JMS

Journées de méthodologie statistique de l'Insee

2018

Using Extreme Value Theory to test for Outliers

Nathaniel GBENRO(*)(**)

(*)Ecole Nationale Supérieure de Statistique et d'Economie Appliquée
(**) Université de Cergy Pontoise

nathaniel.gbenro@u-cergy.fr

Mots-clés. Extreme Value Theory, Outliers, Hypothesis Tests, etc

Résumé

Dans cet article, il est proposé une procédure d'identification des outliers basée sur la théorie des valeurs extrêmes (EVT). Basé sur un papier de Olmo (2009), cet article généralise d'une part l'algorithme proposé puis d'autre part teste empiriquement le comportement de la procédure sur des données simulées. Deux applications du test sont conduites. La première application compare la nouvelle procédure de test et le test de Grubbs sur des données générées à partir d'une distribution normale. La seconde application utilisent divers autres distributions (autre que normales) pour analyser le comportement de l'erreur de première espèce en absence d'outliers. Les résultats de ces deux applications suggèrent que la nouvelle procédure de test a une meilleure performance que celle de Grubbs. En effet, la première application conclut au fait lorsque les données sont générées d'une distribution normale, les deux tests ont une erreur de première espèces quasi identique tandis que la nouvelle procédure a une puissance plus élevée. La seconde application, quant à elle, indique qu'en absence d'outlier dans un échantillon de données issue d'une distribution autre que normale, le test de Grubbs identifie quasi-systématiquement la valeur maximale comme un outlier ; La nouvelle procédure de test permettant de corriger cet fait.

Abstract

This paper analyses the identification of aberrant values using a new approach based on the extreme value theory (EVT). The aim of this paper is to suggest a new approach in the identification process of aberrant values in large sample. Similar in spirit with the algorithm suggested by Olmo (2009), this paper extend that algorithm test empirically using simulated data. Two empirical evidences have been made. The first one use simulated data from normal distribution and compare the new test and the Pearson's and Grubbs test. The second application, use also simulated data from various non-normal distributions, to analyse the ability of the new test to no reject the null hypothesis. The results suggest that the GEV outlier test best perform than Grubb's test whatever the underline distribution is normal or not.

1 Introduction

There is a wide literature dealing with outliers detection and treatment (Pearson & Sekar 1936) (Grubbs 1950) (Dixon 1950) (RANGER 1990) (Houfi & El Montasser 2009). This growing interest in the study of aberrant values is due to the fact that outliers could have serious impact on descriptive or econometric analysis. Indeed, (Houfi & El Montasser 2009) has shown that this kind of observation can lead statistical and econometric analysis to wrong result. Hence, it's important to identify this kind of data in order to either reject them, either correct them before analysis.

The outliers's issues is inevitable for those who have to analyze big dataset every day. Indeed, the data used for analysis are not, most of the time, cleaned from all error, especially when working on micro economic data (large N) like household data. For example, it is the case of collection error, entry error or sampling error. Even when the administrator of a data collection have try to clean every error from the dataset, it remains sometimes this kind of influential data. For those reasons, it is important to identify and eventually correct the outliers.

In spite of a wide literature on outliers, there is not a unique definition of that concept. It seems difficult to give an unique and universal definition. Indeed, in the literature, the definition of aberrant values is link to the hypothesis that should be verify. Although, all definitions agree on the fact that aberrant data are differents and move away from the others observations (Dixon 1950) (Grubbs 1950) (RANGER 1990). The most popular definition in the literature states that aberrant values are those with a cumulative density function (CDF) which move away from the CDF of others. That means that given $(X_1, X_2, \ldots, X_n)$, a series of variables, and $(x_1, x_2, \ldots, x_n)$ the observed values, the value x_m is consider as an aberrant value if the CDF of X_m recorded F_m deviate from F, the CFD of the others observations ; For instance, $F_m = (1-\epsilon)F + \epsilon G$, where $\epsilon \in]0, 0.5]$ and $G \neq F$ (two CDF).

In the literature, there are a lot of methods to address the identification and the treatment of outliers. (Chandola et al. 2009) and (Cousineau & Chartier 2015) made a survey which tries to provide a structured and comprehensive overview of the research on outlier's detection. Theses methods can be categorized into two broads groups : univariate and multivariate methods. In the present paper, we will focus on univariate approaches. The univariate methods include the following :

- statistics related to ratio excess / spread (Dixon 1950) ;
- statistics linked to the ratio amplitude / spread (Dixon 1950) ;
- statistics linked to the ratio gap / spread, like the classic test of Grubbs (Grubbs 1950) ;
- statistics linked to the extreme / position report (Thompson 1935) (Pearson & Sekar 1936) (Grubbs 1950) ;
- statistics linked to the ratio of sums of squares (Dixon 1950) ;
- statistics linked to higher order moments ;
- Shapiro-Wilks W statistic (Shapiro et al, 1968 ; Royston. 1982).

Many other methods have been develop to identify outliers in several cases : box plot, modified central tendency indicators, multivariate outliers methods detection, regression case and so on. More details can be obtain in (Nikulin & Zerbet 2002). In this paper, we will focus on two of these tests which are based on the maximum statistic :(Pearson & Sekar 1936) and (Grubbs 1950). The latter are one of the most used methods in practice. As most of the outlier's methods detection, they based on the fact that the data generator process is gaussian. When the underlying data are not normal, they are not useful anymore. (Olmo 2009) suggest a new test based on Extreme

Value Theory (EVT). This test allows to deal with non normal data. However, the suggested test is restricted to some family of distributions.In this paper, we extend it to all distributions and made empirical applications.

This paper have two mains contributions. The first contribution is to suggest a new procedure to test for outliers whatever the underlying distribution (either normal or not). Using some properties of EVT, we are able to use only one critical value for the test. The second contribution is an empirical exercice. Two empirical applications have been made. The first application compare the new EVT based test and one popular outliers test : the Grubbs test. We shown that the new EVT based test perform well than the Grubbs's test and more there is outliers, more the new test perform well than Grubbs's test. We also shown that when the underlying distribution isn't normal, the type I error is still under control of the new EVT based test.

The paper is organized as follows. In the following section, we review background material of (Pearson & Sekar 1936) and (Grubbs 1950) tests. Section 3 summarizes the main results of EVT used to compute our new test's statistic. In section 4, we present test and give some properties. Two empirical evidences have been done in last section 5.

2 The motivation

In this section, we present Pearson's and Grubbs's test since there are the more popular outliers test. We highlight the fact that those test tests are based on the gaussian assumption. Indeed, when the underlying distribution isn't normal, these tests reject systematically the null hypothesis.

2.1 (Pearson & Sekar 1936) and (Grubbs 1950)

Let's consider a sample $x_1, x_2, \ldots, x_n$ generated from a **normal distribution** with mean μ_i and variance σ^2. The ordered sample is noted $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. We like to know if $x_{(n)}$ is an outlier observation or not. Let's start by the definition of outlier of Pearson test given by (Nikulin & Zerbet 2002).

Definition 1. *Given a sample generated from a normal distribution with mean* μ_i *and variance* σ^2*. We define :* $H_0 : \mu_1 = \mu_2 = \ldots = \mu_{m-1} = \mu_{m+1} = \ldots = \mu_n = \mu, \quad \mu_m = \mu + d$
and $H_{a1} = d \neq 0, \quad H_{a2} = d < 0, \quad H_{a3} = d > 0.$
When the alternative hypothesis is H_{a1}*, this is the Grubbs test,* H_{a2} *and* H_{a3} *are the test for the maximum and the minimum.*

It is important to understand the null hypothesis and the alternative. The null hypothesis is the fact that the sample is an iid sample and the alternative is the fact that one observation has different distribution. So the test stipulate that under the null hypothesis, the observation is not an outlier and under the alternative it's an outlier. This is equivalent to say that, the observation is an outlier iff the δ didn't behave like it must if H_0 was true. This same approach will be use to define an outlier in the general case where the CDF F is not necessary a normal distribution.

