Adv Data Anal Classif (2018) 12:785–822
https://doi.org/10.1007/s11634-017-0305-y

REGULAR ARTICLE

Outlier detection in interval data

A. Pedro Duarte Silva[1] · **Peter Filzmoser**[2] · **Paula Brito**[3]

Received: 8 October 2016 / Revised: 7 November 2017 / Accepted: 23 November 2017 /
Published online: 15 December 2017

Abstract A multivariate outlier detection method for interval data is proposed that makes use of a parametric approach to model the interval data. The trimmed maximum likelihood principle is adapted in order to robustly estimate the model parameters. A simulation study demonstrates the usefulness of the robust estimates for outlier detection, and new diagnostic plots allow gaining deeper insight into the structure of real world interval data.

Keywords Outliers · Robust statistics · Interval data · Mahalanobis distance

Mathematics Subject Classification 62-07 (Data Analysis) · 62F35 (Robustness and adaptive procedures) · 62H86 (Multivariate analysis and fuzziness)

1 Introduction

It is often the case that multivariate datasets include atypical data points, i.e. points that deviate from the main pattern. Such data units are usually called *outliers*. Outlier detection is important for two main reasons: on the one hand, outlying data points may be interesting on their own, since they can reveal nonconforming phenomena; on the other hand, the results of usual multivariate methods can be heavily influenced by outliers. Outlierdetection may however be a tricky endeavor as some (false) outliers

✉ A. Pedro Duarte Silva
psilva@porto.ucp.pt

[1] Católica Porto Business School, & CEGE, Universidade Catolica Portuguesa, Rua Diogo Botelho, Porto, Portugal

[2] Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria

[3] Faculdade de Economia & LIAAD-INESC TEC, Universidade do Porto, Porto, Portugal

are often reported, even for data sets with none if no distinction is made between outliers and distribution extremes (Filzmoser et al. 2005). Moreover, true outliers may not stand out as expected since they may affect estimates of location and scatter so much that they no longer look atypical–this is known as "masking effect" in the literature (Hubert et al. 2008). Also, usual rules based on an appropriate cut-off value for robustified Mahalanobis distances [see, for instance, Filzmoser et al. (2005) and Rousseeuw and Zomeren (1990)] suffer from the drawback that they are independent from the sample size.

In this paper we address the problem of outlier detection in multivariate interval data, i.e., data where for each data unit and variable an interval rather than a single real value is recorded. Interval data may occur in different situations, e.g. when describing ranges of variable values like daily stock prices or temperature ranges. Another common and increasingly interesting source of interval data is the aggregation of large data bases, when interval observations result from aggregating the real values describing the individual units.

A common approach for multivariate outlier detection measures outlyingness by Mahalanobis distances. A point i is considered an outlier if its distance $D^2_{m,C}(i)$ from an appropriate estimate of the multivariate mean m is above a threshold, where $D^2_{m,C}(i) = (x_i - m)^t C^{-1}(x_i - m)$ and C is an estimate of the covariance matrix. Under the assumption of d-variate normality, $D^2_{m,C}(i)$ follows approximately a Chi-square distribution with d degrees of freedom. Then the threshold for outlier labeling may be an upper quantile of χ^2_d, e.g., the 97.5% quantile.

However, if m and C are chosen to be the classical sample mean vector and covariance matrix this procedure is not reliable, as $D^2_{m,C}(i)$ may be strongly affected by atypical observations. Therefore, the Mahalanobis distances should be computed with robust estimates of location and scatter. Rousseeuw and Zomeren (1990) use these robust Mahalanobis distances (RD's) for multivariate outlier detection. Many robust estimators for location and covariance have been proposed in the literature.

The minimum covariance determinant (MCD) estimator (Rousseeuw 1984, 1985) uses a subset of the original sample, consisting of the h points in the dataset for which the determinant of the sample covariance matrix is minimal; this is very frequently used in practice since a computationally fast algorithm is available (Rousseeuw and Driessen 1999). Since the Chi-square approximation to the distribution of Mahalanobis distances based on the MCD estimator may not work well, even for moderately large samples, finite samples approximations have been proposed (Cerioli 2010).

Trimmed likelihood estimators (Hadi and Luceño 1997) are also based on a sample subset. In this case, the subset of h points is obtained by maximizing the trimmed likelihood, keeping the multivariate observations that contribute most to the likelihood function. For multivariate Gaussian data, the two approaches, the MCD method and the trimmed likelihood method, lead to the same estimators of covariance (Hadi and Luceño 1997).

In either case, the proportion of data points to be used needs to be specified *a priori*. The choice $h \approx n/2$, where n is the sample size, leads to the highest breakdown point but to low efficiency, while a larger value for h reduces the breakdown point and increases the efficiency; a trade-off is the choice $h \approx 0.75 \cdot n$ (Hubert et al. 2008).

While in classical statistics and multivariate data analysis the statistical units under analysis are single individuals, each of which taking a single value for each (numerical or categorical) variable, it often happens that the data under analysis do not correspond to single observations, but rather to sets of values, either related to groups of units gathered on the basis of some common properties, or observed repeatedly over time. The classical data-array model is then somehow restricted to take into account the variability inherent to such data. In those situations, data is commonly summarized by central statistics, e.g., averages, medians or modes, and other relevant information may be disregarded. This is the case when we are facing huge sets of data, recorded in very large databases (often called nowadays "Big Data"), and elements of interest are not the individual records but rather some second-level entities. For instance, in a database of individual phone calls, we are surely more interested in describing the behavior of some person (or some pre-defined class or group of persons) rather than each call by itself. The analysis requires then that the calls data for each person (or group) be somehow aggregated to obtain the information of interest, however data can no longer be properly described by the usual numerical and categorical variables without an unacceptable loss of information.

Defining appropriate variable types, which may assume new forms of realizations-multiple, possibly weighted, values for each case - that take into account the data intrinsic variability, Symbolic Data Analysis (Billard and Diday 2003; Bock and Diday 2000; Brito 2014; Diday and Noirhomme-Fraiture 2008) provides a framework where the variability observed is explicitly considered in the data representation.

We focus here on interval-valued data, i.e., where for each entity under analysis an interval is observed. Many methods have to this day been developed for the analysis of interval-valued data, ranging from univariate statistics to multivariate methods such as Clustering [see, e.g., De Carvalho et al. (2006) and De Carvalho and Lechevallier (2009)], Principal Component Analysis [see, e.g. Douzal-Chouakria et al. (2011) and Le-Rademacher and Billard (2012)], Discriminant Analysis (Duarte Silva and Brito 2015; Ramos-Guajardo and Grzegorzewski 2016), Regression Analysis (Dias and Brito 2017; Lima Neto and De Carvalho 2008, 2010; Lima Neto et al. 2011), etc. For a survey the reader may refer to Brito (2014) and Noirhomme-Fraiture and Brito (2011). Most methodologies rely however on non-parametric exploratory approaches, nevertheless recent approaches based on parametric models have been proposed-see Brito and Duarte Silva (2012), Le-Rademacher and Billard (2011) and Lima Neto et al. (2011).

In this paper we address the problem of outlier detection in interval data, using the modeling proposed in Brito and Duarte Silva (2012). In the particular case of interval data, outlyingness may be caused by different reasons. At the univariate level, an observation may be considered as an outlier due to its MidPoint, to its Range or to both, or still to a particular relation between them resulting in an outlying interval. Furthermore, from a multivariate perspective, it may be important to distinguish outliers that stand out by their MidPoints, by their Ranges, or both, or by the global relation between all MidPoints and Ranges.

In Li et al. (2006) the authors present an algorithm for outlier detection in interval data, using a distance-based approach. However, their method fixes *a priori* the number of data points to be flagged as outliers, and seems therefore unable to distinguish between outlier-free datasets and contaminated ones. Viattchenin (2012) proposes a heuristic approach based on possibilistic clustering. This approach relies on an exhaustive search over all possible partitions of the given dataset in k clusters, for each fixed k, and therefore is only feasible for small datasets. To the best of our knowledge, the methodology we present here is the first statistical approach to this problem, rooted on classical robustness theory.

The remainder of the paper is organized as follows. Section 2 briefly introduces interval data, and parametric models for interval-valued variables. In Sect. 3 we address the problem of robust parameter estimation and its application to robust outlier detection in interval data. Section 4 presents a simulation study comparing alternative methods for outlier detection. An application is discussed in Sect. 5, where we illustrate the main issues arising in this context. Section 6 concludes the paper putting in evidence its main contributions.

2 Models for interval data

Let $S = \{s_1, \ldots, s_n\}$ be the set of n entities under analysis.

We are in the presence of interval data when for each $s_i \in S$ an interval of $\mathbb{R}$ is recorded for each variable. Formally, an interval-valued variable [see Noirhomme-Fraiture and Brito (2011)] is defined by an application

$$Y : S \to T \text{ such that}$$
$$s_i \to Y(s_i) = [l_i, u_i]$$

where T is the set of intervals of an underlying set $O \subseteq \mathbb{R}$. Let $I = [I_{ij}]$ be an $n \times p$ matrix containing the values of p interval variables on S. Each $s_i \in S$ may then be represented by a p-dimensional vector of intervals, $I_i = (I_{i1}, \ldots, I_{ip})$, $i = 1, \ldots, n$, with $I_{ij} = [l_{ij}, u_{ij}]$, $j = 1, \ldots, p$ (see Table 1).

The value of each interval-valued variable Y_j for each $s_i \in S$ is usually defined by the lower and upper bounds l_{ij} and u_{ij} of $I_{ij} = Y_j(s_i)$; alternatively $Y_j(s_i)$ may also be represented by the MidPoint $c_{ij} = \dfrac{l_{ij} + u_{ij}}{2}$ and Range $r_{ij} = u_{ij} - l_{ij}$ of I_{ij}.

2.1 Parametric models for interval data

In Brito and Duarte Silva (2012), parametric models for interval data, relying on Multivariate Normal or Skew-Normal distributions for the MidPoints and Log-Ranges

Table 1 Matrix I of interval data

	Y_1	…	Y_j	…	Y_p
s_1	$[l_{11}, u_{11}]$	…	$[l_{1j}, u_{1j}]$	…	$[l_{1p}, u_{1p}]$
…	…		…		…
s_i	$[l_{i1}, u_{i1}]$	…	$[l_{ij}, u_{ij}]$	…	$[l_{ip}, u_{ip}]$
…	…		…		…
s_n	$[l_{n1}, u_{n1}]$	…	$[l_{nj}, u_{nj}]$	…	$[l_{np}, u_{np}]$

Table 2 Matrix of MidPoints and Log-Ranges

	Y_1^C	$\dots$	Y_j^C	$\dots$	Y_p^C	$Y_1^{R^*}$	$\dots$	$Y_j^{R^*}$	$\dots$	$Y_p^{R^*}$
s_1	c_{11}	$\dots$	c_{1j}	$\dots$	c_{1p}	r^*_{11}	$\dots$	r^*_{1j}	$\dots$	r^*_{1p}
$\dots$	$\dots$		$\dots$		$\dots$	$\dots$		$\dots$		$\dots$
s_i	c_{i1}	$\dots$	c_{ij}	$\dots$	c_{ip}	r^*_{i1}	$\dots$	r^*_{ij}	$\dots$	r^*_{ip}
$\dots$	$\dots$		$\dots$		$\dots$	$\dots$		$\dots$		$\dots$
s_n	c_{n1}	$\dots$	c_{nj}	$\dots$	c_{np}	r^*_{n1}	$\dots$	r^*_{nj}	$\dots$	r^*_{np}

of the interval-valued variables have been proposed. The Gaussian model has the advantage of allowing for the application of classical inference methods, and is the model considered in this paper.

The model consists in assuming a joint multivariate Normal distribution for the MidPoints C and the logs of the Ranges R, $R^* = ln(R)$, with $\mu = \left[\mu_C^t \ \mu_{R^*}^t\right]^t$ and $\Sigma = \begin{pmatrix} \Sigma_{CC} & \Sigma_{CR^*} \\ \Sigma_{R^*C} & \Sigma_{R^*R^*} \end{pmatrix}$ where μ_C and μ_{R^*} are p-dimensional column vectors of the mean values of, respectively, the MidPoints and Log-Ranges, and Σ_{CC}, Σ_{CR^*}, Σ_{R^*C} and $\Sigma_{R^*R^*}$ are $p \times p$ matrices with their variances and covariances.

With this parametrization, the data in matrix I in Table 1 is equivalently represented by the matrix shown in Table 2:

We remark that the MidPoint c_{ij} and the Range r_{ij} of the value of an interval-valued variable $I_{ij} = Y_j(s_i)$ relate to the same variable, so that the link that might exist between them should be appropriately taken into account. This is done by considering particular configurations of the global variance-covariance matrix (Table 3). The following cases are of particular interest, and have been addressed:

1. Non-restricted case: allowing for non-zero correlations among all MidPoints and Log-Ranges;
2. Interval-valued variables Y_j are uncorrelated, but for each variable, the MidPoint may be correlated with its Log-Range;
3. MidPoints (Log-Ranges) of different variables may be correlated, but no correlation between MidPoints and Log-Ranges is allowed;
4. All MidPoints and Log-Ranges are uncorrelated, both among themselves and between each other.

Table 3 summarizes the different cases considered in this paper.

Table 3 Different cases for the variance-covariance matrix

Case	Characterization	Σ
C1	Non-restricted	Non-restricted
C2	Y_j's non correlated	Σ_{CC}, $\Sigma_{CR^*} = \Sigma_{R^*C}$, $\Sigma_{R^*R^*}$ all diagonal
C3	C's non-correlated with R^*'s	$\Sigma_{CR^*} = \Sigma_{R^*C} = 0$
C4	All C's and R^*'s are non-correlated	Σ diagonal

It should be remarked that in Cases C2, C3 and C4, Σ can be written as a block diagonal matrix, after a possible rearrangement of rows and columns; therefore maximum likelihood estimates under these cases can be obtained directly from the classical non-restricted estimates. Testing for the different models/configurations can be done in a straightforward manner, using the likelihood-ratio approach.

