Marmolejo-Ramos *et al. Journal of Statistical Distributions and Applications* (2015) 2:8
DOI 10.1186/s40488-015-0031-y

Journal of Statistical Distributions and Applications
a SpringerOpen Journal

RESEARCH Open Access

Automatic detection of discordant outliers via the Ueda's method

Fernando Marmolejo-Ramos[1†*], Jorge I. Vélez[2,3†] and Xavier Romão[4]

*Correspondence:
fernando.marmolejo.ramos@pshychology.su.se
†Equal contributors
[1]Gösta Ekman Laboratory, Department of Psychology, Stockholm University, Stockholm, Sweden
Full list of author information is available at the end of the article

Abstract

The importance of identifying outliers in a data set is well known. Although various outlier detection methods have been proposed in order to enable reliable inferences regarding a data set, a simple but less known method has been proposed by Ueda (1996/2009). Since this new method, called Ueda's method, has not been systematically analysed in previous research, a simulation study addressing its performance and robustness is presented. Although the method was derived assuming that the underlying data is normally distributed, its performance was analysed using data from various outlier-prone distributions commonly found in several research fields. The results obtained enable us to define the strengths and weaknesses of the method along with its limits of applicability. Furthermore, an unforeseen field of application of the method, which requires further studies was also identified.

Keywords: Ueda's method; Outliers; Skewed distributions; AIC; Normal distribution; Statistical simulation

Classification codes: 97K80; 68U20

1 Introduction

Identifying outliers is essential to the data analyst in order to make reliable inferences on the data at hand. Various methods have been proposed for the identification of outliers and the logic of those methods depends directly on how outliers are defined (Aggarwal 2013; Chandola et al. 2009). In cognitive science, specifically in experimental psychology and neuroscience, outliers are accommodated via data transformations (e.g., Box-Cox transformation) or eliminated via truncation (e.g. observations below and/or above the extremes of a pre-set range of values are removed). An approach frequently used is the z-value test for outliers (Songwon, 2006) in which observations are converted to z-scores in order to see which observations fall a pre-set number of standard deviations (SDs) away from the mean. Although there have been proposed SD values to be used given specific sample sizes (see Van Selst and Jolicoeur 1994), in practice researchers use ± 2, ± 2.5 and ± 3SDs as benchmarks regardless (examples of the usage of these benchmarks can be found in Bertels et al. 2010; Havas et al. 2007; Otte et al. 2011; a comprehensive simulation study comparing these and other methods can be found in Marmolejo-Ramos et al.2015). The z-score approach implicitly assumes that the data comes from a normal distribution, thus forcing the actual distribution of the data set to adopt a bell shape.

Many outlier detection methods are available for univariate data sets (e.g. see Hayes and Kinsella 2003; Thode 2002; Verma 1997; Verma et al. 2014). Among these, the method proposed by Ueda (Ueda 1996/2009) which still assumes the underlying data to be normally distributed and based on the Akaike's Informatin Criterion (AIC) (Kitagawa 1979), presents interesting features. However, this method has not been systematically studied and, to be best of the authors' knowledge, it has only been featured in just a handful of studies involving genetic data (Kadota et al. 2003a,b; Tsuyuzaki et al. 2013).

The aim of this paper is to study the performance and robustness of the Ueda's method to detect outliers via computer simulations, as well as determine its applicability to other types of data. It should be noted that, in the context of the proposed study, outliers are not just seen as errors in the data, as considered in other situations (see Hoaglin et al. 1983 for an example). Instead, outliers are generally viewed as data values which are numerically distant from the bulk of the sample, thus requiring particular consideration given their significance. In the first section, the Ueda's method is described and a modification of it, aimed at automating the selection of the number of candidate outliers, is proposed. In addition, some examples are provided to illustrate this modification. In order to study the performance of the Ueda's method, various outlier-prone distributions commonly found in actual research are used. For comparative purposes, a 95 % confidence level is used (i.e., the nominal type I error probability is 5 %) and the normal distribution is also included in the study. Finally, and based on the results obtained, potential applications of and research topics involving the Ueda's method are discussed in the context of applied and statistical research.

1.1 Background

Let X be a random variable with probability distribution $f_X(x)$, and let $x_1, x_2, \ldots, x_n, x_{n+1}, \ldots, x_{n+s-1}, x_N$ be a random sample of size $N = n + s$ from this distribution, with n and $s = \{1, 2, 3, \ldots\}$ being the number of regular and outlying observations, respectively.

The Ueda's method is a simplified version of the use of the Akaike Information Criterion (AIC) proposed by Kitagawa (Kitagawa 1979). The aim of the AIC, estimated as

$$AIC = 2k - 2\ln(L), \tag{1}$$

is to provide a measure for model selection. In the expression above, k is the number of independently adjusted parameters and L is the maximum likelihood value for n number of regular observations. Thus, the best probabilistic model for the regular observations is that maximising

$$L = \prod_{i=1}^{n} f_X(x_i), \tag{2}$$

where n represents the regular observations in a given univariate data set and $f_X(x_i)$ represents the value of the density function for the ith observation ($i = 1, 2, \ldots, n$). In the normal distribution, the AIC becomes:

$$AIC = 2n\ln\hat{\sigma} - 2\ln n! + 2s \tag{3}$$

where

$$\hat{\sigma} = N^{-1/2} \sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{4}$$

is the standard deviation of the full sample, and $\bar{x}$ is the sample mean.