In the original paper of (Thompson 1935), the test's statistic is $\tau = \frac{\delta}{s}$ (a studentized statistic), where :

$$\delta = x_{(n)} - \bar{x} \quad \text{and} \quad s = \sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{2.1}$$

(Pearson & Sekar 1936) shown that when changing the alternative hypothesis or when there are more than one outlier, Thompson statistic is not efficient. For this purpose, (Pearson & Sekar 1936) propose the Thompson statistic for testing only the most extreme (either the smallest or the highest observation) as outlier. (Pearson & Sekar 1936) tabulate the critical values in the case of normal distribution either the parameters are know or not [1] without derivate the exact distribution of τ.

Example 1 : Barnett et Lewis (1994) Given a sample of size 10, $x_1 = 1.74$, $x_2 = 1.46$, $x_3 = 1.28$, $x_4 = -0.02$, $x_5 = -0.40$, $x_6 = 0.02$, $x_7 = 3.89$, $x_8 = 1.35$, $x_9 = -0.10$ and $x_{10} = 1.71$.

If we want to test if the largest observation is an outlier according to Pearson's test, we will compute τ statistic. If we suppose that the underlying distribution is normal, we can use Pearson approach. We compute $\tau_{(n)}$ as earlier using the empirical estimator of the mean and variance, $\tau_{(n)} = \frac{x_{(n)}-\bar{x}}{\sigma}$ where $\bar{x}$ and σ are the empirical mean and standard deviation of the sample. x_7=3.89 is found to be an outlier since $\tau_{(n)}$ is greater than 2.29 [2].

Since (Pearson & Sekar 1936) tabulate the critical values without deriving the exact distribution, (Grubbs 1950) derive the exact distribution of the statistic τ and expand the test to the bilateral case. Our main concern in this paper is to test for outlier even when the underlying distribution is not normal. In the next sub section, we shown empirically that Grubbs's or Pearson's test aren't applicable when the underlying distribution isn't normal.

2.2 Sample from non normal parent

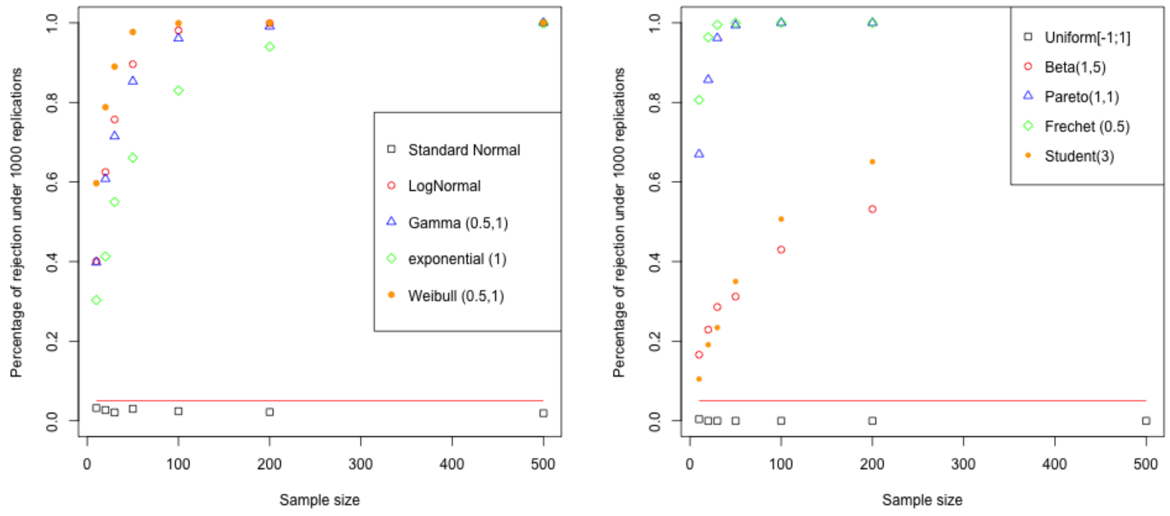
In practice, the distribution of a given sample is unknown although, there is a lot of adequation test like QQ plot, PP plot, Jacques Bera test, Kolmogorov-Smirov test. Grubb's test can't be used when the underlying distribution isn't normal. We test empirically this fact by simulating 1 000 samples of different size (N=10, 20, 30, 50, 100, 200, 500) from different parents [3]. For each sample, we test if the maximum $x_{(n)}$ is an outlier using Grubb's test. We report on y-axis the percentage of the number of times that the maximum have been identify as outlier. According to test theory, since the sample is generated from the same parent, the Grubbs's test should not reject the null hypothesis or to be exact, the rejection rate shouldn't exceed the type I error. So, if the test is suitable, the type I error (5%) should be the same for all sample i.e the null hypothesis should be reject less than 5% (see Figure 1a).

In most cases of samples drawn from normal distribution, the test didn't reject the null hypothesis more than the rate of the type I error. It's seem also that for uniform distribution, the test steel perform well by not identifying the highest observation as outlier (whatever the sample size). On the other hand, we can said that the test isn't suitable for most distributions like exponential, gamma, Paretto, and so on. Indeed, for majority of non normal distributions, Grubbs test identify the maximum as an outliers particularly when the sample size become large.

To overcome, the case of non normal distribution, a naive statistic which could be used is $U_+ = \max x_i$. we can notice that the general distribution of U_+ can be obtained simply when F is known. U_+'s CFD is given by :

$$P(U_+ \leq x) = F^n(x) \tag{2.2}$$

1. either the mean and the variance are know, either the mean is know and the variance are unknown or the mean is unknown and the variance is know, the both are unknown
2. See the critical value's table of Pearson
3. normal, log-normal, gamma, exponential, Weibull, uniform, beta, Pareto and Frechet



(a) Group 1 (b) Group 2

FIGURE 1 – Application of Grubbs criterion under normal and no normal population. The red line indicate the 5% level. The data have been simulated from different parent which are normal, log-normal, gamma, exponential, Weibull, uniform, beta, Pareto and Frechet. Two groups have been created. The first group contain distributions which have an exponential right tail while group 2 contains distributions with polynomial decaying and those with finite right endpoint. More details on the two groups is given later in the paper. We simulate 1 000 samples of different size (N=10, 20, 30, 50, 100, 200, 500). For each sample, we test if the maximum $x_{(n)}$ is an outlier using Grubbs's test. Since the sample is generate from the same parent, the Grubbs's test should not reject the null hypothesis more than 0.05.

The advantages of Grubbs's approach is the fact that the test can be computed even when the parameters are unknown while this naive approach suppose the knowledge of the underlying distribution. Actually, another problem with this naive approach is the fact that the distribution of U_+ is degenerated asymptotically. To overcome this fact and to generalized the approach of Pearson and Grubb to non normal distributions, we will propose another approach based on the extreme values theory. This approach use a normalized version of the above statistic. We present in the next section some tools from extreme value theory which would be use for the new approach.