This modeling has been implemented in the R-package MAINT. Data (Duarte Silva and Brito 2017), available on CRAN. MAINT.Data introduces a data class for representing interval data and includes functions for modeling and analysing these data. In particular, maximum likelihood estimation and statistical tests for the considered configurations are addressed. Methods for (M)ANOVA and Discriminant Analysis of this data class are also provided.

3 Robust parameter estimation and outlier detection

The identification of outliers is based on robust Mahalanobis distances from each data point to the mean. These values are then compared with the 97.5% quantile of an appropriate distribution. Traditionally, the χ^2_{2p} distribution (where p is the number of interval variables and $2p$ the total number of MidPoints and Log-Ranges) is used. However, and since the Chi-square approximation to the distribution of Mahalanobis distances based on the MCD estimator may not work well, even for moderately large samples, finite samples approximations have been proposed (Cerioli 2010). This will be detailed below in Sect. 3.2.

3.1 Trimmed maximum likelihood estimation

The Trimmed Log-Likelihood (lnTL) estimator is a special case of the weighted trimmed likelihood estimator, defined in Hadi and Luceño (1997) and Vandev and Neykov (1998). The lnTL estimator is defined as

$$\hat{\theta}_{lnTL} := \arg \min_{H;\theta\in\Theta^q} \sum_{i\in H}(-\log\varphi(y_i;\theta)), \tag{1}$$

for an index set $H \subset \{1,\ldots,n\}$ of size h to be determined, with the unknown parameter $\theta \in \Theta^q \subset \mathbb{R}^q$, and where $y_i \in \mathbb{R}^d$ for $i = 1,\ldots,n$ are i.i.d. observations with probability density $\varphi(y;\theta)$. Here, h is the trimming parameter, and with the help of trimming it is possible to remove those $n-h$ observations whose values would be highly unlikely to occur if the fitted model was true.

Problem (1) is infeasible to solve for larger data sets, with hundreds of observations or even more. However, Neykov and Müller (2003) proposed a fast algorithm to find an approximative solution.

3.2 Robust model estimation for interval data

The trimmed Maximum Likelihood principle can be readily adapted to the problem of robust parameter estimation for the probabilistic models proposed in Brito and Duarte Silva (2012). In particular, in the case of the Gaussian models to be discussed in this paper, the diagonal by blocks structure of the restricted covariance matrix always implies that trimmed likelihood maximization is equivalent to the minimization of the determinant of the restricted trimmed sample covariance matrix.

That can be seen by noting that for all configurations considered the trimmed log-likelihood

$$ln\ TL(\mu, \Sigma) = -\frac{h}{2}\left(2p\ ln(2\pi) + ln\,|\Sigma| + tr\,\tilde{\Sigma}\Sigma^{-1} + (\tilde{\mu} - \mu)^t\ \Sigma^{-1}\ (\tilde{\mu} - \mu)\right) \tag{2}$$

where h is the number of observations kept in the trimmed sample, and $\tilde{\mu} = \frac{1}{h}\sum_{i=1}^{h} X_i$, $\tilde{\Sigma} = \frac{1}{h}\sum_{i=1}^{h}(X_i - \tilde{\mu})(X_i - \tilde{\mu})^t$ are respectively the trimmed mean and trimmed sample covariance.

It follows [see Brito and Duarte Silva (2012)] that when Σ is block diagonal with blocks $\Sigma_1, \Sigma_2, \ldots, \Sigma_b$, $ln\ TL$ is maximized by $\hat{\mu} = \tilde{\mu}$ and

$$\hat{\Sigma} = \begin{pmatrix} \tilde{\Sigma}_1 & & \\ & \tilde{\Sigma}_2 & 0 \\ 0 & & \ldots \\ & & & \tilde{\Sigma}_b \end{pmatrix}$$

with $\tilde{\Sigma}_j$ being the block of $\tilde{\Sigma}$ corresponding to Σ_j. The maximal value of $ln\ TL$ reduces to $-\frac{h}{2}\left(2p\ ln(2\pi) + ln|\hat{\Sigma}| + 2p\right)$ so that the trimmed maximum likelihood and minimum trimmed covariance determinant principles remain equivalent in this setting.

A consequence of this equivalence is that known refinements of the traditional Minimal Covariance Determinant Estimator can be readily adapted. In particular, following Hubert et al. (2008), we implemented a one-step re-weighted bias-corrected estimate given by

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} w_i x_i}{m} \tag{3}$$

$$\hat{\Sigma}^1 = \frac{l_{m,2p}\ c^1_{m,h,n,2p,Cf} \sum_{i=1}^{n} w_i (x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)^t}{m} \tag{4}$$

$$m = \sum_{i=1}^{n} w_i \qquad w_i = \begin{cases} 1, \text{ if } & l_{h,2p}\ c_{h,n,2p,Cf}\ D_{\hat{\mu},\hat{\Sigma}}(i) \leq \sqrt{Q_{0.975}} \\ 0, & otherwise \end{cases}$$

where $Q_{0.975}$ is the 97.5% quantile of the $D^2_{\hat{\mu},\hat{\Sigma}}$ distribution. The traditional approach consists in using the Chi-square approximation, so that $Q_{0.975} = \chi^2_{2p,0.975}$, but Hardin and Rocke (2005) proposed using an F distribution with better finite sample properties. In expression (4), $l_{\alpha,2p} = \frac{\alpha/n}{P(\chi^2_{2p+2} \leq \chi^2_{2p;\alpha/n})}(\alpha = m; h)$ are consistency correction factors, $Cf = \{C1, C2, C3, C4\}$ are the alternative configurations of the covariance matrix, and $c_{h,n,2p,Cf}$, $c^1_{m,h,n,2p,Cf}$, are raw ($c_{h,n,2p,Cf}$) and one-step re-weighted

Table 4 Raw finite sample correction factors: $c_{h,n,2p,Cf}$

n	h/n	$p=4$				$p=10$			
		C1	C2	C3	C4	C1	C2	C3	C4
50	0.50	1.2640	1.1574	1.1597	1.1137	1.4446	1.1217	1.2329	1.0917
	0.75	1.1473	1.0918	1.0947	1.0687	1.2732	1.0796	1.1517	1.0611
200	0.50	1.0657	1.0423	1.0426	1.0314	1.1031	1.0356	1.0628	1.0275
	0.75	1.0361	1.0237	1.0236	1.0176	1.0631	1.0216	1.0385	1.0168

Table 5 Re-weighted finite sample correction factors: $c^1_{m,[0.75\,n],n,2p,Cf}$

n	m/n	$p=4$				$p=10$			
		C1	C2	C3	C4	C1	C2	C3	C4
50	0.750	1.1655	1.1019	1.1258	1.1196	1.2301	1.1154	1.1531	1.1341
	0.975	1.0462	1.0488	1.0419	1.0472	1.0979	1.0610	1.0653	1.0598
200	0.750	1.1214	1.0598	1.0899	1.0747	1.1316	1.0649	1.0942	1.0790
	0.975	1.0105	1.0106	1.0110	1.0078	1.0188	1.0153	1.0147	1.0115

($c^1_{m,h,n,2p,Cf}$) finite-sample bias-correction factors, found by a simulation and interpolation procedure, along the lines proposed in Pison et al. (2002), and described in the Appendix.

Illustrative values of finite sample correction factors are given in Table 4 (raw correction factors) and Table 5 (re-weighted correction factors).

The final rule for outlier identification compares the robust Mahalanobis distances, based on covariance estimates obtained as discussed above, with the 97.5% quantile of either the χ^2_{2p} distribution or using the approximations [see Cerioli (2010)]:

$$D^2_{\hat{\mu}_1,\hat{\Sigma}_1} \sim \frac{(m-1)^2}{m} Beta\left(p, \frac{m-2p-1}{2}\right), \quad if \ \ w_i = 1 \tag{5}$$

$$D^2_{\hat{\mu}_1,\hat{\Sigma}_1} \sim \frac{m+1}{m}\frac{(m-1)2p}{m-2p} F\left(2p, m-2p\right), \quad if \ \ w_i = 0 \tag{6}$$

3.3 Choice of the trimming parameter

The trimmed maximum likelihood estimation procedure, described in the previous section, relies on the choice of the value of the trimming parameter h. There is no consensus on how to choose this parameter. One has to consider the trade-off between the potential outlier influence and the efficiency of the resulting estimators. One extreme consists in maximizing the breakdown point, which leads to roughly half the sample. The other extreme is the maximum likelihood estimation, which maximizes efficiency but is also the most sensitive to outliers. Some authors (Hubert et al. 2008) suggest using 75% of the sample, which they argue, usually results in a reasonable compromise. Ideally, one should use the true number of regular observations, this is however unknown in practice.

These considerations lead us to proposing a two-step approach. In the first step, the outlier detection procedure is run to get an estimate of the outlier proportion. The second step repeats the procedure fixing the trimming parameter at the obtained value.

To assess the good founding of this proposal, a small simulation study has been conducted, controlling for the sample size, the outlier proportion, the contamination level and the data covariance structure. This simulation considered the following alternatives: maximum likelihood, maximization of the breakdown point, setting the trimming parameter to 75% of the sample, the two-step approach with 50% and 75% in the first step, and the trimming parameter chosen by the Bayesian Information Criterion (BIC). The performance measure used to compare these methods is the *F-measure*, defined as the harmonic mean of Precision and Recall [see, for instance, Rijsbergen (1979)], where Precision is the fraction of identified data points that are indeed outliers, while Recall is the fraction of true outliers that are identified as outliers by the method. This measure is computed as:

$$F = \frac{2 \times \text{true positives}}{2 \times \text{true positives} + \text{false positives} + \text{false negatives}} \tag{7}$$

The higher the F-measure the better the method's performance.

The results showed that depending on the data condition either the maximum likelihood, the method that maximizes the breakdown point, or the two step procedure with 75% in the first step, provide the best results. These three methods will be further compared in a more extensive simulation study to be described in Sect. 4.

Results of this preliminary study are available from the authors upon request.

Furthermore, we found that these procedures scale well and we were able to apply them to problems with up to $n = 1000$ observations and $p = 50$ interval-valued variables in less than 10 minutes of computer time.

4 Simulation studies

To better understand the factors affecting the relative performance of the methods under comparison, we performed a controlled simulation experiment.

4.1 Experimental design

We considered a factorial design with the following six factors:

- Number of interval variables (**NV** - 2 levels): $p = 4$ and $p = 10$.
- Sample size (**SS** - 3 levels): Total number of sample observations, set at $n = 50, n = 100, n = 200$.
- Data Generating Process (**DGP** - 2 levels): MidPoints and Log-Ranges jointly Normally distributed or generated from linear transformations of Gaussian and Uniform variables.

Table 6 Methods with the largest F-measure, for the considered data settings–$p = 4$

	Gaussian data			Gaussian and Uniform data		
	Noise = 0.05	Noise = 0.1	Noise = 0.2	Noise = 0.05	Noise = 0.1	Noise = 0.2
n	Contamination Level = 3; Case = C1					
50	TStp75n	0.50n	0.50n	TStp75n	0.50n	0.50n
100	0.50n	0.50n	0.50n	TStp75n	0.50n	0.50n
200	0.50n	0.50n	0.50n	0.50n	0.50n	0.50n
n	Contamination Level = 3; Case = C2					
50	Fs0.50n	0.50n	0.50n	0.50n	0.50n	0.50n
100	0.50n	0.50n	0.50n	0.50n	0.50n	0.50n
200	0.50n	0.50n	0.50n	0.50n	0.50n	0.50n
n	Contamination Level = 3; Case = C3					
50	TStp75n	0.50n	0.50n	TStp75n	0.50n	0.50n
100	0.50n	0.50n	0.50n	TStp75n	0.50n	0.50n
200	0.50n	0.50n	0.50n	TStp75n	0.50n	0.50n
n	Contamination Level = 3; Case = C4					
50	Fs0.50n	0.50n	0.50n	Fs0.50n	Fs0.50n	0.50n
100	TStp75n	0.50n	0.50n	Fs0.50n	0.50n	0.50n
200	0.50n	0.50n	0.50n	Fs0.50n	0.50n	0.50n
n	Contamination Level = 6; Case = C1					
50	FsTStp75n	TStp75n	0.50n	FsTStp75n	TStp75n	0.50n
100	FsTStp75n	FsTStp75n	0.50n	FsTStp75n	TStp75n	0.50n
200	FsTStp75n	FsTStp75n	0.50n	FsTStp75n	FsTStp75n	0.50n
n	Contamination Level = 6; Case = C2					
50	Alln(MLE)	TStp75n	0.50n	Alln(MLE)	TStp75n	0.50n
100	Alln(MLE)	FsTStp75n	0.50n	Alln(MLE)	FsTStp75n	0.50n
200	FsTStp75n	FsTStp75n	0.50n	FsTStp75n	Fs0.50n	0.50n
n	Contamination Level = 6; Case = C3					
50	FsTStp75n	TStp75n	0.50n	FsTStp75n	TStp75n	0.50n
100	FsTStp75n	FsTStp75n	0.50n	FsTStp75n	FsTStp75n	0.50n
200	FsTStp75n	FsTStp75n	0.50n	FsTStp75n	FsTStp75n	0.50n
n	Contamination Level = 6; Case = C4					
50	Alln(MLE)	TStp75n	0.50n	Alln(MLE)	TStp75n	0.50n
100	Alln(MLE)	Fs0.50n	TStp75n	FsTStp75n	Fs0.50n	0.50n
200	FsTStp75n	FsTStp75n	0.50n	FsTStp75n	Fs0.50n	0.50n
n	Contamination Level = 12; Case = C1					
50	Alln(MLE)	FsTStp75n	Fs0.50n	Alln(MLE)	FsTStp75n	Fs0.50n
100	FsTStp75n	FsTStp75n	Fs0.50n	FsTStp75n	FsTStp75n	FsTStp75n
200	FsTStp75n	FsTStp75n	FsTStp75n	FsTStp75n	FsTStp75n	FsTStp75n

