DEUTSCHE
BUNDESBANK
EUROSYSTEM

Discussion Paper

Deutsche Bundesbank
No 02/2013

A distribution-free test for outliers

Bertrand Candelon
(Maastricht University)

Norbert Metiu
(Deutsche Bundesbank)

Deutsche Bundesbank, Wilhelm-Epstein-Straße 14, 60431 Frankfurt am Main,
Postfach 10 06 02, 60006 Frankfurt am Main

Tel +49 69 9566-0

Please address all orders in writing to: Deutsche Bundesbank,
Press and Public Relations Division, at the above address or via fax +49 69 9566-3077

Internet http://www.bundesbank.de

ISBN 978–3–86558–881–4 (Printversion)
ISBN 978–3–86558–882–1 (Internetversion)

Non-technical summary

Statistical data analysis usually begins with the collection of observations from a certain population. However, the sampling process is subject to numerous sources of error. Therefore, the data collected may contain some unusually small or large observations, so-called outliers. Determining whether a data set contains one or more outliers is a challenge commonly faced in applied statistics. This is a particularly difficult task if the properties of the underlying population are not known. Nevertheless, in many empirical analyzes, the assumption that the data come from a particular population is too restrictive or unrealistic.

This paper develops a statistical test for outliers in data drawn from an unknown population. Our methodology relies on a nonparametric bootstrap procedure. Simulation experiments show that the proposed test detects outliers correctly whatever the underlying population, even for relatively small samples. Consequently, our test could be instrumental in a wide range of empirical applications where few observations are available and the underlying population is unknown.

The empirical performance of the test is illustrated by means of two examples in the fields of aeronautics and macroeconomics. Specifically, our second example investigates annual inflation rates for Germany from 1428 to 2010. The nearly six centuries of monetary history under study witnessed several episodes of hyperinflation, which constitute potential outliers in the sense, that these observations do not belong to the same population as the others. Indeed, we find six outliers in the sample. The earliest outlier occurred in 1621, at the peak year of the *Kipper- und Wipperzeit* (Tipper and Seesaw Time), a monetary crisis in the Holy Roman Empire between 1619 and 1623, which was characterized by hyperinflation through debasement of commodity (gold, silver, and copper) money. Furthermore, we find that the infamous hyperinflation in the early 20th century, and its collapse, was characterized by outlying annual inflation rates in 1917, 1920, 1922, 1923, and 1924, whereas other historical periods with high inflation/deflation are not identified as outliers.

Nicht-technische Zusammenfassung

Die Analyse statistischer Daten beginnt in der Regel mit der Erhebung von Beobachtungen aus einer bestimmten Grundgesamtheit. Der Prozess der Stichprobenerhebung unterliegt jedoch einer Reihe von Fehlerquellen. Aus diesem Grund können die erfassten Daten einige ungewöhnlich niedrige oder hohe Werte enthalten, die sogenannten Ausreißer. Die Frage, ob ein Datensatz einen oder mehrere Ausreißer enthält, ist in der angewandten Statistik ein weitverbreitetes Problem. Sie erweist sich als besonders schwierig, wenn die Merkmale der zugrunde liegenden Grundgesamtheit nicht bekannt sind. Gleichwohl ist in vielen empirischen Untersuchungen die Annahme, dass die Daten aus einer bestimmten Grundgesamtheit stammen, zu restriktiv bzw. unrealistisch.

In der vorliegenden Arbeit wird ein statistischer Test entwickelt, um Daten aus einer unbekannten Grundgesamtheit auf Ausreißer hin zu untersuchen. Dabei kommt ein nichtparametrisches Bootstrap-Verfahren zur Anwendung. Simulationsexperimente zeigen, dass der vorgeschlagene Test Ausreißer korrekt anzeigt, und zwar ungeachtet der zugrunde liegenden statistischen Masse und sogar für vergleichsweise kleine Stichproben. Der Test könnte also für eine Vielzahl empirischer Anwendungen, bei denen nur wenige Beobachtungen verfügbar sind und die Grundgesamtheit nicht bekannt ist, hilfreich sein.