3 Statistic From Extreme Value Theory

3.1 Generality on EVT

Given an iid ordered sample of size n from distribution function F (at least for n-1 observations), $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, our aims is this section is to find a statistic like Grubbs's statistic, which will test either the highest $x_{(n)}$ (resp the lowest $x_{(1)}$) observation is an outlier or not whatever F is normal or not. This test will be a generalized Grubbs statistic. Keep remind that, Grubbs's statistic is a studentized statistic. The normalization is done by centering each observation with the empirical mean $\bar{x}$ and by scaling with the empirical standard deviation σ. So we wish to derive a statistic, γ_n which have the following form :

$$\gamma_n = \frac{x_{(n)} - a_n}{b_n} \tag{3.1}$$

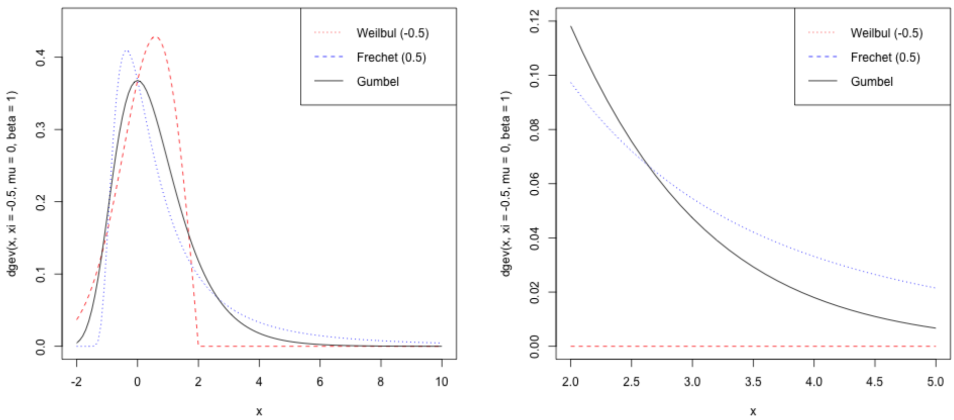
where a_n (resp $b_n > 0$) will play the role of the mean $\bar{x}$ (resp the standard deviation σ). The generalized extreme values (GEV) theory guaranty under some regularity conditions that with a well choices series a_n and b_n, it exist a non degenerated distribution G such that :

$$P(\frac{x_{(n)} - a_n}{b_n} \leq y) = F^{(n)}(yb_n + a_n) \underset{n \to +\infty}{\longrightarrow} G_\epsilon(y) \tag{3.2}$$

Only three limit distributions are possibles according to the value of ϵ : Gumbel distribution ($\epsilon = 0$, type I), Frechet distribution ($\epsilon > 0$, type II) or Weibull distribution ($\epsilon < 0$, type III). The CDF of these functions are given by :

$$G_\epsilon(x) = \begin{cases} exp[-(1 + \epsilon\frac{x-\mu}{\sigma})_+^{-\frac{1}{\epsilon}}] & \text{si } \epsilon \neq 0 \\ exp[-exp(-\frac{x-\mu}{\sigma})] & \text{si } \epsilon = 0 \end{cases} \tag{3.3}$$

where $x_+ = \max(x, 0)$, μ and σ are location and scale parameters. The GEV distribution is characterized by the shape parameter ϵ, which is known as the extreme value index (EVI). This index characterize the right tail of the distribution. When $\epsilon > 0$, the distribution has a heavy tail ; when $\epsilon < 0$, the distribution is a short tail distribution while a null value of ϵ refer to an exponential tail distribution. The Weibull distribution have a finite endpoint as shown by figure 2a while Gumbel and Frechet have an infinity endpoint. The endpoint of a distribution F is define as $x_F = sup\{x \text{ s.t } F(x) < 1\}$. In the right tail outliers identification, it's more likely that the maximum belongs to either Gumbel either Frechet distribution while in left tail outlier cases, it is Gumbel and Weibull distributions which are more likely to describe the extreme behavior.



(a) Limit distributions (b) View of distribution's tail

FIGURE 2 – Three cases of extreme value distribution with $\mu = 0$ and $\sigma = 1$

When the sample is generated from CDF F and the standardized sample maximum γ_n (see 3.1) converges to G_ϵ, we will say that F belongs to the maximum domain attraction of G_ϵ, $F \in D(G_\epsilon)$. It shown that the Frechet domain includes distributions such as the Pareto type, Burr, inverse gamma, log gamma, Student, and Frechet ; The Weibull domain includes distributions such as beta, power-law with an upper bound, reversed-Burr, and reversed-Weibull ;

The Gumbel domain includes many commonly used distributions such as exponential, Weibull, gamma, logistic, normal, and log-normal (Embrechts et al. 2013). So, to compute a statistic like the one in equation 3.1, one need to know two parameters : the normalized parameters and the EVI. The next section deal with the estimation of the EVI while section 4 defined the norming constants used in our test.

3.2 EVI estimators and the Generalized Pareto Distribution (GPD)

Given a sample, the analyst doesn't know the underlying distribution which generate the observations. So, he doesn't know also which of the three limiting distributions will best fit the maximum behavior. To select between the three distributions, he have to estimate the tail index ϵ. Many estimators have been suggested in literature. A well known estimator is Hill estimator which is given by (3.4). This estimator is suitable when the EVI is non negative.

$$\epsilon_n^H = \frac{1}{n_{u_n}} \sum_{i=0}^{n_{u_n}-1} ln(x_{(n-i)}) - ln(u_n) \tag{3.4}$$

where u_n is a given threshold and where $n_{u_n} = \# \{x_i : x_i > u_n\}$ [4]. But, as we said before, Hill EVI estimator is suitable for a non negative index. One more general EVI estimator is Dekkers-Einmahl-de Haan estimator which can be used whatever the sign of ϵ. This latter is defined as :

$$\epsilon_n^{DEH} = 1 + H_n^{(1)} + \frac{1}{2}\left(\frac{(H_n^{(1)})^2}{H_n^{(2)}} - 1\right)^{-1} \tag{3.5}$$

where

$$H_n^{(1)} = \frac{1}{k}\sum_{i=0}^{k-1}\left(ln(x_{(n-i)}) - ln(x_{(n-k)})\right)$$

and

$$H_n^{(2)} = \frac{1}{k}\sum_{i=0}^{k-1}\left(ln(x_{(n-i)}) - ln(x_{(n-k)})\right)^2$$

Given the k highest observations, $x_{(n-k)}$ plays the role of a threshold like the one defined in the Hill estimator. Many other estimators can be found in the literature like generalized Hill estimators, Pickands estimators, bootstrap based estimators (see (Embrechts et al. 2013) for a good summary). The ideas behind these estimators can be found in figure 2b. It seem like above a threshold, we can make differences between the three limiting distributions. This fact is confirm by Balkema-de- Haan (1974) and (Pickands III 1975) theorem which established that :

$$\lim_{u_n \to +\infty} \left[P(X \leq a | X > u_n) - G_{\epsilon,\sigma_{u_n}}(a - u_n)\right] \to 0 \tag{3.6}$$

where :

$$G_{\epsilon,\sigma_{u_n}}(a - u_n) = \begin{cases} 1 - (1 + \epsilon\frac{a-u_n}{\sigma_{u_n}})_+^{-\frac{1}{\epsilon}} & \text{si } \epsilon \neq 0 \\ 1 - exp(-\frac{a-u_n}{\sigma_{u_n}}) & \text{si } \epsilon = 0 \end{cases} \tag{3.7}$$

4. $\#A$ denote the cardinal of the set A.

$G_{\epsilon,\sigma_{u_n}}$ is the Generalized Pareto Distribution (GPD). This theorem says that above a threshold, we can approximate the behavior of X by GPD whatever the parent F. The connection between the GPD and GEV was first established by (Pickands III 1975). This distribution is related to GEV distribution through the formula $G_{\epsilon,\sigma_{u_n}}(x) = 1 - ln(G_\epsilon(x))$ with G_ϵ the GEV distribution with location equal to u_n and scale parameter to σ_{u_n}. Starting from the GEV distribution G_ϵ with location μ and scale parameter σ, the equivalent GPD is given with the same EVI and scale parameter $\sigma_{u_n} = \sigma + \epsilon(\mu - u_n)$ (Dey et al. 2016).

