \) in order to better observe the changes in the behaviour of the statistic. We choose the median, \( 5^{th} \) percentile, and \( 95^{th} \) percentile points as indicators. This procedure was repeated 10000 times. The values obtained are given in the figures 1 below. We also provide in figures 2 moving averages of \( R \)-statistic throughout a sample of \( n = 1000 \).
Figure 1: Distribution of the $R$ statistic across 10,000 samples for fixed $m = 50$, 500, and 950.
Figure 2: Samples of \( n=1000 \) and corresponding \( R \) statistic from exponential distribution and identified outliers model.

### 6.1 Automatic selection of \( \kappa \)-threshold

Ideally, we would like to define the cut-off point with respect to the derivatives of the distribution function of our statistic, which would then allow us to define a theoretical cut-off point for any sample. However, there is no established well-defined notion for an elbow point. The literature of elbow detection\(^3\) is therefore algorithmically focused [19].

The elbow detection literature is based on the following pointwise definition of the curvature of a function. Most works than define the elbow as the point of maximum curvature, and use algorithmic means to calculate it.

**Definition 6.1 (Curvature of a function [20])** For any continuous function \( f \), there exists a standard closed-form \( K_f(x) \) that defines the curvature of \( f \) at any point as a function of its first and second derivative:

\[
K_f(x) = \frac{f''(x)}{(1 + f'(x)^2)^{3/2}}
\]

However, the definition 6.1 of curvature does extend easily to discrete data, instead [20], [21], and [19] use Menger curvature, which is defined for three points as the curvature of the circle circumscribed by those points. There are also angle-based [20] and exponentially weighted moving average (EWMA) based methods [22] which use the differences between successive points and EWMA smoothing to check deviations from arrival times respectfully.

We will use the kneedle detection algorithm for estimating the ”elbow point” of our statistic [20]. The kneedle algorithm uses dynamic first derivative threshold, in combination

\(^3\)Also called knee or kneecap detection.
with the IsoData [23] to find the elbows of a discrete data. It can work on discrete datasets and has a sensitivity parameter which can be fine-tuned for how sensitively a knee is to be detected. While the smaller values of the sensitivity parameter respond to quick change, the larger values are more robust.

In the Figure 3 we revisit our simulation studies and show the simulation studies and show the chosen cut-off points using the kneedle algorithm, choosing the sensitivity parameter as 5.0.

**Half-Normal Distribution** | **Exponential Distribution** | **Identified Outliers Model**

![Graph 1](image1.png) | ![Graph 2](image2.png) | ![Graph 3](image3.png)

Figure 3: Simulation studies in figure 2 revisited with \( \kappa \)-threshold chosen automatically.

### 6.2 Application II: Distinguishing two Pareto tails

Now that we have a method for selecting a \( \kappa \) value, we conduct further simulation tests based on distinguishing two Pareto tails in order to show the efficacy of our statistic. Our goal is to find a threshold \( \kappa \) beyond which > .90 of the observations belong to the ”heavier” tail. We set up our experiment as follows. For given two tail indices \( \alpha_1 < \alpha_2 \), we first determine a \( \kappa \)-threshold value for the \( \alpha_1 \)-indexed Pareto distribution. Then we sample a \( N = 1000000 \) observations from both of the distributions and calculate our statistic across the entire sample. We report the percentage of observations from \( \alpha_2 \) in the samples above the predetermined threshold. We repeat this experiment a 1000 times in order to build a confidence interval on our results.

A key problem in our experiments is the selection \( \alpha_1 \) and \( \alpha_2 \) values, it has been established that as the two indices get closer it becomes numerically difficult to differentiate between the tails, even if one of the tails are outside of the Levy-stable regime [24]. We start our experiments with indices \( \alpha_1 = 1.5 \) and \( \alpha_2 = 2.5 \) and move \( \alpha_2 \) closer every iteration.

The results of the experiments are given in the table 1 below. When the tail indices are far apart our statistic produces an ideal change point for differentiating between the two tails. It is clear that as the two tail indices get closer our statistic becomes unable to distinguish the difference between the two tails. However, it remains consistent with variance \( \sim 10^{-5} \).
across all samples.

| \( \alpha_1 \) | \( \alpha_2 \) | -threshold | Expected percentage of \( \alpha_2 \) above \( \kappa \) | Variance |
|-------|-------|---------|------------------|--------|
| 1.5   | 2.5   | 2.745   | 0.95             | \( 1.88 \times 10^{-5} \) |
| 1.5   | 2.3   | 2.745   | 0.924            | \( 4.06 \times 10^{-5} \) |
| 1.5   | 2.1   | 2.745   | 0.85             | \( 1 \times 10^{-4} \)    |
| 1.5   | 1.9   | 0.78    | 0.95             | \( 7.37 \times 10^{-5} \) |
| 1.5   | 1.7   | 2.745   | 0.65             | \( 3.74 \times 10^{-5} \) |

Table 1: Simulation results of two Pareto tails.

7 Results and Discussions

In this work we stray away from an all-encompassing definition of an outlier, rather focus on a definition for the one dimensional case in terms of order statistics. In a sense our definition tries to approach the problem from the point of view of investigating data points “that arouse suspicions that it was generated by a different mechanism” [4].

There are a few good reasons motivating our decision. Similar to the case in extreme value theory, it is difficult to order random vectors and the limit cases are not as intuitive [25]. Furthermore, any multivariate definition must take into consideration cases where two or more random variables are not independent. Since the requirement of independence is too stringent for applied studies as conditional outliers are all too common an occurrence in real life [26].