Table 6 continued

	Gaussian data			Gaussian and Uniform data		
	Noise = 0.05	Noise = 0.1	Noise = 0.2	Noise = 0.05	Noise = 0.1	Noise = 0.2
n			Contamination Level = 12; Case = C2			
50	Alln(MLE)	FsTStp75n	TStp75n	Alln(MLE)	FsTStp75n	TStp75n
100	FsTStp75n	FsTStp75n	TStp75n	FsTStp75n	FsTStp75n	TStp75n
200	FsTStp75n	FsTStp75n	Fs0.50n	FsTStp75n	FsTStp75n	Fs0.50n
n			Contamination Level = 12; Case = C3			
50	Alln(MLE)	FsTStp75n	TStp75n	Alln(MLE)	FsTStp75n	TStp75n
100	FsTStp75n	FsTStp75n	FsTStp75n	FsTStp75n	FsTStp75n	TStp75n
200	FsTStp75n	FsTStp75n	FsTStp75n	FsTStp75n	FsTStp75n	FsTStp75n
n			Contamination Level = 12; Case = C4			
50	Alln(MLE)	FsTStp75n	TStp75n	Alln(MLE)	FsTStp75n	TStp75n
100	FsTStp75n	FsTStp75n	TStp75n	FsTStp75n	FsTStp75n	TStp75n
200	FsTStp75n	FsTStp75n	Fs0.50n	FsTStp75n	FsTStp75n	Fs0.50n

- True configuration (**TConf** - 4 levels): Case of true covariance of MidPoints and Log-Ranges. Set at the levels C1 (unrestricted), C2 (Uncorrelated Interval Variables), C3 (MidPoints uncorrelated with Log-Ranges) and C4 (all MidPoints and Log-Ranges uncorrelated with each other).
- Noise level (**NL** - 4 levels): Percentage of true outliers, set at $Noise = 0.0\%$, 5.0%, 10.0%, 20.0%.
- Contamination level (**CL** - 3 levels): Mahalanobis distance between the mean vectors ($2p$-dimensional) of the regular and the outlying observations, set at 3.0, 6.0, 12.0.
- Method (**M** - 6 levels): Maximum likelihood (Alln (MLE)), maximum trimmed likelihood with 50% trimming, with quantiles from the Chi-square (0.50n) or from the F or Beta distributions (Fs0.50n), two-step maximum trimmed likelihood with 75% trimming in the first step, with quantiles from the Chi-square (TStp75n) or from the F or Beta distributions (FsTStp75n), the distance-based approach by Li et al. (2006), with four neighbors and the true number of outliers (LLLk4).

For each data condition with true outliers, defined by a combination of factors **NV**, **SS**, **DGP**, **TConf**, **NL** and **CL** we generated 1000 independent samples, and applied the six outlier detection methods. In each case, we computed the Precision, Recall and F-measure (see (7)), as defined in Sect. 3.3. For the conditions with no outliers, the distance-based approach by Li et al. (2006) was not applied; for all other methods we computed the proportion of observations incorrectly flagged as outliers.

Table 7 Methods with the largest F-measure, for the considered data settings–$p = 10$

	Gaussian data			Gaussian and Uniform data		
	Noise = 0.05	Noise = 0.1	Noise = 0.2	Noise = 0.05	Noise = 0.1	Noise = 0.2
n			Contamination Level = 3; Case = C1			
50	TStp75n	0.50n	0.50n	TStp75n	0.50n	0.50n
100	TStp75n	0.50n	0.50n	TStp75n	0.50n	0.50n
200	0.50n	0.50n	0.50n	0.50n	0.50n	0.50n
n			Contamination Level = 3; Case = C2			
50	Fs0.50n	Fs0.50n	Fs0.50n	FsTStp75n	Fs0.50n	Fs0.50n
100	Fs0.50n	0.50n	0.50n	Fs0.50n	Fs0.50n	Fs0.50n
200	0.50n	0.50n	0.50n	Fs0.50n	0.50n	0.50n
n			Contamination Level = 3; Case = C3			
50	TStp75n	0.50n	0.50n	TStp75n	0.50n	0.50n
100	0.50n	0.50n	0.50n	0.50n	0.50n	0.50n
200	0.50n	0.50n	0.50n	0.50n	0.50n	0.50n
n			Contamination Level = 3; Case = C4			
50	Fs0.50n	Fs0.50n	Fs0.50n	FsTStp75n	Fs0.50n	Fs0.50n
100	Fs0.50n	Fs0.50n	Fs0.50n	Fs0.50n	Fs0.50n	Fs0.50n
200	Fs0.50n	Fs0.50n	Fs0.50n	Fs0.50n	Fs0.50n	Fs0.50n
n			Contamination Level = 6; Case = C1			
50	TStp75n	TStp75n	0.50n	TStp75n	TStp75n	0.50n
100	Fs0.50n	TStp75n	0.50n	TStp75n	TStp75n	0.50n
200	Fs0.50n	0.50n	0.50n	Fs0.50n	0.50n	0.50n
n			Contamination Level = 6; Case = C2			
50	Alln(MLE)	TStp75n	0.50n	TStp75n	0.50n	0.50n
100	TStp75n	TStp75n	0.50n	TStp75n	0.50n	0.50n
200	TStp75n	TStp75n	0.50n	TStp75n	0.50n	0.50n
n			Contamination Level = 6; Case = C3			
50	Fs0.50n	TStp75n	0.50n	Fs0.50n	TStp75n	0.50n
100	TStp75n	TStp75n	0.50n	TStp75n	TStp75n	0.50n
200	FsTStp75n	0.50n	0.50n	TStp75n	0.50n	0.50n
n			Contamination Level = 6; Case = C4			
50	Alln(MLE)	TStp75n	0.50n	TStp75n	0.50n	0.50n
100	Alln(MLE)	TStp75n	0.50n	TStp75n	0.50n	0.50n
200	Alln(MLE)	TStp75n	0.50n	TStp75n	TStp75n	0.50n
n			Contamination Level = 12; Case = C1			
50	FsTStp75n	TStp75n	0.50n	FsTStp75n	TStp75n	0.50n
100	FsTStp75n	Fs0.50n	Fs0.50n	FsTStp75n	Fs0.50n	Fs0.50n
200	FsTStp75n	FsTStp75n	Fs0.50n	FsTStp75n	FsTStp75n	Fs0.50n

Table 7 continued

	Gaussian data			Gaussian and Uniform data		
	Noise = 0.05	Noise = 0.1	Noise = 0.2	Noise = 0.05	Noise = 0.1	Noise = 0.2
n	Contamination Level = 12; Case = C2					
50	Alln(MLE)	Alln(MLE)	TStp75n	Alln(MLE)	Alln(MLE)	TStp75n
100	Alln(MLE)	TStp75n	TStp75n	Alln(MLE)	TStp75n	TStp75n
200	Alln(MLE)	TStp75n	TStp75n	Alln(MLE)	TStp75n	TStp75n
n	Contamination Level = 12; Case = C3					
50	Alln(MLE)	TStp75n	0.50n	Alln(MLE)	TStp75n	0.50n
100	FsTStp75n	FsTStp75n	Fs0.50n	FsTStp75n	FsTStp75n	Fs0.50n
200	FsTStp75n	FsTStp75n	TStp75n	FsTStp75n	FsTStp75n	TStp75n
n	Contamination Level = 12; Case = C4					
50	Alln(MLE)	TStp75n	0.50n	Alln(MLE)	TStp75n	0.50n
100	Alln(MLE)	TStp75n	TStp75n	Alln(MLE)	TStp75n	TStp75n
200	Alln(MLE)	TStp75n	TStp75n	TStp75n	TStp75n	TStp75n

4.2 Data generation

Data were generated as follows:

For each observation $s_i, i = 1, \ldots, n$, we first generated a $2p$ dimension random vector, $V = [V_1 V_2]$, with independent components.

For Case 4, where MidPoints and Log-Ranges are uncorrelated both between themselves and between each other, the V_1 variables define the intervals' MidPoints C, and the V_2 variables the Log-Ranges R^*.

In Case 3, MidPoints (Log-Ranges) of different variables may be correlated, but no correlation between MidPoints and Log-Ranges is allowed. Then, the MidPoints C are given by linear combinations of the V_1 components, $C = L_1\ V_1^t$; likewise, the Log-Ranges R^* are given by linear combinations of the V_2 components, $R^* = L_2\ V_2^t$. The coefficients in matrices L_1 and L_2 are obtained from independently generated Uniform values, which are then normalized to ensure that L_1 and L_2 are orthonormal.

Case 1 allows for non-zero correlations among all MidPoints and Log-Ranges. Therefore, MidPoints C are defined as linear combinations of both V_1 and V_2 components, $C = [L_1|L_3]\ [V_1^t V_2^t]^t$. Log-Ranges R^* are then given by linear combinations of the V_2 components, $R^* = L_2\ V_2^t$. $L = [L_1|L_3]$ and L_2 are orthonormal matrices obtained as in Case 3.

Finally, in Case 2 interval-valued variables Y_j are uncorrelated, but for each variable, the MidPoint may be correlated with its Log-Range. Accordingly, C and R^* are generated in the same way as in Case 1, but placing all required zeros in matrices L_1, L_2, L_3 to ensure that MidPoints and Log-Ranges of different interval variables have null correlations.

In the first setup, both V_1 and V_2 are Gaussian, whereas in the second setup V_1 is multivariate Normal and V_2 is a vector of independent Uniform variables. In all cases,

Table 8 F-measure obtained by the best methods listed in Tables 6 and 7, for the corresponding data settings–$p=4$: mean (standard error)

	Gaussian data			Gaussian and Uniform data		
	Noise = 0.05	Noise = 0.1	Noise = 0.2	Noise = 0.05	Noise = 0.1	Noise = 0.2
n			Contamination Level = 3; Case = C1			
50	0.2901 (0.0079)	0.3330 (0.0068)	0.3380 (0.0064)	0.3427 (0.0092)	0.3373 (0.0066)	0.3367 (0.0063)
100	0.3413 (0.0067)	0.3289 (0.0061)	0.2083 (0.0046)	0.4019 (0.0080)	0.3581 (0.0065)	0.2058 (0.0049)
200	0.3535 (0.0050)	0.3143 (0.0045)	0.1595 (0.0030)	0.4107 (0.0058)	0.3263 (0.0049)	0.1382 (0.0029)
n			Contamination Level = 3; Case = C2			
50	0.3875 (0.0098)	0.4191 (0.0078)	0.3274 (0.0063)	0.4287 (0.0108)	0.4488 (0.0081)	0.3239 (0.0064)
100	0.4057 (0.0069)	0.4133 (0.0057)	0.2804 (0.0044)	0.4620 (0.0075)	0.4204 (0.0058)	0.2816 (0.0045)
200	0.4005 (0.0048)	0.3996 (0.0041)	0.2739 (0.0031)	0.4528 (0.0052)	0.4281 (0.0041)	0.2621 (0.0033)
n			Contamination Level = 3; Case = C3			
50	0.3664 (0.0097)	0.4012 (0.0078)	0.3118 (0.0066)	0.4081 (0.0107)	0.4149 (0.0079)	0.3044 (0.0068)
100	0.3719 (0.0066)	0.3747 (0.0058)	0.2481 (0.0045)	0.4207 (0.0077)	0.3807 (0.0061)	0.2203 (0.0045)
200	0.3932 (0.0052)	0.3743 (0.0043)	0.2127 (0.0031)	0.4251 (0.0055)	0.3689 (0.0044)	0.1821 (0.0030)
n			Contamination Level = 3; Case = C4			
50	0.3951 (0.0099)	0.4155 (0.0076)	0.3201 (0.0060)	0.4648 (0.0110)	0.4222 (0.0070)	0.2948 (0.0059)
100	0.4129 (0.0072)	0.4241 (0.0056)	0.3112 (0.0042)	0.4659 (0.0069)	0.4217 (0.0057)	0.2917 (0.0043)
200	0.4083 (0.0049)	0.4163 (0.0040)	0.2984 (0.0031)	0.4592 (0.0050)	0.4273 (0.0041)	0.2859 (0.0031)
n			Contamination Level = 6; Case = C1			
50	0.8376 (0.0063)	0.8286 (0.0047)	0.7974 (0.0064)	0.8955 (0.0057)	0.8864 (0.0037)	0.8353 (0.0057)
100	0.8820 (0.0033)	0.9293 (0.0040)	0.9323 (0.0030)	0.9390 (0.0026)	0.9423 (0.0020)	0.9520 (0.0022)
200	0.8816 (0.0020)	0.9470 (0.0014)	0.9616 (0.0013)	0.9370 (0.0016)	0.9677 (0.0012)	0.9681 (0.0012)

Table 8 continued

	Gaussian data			Gaussian and Uniform data		
	Noise = 0.05	Noise = 0.1	Noise = 0.2	Noise = 0.05	Noise = 0.1	Noise = 0.2
n	Contamination Level = 6; Case = C2					
50	0.8916 (0.0044)	0.9364 (0.0021)	0.9606 (0.0014)	0.9435 (0.0036)	0.9602 (0.0019)	0.9694 (0.0015)
100	0.8907 (0.0030)	0.9485 (0.0015)	0.9708 (0.0009)	0.9439 (0.0022)	0.9714 (0.0012)	0.9807 (0.0007)
200	0.8790 (0.0019)	0.9510 (0.0010)	0.9752 (0.0006)	0.9407 (0.0015)	0.9742 (0.0008)	0.9815 (0.0005)
n	Contamination Level = 6; Case = C3					
50	0.8627 (0.0047)	0.9136 (0.0025)	0.9358 (0.0021)	0.9163 (0.0047)	0.9424 (0.0022)	0.9503 (0.0024)
100	0.8928 (0.0028)	0.9468 (0.0020)	0.9640 (0.0012)	0.9436 (0.0023)	0.9677 (0.0016)	0.9697 (0.0012)
200	0.8855 (0.0020)	0.9521 (0.0011)	0.9718 (0.0007)	0.9435 (0.0015)	0.9741 (0.0008)	0.9762 (0.0007)
n	Contamination Level = 6; Case = C4					
50	0.9068 (0.0039)	0.9459 (0.0020)	0.9675 (0.0014)	0.9516 (0.0034)	0.9658 (0.0018)	0.9757 (0.0013)
100	0.8959 (0.0030)	0.9484 (0.0013)	0.9746 (0.0009)	0.9366 (0.0022)	0.9714 (0.0012)	0.9782 (0.0008)
200	0.8724 (0.0019)	0.9522 (0.0010)	0.9753 (0.0006)	0.9359 (0.0016)	0.9742 (0.0008)	0.9791 (0.0006)
n	Contamination Level = 12; Case = C1					
50	0.9123 (0.0035)	0.9736 (0.0014)	0.9937 (0.0007)	0.9476 (0.0030)	0.9848 (0.0011)	0.9971 (0.0004)
100	0.9020 (0.0025)	0.9699 (0.0011)	0.9941 (0.0004)	0.9503 (0.0020)	0.9839 (0.0008)	0.9992 (0.0001)
200	0.8936 (0.0018)	0.9663 (0.0008)	0.9960 (0.0002)	0.9450 (0.0014)	0.9857 (0.0006)	0.9985 (0.0001)
n	Contamination Level = 12; Case = C2					
50	0.9195 (0.0035)	0.9596 (0.0017)	0.9931 (0.0023)	0.9579 (0.0026)	0.9800 (0.0012)	0.9986 (0.0004)
100	0.8927 (0.0025)	0.9601 (0.0012)	0.9933 (0.0004)	0.9430 (0.0019)	0.9832 (0.0008)	0.9975 (0.0003)
200	0.8827 (0.0019)	0.9604 (0.0009)	0.9927 (0.0003)	0.9446 (0.0015)	0.9842 (0.0006)	0.9976 (0.0002)