Die empirische Leistungsfähigkeit des Tests wird an zwei Beispielen aus den Bereichen Luftfahrt und Makroökonomie verdeutlicht. Konkret werden dazu im zweiten Beispiel die jährlichen Inflationsraten in Deutschland im Zeitraum von 1428 bis 2010 untersucht. Die hier betrachteten rund sechs Jahrhunderte Währungsgeschichte umfassen mehrere Phasen der Hyperinflation, die insofern potenzielle Ausreißer darstellen, als sie nicht zur selben Grundgesamtheit gehören wie die anderen beobachteten Werte. Tatsächlich finden sich in der Stichprobe sechs Ausreißer. Der erste Ausreißer lässt sich für das Jahr 1621 feststellen, auf dem Höhepunkt der Kipper- und Wipperzeit, einer Krise des Münzwesens im Heiligen Römischen Reich in der Zeit von 1619 bis 1623, in der es zu einer Hyperinflation durch eine Entwertung des Warengeldes (Gold, Silber und Kupfer) kam. Als weiteres Ergebnis lässt sich festhalten, dass die berüchtigte Hyperinflation und der Zusammenbruch Anfang des 20. Jahrhunderts durch Ausreißer bei den jährlichen Preissteigerungsraten in den Jahren 1917, 1920, 1922, 1923 und 1924 gekennzeichnet war, während andere historische Episoden mit hoher Inflation/Deflation nicht als Ausreißer identifiziert werden.

Bundesbank Discussion Paper No 02/2013

A Distribution-Free Test for Outliers*

Bertrand Candelon
Maastricht University

Norbert Metiu
Deutsche Bundesbank

Abstract

Determining whether a data set contains one or more outliers is a challenge commonly faced in applied statistics. This paper introduces a distribution-free test for multiple outliers in data drawn from an unknown data generating process. Besides, a sequential algorithm is proposed in order to identify the outlying observations in the sample. Our methodology relies on a two-stage nonparametric bootstrap procedure. Monte Carlo experiments show that the proposed test has good asymptotic properties, even for relatively small samples and heavy tailed distributions. The new outlier detection test could be instrumental in a wide range of statistical applications. The empirical performance of the test is illustrated by means of two examples in the fields of aeronautics and macroeconomics.

Keywords: Bootstrap, mode testing, nonparametric statistics, outlier detection.

JEL classification: C14.

*Contact address: Department of Economics, Maastricht University, PO Box 616, 6200 MD, Maastricht, The Netherlands. Research Centre, Deutsche Bundesbank, Wilhelm-Epstein-Straße 14, 60431 Frankfurt am Main, Germany. E-Mail: b.candelon@maastrichtuniversity.nl, norbert.metiu@bundesbank.de. The authors thank Don Harding, Jan Piplack, participants of the 2009 Econometric Society Australasian Meeting in Canberra, the 26th Symposium on Money, Banking and Finance in Orléans, the Third Methods in International Finance Network Annual Workshop in Luxembourg, and seminars at the University of Rome "Tor Vergata" and Maastricht University for comments and suggestions.

1 Introduction

Determining whether a data set contains one or more outliers is a challenge commonly faced in applied statistics. This is a particularly difficult task if the underlying data generating process (DGP) is unknown, since the corresponding probability density function (pdf) can have a variety of shapes in its tails. Nevertheless, the assumption of a particular DGP is often too restrictive or unrealistic. To tackle this issue, a distribution-free test for outliers in large samples has been proposed by Walsh (1959). However, the problem of nonparametric rejection of outliers is exacerbated in finite samples, i.e., when the number of observations is relatively small.[1] This highlights the need for a distribution-free test for outliers in small samples, and our objective is to fill this gap.

Building on earlier work by Singh and Xie (2003) and Silverman (1981), we propose a novel nonparametric test for multiple outliers in data drawn from an unknown DGP. Besides, a sequential algorithm is proposed in order to identify the outlying observations in the sample. The new outlier test relies on a two-stage nonparametric bootstrap procedure. Monte Carlo experiments show that the test has good asymptotic properties, even for relatively small samples and heavy tailed distributions. Our simulations also reveal the importance of a scaling parameter for finite sample performance. We believe that the proposed test could be instrumental in a wide range of statistical applications where few observations are available and the underlying DGP is unknown.

The remaining sections are organized as follows. Section 2 introduces the elementary statistical definitions and describes the new outlier test. In Section 3, we explore the asymptotic properties of the test in a Monte Carlo experiment. Section 4 presents two empirical examples using aeronautic and macroeconomic data. Finally, concluding remarks are contained in Section 5.

2 The Bootlier test

Consider a sample of observations labeled $\mathbf{Y} = [Y_1, ..., Y_n]$, where $n = 1, 2, ..., N$. The cumulative distribution function (cdf) that generates this sequence, $F_Y(.)$, is unknown. If any observation Y_i is not distributed according to $F_Y(.)$, it is considered as an outlier. We develop a two-step method to test for outliers in $\mathbf{Y}$, where the sample is potentially small and the underlying distribution $F_Y(.)$ is not known. First, we construct a test for the *presence* of one or more outliers, building on two established statistical methods (Singh and Xie, 2003; Silverman, 1981). Second, we use a sequential algorithm to determine the number of outlying observations and to *locate* them in $\mathbf{Y}$.