A particular strong point of our method is the ability to select cut-off points for discrete set of points without the need of a priori information. As a result, our definition is innately compatible with any definition which produces outlier scores. For the multivariate case, since much of the literature is applied, we recommend the reader to first use the method which suits them best. Our definition, together with the \( \kappa \)-threshold selection, can be used afterwards to select a natural cut-off point.

We calculate the case \( S_{m-1}/T_{n-m} \) and approximate for the defined \( R \)-statistic \( S_m/T_{n-m} \). From the approximation, we can see that concentration of \( R \)-statistic depends on the tail of the random variable. Our simulation studies in 6 also confirmatory. In particular, in 1 it is considerably more difficult to find a cut-off point for the half-normal and exponential distributions compared to the identified outliers model. Nevertheless, we see in section 6.2 that even in the cases where the tail index is not in Lévy stable region, our statistic still produces results with very small variance with varying degrees of success.

For future work, we may choose the possible cut-off point by using the \( R \)-statistic on moving blocks first, then selecting the candidate blocks based on results. Finally, we can find the threshold value by looking at the block with the most dramatic increase. Identify potential blocks for the cut-off point is also helpful for the cases when a practitioner wishes to pick the value the \( \kappa \)-threshold by hand.
References

[1] C. C. Aggarwal, Outlier Analysis. Springer International Publishing, 2015.

[2] V. Barnett and T. Lewis, Outliers in statistical data. Wiley, 1978.

[3] F. E. Grubbs, “Procedures for detecting outlying observations in samples,” Technometrics, vol. 11, no. 1, pp. 1–21, 1969.

[4] D. M. Hawkins, Identification of outliers. Springer, 1980, vol. 11.

[5] L. B. Klebanov, A. V. Kakosyan, and A. Karlova, “Outliers, the law of large numbers, index of stability and heavy tails,” arXiv preprint arXiv:1612.09265, 2016.

[6] N. Markovich, Nonparametric analysis of univariate heavy-tailed data: research and practice. John Wiley & Sons, 2008, vol. 753.

[7] R.-D. Reiss, Approximate Distributions of Order Statistics: With Applications to Nonparametric Statistics. London: Springer-Verlag, 1989.

[8] V. B. Nevzorov, Records: Mathematical Theory, ser. Europe and Central Asia Environmentally and Socially Sustain. American Mathematical Society, 2001.

[9] H. A. David and H. N. Nagaraja, Order statistics. John Wiley & Sons, 2003.

[10] B. Arnold, N. Balakrishnan, and H. Nagaraja, A First Course in Order Statistics, ser. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 2008. [Online]. Available: https://books.google.com.tr/books?id=gUD-S8USIDwC

[11] J. Crow and M. Kimura, An Introduction to Population Genetics Theory. Burgess Publishing Company, 1970.

[12] P. M. Burrows, “Expected selection differentials for directional selection,” Biometrics, pp. 1091–1100, 1972.

[13] D. Falconer, Introduction to Quantitative Genetics, ser. Always Learning. Pearson Education, 1996.

[14] H. N. Nagaraja, “Some finite sample results for the selection differential,” Annals of the Institute of Statistical Mathematics, vol. 33, no. 3, pp. 437–448, 1981.

[15] H. Nagaraja, “Some nondegenerate limit laws for the selection differential,” The Annals of Statistics, pp. 1306–1310, 1982.

[16] V. B. Nevzorov, “Representations of order statistics, based on exponential variables with different scaling parameters,” Journal of Soviet Mathematics, vol. 33, no. 1, pp. 797–798, 1986.

[17] H. Jasinelewicz and W. Kordecki, “Convolutions of erlang and of pascal distributions with applications to reliability,” Demonstratio mathematica, vol. 36, no. 1, pp. 231–238, 2003.
[18] K. Balakrishnan, *Exponential distribution: theory, methods and applications*. Routledge, 2019.

[19] M. Antunes, D. Gomes, and R. L. Aguiar, “Knee/elbow estimation based on first derivative threshold,” in 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService), 2018, pp. 237–240.

[20] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, “Finding a "kneedle" in a haystack: Detecting knee points in system behavior,” in 2011 31st International Conference on Distributed Computing Systems Workshops, 2011, pp. 166–171.

[21] X. Tolsa, “Principal values for the cauchy integral and rectifiability,” Proceedings of the American Mathematical Society, vol. 128, no. 7, pp. 2111–2119, 2000.

[22] J. R. Albrecht, C. Tuttle, A. C. Snoeren, and A. Vahdat, “Loose synchronization for large-scale networked systems.” in USENIX Annual Technical Conference, General Track, 2006, pp. 301–314.

[23] T. Ridler, S. Calvard et al., “Picture thresholding using an iterative selection method,” IEEE trans syst Man Cybern, vol. 8, no. 8, pp. 630–632, 1978.

[24] R. Weron, “Levy-stable distributions revisited: tail index â‰¥ 2 does not exclude the levy-stable regime,” International Journal of Modern Physics C, vol. 12, no. 02, pp. 209–223, 2001.

[25] L. De Haan, A. Ferreira, and A. Ferreira, *Extreme value theory: an introduction*. Springer, 2006, vol. 21.

[26] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM computing surveys (CSUR), vol. 41, no. 3, pp. 1–58, 2009.