Ueda (Ueda 1996/2009) noticed that $n^{-1} \ln n! \approx 1$ for $n \sim 5-9$, ≈ 2 for $n \sim 10-28$, ≈ 3 for $n \sim 29-82$, and so forth (see Fig. 1a). By considering the term $2s - 2\ln n!$ in the right side of Eq. (3) and as $n = N - s$ and $n^{-1} \ln n! \approx 1$ for $n \sim 5-9$, it follows that

$$\begin{aligned} -2\ln n! + 2s &\approx 2s - 2n \\ &= 2s - 2(N-s) \\ &= 4s - 2N \\ &= 4s - \text{constant} \end{aligned} \tag{5}$$

This same procedure can be used for other values of n in order to obtain $6s -$ constant, $8s -$ constant, and so on (see Table two in Ueda 1996/2009). Note that the resulting values $4s, 6s, 8s, \ldots$ are correction terms that depend on n, but that are inaccurate since the total number of samples N is constant. To ameliorate the problem, Ueda used the correction factor $\sqrt{2}\, s\, \frac{\ln n!}{n}$ which still depends on n, but includes the number of outliers s in the full sample (see Fig. 1b). Based on this, the proposed test statistic to identify outliers using this method is given by (Ueda 1996/2009):

$$\begin{aligned} U_t &= \frac{1}{2} AIC \\ &\approx n \ln \hat{\sigma} - \sqrt{2}\, s\, \frac{\ln n!}{n} \\ &= (N-s) \ln \hat{\sigma} - \sqrt{2}\, s\, \frac{\ln n!}{n}. \end{aligned} \tag{6}$$

In Eq. (6), the term $n \ln \hat{\sigma}$ provides an index of the predictability of true outliers from a set of candidate outliers (i.e., particularly from both ends of the tails of a distribution), whereas the term $\sqrt{2}\, s\, \frac{\ln n!}{n}$ penalises the model according to the number of parameters (i.e., the number of outliers s). Specifically, low values in

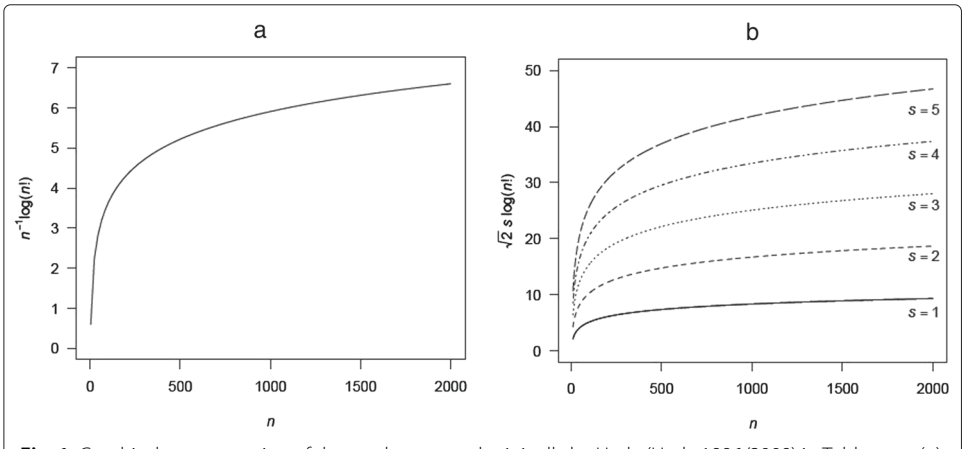

Fig. 1 Graphical representation of the results reported originally by Ueda (Ueda 1996/2009) in Tables one (**a**) and three (**b**). Here, $N = n + s$ is the sample size, n is the number of non outlying observations, and s represents the number of candidate outliers

the first term indicate that true outliers have been found, whereas high values in the second term signal unreliability of the model due to having many parameters (Kadota et al. 2003a,b). Thus, the lowest AIC associated with a combination of candidate outliers signals the best model that detects a set of potential outliers.

1.2 Calculating U_t

Based on Ueda (1996/2009), the following procedure is suggested to determine outliers in a full sample:

1. Order the full sample to obtain $x^{\text{ord}} = \left(x_{(1)}, x_{(2)}, \ldots, x_{(N)}\right)$, with $x_{(i)}$ being the ith order statistic ($i = 1, 2, \ldots, N$).
2. Calculate the z-score for each $x_{(i)}$ as
$$z_{(i)} = \frac{x_{(i)} - \bar{x}}{\hat{\varsigma}} \tag{7}$$
with
$$\hat{\varsigma} = \sqrt{\frac{\sum_{i=1}^{N}(x_{(i)} - \bar{x})^2}{N-1}}$$
and $\bar{x}$ the sample mean ($i = 1, 2, \ldots, N$).
3. Provided a number of steps $s \geq 0$, calculate the test statistic U_t as in Eq. (6). Observe that $s = 0$ means that no outliers are present in the data, and any other value of s implies otherwise. Furthermore, as the original sample is ordered, $s > 0$ also implies that the outliers are at either end of the sample (e.g. on the lower or upper tail of the distribution); which end of the ordered sample is tackled first, heavily depends on how the calculation of the test statistic is performed (see § 1.3).
4. Suppose that $s = 1$. It could be the case that the outlier is located either at the beginning of the ordered sample, i.e. $x_{(1)}$ is the outlier, or at the end of it, i.e. $x_{(N)}$ is the outlier. If the former, then the test statistic U_t is calculated as in Eq. (6) after removing $x_{(1)}$ from the ordered sample. If the latter, $x_{(N)}$ is removed and U_t subsequently calculated.
5. Step 4 continues for other values of $s > 1$. As a result of this process, a collection of U_t values is obtained.

1.3 On the number of steps

The previous section described how the test statistic U_t is calculated provided a value of $s \geq 0$. One question that remains to be answered is how many values of U_t need to be calculated. Let C be such a number and suppose that up to s outliers are to be detected in the sample. Then, from the collection of U_t values obtained in step 5 of § 1.2, the following matrix is constructed:

$$U_t = \begin{bmatrix} U_{00} & U_{01} & U_{02} & \cdots & U_{0m} \\ U_{10} & U_{11} & U_{12} & \cdots & U_{1m} \\ U_{20} & U_{21} & U_{22} & \cdots & U_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ U_{m0} & U_{m1} & U_{m2} & \cdots & U_{mm} \end{bmatrix}_{m \times m} \tag{8}$$

where $m = s + 1$, U_{00} is the U_t statistic when no outliers are detected, and U_{ij} is the U_t statistic when i and j outliers are detected in the lower and upper tails of x^{ord},

respectively ($i, j = 1, 2, \ldots, m$). This is equivalent to removing $x_{(i)}, x_{(i-1)}, \ldots, x_{(1)}$ and $x_{(N-j)}, x_{(N-j+1)}, \ldots, x_{(N)}$ from x^{ord}.