2.1 A bootstrap based outlier detection plot

A characterizing property of bootstrap resampling is that, when there is an outlier in a data set, it is contained in only a subset of bootstrap resamples. The outlier causes a significant increase in the sample mean of the bootstrap resample, which makes the bootstrap histogram of the sample mean a mixture distribution with more than one mode. Exploiting this feature, Singh and Xie (2003) propose a graphical tool denoted

[1] See Barnett and Lewis (1994) for an overview of outlier detection and finite sample.

Bootstrap Based Outlier Detection Plot (or simply 'Bootlier plot'), which can suggest the presence of at least one (but possibly more) outlier(s) in a sample drawn from an unknown distribution. However, unless the outlier(s) is (are) very severe, the multimodality of the bootstrap histogram is not quite visible. Therefore, Singh and Xie (2003) propose bootstrapping a statistic termed 'mean - trimmed mean' (MTM), and inspecting the modality of the MTM's density for outliers.

The Bootlier plot is obtained as follows. Let $\mathbf{Y}^b = [Y_1^b, ..., Y_n^b]$ $(b = 1, 2, ..., B)$ denote the bootstrap counterpart of $\mathbf{Y}$. First, consider the k-trimmed mean of the bth bootstrap resample $\mathbf{Y}^b$, which is computed by taking the mean after removing the k smallest and largest observations from $\mathbf{Y}^b$:

$$\bar{\mathbf{Y}}^b(k) = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} Y_{(i)}^b, \tag{1}$$

where $Y_{(i)}^b$ are the (ascending) order statistics and k is some trimming value.[2] The MTM of the bth bootstrap resample, M^b, is the difference between the arithmetic mean and the k-trimmed mean:

$$M^b = \frac{1}{n} \sum_{i=1}^{n} Y_i^b - \bar{\mathbf{Y}}^b(k). \tag{2}$$

By construction, the pdf of M^b, $f_M(.)$, is very sensitive to unusually small or large observations – outliers – in the sample $\mathbf{Y}$. In particular, Singh and Xie (2003) show that, in the presence of outlier(s), the limiting bootstrap distribution of M^b can be expressed as a mixture of normal distributions. Therefore, if there is a minimum amount of separation between the outliers and the remainder of the sample, then the mixture density – the Bootlier plot – will be characterized by several modes.[3] Hence, $f_M(.)$ typically exhibits one mode associated with $F_Y(.)$ and at least another mode corresponding to the outlier(s). Consequently, testing for the presence of outlier(s) in $\mathbf{Y}$ is equivalent to testing for the modality of the probability density function $f_M(.)$. If $f_M(.)$ is unimodal, the sample $\mathbf{Y}$ is free of outliers, while if $f_M(.)$ is multimodal, the presence of outliers in $\mathbf{Y}$ is confirmed. However, note that the number of modes does not necessarily match the number of outliers. Typically, several outliers of the same magnitude will be located around the same pole and only one mode will appear in the probability density function $f_M(.)$. It is thus not correct to associate a given number of outliers with the same number of modes in the density $f_M(.)$.

Nevertheless, determining the modality of the bootstrap density absent an assumption for the functional form of $F_Y(.)$ is not straightforward. Singh and Xie (2003) introduce a 'Bootlier index' as a rule-of-thumb tool in order to determine the degree of bumpiness of the density function, but they do not provide a statistical framework to test for multimodality and thus for the presence of outliers. They state that "Formal tests for outliers can be constructed with the Bootlier index as test statistic under a distributional assumption" (page 543). Our approach is different. In what follows we consider a distribution-free test for the null of no outliers.

[2] Following Singh and Xie (2003), we set $k = 2$ in our applications.

[3] We explore the degree of separation in a Monte Carlo experiment by introducing outliers of different magnitudes into samples drawn from a known distribution.

2.2 Testing for multimodality

A formal test for the presence of outlier(s) in $\mathbf{Y}$ can be formulated from the following hypotheses:

- H_0: $f_M(.)$ has precisely one mode (and no local minimum) in the interior of a given closed interval $\Im$;
- H_1: $f_M(.)$ has more than one mode in $\Im$.

H_0 is equivalent to the null hypothesis that there are no outliers in $\mathbf{Y}$, while H_1 corresponds to the alternative that there is one or more outliers. In order to test these hypotheses, we couple the Bootlier plot with a distribution-free test for multimodality proposed by Silverman (1981), which is based on the property that the kernel density estimator is a consistent nonparametric estimator of a pdf.