Table 8 continued

	Gaussian data			Gaussian and Uniform data		
	Noise = 0.05	Noise = 0.1	Noise = 0.2	Noise = 0.05	Noise = 0.1	Noise = 0.2
n	Contamination Level = 12; Case = C3					
50	0.9171 (0.0034)	0.9723 (0.0015)	0.9968 (0.0013)	0.9603 (0.0025)	0.9880 (0.0010)	0.9999 (0.0001)
100	0.9078 (0.0024)	0.9699 (0.0011)	0.9939 (0.0004)	0.9511 (0.0019)	0.9875 (0.0007)	0.9985 (0.0002)
200	0.8927 (0.0019)	0.9649 (0.0009)	0.9939 (0.0003)	0.9501 (0.0014)	0.9863 (0.0006)	0.9981 (0.0002)
n	Contamination Level = 12; Case = C4					
50	0.9235 (0.0032)	0.9574 (0.0017)	0.9936 (0.0018)	0.9589 (0.0026)	0.9807 (0.0012)	0.9983 (0.0008)
100	0.8810 (0.0025)	0.9629 (0.0011)	0.9929 (0.0004)	0.9447 (0.0020)	0.9825 (0.0008)	0.9980 (0.0002)
200	0.8822 (0.0019)	0.9600 (0.0009)	0.9929 (0.0003)	0.9424 (0.0014)	0.9836 (0.0006)	0.9978 (0.0002)

Table 9 F-measure obtained by the best methods listed in Tables 6 and 7, for the corresponding data settings–$p = 10$: mean (standard error)

	Gaussian data			Gaussian and Uniform data		
	Noise = 0.05	Noise = 0.1	Noise = 0.2	Noise = 0.05	Noise = 0.1	Noise = 0.2
n	Contamination Level = 3; Case = C1					
50	0.1573 (0.0059)	0.2065 (0.0038)	0.2764 (0.0039)	0.1525 (0.0059)	0.2097 (0.0037)	0.2750 (0.0038)
100	0.2056 (0.0050)	0.2382 (0.0037)	0.2641 (0.0032)	0.2306 (0.0052)	0.2446 (0.0041)	0.2580 (0.0037)
200	0.2072 (0.0040)	0.2025 (0.0035)	0.1432 (0.0023)	0.2317 (0.0046)	0.1901 (0.0037)	0.1116 (0.0021)
n	Contamination Level = 3; Case = C2					
50	0.2591 (0.0080)	0.2980 (0.0064)	0.2519 (0.0046)	0.3055 (0.0094)	0.3276 (0.0068)	0.2361 (0.0045)
100	0.2650 (0.0059)	0.2773 (0.0051)	0.2011 (0.0037)	0.3076 (0.0064)	0.2762 (0.0049)	0.1688 (0.0031)
200	0.2630 (0.0047)	0.2584 (0.0036)	0.1868 (0.0025)	0.2852 (0.0050)	0.2584 (0.0039)	0.1515 (0.0024)
n	Contamination Level = 3; Case = C3					
50	0.1961 (0.0067)	0.2853 (0.0052)	0.3167 (0.0048)	0.2280 (0.0074)	0.2872 (0.0055)	0.3139 (0.0051)
100	0.2389 (0.0053)	0.2702 (0.0050)	0.2021 (0.0038)	0.2721 (0.0064)	0.2656 (0.0053)	0.1738 (0.0038)
200	0.2361 (0.0045)	0.2337 (0.0037)	0.1449 (0.0023)	0.2595 (0.0050)	0.2166 (0.0038)	0.1093 (0.0021)
n	Contamination Level = 3; Case = C4					
50	0.2615 (0.0075)	0.3338 (0.0062)	0.2865 (0.0048)	0.3175 (0.0093)	0.3661 (0.0068)	0.2806 (0.0046)
100	0.2747 (0.0057)	0.2926 (0.0046)	0.2169 (0.0031)	0.3144 (0.0067)	0.2958 (0.0052)	0.1978 (0.0032)
200	0.2700 (0.0043)	0.2752 (0.0034)	0.1971 (0.0023)	0.3023 (0.0048)	0.2736 (0.0035)	0.1702 (0.0022)
n	Contamination Level = 6; Case = C1					
50	0.4327 (0.0068)	0.3522 (0.0097)	0.3115 (0.0046)	0.4411 (0.0065)	0.3854 (0.0101)	0.3210 (0.0049)
100	0.6526 (0.0097)	0.5851 (0.0089)	0.3868 (0.0070)	0.6142 (0.0043)	0.6765 (0.0088)	0.4176 (0.0079)
200	0.7873 (0.0046)	0.7984 (0.0027)	0.6865 (0.0108)	0.8405 (0.0044)	0.8578 (0.0027)	0.7611 (0.0096)

Table 9 continued

	Gaussian data			Gaussian and Uniform data		
	Noise = 0.05	Noise = 0.1	Noise = 0.2	Noise = 0.05	Noise = 0.1	Noise = 0.2
n	Contamination Level = 6; Case = C2					
50	0.8200 (0.0059)	0.8999 (0.0037)	0.9130 (0.0026)	0.9076 (0.0049)	0.9231 (0.0027)	0.9331 (0.0024)
100	0.8363 (0.0035)	0.9030 (0.0025)	0.9190 (0.0018)	0.9126 (0.0029)	0.9384 (0.0018)	0.9363 (0.0016)
200	0.8351 (0.0024)	0.9134 (0.0015)	0.9214 (0.0012)	0.9152 (0.0021)	0.9429 (0.0013)	0.9329 (0.0012)
n	Contamination Level = 6; Case = C3					
50	0.6362 (0.0085)	0.7481 (0.0086)	0.6516 (0.0083)	0.7092 (0.0086)	0.8095 (0.0074)	0.7107 (0.0083)
100	0.7278 (0.0037)	0.8235 (0.0048)	0.8546 (0.0048)	0.8275 (0.0038)	0.8693 (0.0040)	0.8815 (0.0042)
200	0.8314 (0.0033)	0.8736 (0.0019)	0.8820 (0.0029)	0.8816 (0.0025)	0.9102 (0.0017)	0.8955 (0.0026)
n	Contamination Level = 6; Case = C4					
50	0.8375 (0.0058)	0.9112 (0.0034)	0.9179 (0.0025)	0.8967 (0.0051)	0.9290 (0.0027)	0.9315 (0.0024)
100	0.8480 (0.0036)	0.9110 (0.0023)	0.9201 (0.0017)	0.9126 (0.0027)	0.9411 (0.0018)	0.9294 (0.0016)
200	0.8464 (0.0025)	0.9124 (0.0014)	0.9231 (0.0012)	0.9107 (0.0020)	0.9444 (0.0012)	0.9310 (0.0012)
n	Contamination Level = 12; Case = C1					
50	0.9866 (0.0020)	0.9280 (0.0077)	0.4829 (0.0087)	0.9938 (0.0014)	0.9567 (0.0059)	0.5374 (0.0090)
100	0.9686 (0.0016)	0.9783 (0.0015)	0.9760 (0.0059)	0.9882 (0.0009)	0.9877 (0.0018)	0.9873 (0.0044)
200	0.9289 (0.0016)	0.9827 (0.0006)	0.9946 (0.0002)	0.9717 (0.0011)	0.9941 (0.0004)	0.9983 (0.0001)
n	Contamination Level = 12; Case = C2					
50	0.9608 (0.0024)	0.9944 (0.0007)	0.9969 (0.0004)	0.9935 (0.0010)	0.9986 (0.0010)	0.9995 (0.0002)
100	0.9712 (0.0015)	0.9625 (0.0013)	0.9956 (0.0003)	0.9966 (0.0005)	0.9877 (0.0008)	0.9987 (0.0002)
200	0.9727 (0.0012)	0.9556 (0.0010)	0.9927 (0.0003)	0.9943 (0.0008)	0.9845 (0.0006)	0.9980 (0.0002)

Table 9 continued

	Gaussian data			Gaussian and Uniform data		
	Noise = 0.05	Noise = 0.1	Noise = 0.2	Noise = 0.05	Noise = 0.1	Noise = 0.2
n			Contamination Level = 12; Case = C3			
50	0.9253 (0.0033)	0.9991 (0.0004)	0.8460 (0.0015)	0.9698 (0.0022)	0.9992 (0.0005)	0.8679 (0.0017)
100	0.9111 (0.0022)	0.9692 (0.0011)	0.9948 (0.0003)	0.9587 (0.0017)	0.9886 (0.0007)	0.9983 (0.0002)
200	0.8998 (0.0017)	0.9657 (0.0008)	0.9954 (0.0002)	0.9536 (0.0013)	0.9856 (0.0005)	0.9989 (0.0001)
n			Contamination Level = 12; Case = C4			
50	0.9716 (0.0021)	0.9749 (0.0014)	0.9775 (0.0010)	0.9933 (0.0010)	0.9941 (0.0007)	0.9932 (0.0005)
100	0.9705 (0.0016)	0.9646 (0.0013)	0.9949 (0.0003)	0.9874 (0.0010)	0.9875 (0.0008)	0.9988 (0.0002)
200	0.8843 (0.0021)	0.9534 (0.0010)	0.9931 (0.0003)	0.9412 (0.0015)	0.9835 (0.0006)	0.9982 (0.0002)

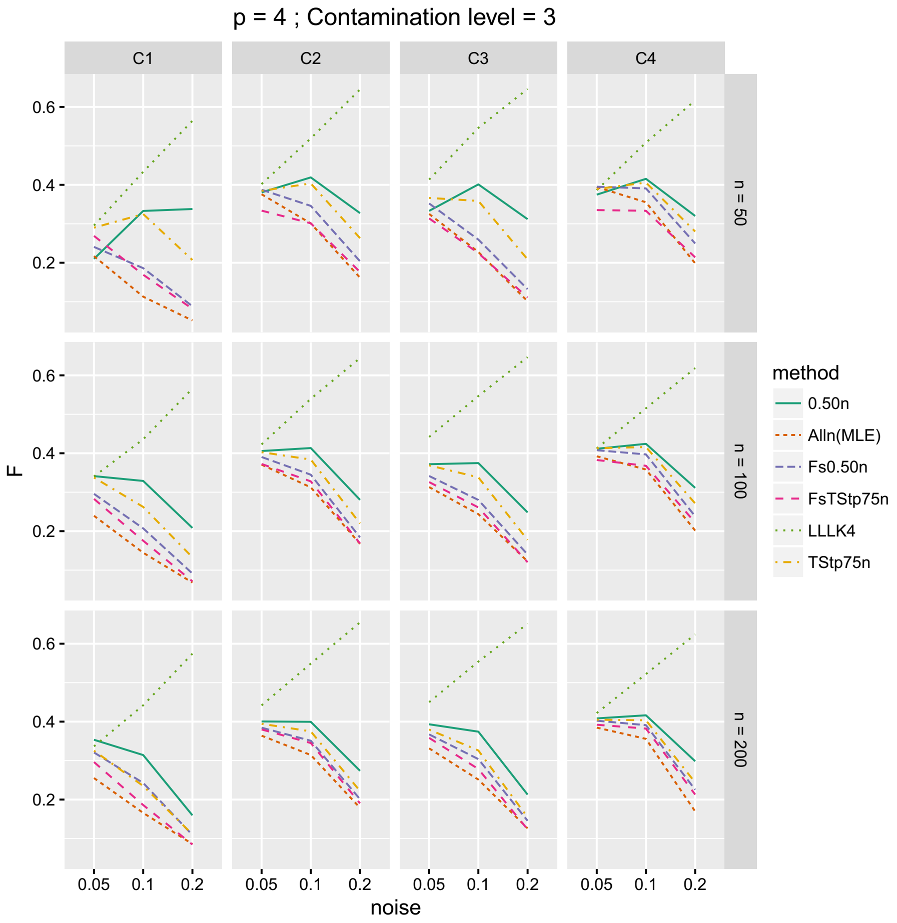

Fig. 1 Performance (F-measure) as a function of the proportion of outlying observations (*Noise*), covariance configuration (C1, C2, C3, C4) and number of observations (n), for Mahalanobis distance = 3, with $p = 4$ interval variables, generated from Gaussian distributions

variances are set to unity. For the regular observations, all components of V_1 and V_2 are generated from distributions with null expected value, while for the outlying observations the expected values are set to ensure the desired level of Mahalanobis distance.