Coming back to the estimation of the EVI ϵ, the estimators presented above which are given by formula (3.4) or (3.5) need the knowledge of the threshold u_n (resp $x_{(n-k)}$). Extreme value theory provide many estimators of the threshold. We present in the following subsection one of the most used methods in the literature.

3.3 Some threshold selection methods

There are many tools and methods to select threshold in the extreme theory literature. We present the most used in the literature which are : the excess function plot, the Pickands method and the Reiss-Thomas method.

The excess function is based on the expectation value of the GPD. Given a threshold x, when the EVI is less than one ($\epsilon < 1$) it's easy to check that the expectation value of $G_{\epsilon,\sigma}$ is given by :

$$e_x = \frac{\sigma}{1-\epsilon} + \frac{\epsilon}{1-\epsilon}x \tag{3.8}$$

This expectation value is a linear function of the threshold x. So, the ideas is to plot the empirical excess plot which is given by equation (3.9) against the index i. Two informations can be read from the graph : (i) the type of the limiting distribution and (ii) the threshold. The threshold is the index at which the excess function become linear. In practice, it must exist many index at which the excess plot become linear. The threshold is chosen as the greatest since the GPD hold for a large threshold. The second information is read from the trend of the function. A constant trend (a horizontal line) implied that the underlying function belongs to Gumbel attraction domain. A downward trend suggest a short tail distribution (Weibull attraction domain) while a upward trend is an indication of a heavy tail distribution (F belongs to Frechet attraction domain). This is due to equation (3.8) slope which is equal to $\frac{\epsilon}{1-\epsilon}$.

$$e_{[i]} = \frac{1}{n-i}\sum_{j=0}^{n-i-1} x_{(n-j)} - x_{(i)} \tag{3.9}$$

The second method is described by (Pickands III 1975). Pickands suggest to choose the threshold equal to $x_{\lfloor k \rfloor}$ where $k = n - 4 \times M + 1$. M is the solution to the following optimization problem :

$$\underset{1\leq j\leq \lfloor \frac{n}{4} \rfloor}{Argmin}\ sup_{0\leq x<\infty}|\hat{S}_j(x) - \hat{G}_j(x)| \tag{3.10}$$

where $\lfloor . \rfloor$ denoted the integer part, $\hat{G}_j$ is an estimate of the GPD based on the median and the third quartile of the sample and $\hat{S}_j(x)$ which is the upper tail is defined as :

$$\hat{S}_j(x) = \frac{1}{4j}\sum_{m=1}^{4j-1} 1[x_{(n-m+1)} - x_{(n-4l+1)} \geq x] \tag{3.11}$$

The last method to determine the threshold which we will present is Reiss and Thomas (1997) (Neves & Alves 2004). This method is an automatic selection procedure. The threshold is defined as $x_{\lfloor n-k+1 \rfloor}$ where k is the solution to the following optimization problem :

$$\underset{2\leq k\leq n}{Argmin} \frac{1}{k-1}\sum_{i=1}^{k} i^{\delta}(\epsilon_i - \epsilon_k) \tag{3.12}$$

ϵ_i is an estimation of the EVI. (Neves & Alves 2004) suggest that an optimal value for δ is 0.35 when $\epsilon < 0$ and 0.4 when $\epsilon \geq 0$. A good summary of others methods can be found in (Scarrott & MacDonald 2012).

4 Hypothesis Test and Test's procedure

4.1 Hypothesis Test and properties

In this section, we present the hypothesis, their properties and the test's procedure. We based on similar formulation of the hypothesis like in Pearson & Sekar (1936). As did, (Pearson & Sekar 1936) and (Grubbs 1950), we will define outlier based on the fact that all observations have the same distribution.

Definition 2. *Let $x_{(n)}$ (resp. $x_{(1)}$) be the maximum (resp. the minimum) of a sample of size n, with $X_i \sim F_i$. The hypothesis test for identifying $x_{(n)}$ or $x_{(1)}$ as outlier is specify by :*

$$\begin{cases} H_0 : F_i = F \ \forall \ i \in 1:n \\ H_a : \exists \ m \ s.t \ F_m \neq F \end{cases}. \tag{4.1}$$

This test is a two tailed test since under the alternative. The outlier can either be on the right side either on the left side of the distribution or on the both side. We can also define the one tail version of the above test.

Definition 3. *Let $x_{(n)}$ (resp. $x_{(1)}$) be the maximum (resp. the minimum) of a sample of size n, with $X_i \sim F_i$. The hypothesis test of identifying :*

1. *$x_{(n)}$ as an outlier is :*

$$\begin{cases} H_0 : F_i = F \ \forall \ i \in 1:n \\ H_a : \exists \ m \ and \ x^* \ s.t \ \ \forall y \geq x^* \ F_m(y) < F(y) \end{cases} \tag{4.2}$$

2. *$x_{(1)}$ as an outlier is :*

$$\begin{cases} H_0 : F_i = F \ \forall \ i \in 1:n \\ H_a : \exists \ m \ and \ x^* \ s.t \ \ \forall y \leq x^* \ F(y) < F_m(y) \end{cases} \tag{4.3}$$

The definition (4.2) states that under alternative hypothesis, it is more likely to have a large value of $x_{(n)}$ while the definition (4.3) states that under alternative hypothesis, it is more likely to have a small value of $x_{(1)}$. To test hypothesis given by (4.2), we consider the following statistics γ_n^0 for $F \in D(G_0)$, γ_n^{ϵ} for $F \in D(G_\epsilon)$ with $\epsilon > 0$ and $\bar{\gamma}_n^{\epsilon}$ for $F \in D(\bar{G}_\epsilon)$ with $\epsilon < 0$.

$$\gamma_n^0 = \frac{x_{(n)} - a_n}{b_n}, \ \gamma_n^{\epsilon} = ln\left(\frac{x_{(n)} - a_n}{b_n}\right)^{\frac{1}{\epsilon}} \ \text{and} \ \bar{\gamma}_n^{\epsilon} = ln\left(-\frac{x_{(n)} - a_n}{b_n}\right)^{-\frac{1}{\epsilon}} \tag{4.4}$$

The critical value at α significance level for these statistics is given by $\Lambda_\alpha = -ln(-ln(1-\alpha))$ [5] obtained from the inverse of the Gumbel extreme value distribution (Olmo 2009). The power of this test have been study by (Olmo 2009) and the main results is given in the following proposition.

Proposition 1. *Let $x^0_{(n)}$ be the maximum of a sample of size n, under hypothesis $X_i \sim F$ and $x_{(n)}$ as in (4.2). We define $\eta = x_{(n)} - x^0_{(n)}$ if $F \in D(G_0)$, $\eta = \frac{x_{(n)}}{x^0_{(n)}} - 1$ if $F \in D(G_\epsilon)$ with $\epsilon > 0$ and $\eta = \frac{x_{(n)}-x_F}{x^0_{(n)}-x_F} - 1$ if $F \in D(\bar{G}_\epsilon)$ with $\epsilon < 0$. We have the following results (See (Olmo 2009)) :*

1. *If $F \in D(G_0)$, then the hypothesis (4.2) is consistent under the alternative hypothesis H_a iff $b_n = o(\eta)$.*
2. *If $F \in D(G_\epsilon) \cup D(\bar{G}_\epsilon)$, then the hypothesis (4.2) is consistent under the alternative hypothesis H_a iff $\eta^{-1} = o(1)$.*