For the MTM statistics $M^1, ..., M^B$ drawn from density $f_M(.)$, the kernel density estimate at any point x is expressed as:

$$\hat{f}(x,h) = \frac{1}{bh}\sum_{b=1}^{B} K(\frac{x - M^b}{h}), \tag{3}$$

where h is a bandwidth (or smoothing parameter) and $K(.)$ is a kernel function. Without loss of generality, $K(.)$ is chosen to be the standard normal density function following Silverman (1981). For a large class of kernel functions – including the standard normal –, the number of modes of the kernel density is a right-continuous decreasing function of the bandwidth h. Thus, for a sufficiently large bandwidth, $\hat{f}(., h)$ has a single mode in the interior of the given closed interval $\Im$. Furthermore, there is a narrowest bandwidth h_{crit}, for which the kernel density estimated with this bandwidth, $\hat{f}(., h_{crit})$, is unimodal. This is the so-called critical bandwidth defined as $h_{crit} = inf(h;\ \hat{f}(., h)$ has precisely one mode in $\Im)$. The critical bandwidth is larger for a multimodal density function than for a unimodal one, since for a multimodal density a larger bandwidth is required to smooth out multiple modes. Using this property, Silverman (1981) proposes a bootstrap procedure to test for the multimodality of any pdf.

Coupling the Bootlier plot with Silverman's test provides a distribution-free test for the presence of outliers in a sample drawn from an unknown pdf. We refer to this method as the 'Bootlier test'. The testing procedure can be summarized as follows:

1. Draw a large number $b = 1, 2, ..., B$ of random samples from $\mathbf{Y}$ with replacement, and for each resample $\mathbf{Y}^b$ compute the mean-trimmed mean statistic M^b.
2. Obtain the kernel density estimate in Equation 3 for the mean-trimmed mean statistics $M^1, ..., M^B$, denoted $\hat{f}_M(., h)$.
3. Estimate the critical bandwidth $\hat{h}_{crit}$ of the density $\hat{f}_M(., h)$, and re-estimate the kernel density with the critical bandwidth, that is, $\hat{f}_M(., \hat{h}_{crit})$.
4. Silverman (1981) bootstrap algorithm:

(a) Let $M^{1*}, ..., M^{B*}$ denote a bootstrap resample drawn from the distribution with density $\hat{f}_M(., \hat{h}_{crit})$.[4]

(b) Obtain the kernel density estimate in Equation 3 for the bootstrap mean-trimmed mean statistics $M^{1*}, ..., M^{B*}$, denoted $\hat{f}_{M^*}(., h)$.

(c) Estimate the bootstrap critical bandwidth $\hat{h}^*_{crit}$ of the bootstrap density $\hat{f}_{M^*}(., h)$.

(d) Repeat steps (a) - (c) a large number of times.

5. The null hypothesis of unimodality (no outliers in $\mathbf{Y}$) is rejected if $Prob(\hat{h}^*_{crit} \leq \lambda_\alpha \hat{h}_{crit}) \geq 1 - \alpha$, where α is the nominal size (usually 5%) and λ_α is a scaling parameter which ensures that the empirical size corresponds to the nominal size.

2.3 Identification of outliers

The Bootlier test can be implemented to test for the presence of outliers in a data set. A multimodal pdf of the MTM statistics points to the existence of a single or multiple outliers in the sample. However, the test does not indicate which observations are outliers. To locate the outlying observations, we build subsamples by sequentially canceling observations from the tails of the original sample ordered in ascending order, and we perform the Bootlier test on each ordered subsample until the null hypothesis of unimodality cannot be rejected for a particular subset of observations. The data points not contained in this subset are the outliers.

Formally, the sequential algorithm can be summarized as follows. Consider the ascending order statistics $\mathbf{Y}_{(i)} = [Y_{(1)}, Y_{(2)}, ..., Y_{(n-1)}, Y_{(n)}]$. First, we test for the presence of outliers in $\mathbf{Y}_{(i)}$ using the Bootlier test. If the unimodality null hypothesis is rejected, then $\mathbf{Y}$ contains one or more outliers and these must be located in the upper and/or lower tails of $\mathbf{Y}_{(i)}$. We sequentially cancel observations from the tails, i.e., we take the following subsamples: $[Y_{(1)}, ..., Y_{(n-1)}]$, $[Y_{(2)}, ..., Y_{(n)}]$, $[Y_{(1)}, ..., Y_{(n-2)}]$, $[Y_{(2)}, ..., Y_{(n-1)}]$, $[Y_{(3)}, ..., Y_{(n)}]$, $[Y_{(1)}, ..., Y_{(n-3)}]$, etc., and we perform the Bootlier test for each subsample until we cannot reject the null hypothesis, and we find the largest subsample which exhibits unimodality. The observations within this ($\tilde{n}$-dimensional) subset $\mathbf{Y}_{\tilde{n} \leq n}$ are identically distributed, while the complement set $\mathbf{Y}_{\tilde{n} \neq n}$ contains the outliers. The number of outliers is $n - \tilde{n}$.[5]