From U_t it is clear that the number of calculations to detect up to s outliers in x^{ord} is, at the most, m^2. However, this number can be reduced to $m^2 - 2$ if U_{00} and U_{mm}, included for convenience and comparison purposes, are not calculated. Hence, $m^2 - 2 \leq C \leq m^2$.

To illustrate this, consider the squared upper-left block of U_t whose entries are U_{00}, U_{10}, U_{01} and U_{11}, and which correspond to the test statistic in Eq. (6) when no outliers are detected and when $x_{(1)}$, $x_{(N)}$, and both $x_{(1)}$ and $x_{(N)}$ are removed from the full sample, respectively.

1.4 Automatic detection of the number of discordant outliers

In the original version of the Ueda's method, the analyst sets the number of expected outliers in the data set in advance, i.e. the value of s following the notation considered herein. Carling (2000) showed that, for several probability distributions and various sample sizes, approximately 5 % to 50 % of the total number of observations N can be, in general, set as the minimum and maximum reasonably expected number of potential outliers, respectively.

In the automation process for selecting the number of candidates outliers s in a sample of size N, s is constrained to be $1 \leq s \leq s_{\max}$, with $s_{\max}$ being the maximum number of outliers to be detected. In this inequality, the minimum value of s is set to 1 such that at least one outlier is detected by the Ueda's method in a full sample.

Because of how the U_t statistic is calculated (see § 1.2), it is straightforward to show that

$$s_{\max} = \begin{cases} \frac{1}{2}(N-1) & \text{if } n \text{ is odd,} \\ \\ \frac{N}{2} - 1 & \text{if } n \text{ is even.} \end{cases} \tag{9}$$

Based on the value of $s_{\max}$, the U_t matrix is then constructed as in §1.3. It can be seen that the number of outliers in the sample s corresponds to the sum of the indexes of the entry of U_t for which the U_t statistic is minimum, and the location of these outliers in x^{ord} is given by the entry's indexes. For instance, if the minimum value of U_t is U_{12}, then the number of outliers in the sample is $s = 1 + 2 = 3$ and the values will be given by $x_{(1)}$, $x_{(N-1)}$ and $x_{(N)}$ in x^{ord}.

1.5 Examples

This section presents three examples illustrating the use of the Ueda's method and the proposed implementation to automatically detect candidate outliers. Here, one normal, one positively- and one negatively-skewed distributions are shown. In order to have control on the outliers to be found via the Ueda's method, some observations were added to each one of these distributions to play the role of outliers. In the three scenarios, the outlying observations were sampled from normal distributions and placed on the right tails of the normal and positively-skewed distributions, and on the left tail of the negatively-skewed distribution.

Example 1. Normally distributed data. Consider a random sample $x_1, x_2, \ldots, x_{25}$ from a $N(300, 10^2)$ distribution, and a second random sample $y_1, y_2, \ldots, y_5$ of outlier observations from a $N(400, 5^2)$ distribution. Then, the full sample of size $N = n + s = 30$ will

be given by $x_1, x_2, \ldots, x_{25}, y_1, y_2, \ldots, y_5$. Although the true number of outliers is $s = 5$, $s_{\max}$ is set to 10 (see Appendix 6.1) so up to 10 observations are automatically tested for their potentiality to be outliers using the Ueda's method. As shown in Fig. 2a, the number of outliers detected in the full sample is $s = 5$.

Example 2. Positively-skewed distributed data. One hundred observations were generated from an Ex-Gaussian distribution with parameters $\mu = 100$, $\sigma = 10$ and $\nu = 10$ and 20 outlying observations were generated from a $N(200, 20^2)$. Using the Ueda's method with $s_{\max} = 25$, a total of 22 outliers are detected. Fig. 2b depicts the U_t statistic as a function of $s_{\max}$. Note that the minimum value of U_t is reached when $s_{\max} = 22$, which suggests that the number of outliers in the full sample of size $N = 120$ is $s = 22$ and not $s = 20$ as initially introduced. This example shows that the Ueda's method detects those outlier observations initially introduced in addition to a few more that were not thought to be outliers in the full sample. Closer inspection reveals that these two additional observations are located in the upper tail of the distribution and may, as the Ueda's method shows, be signalled as potential outliers (see Fig. 2b).

Example 3. Negatively-skewed distributed data. Consider a random sample of size $n = 50$ from a Beta distribution with parameters $\alpha = 5$ and $\beta = 2$. In this case, three observations generated from a $N(0.1, 0.02^2)$ distribution are added for the Ueda's method to detect them in the combined sample of size $N = 53$ using $s_{\max} = 5$. As shown in Fig. 2c, the number of outliers detected is the one introduced.

In the case of the normal distribution, the outlying observations were readily captured by the Ueda's method. However, examples 2 and 3 represent cases of distributions prone to having outliers (in the sense they were defined in Section 1). That is, these types of