4.3 Results

Tables 6 and 7 present the method resulting in the largest F-measure (see (7)) among all considered methods (except the distance-based approach by Li et al. (2006)), for the data settings with true outliers and Tables 8 and 9 gather the corresponding F values, for all methods except the distance-based approach by Li et al. (2006). In each case the

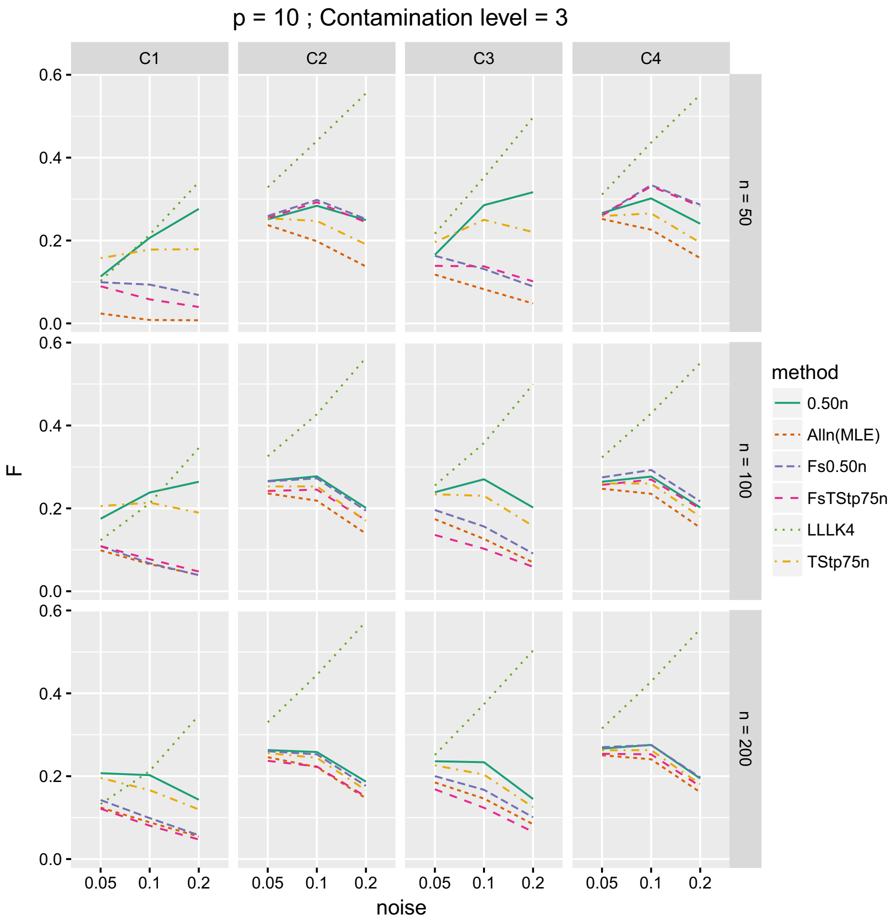

Fig. 2 Performance (F-measure) as a function of the proportion of outlying observations (*Noise*), covariance configuration (C1, C2, C3, C4) and number of observations (n), for Mahalanobis distance = 3, with $p = 10$ interval variables, generated from Gaussian distributions

mean F value of the 1000 replications is given, along with the corresponding standard error. Our experiments showed that Li et al. (2006) method (denoted as LLLK4 in Figs. 1, 2, 3, 4, 5, 6, 7, 8 and 9) which requires the knowledge of the true number of outliers, works the best when the contamination level is either low or highest, but not in intermediate situations–see Figs. 1, 2, 3, 4, 5 and 6. As the number of outliers is never known in practice, and its discovery is one of the challenges, comparisons with this method are not fair and its performance may only be understood as a theoretical benchmark.

The results are very similar for data generated only from Gaussian or from Gaussian and Uniform distributions, so that we present graphical representations only for the former case, in Figs. 1, 2, 3, 4, 5 and 6.

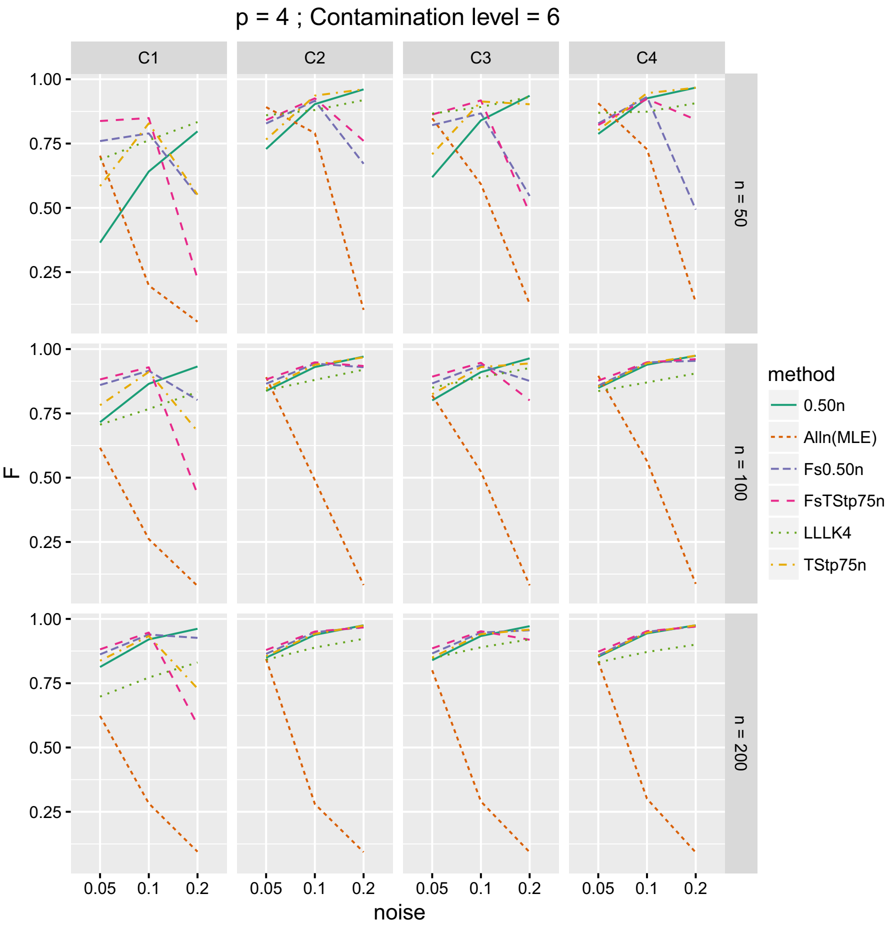

Fig. 3 Performance (F-measure) as a function of the proportion of outlying observations (*Noise*), covariance configuration (C1, C2, C3, C4) and number of observations (n), for Mahalanobis distance = 6, with $p = 4$ interval variables, generated from Gaussian distributions

Contamination level = 3

With four interval-valued variables, the distance-based approach by Li et al. (2006) is by far the best, except in some cases with non-restricted covariances with 5% outliers. Among the methods not using the knowledge of the true number of outliers, the method that uses 50% of the sample with quantiles from the Chi-square distribution (denoted as 0.50n method) is often the best one, sometimes tied with the two-step procedure with Chi-square quantiles. When the number of interval variables is set to 10 the pattern is similar, but for unrestricted covariance matrix (C1), with either 5% or 10% outliers, both the 0.50n method and the two-step approach with Chi-square quantiles may perform better than the distance-based one.

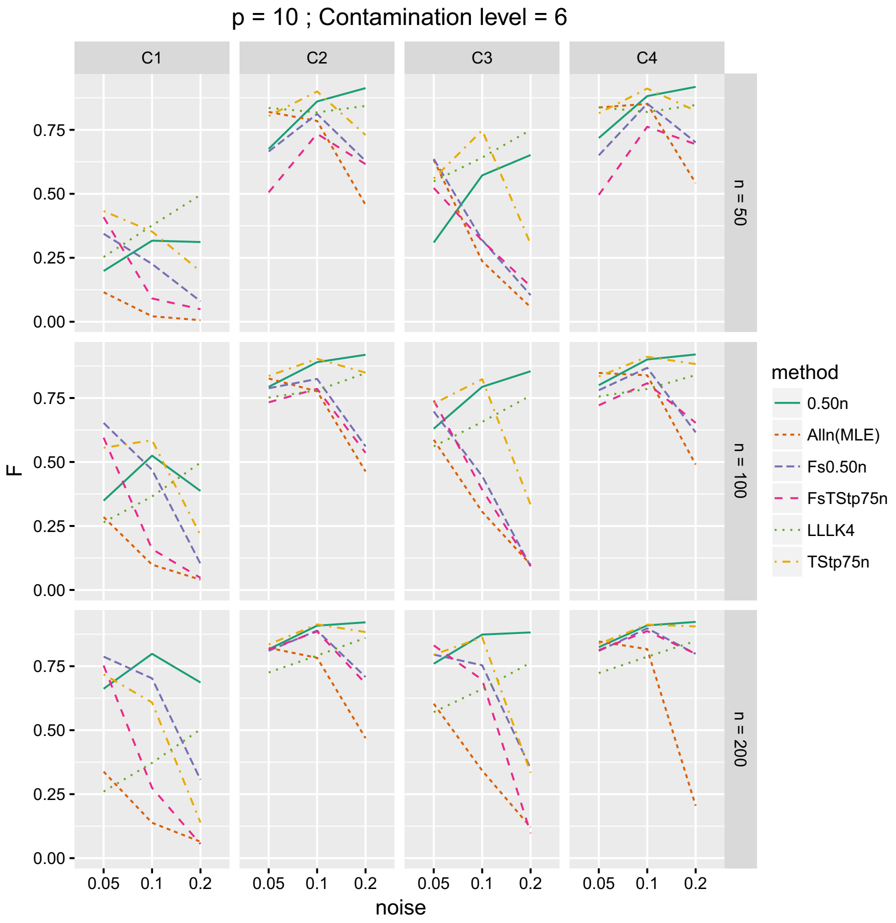

Fig. 4 Performance (F-measure) as a function of the proportion of outlying observations (*Noise*), covariance configuration (C1, C2, C3, C4) and number of observations (n), for Mahalanobis distance = 6, with $p = 10$ interval variables, generated from Gaussian distributions

Contamination level = 6

Maximum likelihood (denoted Alln(MLE)) is almost always (with a few exceptions when $p = 10$) the worse method when the proportion of outliers is equal or higher than 10%, with a remarkable bad performance in configurations involving a large number of parameters. For $p = 4$, and small samples relatively to the number of parameters, the performance of the methods using quantiles from the F or the Beta distributions deteriorates as the true proportion of outliers increases. When we increase the number of interval variables to 10, for large outlier proportions the 0.50n method (using half the sample with Chi-square quantiles) performs better than the other parametric methods. When there are only 5% outliers, the FS0.50n method and the two-step approach with using quantiles from the Chi-square distribution are usually the most competitive methods.

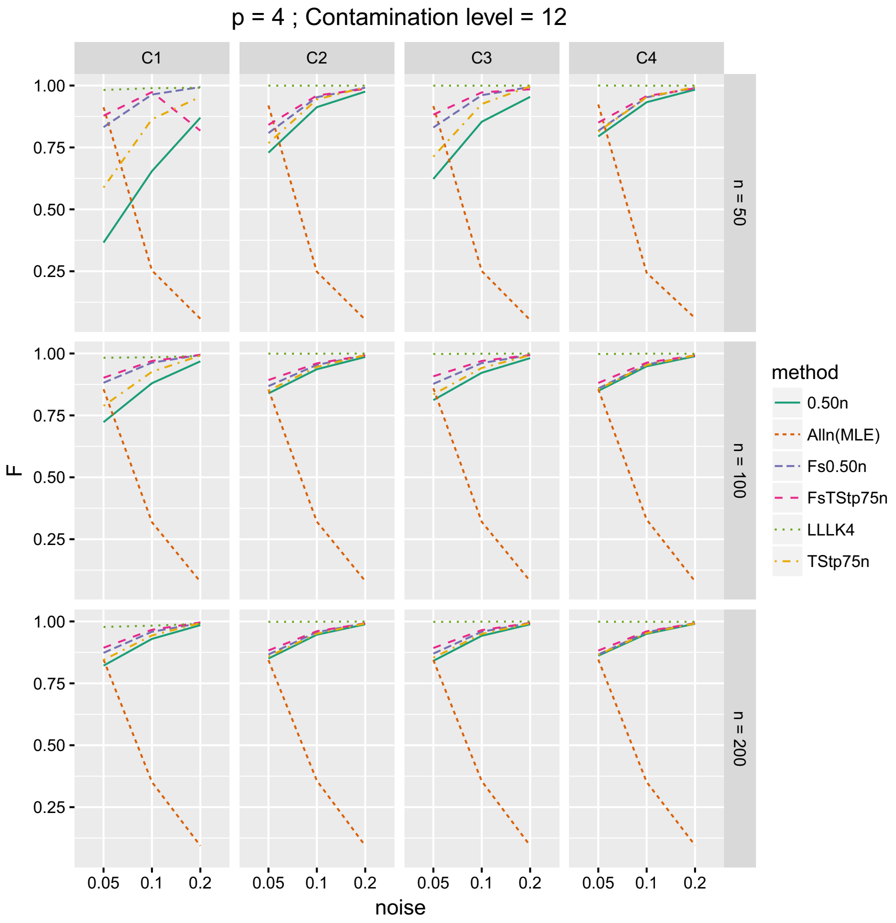

Fig. 5 Performance (F-measure) as a function of the proportion of outlying observations (*Noise*), covariance configuration (C1, C2, C3, C4) and number of observations (n), for Mahalanobis distance = 12, with $p = 4$ interval variables, generated from Gaussian distributions

Contamination level = 12

When $p = 4$, maximum likelihood always performs badly when the proportion of outliers is equal or higher than 10%. For small samples (n = 50), both the 0.50n and the two-step methods (using quantiles from the Chi-square distribution) do not perform well. For a larger dimension ($p = 10$), and for medium or large samples, all methods perform well for the most restricted covariance configurations (C2 and C4), except maximum likelihood in most cases. With these configurations, when there are only 5% outliers the maximum likelihood method performs the best, for larger outlier proportions the two-step using Chi-square quantiles is overall the most competitive. For configurations with more parameters (C1 and C3), depending on the particular condition, either the FS0.50n method or the two-step approach using Chi-square quantiles provide the best results.