3 Simulation study

We investigate the finite sample behavior of the Bootlier test in a Monte Carlo experiment. Mammen, Marron, and Fisher (1992) and Hall and York (2001) have studied

[4] In practice we compute bias-corrected resamples following Efron (1979):

$$M^{b*} = \mu_{M^{b*u}} + (M^{b*u} - \mu_{M^{b*u}} + \hat{h}_{crit}\varepsilon_i)(1 + \hat{h}_{crit}/\sigma^2_{M^{b*u}})^{-\frac{1}{2}}$$

where ε_i is an *i.i.d* variable drawn from the distribution $K(.)$, $\mu_{M^{b*u}}$ is the sample mean of M^{b*u}, $\sigma^2_{M^{b*u}}$ denotes the variance of M^{b*u}, and the superscript u stands for uncorrected values of the bootstrap resample.

[5] For example, if the density of MTMs corresponding to the $((n-2) \times 1)$ vector $[Y_{(2)}, ..., Y_{(n-1)}]$ is unimodal, then these observations are drawn from the same (unknown) distribution, and the observations $Y_{(1)}$ and $Y_{(n)}$ are outliers (and $\tilde{n} = 2$).

the asymptotic properties of the test proposed by Silverman (1981). The test is found to be conservative, as the true probability that it incorrectly rejects the null hypothesis of unimodality lies below the nominal level when the scaling parameter λ_α equals 1. Furthermore, Fisher and Marron (2001) show that problems arise when the underlying distribution is heavy tailed. Therefore, we correct for the downward bias in the empirical size of the multimodality test by calibrating λ_α such that the empirical size is close to the nominal size.

Two cdfs are considered, which have different shapes in their tails.[6] Samples of size n are generated from the standard normal distribution and from the Student-$t(n-1)$ distribution with $n-1$ degrees of freedom, which has heavy mass on the tails. 1,000 Monte Carlo replications of the Bootlier test are performed. For each Monte Carlo replication, the $f_M(.)$ density of the MTM statistics is estimated from 10,000 bootstrap draws from the sample $\mathbf{Y}$, and Silverman's test is performed with 1,000 bootstrap replications of M^{b*} drawn from the distribution with density $\hat{f}_M(., h_{crit})$. The asymptotic sample size is set to $n = 100$ for ease of computer time. To explore the small sample performance of the test, the second sample size is set to $n = 10$. The simulations are performed using the Statistics Toolbox in MATLAB.

First, we simulate under the null hypothesis of no outliers, in order to assess the size properties of the Bootlier test. Table 1 reports the rejection frequencies of the null hypothesis of no outlier. The rejection frequencies confirm the findings of Hall and York (2001) and Fisher and Marron (2001): Silverman's modality test is undersized when the scaling parameter equals $\lambda_\alpha = 1$. Moreover, the size bias increases as the sample size shrinks.

We calibrate λ_α such that the test achieves its nominal size in both large and small samples. Thus, Table 1 also shows the optimal – size adjusted – λ_α^{opt}, which ensures an empirical size close to the nominal size of 5%. Overall, we find that λ_α^{opt} is close to the value obtained by Hall and York (2001).[7] The scaling parameters are similar for both distributions when $n = 10$. Moreover, for large n, they converge to 1, although the convergence is slower for the fat tailed distribution.

Table 1: Monte Carlo simulation: Size

	N(0,1) distribution		Student-t distribution	
Sample Size	Rejec. Freq.	λ_α	Rejec. Freq.	λ_α
$n = 10$	0.00	1	0.00	1
	0.05	λ_α^{opt}=1.137	0.05	λ_α^{opt}=1.134
$n = 100$	0.01	1	0.04	1
	0.05	λ_α^{opt}=1.021	0.05	λ_α^{opt}=1.070

Note: The top panel reports the rejection frequencies of the null hypothesis of no outlier for the standard normal distribution, the bottom panel reports the rejection frequencies for the Student-t distribution with $\lambda_\alpha = 1$, and with λ_α^{opt} set at a nominal size of 5%.