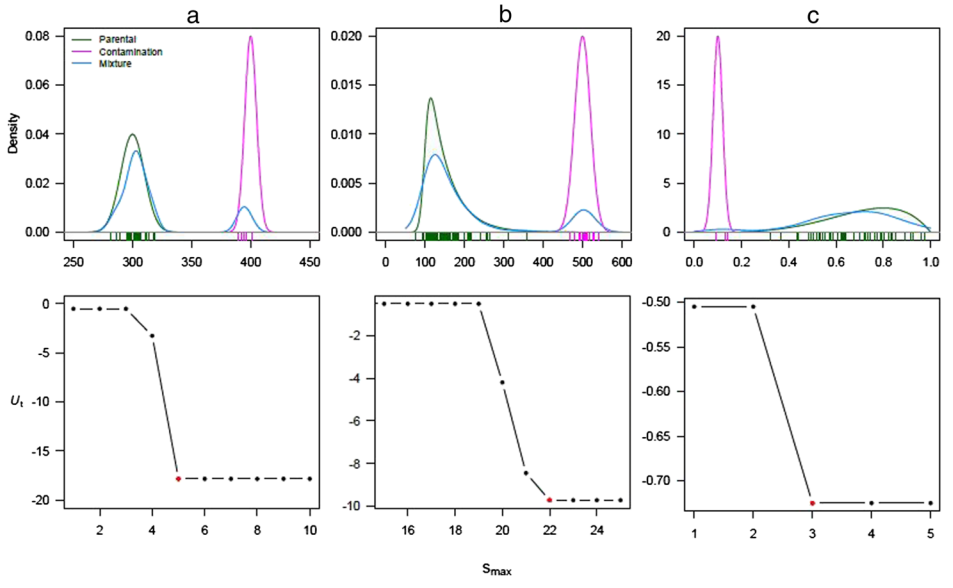

Fig. 2 Results of Ueda's method for examples **a** 1, **b** 2 and **c** 3. *First row:* Probability densities for the non contaminated (green), the contaminating (pink) and full (light blue) samples. *Second row:* U_t statistic as a function of $s_{\max}$. The value of $s_{\max}$ for which U_t is the lowest (red dot) indicates the number of candidate discordant outliers in the full sample

distributions can have outliers as they have elongated tails (see Gleason 1993) accounting for some observations that fall away from where most of the data tend to cluster. Although in the case of the negatively-skewed distribution only the observations placed on the left tail were appropriately caught by the method, this did not happen in the case of the positively-skewed distribution. In that case, two observations that were not part of the added normally-distributed observations were signalled as outliers. This specific case demonstrates that skewed distributions can in fact have outlying observations on the tails and that such observations do occur due to the natural way in which such types of distributions can take place (e.g., positively-skewed data is the norm in human reaction time research). What is interesting is that the Ueda's method seems to be sensitive to these types of naturally-occurring outlying observations. As skewed distributions are prone to having outliers and are typically found in actual research, it is thus essential to determine how the Ueda's method performs in these cases.

2 Simulation study

To study the performance of the Ueda's method, a simulation study was carried out using a similar simulation scheme to that presented by Vélez and Correa (2014), and following the recommendations by Salazar and Baena (2009). In brief, n observations of a particular probability distribution (see Table 1) are generated and a measure of interest is calculated. The following algorithm was implemented in R (R Core Team 2013) to evaluate the performance and robustness of the Ueda's method when *no outliers* are introduced into the data:

1. Draw n random observations from a probability distribution $f_X(x|\theta)$, where θ is the parameter vector (see Table 1).
2. Apply the Ueda's method and determine the number of outliers detected. Denote this number as $\hat{s}$.
3. Repeat steps 1 and 2, B times, and calculate $R = N_0/B$, where N_0 is the number of times in the B samples of size n that $\hat{s} = 0$, i.e. no outliers are detected.

After generating $n = \{10, 20, 30, 50, 100, 200, 500\}$ observations from each probability distribution in Table 1, R was calculated. These distributions and sample sizes were chosen because of what is often seen in real-world applications. Indeed, the Ex-Gaussian, Gamma, Weibull and Lognormal distributions are positively skewed distributions used

Table 1 Probability distributions considered in this study. The probability density function $f(\cdot)$ is shown in the second column, and the parameter vector θ defining each distribution in the third column

Distribution	$f(\cdot)$	θ
Normal	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ, σ
Student's t	$\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\nu/2)\sqrt{\nu\pi}} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	ν
Tukey λ	See text in §2	λ
Beta	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	α, β
Ex-Gaussian	$\frac{1}{\nu\sqrt{2\pi}} e^{\frac{\sigma^2}{2\nu^2} - \frac{x-\mu}{\nu}} \cdot \int_{-\infty}^{[(x-\mu)/\sigma]-\sigma/\nu} e^{-\frac{y^2}{2}} dy$	μ, σ, ν
Gamma	$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}$	α, β
Weibull	$\frac{\beta}{\alpha^\beta} x^{\beta-1} e^{-(\frac{x}{\alpha})^\beta}$	α, β
Lognormal	$\frac{1}{x\sqrt{2\pi}\sigma} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$	μ, σ

in cognitive science to model behavioural and neurological data (e.g., Leiva et al. 2015). The Beta distribution has been used to model soil data (Haskett et al. 1995) and rates and proportions (Ferrari and Cribari-Neto 2004), the Student's t has been used to fit share price changes (Praetz 1972), and the Tukey-λ has been fitted to solar radiation data (Öztürk and Dale 1982). In all simulation scenarios, a total of $B = 10,000$ replicates were used. In what follows, it is described how the simulation scenarios were constructed for each probability distribution *f*.

Every single probability distribution f in Table 1 is defined by a set of parameters θ. Provided *f*, a simulation scenario is defined as a combination of n and specific values of θ for that *f*. For instance, in the Normal distribution, $\theta = (\mu, \sigma)$. Without loss of generality, μ was fixed at 0 and $\sigma = \{0.5, 1, 1.5, \ldots, 5\}$ for each sample size, to obtain a total of 70 simulation scenarios. In the Student's t-distribution, 350 scenarios were studied and each of them was defined by the combination of the sample size n and $\nu = \{1, 2, \ldots, 50\}$.