Proof. The proof for $F \in D(G_0) \cup D(G_\epsilon)$ have been done by (Olmo 2009). We extend the proof for the case $D(\bar{G}_\epsilon)$. Note that for distributions belong to Weibull attraction domain, the normalization coefficients can be choose equal to : $a_n = x_F$ and $b_n = x_F - F^{-1}(1-\frac{1}{n})$. So, after some algebra $\bar{\gamma}^\epsilon_n$ can be rewritten and we get :

$$\begin{aligned}
\bar{\gamma}^\epsilon_n &= -\frac{1}{\epsilon}ln(x^0_{(n)} - x_F) - \frac{1}{\epsilon}ln[F^{-1}(1-\frac{1}{n}) - x_F] - \frac{1}{\epsilon}ln(1+\eta) \\
&= \bar{\gamma}^\epsilon_{n,H_0} - \frac{1}{\epsilon}ln(1+\eta)
\end{aligned}$$

Now we note that,

$$\begin{aligned}
P(\bar{\gamma}^\epsilon_n > \Lambda_\alpha) &= P(\bar{\gamma}^\epsilon_{n,H_0} > \Lambda_\alpha + \frac{1}{\epsilon}ln(1+\eta)) \\
&= (1-(1-\alpha)^{(1+\eta)^{-\frac{1}{\epsilon}}})
\end{aligned}$$

where $(1-\alpha)^{(1+\eta)^{-\frac{1}{\epsilon}}} \underset{n\to\infty}{\longrightarrow} 0$ iff $\eta^{-1} = o(1)$ and given $\epsilon < 0$.

In this section, we had presented the properties of the hypothesis test of identifying outliers. The next section present the algorithm for implementing the test.

4.2 Test Procedure

The test statistic of hypothesis (4.2) required the knowledge of the norming constants a_n and b_n. If the CDF F belongs to Frechet or Weibull max-domain of attraction, we will also need to know the extreme value index ϵ to compute the statistic γ^ϵ_n and $\bar{\gamma}^\epsilon_n$. When the data generator process (DGP) is known, the parameters (a_n and b_n) can be estimated directly. These estimations have been computed for several distributions in Embrechts et al. (2013). For example, in the case of normal distribution, the norming constants are given by (4.5). The norming constant have been also derived for several distributions like : Uniform, Weibull, Gamma, Log-normal, etc.

$$b_n = (2ln(n))^{-1/2} \text{ and } a_n = b_n^{-1} - \frac{ln(4\pi) + ln(ln(n))}{2b_n^{-1}} \tag{4.5}$$

But in practice, the DGP is unknown. So the direct estimation can't be used, To overcome this situation, we derive a semi parametric estimator for the norming constants using a suggested general formula given by (Embrechts et al. 2013). The general suggested norming constants are :

5. We can use the same critical value because of the fact that : if Y $\sim G_\epsilon$, $\epsilon > 0$ and Z $\sim G_\epsilon$, $\epsilon < 0$ then $\frac{1}{\epsilon}ln(Y)$ and -$\frac{1}{\epsilon}ln(-Z) \sim G_0$.

Max-Domain	a_n	b_n
Gumbel	$F^{-1}(1-\frac{1}{n})$	$\gamma(a_n)$
Frechet	0	$F^{-1}(1-\frac{1}{n})$
Weibull	x_F	$x_F - F^{-1}(1-\frac{1}{n})$

TABLE 1 – Norming constants

where $\gamma(.)$ is an auxiliary function of distributions belong to Gumbel max-domain. This function can be treat as the mean excess function define by :

$$\gamma(x) = E[X - x|X > x] \tag{4.6}$$

To determine the norming constants, we have to estimate the endpoint x_F and the quantile $F^{-1}(1-\frac{1}{n})$. (Olmo 2009) have suggested estimator based on the Generalized Pareto distribution (GPD) based on the Balkema-de- Haan (1974) and Pickands III (1975) theorems. We have the following result.

Proposition 2. *BHP theorem (3.6) can be re-parametrize to :*

1. *If F belongs to Gumbel Max domain of attraction, $\exists$ a function $\gamma(t) > 0$ s.t :*

$$\lim_{u_n \to +\infty} [P(X \leq a|X > u_n) - (1 - exp(-\frac{a - u_n}{\gamma(u_n)}))] \to 0 \tag{4.7}$$

2. *If F belongs to Frechet Max domain of attraction, we have the following result :*

$$\lim_{u_n \to +\infty} [P(X \leq a|X > u_n) - (1 - (\frac{a}{u_n})^{-\frac{1}{\epsilon}})] \to 0 \tag{4.8}$$

3. *If F belongs to Weibull Max domain of attraction, we have the following result :*

$$\lim_{u_n \to x_F} [P(X \leq a|X > u_n) - (1 - (\frac{x_F - a}{x_F - u_n})^{\frac{1}{\epsilon}})] \to 0 \tag{4.9}$$

Based on the fact that a nonparametric estimator of $F(u_n)$ is $(1 - \frac{n_{u_n}}{n})$, where $n_{u_n} = \#\{x_i : x_i > u_n\}$, we derive a semi parametric estimator for the quantile $F^{-1}(1-\frac{1}{n})$ as the one proposed by (Olmo 2009). The following corollary specify the estimator of $F^{-1}(1-\frac{1}{n})$ according to Max domain of attraction.

Corrolary 1. *Under regularity assertions of the BHP theorem, and for a well choose threshold u_n, a semi parametric estimator of $F^{-1}(1-\frac{1}{n})$ is :*

1. *If F belongs to Gumbel Max domain of attraction, we have the following estimator :*

$$\hat{F}^{-1}(1 - \frac{1}{n}) = u_n + \gamma(u_n)ln(n_{u_n}) \tag{4.10}$$

2. *If F belongs to Frechet Max domain of attraction, we have the following estimator :*

$$\hat{F}^{-1}(1 - \frac{1}{n}) = n_{u_n}^{\hat{\epsilon}} u_n \tag{4.11}$$

3. *If F belongs to Weibull Max domain of attraction, we have the following estimator :*

$$\hat{F}^{-1}(1 - \frac{1}{n}) = x_F - n_{u_n}^{-\hat{\epsilon}}(x_F - u_n) \tag{4.12}$$

The proof of the proposition 2 is given in appendix A. Now, as we have an estimator for the quantile $F^{-1}(1-\frac{1}{n})$, we look for an estimator of the right endpoint for distributions belong to Type III. A general estimator with good properties (consistency and asymptotic distribution) have been proposed by (Alves & Neves 2014). This estimator is defined by :

$$\hat{x_F} = x_{(n)} + \sum_{i=0}^{k-1} a_{ik}(x_{(n-k)} - x_{(n-k-i)}) \tag{4.13}$$

where

$$a_{ik} = \frac{1}{log(2)} log(\frac{k+i+1}{k+i})$$

We now have all necessary elements to define the test statistics. The statistics given by equation (4.4) can be rewritten as follow :

$$\gamma_n^0 = \frac{x_{(n)} - u_n - \gamma(u_n) ln(n_{u_n})}{\gamma(u_n + \gamma(u_n) ln(n_{u_n}))} \tag{4.14}$$

$$\gamma_n^{\hat{\epsilon_n}} = \frac{1}{\hat{\epsilon_n}} ln(x_{(n)}) - \frac{1}{\hat{\epsilon_n}} ln(u_n) - ln(n_{u_n}) \tag{4.15}$$

$$\bar{\gamma}_n^{\hat{\epsilon_n}} = -\frac{1}{\hat{\epsilon_n}} ln(\sum_{i=0}^{k-1} a_{ik}(x_{(n-k)} - x_{n-k-i})) + \frac{1}{\hat{\epsilon_n}} ln(x_{(n)} - u_n + \sum_{i=0}^{k-1} a_{ik}(x_{(n-k)} - x_{n-k-i})) - ln(n_{u_n}) \tag{4.16}$$

where $n_{u_n} = \#\{x_i : x_i > u_n\}$, u_n a threshold and $\hat{\epsilon_n}$ is an estimate of the EVI. One can use DEH estimator given by (3.5), but when this parameter is non negative, it is better to use the Hill estimator given by (3.4) as it has the lowest mean square error (Embrechts et al. 2013). Many others estimators can be found in the literature (Dey et al. 2016).