[6] Our procedure requires the existence of finite means for the computation of the MTM statistic. Therefore, random variables generated from the Cauchy distribution cannot be considered, since for the latter finite moments do not exist (see Casella and Berger, 2002).

[7] In Equation (4.1), on page 524, they obtain $\lambda_\alpha^{opt} = 1.1294$ for $\alpha = 5\%$.

Next, we turn to the power of the test. Table 2 shows the rejection of the null hypothesis when data is simulated under the presence of an outlier, such that the outlier equals $\hat{\mu} + i\hat{\sigma}$, where $\hat{\mu}$ is the sample mean of the baseline sample (absent outliers), while $\hat{\sigma}$ is the sample standard deviation. The size of the outlier depends proportionally on the value of i (we consider $i = 3.5$, 4, 4.5, and 5). The power is corrected for size distortions, since the simulations are also performed with the optimal scaling parameter λ_α^{opt}.

Table 2: Monte Carlo simulation: Power (size-adjusted)

N(0,1) distribution; Outlier = $\hat{\mu} + i\hat{\sigma}$					
		i=3.5	i=4	i=4.5	i=5
Sample Size	Lambda	Rejec. Freq.	Rejec. Freq.	Rejec. Freq.	Rejec. Freq.
n=10	1	0.22	1.00	1.00	1.00
	λ_α^{opt}	0.96	1.00	1.00	1.00
n=100	1	1.00	1.00	1.00	1.00
	λ_α^{opt}	1.00	1.00	1.00	1.00
Student-t distribution; Outlier = $\hat{\mu} + i\hat{\sigma}$					
		i=3.5	i=4	i=4.5	i=5
Sample Size	Lambda	Rejec. Freq.	Rejec. Freq.	Rejec. Freq.	Rejec. Freq.
n=10	1	0.19	0.98	1.00	1.00
	λ_α^{opt}	0.41	1.00	1.00	1.00
n=100	1	1.00	1.00	1.00	1.00
	λ_α^{opt}	1.00	1.00	1.00	1.00

Note: Rejection frequencies of the null hypothesis of no outlier when the distributions are simulated under the presence of an outlier, with outliers specified as values corresponding to the mean $\hat{\mu}$ plus i times the size of the sample standard deviation $\hat{\sigma}$ (outlier = $\hat{\mu} + i\hat{\sigma}$). The power is corrected for size distortions, as simulations are performed with the optimal λ_α^{opt}.

Table 2 reveals that, when the test is undersized (i.e., $\lambda_\alpha = 1$) and the outlier is small in magnitude relative to the sample mean ($i = 3.5$), the test has moderate power. This result holds for both distributions. However, when the test is correctly sized, it generally attains good power (the only exception being the Student-t distribution when both the sample size and the outlier is small). The frequency of correct rejection of the null hypothesis reaches 100% (for a 5% nominal size) in all cases when the outlier is at least four times the sample standard deviation, irrespective of the distribution considered.

4 Empirical illustration

Example 1

The empirical performance of the Bootlier test is illustrated by means of two examples. First, we revisit an aeronautics data set studied earlier by Dalal, Fowlkes, and Hoadley (1989) and Singh and Xie (2003). Shortly after liftoff on January 28, 1986, the space shuttle *Challenger* disintegrated over the Atlantic Ocean. The explosion occurred due to the leakage of an O-ring that sealed the right solid booster rocket of the vehicle, which allowed pressurized gas from within the rocket to reach the outside, leading to combustion.

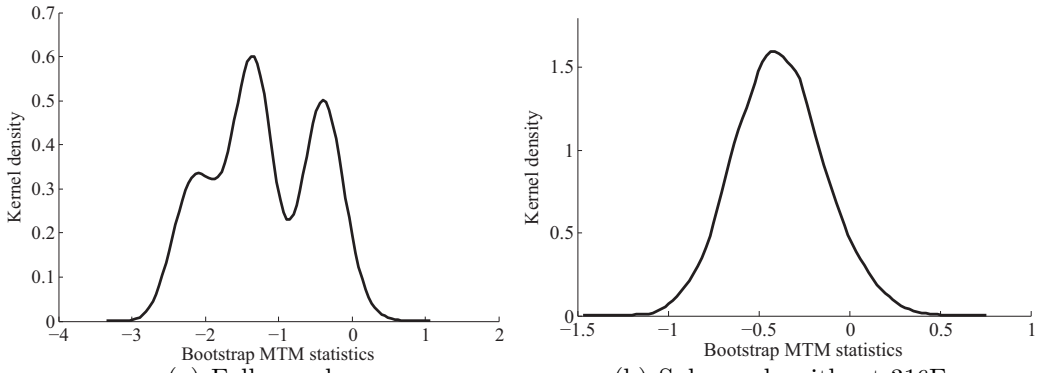
It has been subsequently shown that the O-rings do not seal properly at low temperatures (see Dalal et al., 1989).