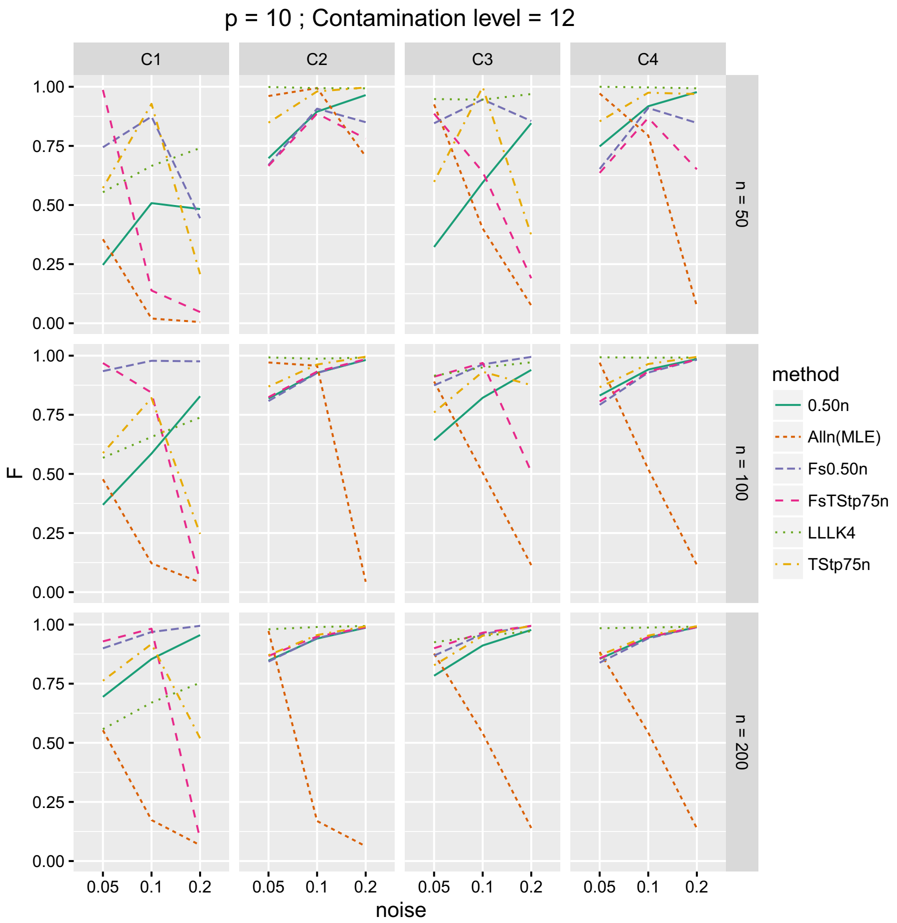

Fig. 6 Performance (F-measure) as a function of the proportion of outlying observations (*Noise*), covariance configuration (C1, C2, C3, C4) and number of observations (n), for Mahalanobis distance = 12, with $p = 10$ interval variables, generated from Gaussian distributions

Non-contaminated case

In the situations where there are no outliers, either maximum likelihood (Alln(MLE)) or the two-step approach with quantiles from the F or Beta distributions (FsTStp75n) are those that flag the lowest number of observations-see Table 10. We note that maximum likelihood performs better in relative terms for lower sample sizes and larger number of parameters to be estimated. This general pattern is valid both for data generated from Gaussian or from Gaussian and Uniform distributions. Furthermore, in all situations, the proportion of wrongly flagged observations is below the nominal significance level of 2.5%-see Table 11; Figs. 8 and 9.

General conclusions

We notice that no method is uniformly best across data conditions. Maximum likelihood is competitive with very low numbers of true outliers, but quickly deteriorates as

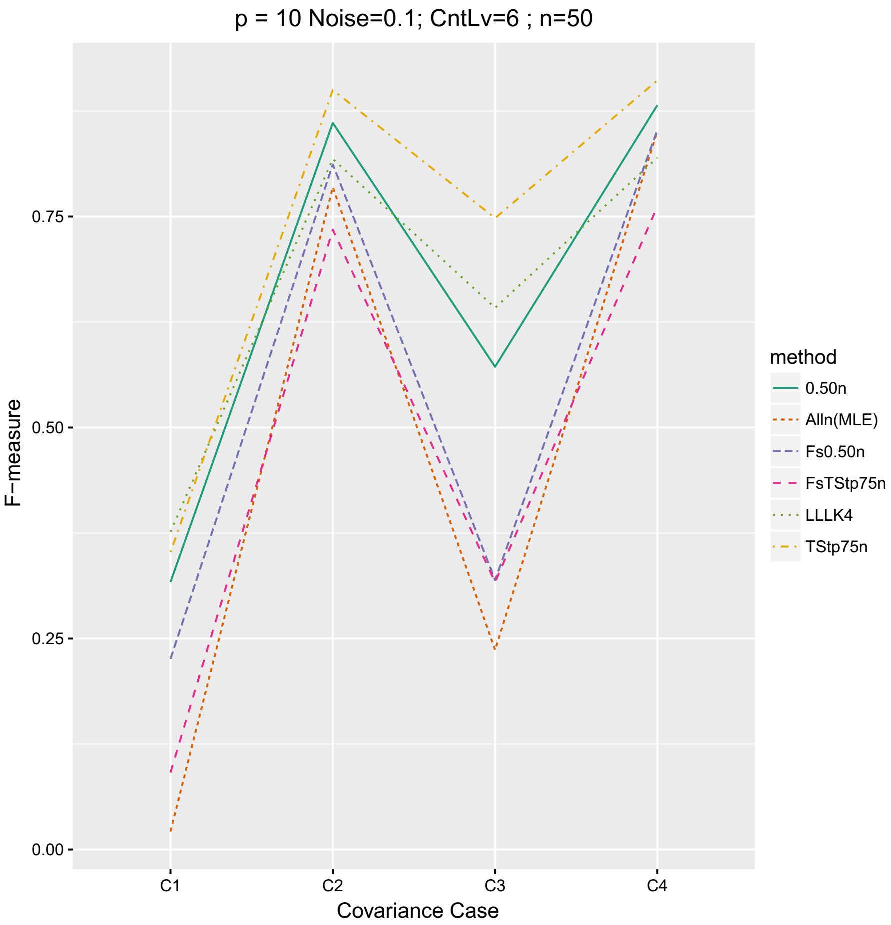

Fig. 7 Performance (F-measure) of the considered methods for the different covariance configurations (C1, C2, C3, C4), with Mahalanobis distance = 6, 10% outliers, n = 50, $p = 10$, data generated from Gaussian distributions

the proportion of outliers increases; the method that uses 50% of the sample provides the best results with high proportions of outliers and low or moderate contamination levels; the two-step method using Chi-square quantiles (TStp75n) is usually the best with moderate or large proportions of outliers and moderate or large contamination levels-see Figs. 1, 2, 3, 4, 5 and 6; this same approach with F or Beta quantiles (FsTStp75n) works best for small proportions of outliers but requires relatively large samples, particularly when there are many parameters to be estimated. This aspect may be relevant for interval data, where the number of parameters grows faster than it is the case for standard real data, unless highly restricted configurations are assumed.

As concerns the different covariance cases, we note that maximum likelihood and the methods using F or Beta quantiles perform better in Cases C2 and C4, where

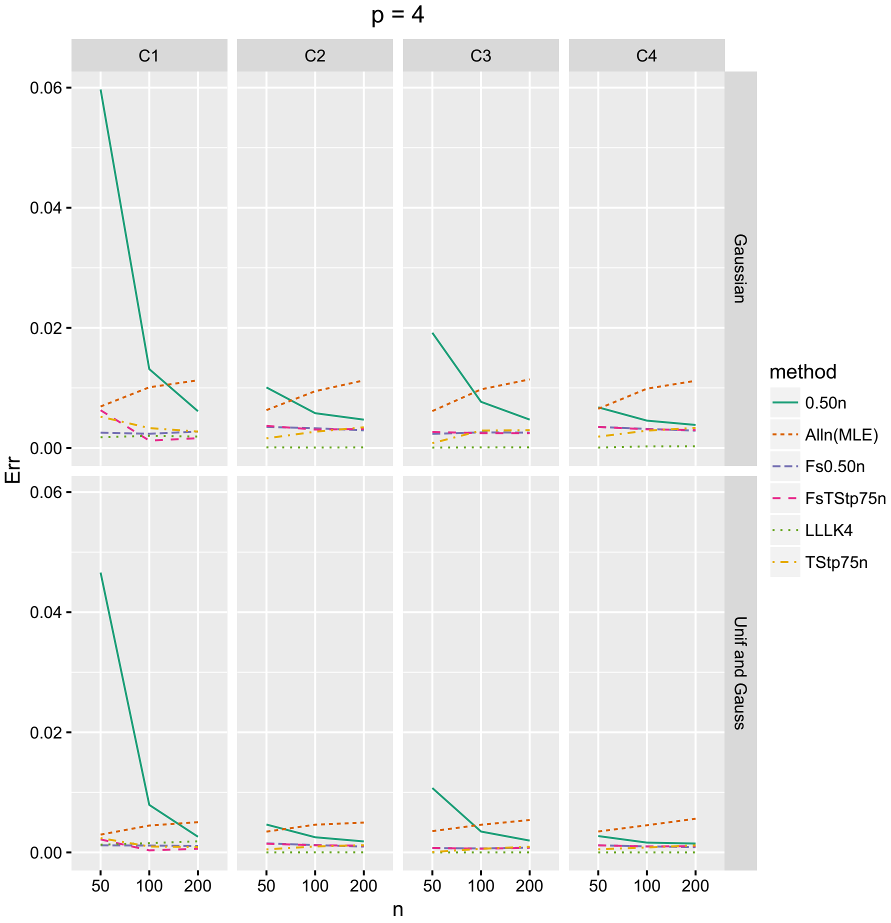

Fig. 8 Percentage of false outliers Err for all methods, for the different covariance configurations (C1, C2, C3, C4) and number of observations (n), with $p = 4$

there are less parameters to estimate, specially when the training samples are small. The two-step methods provide a more stable performance across covariance cases-see Fig. 7.

In practical cases, a preliminary inspection of the data may provide some insights as concerns the proportion and severity of outliers; this information together with the number of variables and the size of the dataset allows identifying appropriate setups to be considered and therefore the corresponding adequate methods.

5 Application

In this section we analyse a set of 27 car models described by four interval-valued variables–*lnPrice, Engine Capacity, Top Speed, Acceleration* (see Table 12). We use

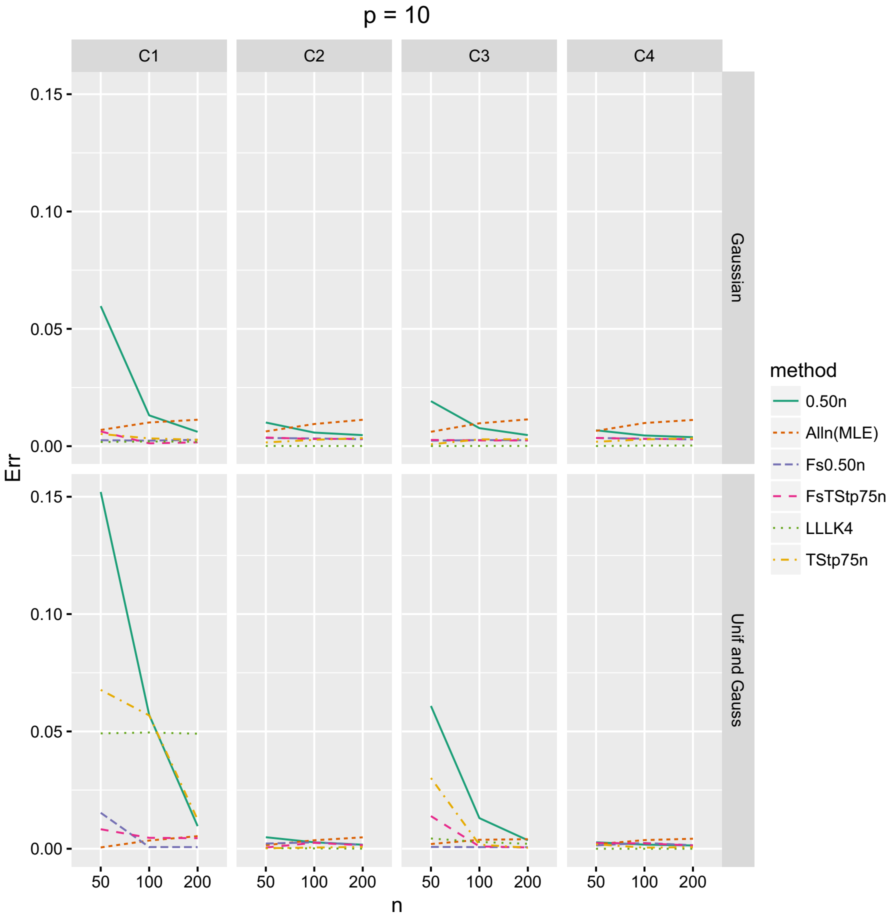

Fig. 9 Percentage of false outliers for all methods, for the different covariance configurations (C1, C2, C3, C4) and number of observations (n), with $p = 10$

the natural logarithm of the Price, since the MidPoints of the original Price variable have a strong positive skewness. The dataset is available in the package MAINT.Data.