To detect iteratively more than one outlier in sample coming from distribution belongs to Frechet attraction domain, (Olmo 2009) suggest an algorithm which is given below. In the algorithm, we note by "m" the number of outliers candidate. For each outlier candidate, the procedure is repeated "t" times in order to filter the effects of estimating ϵ_n from data. The ordered sample is denoted $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. The steps of the procedure are given by algorithm A or B according to the value of ϵ_n. In the case of distribution belongs to Gumbel attraction domain (algorithm B), there is no need of computing $\hat{\epsilon_n}$, and we have adapted the algorithm A and the formula in step 4 is adjusted with γ_n^0 and there is no more k, k_j and t in the algorithm A (we will refer to the algorithm in Gumbel case as algorithm B).

Algorithm A

1. j=k=n ; iter=tol_0=0 ; t ; m.
2. Compute $x_{(j)} = Max(x_{(1)}, x_{(2)}, \ldots, x_{(j)})$
   iter=iter+1
3. Choose k_j as the fraction of j to estimate the EVI.
   $tol_0 = tol_0 + 1$
   $u_n = x_{(k-k_j)}$
   Compute $\hat{\epsilon_j}$ given by (3.5) with $x_{(k-k_j+1)}, \ldots, x_{(k)}$
4. Compute $\gamma_j^{\hat{\epsilon_j}}$ given by (4.15) with $n_{u_n} = \#\{x_i : x_i > u_n\}$
5. If $\gamma_j^{\hat{\epsilon_j}} > \Lambda_\alpha$ then
   Report "$x_{(j)}$ as outlier"
   $j \leftarrow j-1$; $k = j$; $tol_0 = 0$.
   If $iter < m$ then go to step 2 otherwise it's the end.
6. Otherwise (i.e $\gamma_j^{\hat{\epsilon_j}} \leq \Lambda_\alpha$)
   If $tol_0 \leq t$ then $k \leftarrow k-1$ and go to step 3.
   Otherwise if $iter < m$ then go to step 2
   Otherwise it's the end.

Algorithm B

1. j=n ; iter=0 ; m.
2. Compute $x_{(j)} = Max(x_{(1)}, x_{(2)}, \ldots, x_{(j)})$
   iter=iter+1
3. Compute γ_j^0 given by (4.14)
4. If $\gamma_j^0 > \Lambda_\alpha$ then
   Report "$x_{(j)}$ as outlier"
   $j \leftarrow j-1$;
   If $iter < m$ then go to step 2 otherwise it's the end.
5. Otherwise (i.e $\gamma_j^0 \leq \Lambda_\alpha$), it's the end.

We identify three points to discuss in these algorithms. The first point is the choice of the algorithm A or B. Nothing have been said by Olmo (2009) on that. We suggest to use many tools provided by the GEV theory like the excess function plot. If the trend is constant, we use Algorithm B and if the trend is upward we use algorithm A. The case of downward trend have not been treat by Olmo (2009). However, in practice even if the underlying distribution belongs to Gumbel attraction domain, it will be difficult to get a perfect horizontal line (constant trend). So we can compute an ordinary least square on the highest observations and test if the slope of the trend is statistically different from 0. If so, we use algorithm A otherwise we use algorithm B. Another approach is to compute EVI through (3.5). If EVI is non null, we use algorithm A otherwise we use algorithm B.

The second point is related to the choice of k_j in algorithm A. The choice is done arbitrary. We suggest to use a more objective method by computing the minimum between the k given by Pickands III (1975) and that of Reiss and Thomas (1997). So we set $k_j = \max(k_P, k_{RT})$. To compute, k_j according to Reiss and Thomas (1997), we suggested to use δ =0.35 as indicated by (Neves & Alves 2004).

Finally, the last point refer to step 6 of algorithm A. If we have m outliers candidate and we found that $x_{(n-j)}$, $0 \leq j \leq m-1$ is not an outlier, it's does not matter to continue the algorithm, we should stop. So we suggest to drop step 6.b (Otherwise if $iter < m$ then go to step 2) in algorithm A. Based ont those three points, we suggest a modified version of the algorithm and expand it to distributions belong to Weibull Max domain.

Algorithm Start

1. Use the excess function plot
2. Use DEH estimator.

Algorithm B'($\epsilon = 0$)

1. $j, k \leftarrow n$; $b_1, b_2, iter, tol_0 \leftarrow 0$;t ; m. $iter \leftarrow iter + 1$;
2. **While** ($iter \leq m$ and b_1=0) do
3. Compute $x_{(j)} = Max(x_{(1)}, x_{(2)}, \ldots, x_{(j)})$
4. $k_j = j - \min(k_P, k_{RT})$ where k_P is given by (3.10) and k_{RT} bu (3.12) $tol_0 \leftarrow tol_0 + 1$
5. **While** ($tol_0 \leq t$ and b_2=0) do $u_n = x_{(k-k_j)}$
6. Compute γ_j^0 given by (4.14) with $n_{u_n} = n - k + k_j$
7. If $\gamma_j^0 > \Lambda_\alpha$ then
   Report "$x_{(j)}$ as outlier"
   $tol_0 \leftarrow 1$
   $b_2 \leftarrow 1$
8. Otherwise (i.e $\gamma_j^0 \leq \Lambda_\alpha$)
   $tol_0 \leftarrow tol_0 + 1$
   $k \leftarrow k - 1$

   End Do ;
   If $tol_0 = t - 1$ then $b_1 = 1$
   $iter \leftarrow iter + 1$;
   $j \leftarrow j - 1$; $k \leftarrow j$;
   $tol_0 \leftarrow 1$;
   $b_2 \leftarrow 0$;
   End Do ;

Algorithm A'($\epsilon \neq 0$)

1. $j, k \leftarrow n$; $b_1, b_2, iter, tol_0 \leftarrow 0$;t ; m. $iter \leftarrow iter + 1$;
2. **While** ($iter \leq m$ and b_1=0) do
3. Compute $x_{(j)} = Max(x_{(1)}, x_{(2)}, \ldots, x_{(j)})$
4. $k_j = j - \min(k_P, k_{RT})$ where k_P is given by (3.10) and k_{RT} bu (3.12) $tol_0 \leftarrow tol_0 + 1$
5. **While** ($tol_0 \leq t$ and b_2=0) do $u_n = x_{(k-k_j)}$
6. Compute $\hat{\epsilon_j}$ given by (3.5) or (3.4) with $x_{(k-k_j+1)}, \ldots, x_{(k)}$
7. Compute $\gamma_j^{\hat{\epsilon_j}}$ or $\bar{\gamma}_j^{\hat{\epsilon_j}}$ given by (4.15) and (4.16) with $n_{u_n} = n - k + k_j$ according to the sign of $\hat{\epsilon_j}$
8. If $\gamma_j^{\hat{\epsilon_j}} (resp. \bar{\gamma}_j^{\hat{\epsilon_j}}) > \Lambda_\alpha$ then
   Report "$x_{(j)}$ as outlier"
   $tol_0 \leftarrow 1$
   $b_2 \leftarrow 1$
9. Otherwise (i.e $\gamma_j^{\hat{\epsilon_j}} (resp. \bar{\gamma}_j^{\hat{\epsilon_j}}) \leq \Lambda_\alpha$)
   $tol_0 \leftarrow tol_0 + 1$
   $k \leftarrow k - 1$

   End Do ;
   If $tol_0 = t - 1$ then $b_1 = 1$
   $iter \leftarrow iter + 1$;
   $j \leftarrow j - 1$; $k \leftarrow j$;
   $tol_0 \leftarrow 1$;
   $b_2 \leftarrow 0$;
   End Do ;

5 Applications

In this section, we will apply the new test to two situations. First, we will compare the new test to the well know Grubbs's test when the DGP is normal. The second application aims to show the ability of the new test to overcome the limit of Grubbs's test when observations are from non normal parent as mentioned in section 2.