Consider the recorded temperatures at which the O-rings were sealed on all 25 shuttle launches of the *Challenger*, expressed in degrees Fahrenheit:

66 70 69 80 68 67 72 73 70 57 63 70 78 67 53 67 75 70 81 76 79 75 76 58 31.

The last observation is 31^{o}F, the temperature on the day of the explosion. This data point qualifies as a good candidate for being an outlier. Figure 1 shows the Bootlier plot of the full sample, which exhibits 3 modes, and the Bootlier plot of the sample after removing the last observation, which is unimodal. Hence, a visual inspection of the Bootlier plots suggests that 31^{o}F is indeed an outlier. Next, the Bootlier test is performed to obtain formal statistical evidence. The test for the full sample gives a p-value lower than 1%, indicating a clear rejection of the null hypothesis of unimodality (no outliers), while the test on the subsample which does not contain 31^{o}F leads to a p-value of 0.21. Consequently, the null hypothesis cannot be rejected for the latter subsample, and the temperature at which the O-rings were sealed on the day of the accident proves to be an outlier.

Figure 1: Bootlier plots of *Challenger* data



(a) Full sample (b) Subsample without 31^{o}F

Example 2

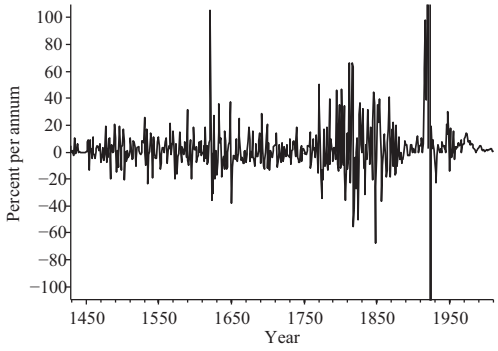
Our second example investigates macroeconomic data. We consider annual inflation rates (annual percent changes in the aggregate price level) for Germany from 1428 to 2010. The data come from Reinhart and Rogoff (2011) and are available on the Internet site of the *American Economic Association*.[8]

Figure 2 plots the time series. In order to obtain a good visualization of the series, the data range is censored in the figure from above at 1920, 1922, and 1923 when the rate of inflation was 291.45, 2715.224, and 2.11E+11 percent per annum respectively, and from below at 1924 when the inflation rate was -200 percent per annum. The nearly six centuries of monetary history under study witnessed several episodes of hyperinflation, which constitute potential outliers. At first glance, three relatively more turbulent periods

[8]See: http://www.aeaweb.org/articles.php?doi=10.1257/aer.101.5.1676.

stand out. First, the Thirty Years' War in the early 17th century, second, the prolonged inflationary period surrounding the industrial revolution from the late 18th to the late 19th century, and third, the infamous hyperinflation of the 1920s. The figure also reveals that the most recent decades are characterized by the historically lowest and least volatile inflation. This period is often described as the "Great Moderation", and it coincides with the adaption of a proactive approach toward inflation in central banking.

Figure 2: Germany: inflation, annual percent change, 1428-2010.



The results of the Bootlier test for the inflation data are reported in Table 3. We find six outliers in the data, and once we remove these from the sample, we obtain a p-value of 0.09 for the Bootlier test. The earliest outlier occurs in 1621, at the peak year of the *Kipper- und Wipperzeit* (Tipper and Seesaw Time), a monetary crisis in the Holy Roman Empire between 1619 and 1623, which was characterized by hyperinflation through debasement of commodity (gold, silver, and copper) money. An exhaustive historical account of the crisis is offered by Kindleberger (1991).