Let $p \in [0, 1]$ and $x = F^{-1}(p)$, with F being the cumulative distribution function of a random variable X. In the Tukey-λ distribution, the quantile and density function are respectively given by (Chalabi et al. 2014)

$$F^{-1}(p|\lambda) = \lambda + \frac{p^{\lambda} - (1-p)^{\lambda}}{\lambda} \tag{10}$$

and

$$f(x|\lambda) = f\left(F^{-1}(p)\right) = \frac{1}{p^{\lambda-1} + (1-p)^{\lambda-1}}. \tag{11}$$

For specific values of λ, the Tukey-λ distribution resembles the characteristics of some known probability distributions. For instance, the values $\lambda = -1$, $\lambda = 0.14$ and $\lambda = 1$ correspond to the Cauchy$(0, \pi)$, $N(0, 1.4636^2)$ and Uniform$(-1, 1)$ distributions, respectively. In the present simulation strategy, λ was varied in the interval [-2, 2] in steps of 0.2, excluding $\lambda = 0$ and including $\lambda = 0.14$. Thus, a total of 147 scenarios were evaluated.

For the Beta(α, β) distribution, the parameters α and β varied in the rectangle $[a, b] \times [a, b]$ with $a = 0.5$ and $b = 5$ using increments of $h = 0.5$ within each margin.

The Ex-Gaussian distribution is defined by $\theta = (\mu, \sigma, \nu)$. To generate observations from this distribution, $\mu = \{200, 300, 400, 500, 600\}$, σ was fixed at 20, and $\nu = \{400, 300, 200, 100, 50\}$.

In the Gamma(α, β) and Weibull(α, β) distributions, $\alpha, \beta = \{2, 4, 6, 8, 10\}$. A similar approach was used for the Lognormal distribution after replacing μ by α and σ by β.

3 Results

As shown in Fig. 3a, the results for the normal distribution indicate that the proportion of no outliers detected in the data by Ueda's method is below the expected 95 % when $n < 100$, and that the standard deviation σ seems to have no effect on this. Even so, the proportion of times the method finds no outliers is >90 % across sample sizes. Conversely, the method tends to find more outliers than expected when the degrees of freedom of the Student's t-distribution are small ($\nu < 10$, Fig. 3b). This is particularly clear when both the degrees of freedom and the sample size are small (in which case the proportion of not finding outliers decreases to ~ 0.75). Graphical inspections of Student's t-distributions corroborate this finding: when $\nu = 1$, the Student's t-distributions presents some observations that fall very far away from the mean, whereas for $\nu = 10$ the observations

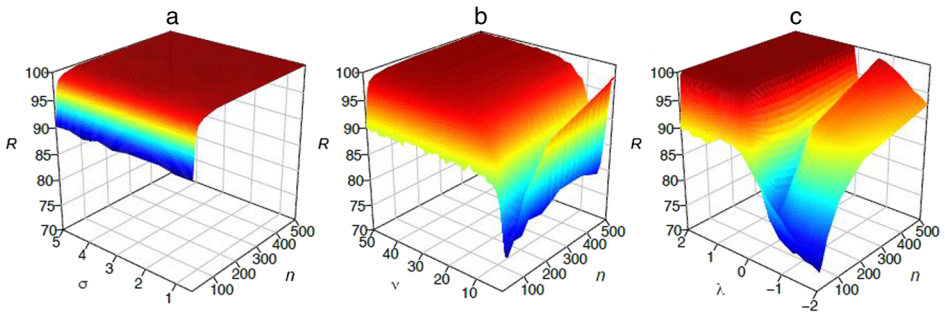

Fig. 3 Results for the **a** Normal, **b** Student's *t*, and **c** Tukey-λ distributions. Higher values of *R* are presented in dark red, and lower values in dark blue. The parameter *n* corresponds to the full sample size

spread out more evenly around the mean. In the Tukey-λ distribution, more outliers tend to be detected when the λ shape parameter decreases ($\lambda < 0$, Fig. 3c). That the Ueda's method finds more outliers in this distribution is accentuated when both n and λ are small. It is worth noting that when λ is large and positive, an uniform distribution is obtained and the Ueda's method does not find outliers. Although uniform distributions are not normal, they are symmetric and observations are spread evenly across the entire data range; these specific results thus suggest that the Ueda's method focuses on the shape and symmetry of the distribution. Overall, compared to the normal distribution case, the Ueda's method tends to find more outliers in the Student's t and Tukey-λ distributions, especially when their parameters generate distributions that are too platykurtic or too leptokurtic.

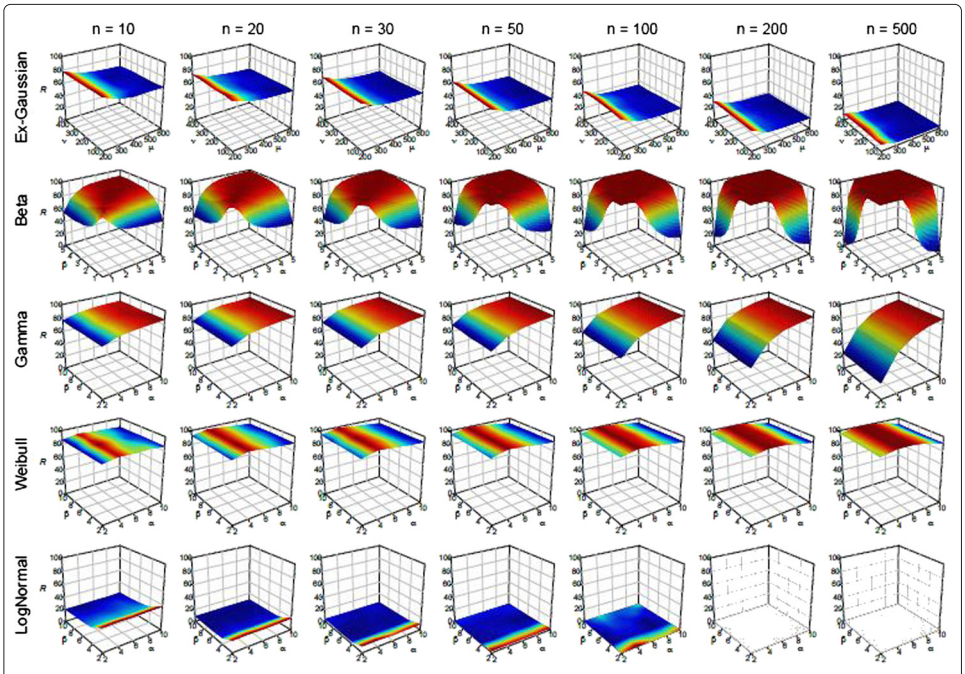

Fig. 4 Results for the Ex-Gaussian, Beta, Gamma, Weibull and Lognormal distributions for fixed sample sizes. Higher values of *R* are presented in dark red, and lower values in dark blue/gray