5.1 Comparing the New GEV based outliers test and Grubbs's test

In this first application, we compare the new test to the well know Grubbs's test when the DGP is normal. For this application, we simulate data coming from normal parent since we want to compare our new approach to Grubbs's test [6]. We consider two cases : a case with noisy data and a case without. The number of noisy observations has been chosen to be equal to ($p \in 0, ..., 3$). The generator of data is given by :

$$Y_i \sim 1_{1 \leq i \leq T-p} N(a, \sigma) + 1_{T-p < i \leq T} N(a + d, \sigma) \quad (5.1)$$

6. Grubbs's test hypothesis rely on the normal distribution

where T is the sample size and p the number of noisy data (outliers). The mean of the underlying distribution "a" and "d" have been chosen to be equal respectively to 0 and 5 and the standard σ to 1. Observations coming from $N(a+d,\sigma)$ were consider like outliers since the parent is different. We compare both tests using algorithm B' with t and m equal to 1. Since the parent is normal, we use the norming constant given by equation (4.5) in computing the test statistic instead of the statistic define in equation (4.14). To compare both tests, we use the sensibility and the specificity indicators. We use bootstrap approach with 1 000 B-samples. Each time, we generate a new sample using formula (5.1) and apply Grubbs's test and GEV outliers test to the maximum value. To compute the specificity, we calculate the percentage of time that each test not identify $x_(n)$ as outlier when p=0. The sensibility have been computed as he percentage of time that each test identify $x_(n)$ as outlier when $p! = 0$.

Table 2 shows the results of the simulations. Columns third and fourth are the type I error that's the percentage of times, the tests identify the highest observation as an outliers while it was not. Columns six until end are the power of the tests under one, two and three outliers in the sample. Two main results can be underline from this table. The first main result is the fact that type I error for Grubbs's test is lower than the new GEV outliers test whatever the sample size is small or large. But both tests have a type I error less than the significance level α. The second result is related to the power of the tests. Both tests have very high power but GEV test's power is greater than that of the Grubb's test. The new test's power increase with the number of outliers in the sample.

N	α (in %)	1-Specificity (p=0)		Sensibility (p=1)		Sensibility (p=2)		Sensibility (p=3)	
		Grubs	EVT	Grubs	EVT	Grubs	EVT	Grubs	EVT
50	1	0.006	0.003	0.81	0.893	0.796	0.991	0.608	0.999
50	5	0.024	0.041	0.919	0.972	0.955	0.997	0.919	1.000
50	10	0.036	0.077	0.939	0.99	0.987	0.999	0.972	1.000
100	1	0.014	0.01	0.807	0.866	0.915	0.975	0.903	0.997
100	5	0.025	0.038	0.923	0.951	0.974	1.000	0.99	0.998
100	10	0.046	0.076	0.935	0.977	0.99	0.999	0.999	1.000
500	1	0.007	0.007	0.768	0.801	0.928	0.951	0.974	0.99
500	5	0.018	0.041	0.859	0.895	0.973	0.987	0.996	1.000
500	10	0.039	0.088	0.903	0.935	0.985	0.994	0.994	1.000
1000	1	0.008	0.011	0.687	0.713	0.913	0.941	0.965	0.98
1000	5	0.019	0.035	0.819	0.865	0.967	0.981	0.992	0.995
1000	10	0.066	0.108	0.875	0.912	0.979	0.993	0.993	0.999

TABLE 2 – Sensibility and Specificity

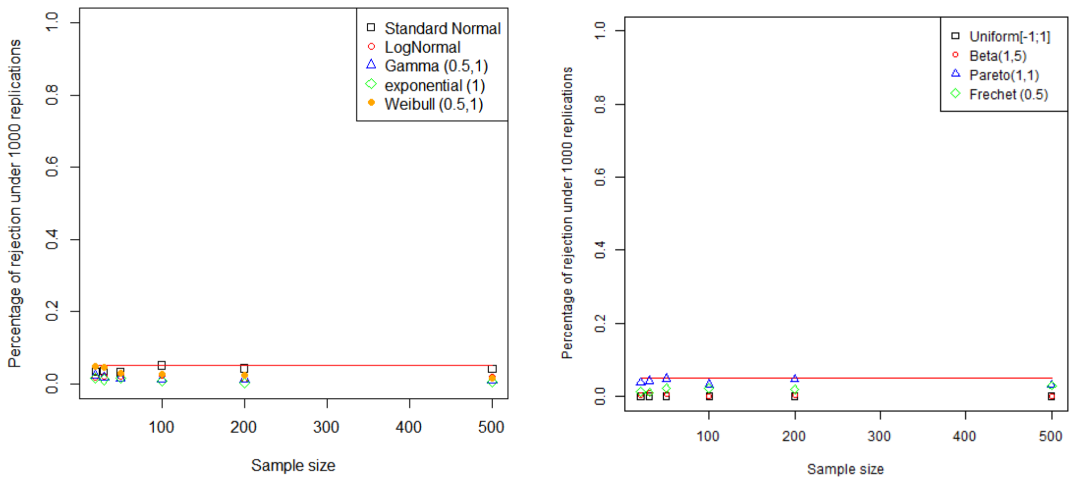
5.2 How GEV outliers test behave under non normal parent

Let's go back to the result of applying Grubb's test to non normal distributions in figure 1b. We have shown that Grubbs's test didn't behave well under non normal parent. In this application, we resume the same application of section 2.2 to non normal distributions. We simulate 1 000 samples of different size (N=20, 30, 50, 100, 200, 500) from various parent [7]. For each sample, we test either the highest observation $x_{(n)}$ is an outlier or not using GEV outliers test. In the example of section 2.2, we use Grubbs's test, but herein you use GEV test. Since the sample is generate from the same parent, the GEV outliers test should not reject the null

7. normal, uniform, gamma, exponential, beta, fisher, khi square and student

hypothesis or to be exact, the rejection rate shouldn't be greater than the significance level (5 % here). The result of the empirical test is given by figure 3.

In contrary to the result obtained using Grubbs's test, the GEV outlier test didn't reject the null hypothesis more than 5% times as it should be expected. This result hold for various distributions as indicated by figure 3. In section 2.2, the type I error vary from values less than 5% to 100%.



(a) Gumbel Max Domain Dist. (b) Frechet and Weibull Max Domain Dist.

FIGURE 3 – Application of GEV outliers test.
The red line indicate the 5% level. The data have been simulated from different parent which are normal, log-normal, gamma, exponential, Weibull, uniform, beta, Pareto and Frechet. Two groups have been created. The first group contain distributions which have an exponential right tail while group 2 contains distributions with polynomial decaying and those with finite right endpoint. We simulate1 000 samples of different size (N=20, 30, 50, 100, 200, 500). For each sample, we test if the maximum $x_{(n)}$ is an outlier using GEV test. Since the sample is generate from the same parent, the GEV outlier test should not reject the null hypothesis more than 0.05.

6 Conclusion

This paper analyses the identification of aberrant values using a new approach : the generalized extreme value outlier test. This test is based on the GEV theory. We suggest a new approach in the identification process of aberrant values in large sample. We first shown that the classical Grubbs's outlier used to reject the null hypothesis when the underlying distributions isn't normal. For most distributions like Log Normal, Gamma, Weibull, exponential, and so on, Grubbs's test used to identify systematically the highest observation as an outlier.