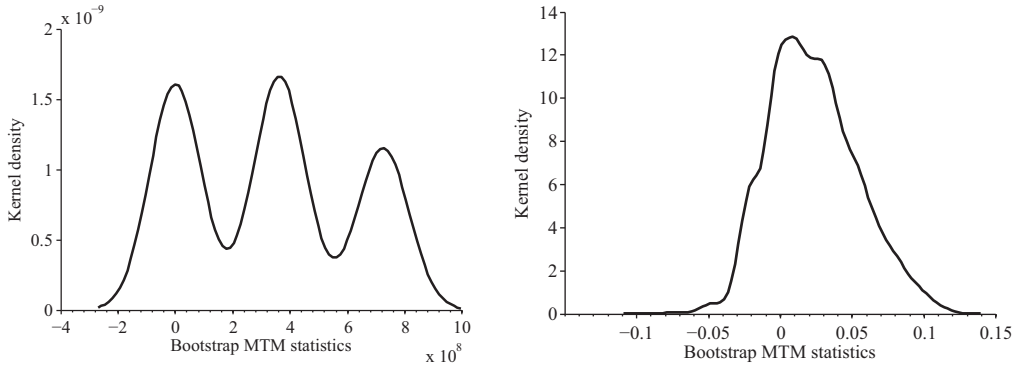
The other five outliers concentrate around the most severe hyperinflation and subsequent collapse in German history in the early 20th century, which has been investigated in a seminal paper by Cagan (1956). Inflation peaked at approximately 211 billion percent per annum in 1923, which is the most outlying observation. Nevertheless, all outliers are relatively severe, given that the sample mean without the outliers is 2.352 percent and the sample standard deviation is 14.250 percent.

Table 3: Outliers in German inflation (percent per annum)

Year	Outlier
1621	105.290
1917	98.182
1920	291.450
1922	2715.224
1923	2.11E+11
1924	-200.000

Figure 3 (a) shows the Bootlier plot for the full sample. Again, the mutimodal pattern is clearly visible, indicating the presence of outliers. Figure 3 (b) shows the Bootlier plot once all outliers are removed. This figure reveals that a mere visual inspection of the Bootlier plot in insufficient and may lead to misleading conclusions regarding outliers.

Figure 3: Bootlier plots of inflation data



(a) Full sample (b) Subsample without outliers

Even though the plot exhibits a minor kink, the modality test cannot statistically reject the unimodality hypothesis at the conventional 5% significance level.

5 Concluding remarks

This paper has introduced a distribution-free test for outliers in statistical data drawn from an unknown DGP. Building on earlier work by Singh and Xie (2003) and Silverman (1981), we construct a test for the presence of one or more outliers, and we propose a sequential algorithm to determine the number of outliers as well as their location in the sample. Monte Carlo experiments show that this new method has good asymptotic properties, even for relatively small samples, whatever shape the underlying pdf may have in its tails.

The empirical performance of the outlier test is illustrated by means of two empirical examples using aeronautic and macroeconomic data. The new test could be instrumental in a wide range of statistical applications where few observations are available and the underlying DGP is unknown. For instance, Candelon, Metiu, and Straetmans (2012) employ the test proposed in this paper to investigate business cycle booms and depressions.

References

Barnett, V. and T. Lewis (1994). *Outliers in Statistical Data* (3^{rd} ed.). Wiley Series in Probability and Statistics. Wiley.

Cagan, P. (1956). The monetary dynamics of hyperinflation. In M. Friedman (Ed.), *Studies in the quantity theory of money*. University of Chicago Press, Chicago.

Candelon, B., N. Metiu, and S. Straetmans (2012). Understanding economic booms and depressions. Mimeo Maastricht University.

Casella, G. and R. L. Berger (2002). *Statistical Inference* (2 ed.). Duxbury Advanced Series. Duxbury Press.

Dalal, S., E. Fowlkes, and B. Hoadley (1989). Risk analysis of the space shuttle: Pre-Challenger prediction of failure. *Journal of the American Statistical Association 84*, 945–957.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics 7*(1), 1–26.

Fisher, N. I. and J. S. Marron (2001). Mode testing via the excess mass estimate. *Biometrika 88*(2), 499–517.

Hall, P. and M. York (2001). On the calibration of Silverman's test for multimodality. *Statistica Sinica 11*, 515–536.

Kindleberger, C. (1991). The economic crisis of 1619 to 1623. *Journal of Economic History 51*, 149–175.

Mammen, E., J. S. Marron, and N. I. Fisher (1992). Some asymptotics for multimodality tests based on kernel density estimates. *Probability Theory and Related Fields 91*, 115–132.

Reinhart, C. and K. Rogoff (2011). From financial crash to debt crisis. *American Economic Review 101*, 1676–1706.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society, B 43*(1), 97–99.

Singh, K. and M. Xie (2003). Bootlier Plot - Bootstrap based outlier detection plot. *Sankhya: The Indian Journal of Statistics 65*(3), 532–559.

Walsh, J. (1959). Large sample nonparametric rejection of outlying observations. *Annals of the Institute of Statistical Mathematics 10*, 223–232.