Figure 4 shows that in positively skewed distributions such as the Ex-Gaussian and LogNormal, the Ueda's method tends to find more outliers as n increases. Additionally, the method has lower chances of not finding outliers in the LogNormal distribution than in the Ex-Gaussian distribution (for instance when $n = 10$, the method is $\sim$2 times more likely to find outliers in the LogNormal distribution than in the Ex-Gaussian distribution). Also, while the Ueda's method tends to detect more outliers as μ increases in the Ex-Gaussian distribution, it also tends to find more outliers as β increases in the LogNormal distribution.

Although not as marked as in the previous cases, Ueda's method has lower chances of not finding outliers as sample sizes increase in the Gamma distribution. However, the α parameter seems decisive in how sensitive Ueda's method becomes. Specifically, the smaller this shape parameter, the more likely it is that the method detects outliers. When α is large, the method detects very few outliers. This result is due to the somewhat bell-shap of the resulting distribution when α is large. However, it is worth noting that the effect of the shape parameter stresses as n increases (note R when $2 < \alpha < 6$ and $n = 10$ as compared to the same α range when $n = 500$). In the case of the two-parameter Weibull distribution, there is a tendency for the Ueda's method to find less outliers as the sample size increases. Interestingly, overall, the Ueda's method has a high likelihood of not finding outliers (approximately $>$0.9 throughout) regardless of n, α and β. It is clear, though, that the Ueda's method is more likely to not find outliers when $3 < \alpha < 5$. Additionally, when the shape parameter in the Weibull distribution (irrespective of the scale parameter) leads to a distribution that takes a normal-like shape, the Ueda's method tends to find very few outliers.

As shown in the second row in Fig. 4, the Ueda's method tends to increase its chances of not finding outliers as the sample size increases, particularly when the parameters of the Beta distribution are the same. However, when the parameters take values whose difference is large (e.g. Beta$(.5, 5)$ and Beta$(5, .5)$), it is more likely that the Ueda's method will detect outliers as n increases. This is likely to occur since the Beta distribution can be positively or negatively skewed in such cases. It is known that a Beta distribution resembles the shape of a normal-distribution when both parameters are large (e.g. Beta$(5, 5)$), thus the Ueda's method seems to treat the resulting Beta distribution as a normal distribution in those cases. However, when both parameters are equally low (e.g. Beta$(.5, .5)$), which generates bimodal Beta distributions, the Ueda's method does not seem to detect many outliers. It is argued here that the Ueda's method tends not to detect many outliers in these bimodal cases since observations can only occur in a range between 0 and 1 and they spread out somewhat evenly around each mode and across the whole distribution (an exception to this case occurs when both shape parameters are close to 0; in that case observations group quite tightly around each mode, thus generating a large gap between the modes).

All in all, these results indicate that the more skewed and asymmetric the distribution, the more likely that the Ueda's method detects outliers and this effect is coupled with an increase in sample size. However, when the parameters of the distribution lead to an approximation of a symmetric bell-shaped distribution (e.g. see Gamma$(10, 10)$, Weibull$(4, 4)$, and Beta$(5, 5)$), the Ueda's method increases the likelihood of not finding outliers and, only in cases like these, an increment in sample size helps the Ueda's method to detect less outliers (see also Supplementary Tables 6.2).

4 Discussion

The main finding from the present simulation study is that the Ueda's method is more likely to detect outliers when a distribution becomes more skewed and asymmetric, and the sample size is increased. When a distribution becomes more bell-shaped-like or more symmetric (even if it is not bell-shaped), the Ueda's method is less likely to detect outliers and such likelihood is coupled with an increase in sample size. These results have implications to applied and statistical research.

Most data in different scientific fields tend to distribute in non-Gaussian forms (see Micceri 1989) and frequently in the form of skewed distributions. According to the results of the simulation study, the Ueda's method is very effective in detecting outliers when applied to negatively- and positively-skewed distributions. However, cases of bimodal (e.g., the waiting times between the start of successive eruptions, see Bowman 1990) and uniform data (the distribution of p-values in genetic association analysis of type 2 diabetes, see Harvard 2007) can occur. Thus, if a graphical assessment of the data indicates it is of bimodal or uniform shape, the usage of the Ueda's method is not recommended. In the case of bimodal distributions, graphical assessments can be complemented with formal tests such as the Gapping (Wainer and Schacht 1978) or the Dip (Hartigan and Hartigan 1985) tests. When dealing with uniform distributions, the test proposed by Cheng and Spiring (1987) can be used. As the simulation results suggest, the Ueda's method tackles the symmetry of these types of distributions and ignores they are not bell-shaped, thus making the method's sensitivity to outliers quite low (see Fig. 4 for the cases of bimodal distributions, e.g. Beta(.5, .5), and Fig. 3c for the case of uniform distributions when $\lambda = 1$).

Given that the Ueda's method focuses on the symmetry of the distribution, this outlier detection method can also serve as a symmetry test. The goal of any symmetry test is to test the null hypothesis that the distribution of a random variable is symmetric about the centre of the distribution (see Hollander 1981). In particular, tests that check for symmetry about the median are known to be powerful, widely used and easy to apply (Lunneborg 2005). One of those tests is the Cabilio and Masaro test which estimates symmetry using a simple computation that requires the sample size, the mean, the median and the standard deviation of the data set (Cabilio and Masaro 1996). Although this test is reportedly superior than other tests, it has low power against Cauchy and Uniform distributions (Lunneborg 2005). The results reported herein indicate that the Ueda's method regards uniform distributions and Cauchy$(0, \pi)$ as symmetric, particularly when the sample size is large (see Fig. 3c). Thus, although the Ueda's method is an outlier detection method, it does perform like a symmetry test in some cases. However, simulation studies are needed to compare the performance of this method against existent symmetry tests (Csörgo and Heathcote 1987; Randles et al. 1980; Yoshizawa 1984; Zheng and Gastwirth 2010) and thus determine whether the Ueda's method could also classify as a test of this sort.