We proceed to outlier detection, based on Mahalanobis distances (from each car model to the mean estimate) and considering the 97.5% quantiles of the relevant distributions. Preliminary likelihood ratio tests reject all restricted configurations against the full model at 1% significance level. BIC values obtained by our robust procedure also point to the unrestricted configuration. To decide on the method to apply, we made a preliminary inspection of the data using classical Mahalanobis distances. Even with this non-robust procedure, it became clear that the dataset had at least 10% severe outliers. Therefore, the most appropriate simulation conditions for this data are the cases with a low number of variables ($p = 4$), small training sample, high proportion of outliers and a large contamination level. For these cases, the conclusions from the simulation study suggest the use of the method based on half of the sample using F and

Table 10 Best methods by data condition (no outliers data)

	Gaussian data			Gaussian and Uniform data		
	n = 50	n = 100	n = 200	n = 50	n = 100	n = 200
Case	$p = 4$					
C1	Alln(MLE)	Alln(MLE)	FsTStp75n	Alln(MLE)	Alln(MLE)	FsTStp75n
C2	Alln(MLE)	FsTStp75n	FsTStp75n	Alln(MLE)	FsTStp75n	FsTStp75n
C3	Alln(MLE)	FsTStp75n	FsTStp75n	Alln(MLE)	FsTStp75n	FsTStp75n
C4	Alln(MLE)	FsTStp75n	FsTStp75n	Alln(MLE)	Alln(MLE)	FsTStp75n
Case	$p = 10$					
C1	Alln(MLE)	Alln(MLE)	FsTStp75n	Alln(MLE)	Alln(MLE)	FsTStp75n
C2	Alln(MLE)	Alln(MLE)	Alln(MLE)	Alln(MLE)	Alln(MLE)	Alln(MLE)
C3	Alln(MLE)	FsTStp75n	FsTStp75n	Alln(MLE)	Alln(MLE)	FsTStp75n
C4	Alln(MLE)	Alln(MLE)	Alln(MLE)	Alln(MLE)	Alln(MLE)	Alln(MLE)

Beta quantiles-see Fig. 5 and Table 6. Maximum likelihood, for which the Chi-square quantiles are used, is also applied for comparison purposes. We also identify outliers using the non-parametric approach of Li et al. (2006), with the authors' recommendation of four neighbors.

For the proposed parametric methods, the results are displayed in Figs. 10, 11, 12, 13, 14, 15 and 16. In these figures, cutoffs from the F distribution are indicated with a dashed line whereas those from the Beta distribution are indicated with a solid line. In Fig. 10 we note that whereas classical distances only identify Skoda Octavia and Honda NSK as outliers, using robust distances three other car models are recognized as outliers, but the Honda NSK is no longer flagged.

Figure 11 shows which car models are identified as outliers when we consider only the MidPoints (respectively, only the Log-Ranges). We observe that three of the car models identified as outliers in the global analysis are also identified as outliers in the separate analysis of MidPoints (Ferrari, Mercedes Class S, Porsche), of which Ferrari is also flagged as an outlier for the Log-Ranges; Skoda Octavia is identified as an outlier for the Log-Ranges but not for the MidPoints.

We then proceed to a variable by variable analysis, again considering MidPoints and Log-Ranges separately or jointly-see Figs. 12, 13, 14, 15 and 16.

Table 13 summarizes the results, and indicates which car models are identified as outliers, when the full (all MidPoints and Log-Ranges) or partial (only MidPoints, only Log-Ranges, individual variables) descriptions are considered. We note that the car models indicated as outliers vary according to the analysis performed, which provide different insights. Interval outliers may not stand out when MidPoints or Log-Ranges are analysed separately, putting in evidence the need for their joint analysis–this may be seen, for instance, with variable Engine Capacity, for which no car model was flagged as outlier either for the MidPoints or the Log-Ranges (see Fig. 13a), but Porsche is flagged in the joint analysis (see Fig. 13b), because for this model the relation between MidPoints and Log-Ranges deviates from the usual pattern. This can be seen in Fig. 14,

Table 11 Percentage of false outliers for best methods

	Gaussian data			Gaussian and Uniform data		
	n = 50	n = 100	n = 200	n = 50	n = 100	n = 200
Case	$p=4$					
C1	0.0131 (4.8e−04)	0.0185 (3.9e−04)	0.0214 (2.9e−04)	0.0070 (3.6e−04)	0.0100 (2.9e−04)	0.0106 (2.1e−04)
C2	0.0205 (5.4e−04)	0.0204 (3.9e−04)	0.0213 (2.9e−04)	0.0102 (4.3e−04)	0.0117 (3.0e−04)	0.0111 (2.2e−04)
C3	0.0179 (5.6e−04)	0.0188 (3.9e−04)	0.0200 (2.8e−04)	0.0081 (3.9e−04)	0.0095 (2.9e−04)	0.0104 (2.2e−04)
C4	0.0209 (5.9e−04)	0.0227 (4.2e−04)	0.0232 (2.9e−04)	0.0107 (4.2e−04)	0.0108 (3.0e−04)	0.0115 (2.3e−04)
Case	$p=10$					
C1	0.0017 (1.8e−04)	0.0109 (3.0e−04)	0.0153 (2.4e−04)	0.0009 (1.4e−04)	0.0056 (2.3e−04)	0.0081 (1.6e−04)
C2	0.0201 (5.7e−04)	0.0220 (4.2e−04)	0.0237 (3.1e−04)	0.0097 (4.1e−04)	0.0105 (2.9e−04)	0.0106 (2.1e−04)
C3	0.0103 (4.3e−04)	0.0143 (3.3e−04)	0.0160 (2.5e−04)	0.0045 (3.1e−04)	0.0077 (2.6e−04)	0.0078 (1.9e−04)
C4	0.0215 (5.9e−04)	0.0235 (4.3e−04)	0.0236 (3.0e−04)	0.0095 (4.1e−04)	0.0107 (3.1e−04)	0.0112 (2.2e−04)

Table 12 ‘Car’ data set with four interval-valued variables

	lnPrice	Engine Capacity	Top Speed	Acceleration
Alfa 145	[10.23, 10.42]	[1370, 1910]	[185, 211]	[8.3, 11.2]
Alfa 156	[10.65, 11.04]	[1598, 2492]	[200, 227]	[8.5, 10.5]
...	...	...	...	...
Porsche	[11.90, 12.41]	[3387, 3600]	[280, 305]	[4.2, 5.2]
Rover 25	[9.98, 10.41]	[1119, 1994]	[160, 185]	[10.7, 15.0]
Passat	[10.59, 11.06]	[1595, 2496]	[192, 220]	[9.6, 12.7]

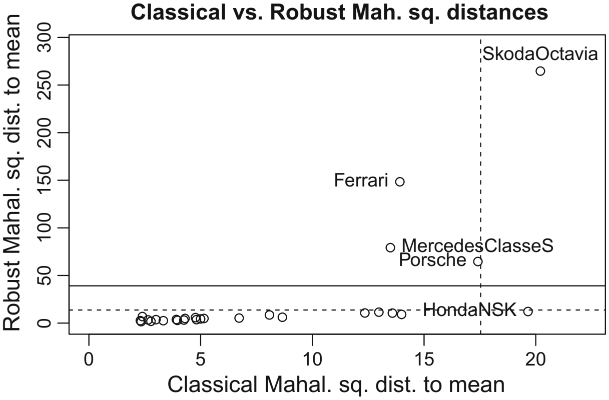

Fig. 10 ‘Car’ dataset: Classical vs robust Mahalanobis distances

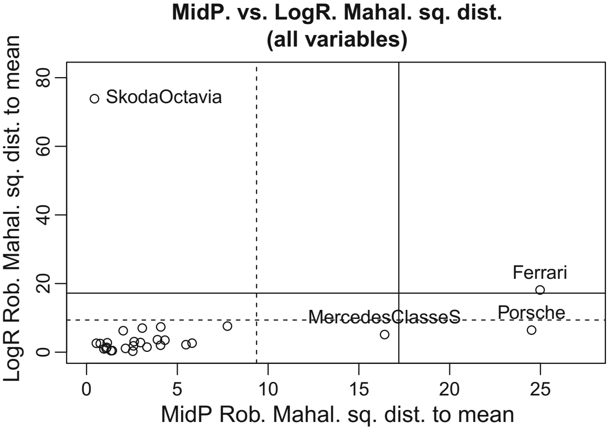

Fig. 11 ‘Car’ dataset: Robust Mahalanobis distances on MidPoints vs Robust Mahalanobis distances on Log-Ranges

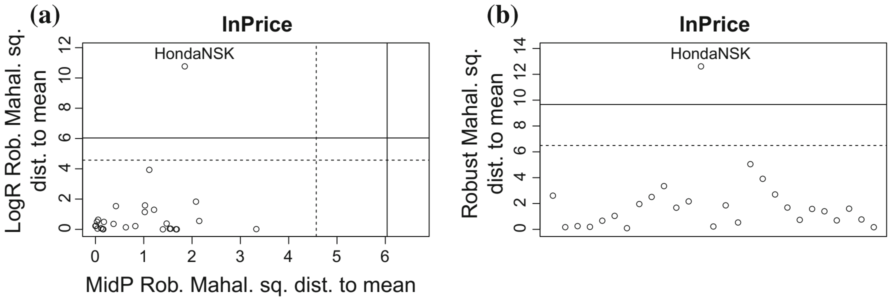

Fig. 12 'Car' dataset: Robust Mahalanobis distances on variable lnPrice: **a** Robust Mahalanobis distances on lnPrice MidPoints vs Robust Mahalanobis distances on lnPrice Log-Ranges; **b** Robust Mahalanobis distances on lnPrice using both MidPoints and Log-Ranges

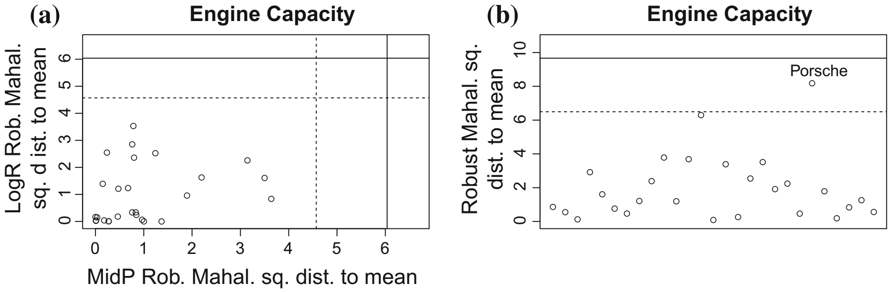

Fig. 13 'Car' dataset: Robust Mahalanobis distances on variable Engine Capacity: **a** Robust Mahalanobis distances on Eng. Cap. MidPoints vs Robust Mahalanobis distances on Eng. Cap. Log-Ranges; **b** Robust Mahalanobis distances on Eng. Cap. using both MidPoints and Log-Ranges

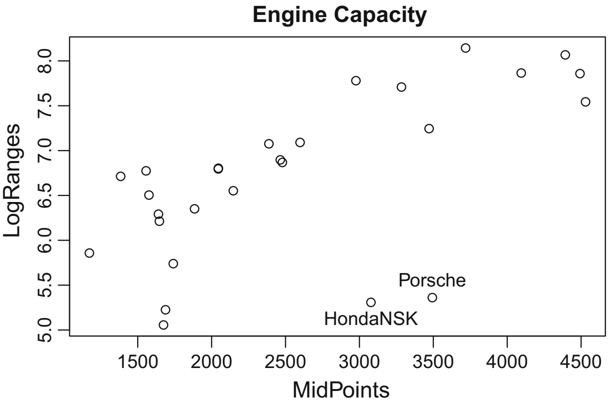

Fig. 14 MidPoints vs. Log-Ranges for Engine Capacity

Honda NSK also deviates from the main pattern but not enough to be flagged using a 97.5% quantile (it would indeed be flagged if we would use the corresponding 95% quantile).

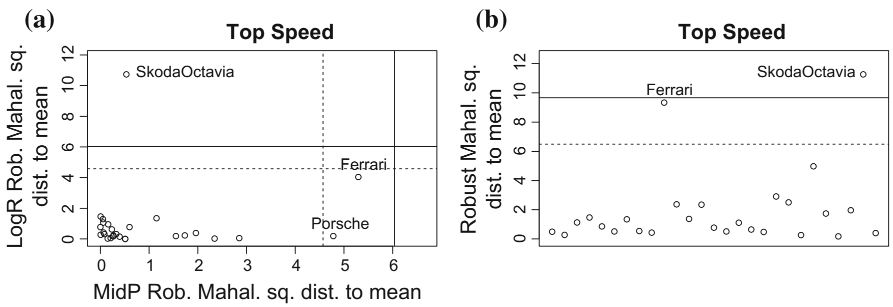

Fig. 15 ‘Car’ dataset: Robust Mahalanobis distances on variable Top Speed: **a** Robust Mahalanobis distances on Top Speed MidPoints vs Robust Mahalanobis distances on Top Speed Log-Ranges; **b** Robust Mahalanobis distances on Top Speed using both MidPoints and Log-Ranges

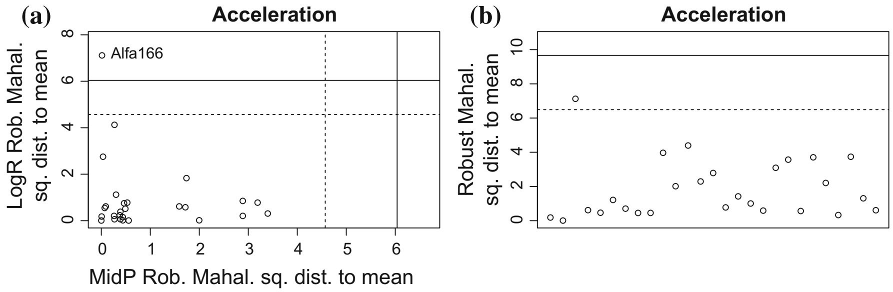

Fig. 16 ‘Car’ dataset: Robust Mahalanobis distances on variable Acceleration: **a** Robust Mahalanobis distances on Acceleration MidPoints versus Robust Mahalanobis distances on Acceleration Log-Ranges; **b** Robust Mahalanobis distances on Acceleration using both MidPoints and Log-Ranges

After removing all six car models identified as possible outliers, the MidPoints and Log-Ranges of the four interval-valued variables pass the Mardia’s, Henze-Zirkler’s and Royston’s tests (Korkmaz et al. 2014) for joint multivariate normality, at the 1% significance level.

The results obtained with the non-parametric approach of Li et al. (2006), using four neighbors, are displayed in Table 14, for different values of the k fixed number of outliers to be identified. We note that while the method identified Ferrari, Porsche and Honda NSK as first outliers, when the fixed number of outliers is increased to $k = 4$ this method indicates the Mercedes Class E, which the applied parametric approach did not flag, Mercedes Class S appears for $k = 5$; for $k = 6$ this method includes Audi A8, also not flagged by the parametric approach.