So, based on the generalized extreme value theory, we develop a new test to identify extreme value. Using bootstrapping method and simulation data, we can underline two results. The first result suggest that the GEV outlier test best perform than Grubb's test even when the underline distribution is normal. The second result is the fact that when the underlying distributions isn't normal, the GEV outlier test seem to reject null hypothesis at a rate less than the significance level. This result is robust to various non normal distributions and the sample size.

As next step, we will investigate the case of very large datasets (N $\gtrsim$ 100 000) as our result is an asymptotic result (as $n \to +\infty$). Also, since the norming constant used in this version of the test is a more general one, we will investigate the specific case (parametric estimation) of normal distribution. In that case, instead of the semi-parametric norming constant, we should be able to use the norming constant given by 4.5. We expect to improve the performance of this test.

Références

Alves, I. F. & Neves, C. (2014), 'A General Estimator for the Right Endpoint', *arXiv preprint arXiv :1412.3972* .

Beirlant, J., Goegebeur, Y., Teugels, J. & Segers, J. (2004), 'Front Matter', *Statistics of Extremes : Theory and Applications* pp. i–xiii.

Benlagha, N., Grun-Réhomme, M. & others (2007), Application de la théorie des valeurs extrêmes en assurance automobile, Technical report, ERMES, University Paris 2.

Bernier, J. (1956), 'Sur l'application des diverses lois limites des valeurs extrêmes au problème des débits de crue', *Houille Blanche* **11**(5).

Burridge, P. & Robert Taylor, A. (2006), 'Additive Outlier Detection Via Extreme-Value Theory', *Journal of Time Series Analysis* **27**(5), 685–701.

Caeiro, F. & Gomes, M. I. (2014), On the bootstrap methodology for the estimation of the tail sample fraction, *in* 'Proceedings of COMPSTAT', pp. 545–552.

Caeiro, F. & Gomes, M. I. (2016), 'Threshold Selection in Extreme Value Analysis', *Extreme Value Modeling and Risk Analysis : Methods and Applications* p. 69.

Chandola, V., Banerjee, A. & Kumar, V. (2009), 'Anomaly detection : A survey', *ACM computing surveys (CSUR)* **41**(3), 15.

Cousineau, D. & Chartier, S. (2015), 'Outliers detection and treatment : A review.', *International Journal of Psychological Research* **3**(1), 58–67.

Dey, D., Roy, D. & Yan, J. (2016), *Extreme Value Modeling and Risk Analysis : Methods and Applications*, CRC Press.

Dixon, W. J. (1950), 'Analysis of extreme values', *The Annals of Mathematical Statistics* **21**(4), 488–506.

Embrechts, P., Klüppelberg, C. & Mikosch, T. (2013), *Modelling Extremal Events : For Insurance and Finance*, Vol. 33, Springer Science & Business Media.

Everitt, B. (2002), *The Cambridge Dictionary of Statistics/BS Everitt*, Cambridge University Press, Cambridge, UK New York :.

Fisher, R. A. & Tippett, L. H. C. (1928), Limiting forms of the frequency distribution of the largest or smallest member of a sample, *in* 'Mathematical Proceedings of the Cambridge Philosophical Society', Vol. 24, Cambridge Univ Press, pp. 180–190.

Gonzalo, J. & Olmo, J. (2004), 'Which extreme values are really extreme ?', *Journal of Financial Econometrics* **2**(3), 349–369.

Grubbs, F. E. (1950), 'Sample criteria for testing outlying observations', *The Annals of Mathematical Statistics* pp. 27–58.

Houfi, M. A. & El Montasser, G. (2009), 'Effets des points aberrants sur les tests de normalité et de linéarité. Applications à la bourse de Tokyo', *Revue Congolaise Economie* **5**(21), 2–38.

Leadbetter, M. R. (1983), 'Extremes and local dependence in stationary sequences', *Probability Theory and Related Fields* **65**(2), 291–306.

Neves, C. & Alves, M. F. (2004), 'Reiss and Thomas, automatic selection of the number of extremes', *Computational statistics & data analysis* **47**(4), 689–704.

Nikulin, M. & Zerbet, A. (2002), 'Détection des observations aberrantes par des méthodes statistiques', *Revue de statistique appliquée* **50**(3), 25–51.

Olmo, J. (2009), 'Extreme value theory filtering techniques for outlier detection'.

Pearson, E. S. & Sekar, C. C. (1936), 'The efficiency of statistical tools and a criterion for the rejection of outlying observations', *Biometrika* **28**(3/4), 308–320.

Pickands III, J. (1975), 'Statistical inference using extreme order statistics', *the Annals of Statistics* pp. 119–131.

Planchon, V. (2005), 'Traitement des valeurs aberrantes : Concepts actuels et tendances générales', *Biotechnologie, agronomie, société et environnement* **9**(1), 19–34.

Rais, S. (2008), 'Outlier detection for the consumer price index', *Statistical Society of Canada Proceedings, Industrial Organization, Finance, and Prices Section, BSMD* .

RANGER, N. (1990), 'R. CLEROUX J.-M. HELBLING', *Revue de statistique appliquée* **38**(1), 5–21.

Scarrott, C. & MacDonald, A. (2012), 'A review of extreme value threshold es-timation and uncertainty quantification', *REVSTAT–Statistical Journal* **10**(1), 33–60.

Thompson, W. R. (1935), 'On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation', *The Annals of Mathematical Statistics* **6**(4), 214–219.

A Proof of proposition 2

(Leadbetter 1983) give necessary and sufficient conditions for the three cdfs in 3.3. These conditions are :

- Type I : F $\in D(G_0)$ iff $\exists\ \gamma(t) > 0$ s.t $\lim\limits_{t \to x_F} \frac{1-F(t+x\gamma(t))}{1-F(t)} = exp(-x)$
- Type II : F $\in D(G_\epsilon)$, $\epsilon > 0$ iff $\lim\limits_{t \to \infty} \frac{1-F(tx)}{1-F(t)} = x^{-\frac{1}{\epsilon}}$
- Type III : F $\in D(G_\epsilon)$, $\epsilon < 0$ iff $\lim\limits_{t \to 0} \frac{1-F(x_F-tx)}{1-F(x_F-t)} = x^{\frac{1}{\epsilon}}$

We also have the following relation :

$$P(X \leq a|X > u_n) = 1 - \frac{1-F(a)}{1-F(u_n)}$$

In the case of type I distribution, we have :

$$\lim_{u_n \to +\infty} [P(X \leq a|X > u_n) - (1 - exp(-\frac{a-u_n}{\gamma(u_n)}))] \to 0$$

Since for a very large value of u_n, we have :

$$\begin{aligned} P(X \leq a|X > u_n) &= 1 - \frac{1-F(a)}{1-F(u_n)} \\ &\sim 1 - exp(-\frac{a-u_n}{\gamma(u_n)}) \end{aligned}$$

Also, for distributions belong to Type II MAx domain, we have this approximation for a very large value of u_n, :

$$\begin{aligned} P(X \leq a|X > u_n) &= 1 - \frac{1-F(a)}{1-F(u_n)} \\ &\sim 1 - (\frac{a}{u_n})^{-\frac{1}{\epsilon}} \end{aligned}$$

The last case is based on the following approximation, when u_n is near the right endpoint x_F :

$$\begin{aligned} P(X \leq a|X > u_n) &= 1 - \frac{1-F(a)}{1-F(u_n)} \\ &\sim 1 - (\frac{x_F - a}{x_F - u_n})^{\frac{1}{\epsilon}} \end{aligned}$$