It is traditional for studies on outliers to use contaminated normal distributions (see Wellmann and Gather 1999 and Jain 1981). However, it is important to bear in mind that using contaminated normal distributions in order to model and explain real data is an inadvisable practice and instead distributions with smooth changes in skewness are recommended (Gleason 1993). In the current study this recommendation was followed by using outlier-prone distributions whose shape was manipulated by stepwise control of

their parameters. Having said that, studying the Ueda's method when dealing with artificially contaminated distributions could help reveal how the method performs in regards to outliers, inliers and masking effects. For example, a recursive outlier detection method known as KUR tends to detect inliers as outliers in one-sided contaminated distributions (Jain 1981). The example shown in Fig. 2b could indicate that the Ueda's method might behave similar to the KUR method; however, this claim is merely speculative and needs to be investigated. Thus, a systematic comparison study of the Ueda's method to other available methods when dealing with contaminated distributions remains to be performed. In order to create smooth elongation in the tails of the distributions, it is here suggested that the contaminated normals be added in a stepwise fashion based on a measure of dispersion (e.g. adding contaminated normals from ± 1.5SD from the mean up to $\pm k$SD in steps of 0.5) and controlling for the proportion of contamination they add to the whole distribution (e.g. the contaminated normal could represent from 5 % up to 40 % of the whole distribution and this could be done in steps of 2.5 %).

5 Conclusions

The Ueda's method is an outlier detection method that focuses on the symmetry and skewness of the distribution in order to detect outlying data points. This method is highly sensitive to outliers when the distribution under analysis is negatively-, positively-skewed or asymmetric about the centre of the distribution and such sensitivity is enhanced by an increase in sample size. However, the sensitivity of the test to outliers decays as the distribution closely resembles a Gaussian shape or is highly symmetric around the centre of the distribution and such sensitivity is lessened by an increase in sample size. Although the Ueda's method models outliers using a density function of the normal distribution, a good fit to a normal distribution plays no role in the method not finding outliers as long as the distribution is either bell-shaped or symmetric. Thus, the Ueda's method should also be studied in the context of symmetry tests.

6 Appendix

6.1 Implementation of Ueda's method in R

This is the R code for the automatic detection of discordant outliers using Ueda's method. Its usage is illustrated by the simulated data in Example 1. It can be seen that the method successfully detects the outliers introduced.

```
## load code
 if(!require(devtools)) install.packages("devtools")

 url <- "http://bit.ly/1dLT9Ez" devtools:::source_url(url)

## example with simulated data
## we introduce five outliers
 set.seed(13)
 x <- rnorm(25, 300, 10) out
 <- rnorm(5, 400, 5) v <- c(x, out)

# detecting up to 10 outliers and plotting the results
 res <- ueda(v, smax = 10)

# data vector after removing outliers
```

```
 res

# show the outliers detected
 <- v[! v detected

# did we detect the actual outliers?
 all.equal(out, detected)
```

6.2 Supplementary Tables

Numerical results for Figs. 3 and 4 can be downloaded from https://www.dropbox.com/s/fvw65z6lsfy2hjh/suppmat.pdf?dl=0 or from the corresponding author.

Additional file

Additional file 1: Supplementary Material.

Acknowledgments
We thank two anonymous reviewers for their comments and suggestions and Rosie Gronthos for proofreading this manuscript. JIV was supported by the Eccles Scholarship in Medical Sciences, the Fenner Merit Scholarship, the Australian National University (ANU) High Degree Research Scholarship, and by grant R42100-2 from the John Curtin School of Medical Research, Canberra, ACT, Australia. JIV thanks the support of Dr. Mauricio Arcos-Burgos from ANU.

Author details
[1]Gösta Ekman Laboratory, Department of Psychology, Stockholm University, Stockholm, Sweden. [2]Arcos-Burgos Group, Department of Genome Sciences, John Curtin School of Medical Research, The Australian National University, Canberra, ACT, Australia. [3]Neuroscience Research Group, University of Antioquia, Medellín, Medellín, Colombia. [4]Department of Civil Engineering, Faculty of Engineering, University of Porto, Porto, Portugal.

Received: 18 May 2015 Accepted: 16 September 2015
Published online: 30 September 2015

References

Aggarwal, C: Outlier Analysis. Springer, New York (2013)
Azzalini, T, Bowman, AW: A look at some data on the old faithful geyser. J. R. Stat. Soc. Series C. **39**(3), 357–366 (1990)
Bertels, J, Kolinsky, R, Morais, J: Emotional valence of spoken words influences the spatial orienting of attention. Acta Psychol. **134**(3), 264–278 (2010)
Cabilio, P, Masaro, J: A simple test of symmetry about an unknown median. Can. J. Stat. **24**(3), 349–361 (1996)
Carling, K: Resistant outlier rules and the non-gaussian case. Comput. Stat. Data Anal. **33**(3), 249–258 (2000)
Chalabi, Y, Scott, DJ, Würtz, D: The Generalized Lambda Distribution as an Alternative to Model Financial Returns (2014). https://www.rmetrics.org/sites/default/files/glambda_0.pdf. Accessed 2014-07-21
Chandola, V, Banerjee, A, Kumar, V: Anomaly detection: a survey. ACM Comput. Surv. **41**(3), 15–58 (2009)
Cheng, SW, Spiring, FA: A test to identify the uniform distribution, with applications to probability plotting and other distributions. IEEE Trans. Reliability. **R-36**(1), 98–105 (1987)
Csörgo, S, Heathcote, CR: Testing for symmetry. Biometrika. **74**(1), 177–184 (1987)
D.G.I. of Broad Institute of Harvard, L. U. MIT, N.I. of BioMedical Research: Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science. **316**(5829), 1331–1336 (2007)
Ferrari, S, Cribari-Neto, F: Beta regression for modelling rates and proportions. J. Appl. Stat. **31**(7), 799–815 (2004)
Gleason, JR: Understanding elongation: the scale contaminated normal family. J. Am. Stat. Assoc. **88**, 327–337 (1993)
Hartigan, JA, Hartigan, PM: The dip test of unimodality. Ann. Stat. **13**(1), 70–84 (1985)
Haskett, JD, Pachepsky, YA, Acock, B: Use of the beta distribution for parameterizing variability of soil properties at the regional level for crop yield estimation. Agric. Syst. **48**(1), 73–86 (1995)
Havas, DA, Glenberg, AM, Rinck, M: Emotion simulation during language comprehension. Psychonomic Bull. Rev. **14**(3), 436–441 (2007)
Hayes, K, Kinsella, T: Spurious and non-spurious power in performance criteria for tests of discordancy. J. R. Stat. Soc. Series D. **52**(1), 69–82 (2003)
Hoaglin, D, Mosteller, F, Tukey, J: Understanding Robust and Exploratory Data Analysis. Wiley, New York (1983)
Hollander, M: Testing for Symmetry. FSU Statistics Report No. M599 (1981). FSU Statistics Report No. M599 http://stat.fsu.edu/techreports/M599.pdf
Jain, RB: Detecting outliers: power and some other considerations. Commun. Stat. - Theory Methods. **A10**, 2299–2314 (1981)
Kadota, K, Tominaga, D, Akiyama, Y, Takahashi, K: Detecting outlying samples in microarray data: a critical assessment of the effect of outliers on sample classification. Chem-Bio Inf. J. **3**(1), 30–45 (2003a)

Kadota, K, Nishimura, S, Bono, H, Nakamura, S, Hayashizaki, Y, Okazaki, Y, Takahashi, K: Detection of genes with tissue-specific expression patterns using akaike?s information criterion procedure. Physiol. Genomics. **12**, 251–259 (2003b)
Kitagawa, G: On the use of aic for the detection of outliers. Technometrics. **21**(2), 193–199 (1979)
Leiva, V, Tejo, M, Guiraud, P, Schmachtenberg, O, Orio, P, Marmolejo-Ramos, F: Modeling neural activity with cumulative damage distributions. Biological Cybernetics. **109**(4), 421–433 (2015)
Lunneborg, CE: Encyclopedia of Statistics in Behavioral Science. In: Everitt, BS, Howell, DC (eds.) John Wiley & Sons, Ltd, Chichester, (2005)
Marmolejo-Ramos, F, Cousineau, D, Benites, L, Maehara, R: On the efficacy of procedures to normalise ex-gaussian distributions. Fronti. Psychol. **5**(1548), 10–3389201401548 (2015)
Micceri, T: The unicorn, the normal curve, and other improbable creatures. Psychol. Bull. **105**(1), 156–166 (1989)
Otte, E, Habel, U, Schulte-Rüter, M, Konrad, K, Koch, I: Interference in simultaneously perceiving and producing facial expressions ? evidence from electromyography. Neuropsychologia. **49**(1), 124–130 (2011)
Öztürk, T, Dale, RF: A study of fitting the generalized lambda distribution to solar radiation data. J. Appl. Meteorol. **21**(7), 995–1004 (1982)
Praetz, PD: The distribution of share price changes. J. Bus. **45**(1), 49–55 (1972)
R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2013). R Foundation for Statistical Computing. http://www.R-project.org/
Randles, R, Fligner, MA, Policello, GE, Wolfe, DA: An asymptotically distribution-free test for symmetry versus asymmetry. J. Am. Stat. Assoc. **75**(369), 168–172 (1980)
Salazar, JC, Baena, A: Análisis y diseño de experimentos aplicados a estudios de simulación. Dyna. **76**(159), 249–257 (2009)
Songwon, S: A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets (2006). MSc Dissertation, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
Thode, H: Testing for Normality. Marcel Dekker, Inc., New York (2002)
Tsuyuzaki, K, Tominaga, D, Kwon, Y, Miyazaki, S: Two-way aic: detection of differentially expressed genes from large scale microarray meta-dataset. BMC Genomics. **14**(Suppl 2:S9) (2013). doi:10.1186/1471-2164-14-S2-S9
Ueda, T: A simple method for the detection of outliers. Electronic J. Appl. Stat. Anal. **2**(1), 67–76 (1996/2009)
Van Selst, M, Jolicoeur, P: A solution to the effect of sample size on outlier elimination. Q. J. Exp. Psychol. **47A**(3), 631–650 (1994)
Vélez, JI, Correa, JC: Should we think of a different Median estimator? Revista Comunicaciones en Estadística. **7**(2), 1–8 (2014)
Verma, SP: Sixteen statistical tests for outlier detection and rejection in evaluation of international geochemical reference materials: example of microgabbro pm-s. Geostandards Newsl. **21**(1), 59–75 (1997)
Verma, SP, González, LD, Rosales-Rivera, M, Quiroz-Ruiz, A: Comparative performance of four single extreme outlier discordancy tests from monte carlo simulations. Sci. World J. **Article ID 746451**, 1–27 (2014). doi:10.1155/2014/746451
Wainer, T, Schacht, S: Gapping. Psychometrika. **43**(2), 203–2012 (1978)
Wellmann, J, Gather, U: A note on contamination models and outliers. Commun. Stat. — Theory Methods. **28**(8), 1793–1802 (1999)
Yoshizawa, CN: Some Tests of Symmetry. PhD Dissertation, Department of Biostatistics, School of Public Health, University of North Carolina at Chapel Hill, North Carolina, USA (1984)
Zheng, T, Gastwirth, JL: On bootstrap tests of symmetry about an unknown median. J. Data Sci. **8**(3), 413–427 (2010)