Summing up, we observe that six car models are identified as possible outliers and would deserve further attention before proceeding to data analysis. They stand out in different complementary analysis. At the univariate level, an observation may be considered as an outlier due to its MidPoint–Ferrari and Porsche are in this case, to its Range, as it is the case for Alfa 166, or to both, or still to a particular relation between them resulting in an outlying interval–as it happens for Porsche or even maybe Honda NSK for Engine Capacity. Furthermore, from a multivariate perspective, it is important

Table 13 Outliers in ‘Car’ data set

	MidPoints and Log-Ranges	MidPoints	Log-Ranges	lnPrice	Engine Capacity	Top Speed	Acceleration
Alfa 166							X
Ferrari	X	X	X			X	
Honda NSK				X			
Mercedes Class S	X	X					
Porsche	X	X			X	X	
Skoda Octavia	X		X			X	

Table 14 Outliers identified by the distance-based method (Li et al. 2006) in the ‘Car’ data set

	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
Ferrari	X	X	X	X	X	X
Porsche		X	X	X	X	X
Honda NSK			X	X	X	X
Mercedes Class E				X	X	X
Mercedes Class S					X	X
Audi A8						X

to distinguish outliers that stand out by their MidPoints–see Mercedes Class S–by their Log-Ranges–as Skoda Octavia–or both-as Ferrari-or still by the global relation between all MidPoints and Log-Ranges.

6 Summary and conclusions

A multivariate outlier detection method has been introduced for interval data, which makes use of parametric modeling of the interval-valued variables according to Brito and Duarte Silva (2012). A joint Multivariate Normal distribution of the MidPoints and Log-Ranges was assumed, with different types of restrictions for the parameters. Maximum likelihood estimation of the parameters, as proposed in Brito and Duarte Silva (2012), leads to non-robust estimates which would not be useful for the purpose of outlier detection, since the estimates themselves are affected by the outliers. This is circumvented by using the weighted trimmed likelihood principle for parameter estimation (Hadi and Luceño 1997). A fast algorithm along the lines of Neykov and Müller (2003) has been implemented in the R package MAINT. Data (Duarte Silva and Brito 2017) which makes parameter estimation for real data sets feasible. Multivariate outlier detection is done by computing Mahalanobis distances for the observations, by plugging in the robust estimates of location and covariance for MidPoints and Log-Ranges. Cutoff values from relevant distributions inform about the outlyingness of the observations. These distributions may be the traditional Chi-square or finite sample approximations following recent proposals (Cerioli 2010; Hardin and Rocke 2005).

The performance of the outlier detection procedure was evaluated by a simulation study, using various different data conditions. In presence of contamination, the advantages of the robust estimators over the non-robust maximum likelihood estimators have been clearly demonstrated. Also in an application it turned out that robust estimation provides interesting new insights, and that diagnostic plots help understanding the outlyingness behavior.

To the best of our knowledge, this is the first statistical outlier investigation for interval data. Further steps may consist in applying multivariate techniques, such as principal component analysis or discriminant analysis. It is straightforward to robustify such methods by plugging in the robust parameter estimates proposed in this paper.

Acknowledgements This work is financed by the ERDF-European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation-COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT - Fundação para a Ciência e Tecnologia (Portuguese Foundation for Science and Technology) as part of projects UID/EEA/50014/2013 and UID/GES/00731/2013.

Appendix

Determination of correction factors

The finite-sample bias-correction factors $c^1_{m,h,n,2p,Cf}$ and $c_{h,n,2p,Cf}$ used in expression (4), are obtained in the following way:

First, based on 1000 independent replications of independently generated standardized Gaussian values, for each combination of $h = \{[0.5n], [0.75n], [0.875n]\}$, $n = \{30, 50, 75, 100, 150, 200, 300, 500\}$, $q = 2p$ with $p = \{1, 2, 3, 4, 5, 7, 10, 15\}$ and covariance configurations $Cf = \{\text{C1, C2, C3, C4}\}$ we found the average of $\tau = |\hat{\Sigma}|^{1/q}$, *i.e.*, the $2p^{th}$ root of the raw consistent-adjusted MCD determinant, which we denote by $avg(\tau)$. Then, for the values of h, n and q included in these simulations $c^*_{h,n,q,Cf} = \frac{1}{avg(\tau)}$ are our first approximations to $c_{h,n,q,Cf}$. In order to find approximations for the remaining parameter values, for each q, Cf and $h = \{0.5, 0.875\}$ we fitted the models

$$\hat{c}^*_{h,n,q,Cf}(n) = 1 + \frac{\gamma_{h,q,Cf}}{n^{\beta_{h,q,Cf}}} \tag{8}$$

and then for each $Cf = \{\text{C1, C2, C3, C4}\}$, $h = \{0.5, 0.875\}$, $r = \{3, 5\}$, $q = 2p$ with $p = \{1, 2, 3, 4, 5, 7, 10, 15\}$ and $n = rq^2$ we fitted

$$\hat{c}^*_{h,n,q,Cf}(q) = 1 + \frac{\eta_{h,r,Cf}}{q^{\kappa_{h,r,Cf}}} \tag{9}$$

Note that $\hat{c}^*_{h,n,q,Cf}(n)$ tends to 1 when n and/or q tend to infinity.

Table 15 Auxiliary model parameters-raw MCD

r	h/n	η				κ			
		C1	C2	C3	C4	C1	C2	C3	C4
3	0.5	− 1.307	− 1.978	− 0.905	− 1.402	1.297	1.906	1.332	1.864
	0.875	− 0.548	− 0.956	− 0.452	− 0.753	1.288	1.990	1.414	1.995
5	0.5	− 0.821	− 1.224	− 0.561	− 0.865	1.286	1.881	1.311	1.838
	0.875	− 0.319	− 0.558	− 0.252	− 0.424	1.274	1.973	1.381	1.964

The final approximation for any n and q is found by first solving the system

$$\begin{aligned} \frac{\eta_{h,3,Cf}}{q^{\kappa_{h,3,Cf}}} &= \frac{\gamma_{h,q,Cf}}{(3q^2)^{\beta_{h,q,Cf}}} \\ \frac{\eta_{h,5,Cf}}{q^{\kappa_{h,5,Cf}}} &= \frac{\gamma_{h,q,Cf}}{(5q^2)^{\beta_{h,q,Cf}}} \end{aligned} \tag{10}$$

in order to $\gamma_{h,q,Cf}$ and $\beta_{h,q,Cf}$, and then setting $c_{h,n,q,Cf} = 1 + \frac{\gamma_{h,q,Cf}}{n^{\beta_{h,q,Cf}}}$.

We did not include $h = 0.75$ or any other h in these models because, as in Pison et al. (2002), we found $c^*_{h,n,q,Cf}$ to be roughly proportional to h so that $c_{h,n,q,Cf}$ for different h values could be found by linear interpolation.

We note that this procedure is identical to the one described in Pison et al. (2002) with the only exception that we have one additional set of model parameters and correction factors for each covariance configuration Cf. In fact, we have found all the auxiliary models for $c^*_{h,n,q,Cf}$ to be well adjusted, but with different parameter values for each configuration Cf, as it can be seen in Table 15.

The authors in Pison et al. (2002) briefly mention that they replicated the same procedure with one step re-weighted instead of raw MCD estimates, in order to find the c^1 one-step re-weighted finite-sample bias-correction factors. We followed their steps but found out that in this case the corresponding $c^{1*}_{h,n,q,Cf}$ approximations were no longer roughly proportional on h, and could have coefficients of determination below 0.05 when regressed on h. This is not that much surprising since the re-weighted MCD uses m instead of h observations to build its final estimate. Therefore, we performed the same simulations as before, but in each replication saved the value of m, and adjusted the following linear regression models (one for each configuration Cf):

$$\tau = \beta^*_{0,n,q,Cf} + \beta^*_{1,Cf}\frac{m}{n} + \beta^*_{2,Cf}\frac{h}{n} \tag{11}$$

where the intercepts $\beta^*_{0,n,q,Cf}$ were found by including dummy variables with all their interactions for all the n and q values used in the simulations.

Then, we adjusted the following models

$$\hat{\beta}^*_{0,n,q,Cf}(n) = 1 - \beta^*_{1,Cf} - \beta^*_{2,Cf} + \frac{\gamma_{q,Cf}}{n^{\beta_{q,Cf}}} \tag{12}$$

Table 16 Auxiliary model parameters–re-weighted MCD

	η		κ		β_1^*	β_2^*
	r = 3	r = 5	r = 3	r = 5		
C1	− 0.280	− 0.220	1.191	1.295	0.435	− 0.075
C2	− 0.516	− 0.274	1.703	1.642	0.204	− 0.021
C3	− 0.321	− 0.234	1.417	1.485	0.318	− 0.045
C4	− 1.283	− 0.477	2.420	2.287	0.275	− 0.015

$$\hat{\beta}^*_{0,n,q,Cf}(q) = 1 - \beta^*_{1,Cf} - \beta^*_{2,Cf} + \frac{\eta_{r,Cf}}{q^{\kappa_{r,Cf}}} \tag{13}$$

ensuring that when n and q tend to infinity $\hat{\beta}^*_{0,n,q,Cf}(n) + \beta^*_{1,Cf} + \beta^*_{2,Cf}$ and $\hat{\beta}^*_{0,n,q,Cf}(q) + \beta^*_{1,Cf} + \beta^*_{2,Cf}$ tend to 1.

We then proceeded as before and found again that all auxiliary models were well adjusted. The estimated values for $\eta_{r,Cf}$, $\kappa_{r,Cf}$, $\beta^*_{1,Cf}$ and $\beta^*_{2,Cf}$ and given in Table 16.

We note that the m coefficient, $\beta^*_{1,Cf}$, is indeed the most important one and is always positive, however the h coefficient, $\beta^*_{2,Cf}$, always negative, is also highly significant. Furthermore, the values in both tables vary considerably according to the covariance configuration, in particular regarding parameter κ which measures the impact of the number of variables in the bias correction factor.

The final $c^1_{m,h,n,q,Cf}$ correction factors are defined by equation

$$c^1_{m,h,n,q,Cf} = \frac{1}{\hat{\tau}} = \frac{1}{\hat{\beta}^*_{0,n,q,Cf}(n) + \beta^*_{1,Cf}\,\frac{m}{n} + \beta^*_{2,Cf}\,\frac{h}{n}} \tag{14}$$

References

Billard B, Diday E (2003) From the statistics of data to the statistics of knowledge: symbolic data analysis. J Am Stat Assoc 98(462):470–487

Bock H-H, Diday E (2000) Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data. Springer, Heidelberg

Brito P (2014) Symbolic data analysis: another look at the interaction of data mining and statistics. WIREs Data Min Knowl Discov 4(4):281–295

Brito P, Duarte Silva AP (2012) Modelling interval data with Normal and Skew-Normal distributions. J Appl Stat 39(1):3–20

Cerioli A (2010) Multivariate outlier detection with high-breakdown estimators. J Am Stat Assoc 105(489):147–156

De Carvalho FAT, Brito P, Bock H-H (2006) Dynamic clustering for interval data based on L_2 distance. Comput Stat 21(2):231–250

De Carvalho FAT, Lechevallier Y (2009) Partitional clustering algorithms for symbolic interval data based on single adaptive distances. Pattern Recogn 42(7):1223–1236

Dias S, Brito P (2017) Off the beaten track: a new linear model for interval data. Eur J Oper Res 258(3):1118–1130

Diday E, Noirhomme-Fraiture M (2008) Symbolic data analysis and the SODAS software. Wiley, Chichester

Douzal-Chouakria A, Billard L, Diday E (2011) Principal component analysis for interval-valued observations. Stat Anal Data Min 4(2):229–246

Duarte Silva AP, Brito P (2017) MAINT.DATA: Model and analyze interval data. R Package,version 1.2.0. http://cran.r-project.org/web/packages/MAINT.Data/index.html

Duarte Silva AP, Brito P (2015) Discriminant analysis of interval data: an assessment of parametric and distance-based approaches. J Classif 32(3):516–541

Filzmoser P (2004) A multivariate outlier detection method. In: S. Aivazian, P. Filzmoser and Yu. Kharin, editors, In Proceedings of the 7th international conference on computer data analysis and modeling, vol 1, 18–22, Belarusian State University, Minsk

Filzmoser P, Reimann C, Garrett RG (2005) Multivariate outlier detection in exploration geochemistry. Comput Geosci 31:579–587

Hadi AS, Luceño A (1997) Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. Comput Stat Data Anal 25(3):251–272

Hardin J, Rocke DM (2005) The distribution of robust distances. J Comput Gr Stat 14:910–927

Hubert M, Rousseeuw PJ, Van Aelst S (2008) High-breakdown robust multivariate methods. Stat Sci 23(1):92–119

Korkmaz S, Goksuluk D, Zararsiz G (2014) MVN: an R package for assessing multivariate normality. R J 6(2):151–162

Le-Rademacher J, Billard L (2011) Likelihood functions and some maximum likelihood estimators for symbolic data. J Stat Plan Inference 141:1593–1602

Le-Rademacher J, Billard L (2012) Symbolic covariance principal component analysis and visualization for interval-valued data. J Comput Gr Stat 21(2):413–432

Li S, Lee R, Lang S-D (2006) Detecting outliers in interval data. In Proceedings of the 44th annual southeast regional conference, ACM, pp 290–295

Lima Neto E, De Carvalho FAT (2008) Centre and range method for fitting a linear regression model to symbolic interval data. Comput Stat Data Anal 52(3):1500–1515

Lima Neto E, De Carvalho FAT (2010) Constrained linear regression models for symbolic interval-valued variables. Comput Stat Data Anal 54(2):333–347

Lima Neto E, Cordeiro GM, De Carvalho FAT (2011) Bivariate symbolic regression models for interval-valued variables. J Stat Comput Simul 81(11):1727–1744

Neykov N, Filzmoser P, Dimova R, Neytchev P (2007) Robust fitting of mixtures using the trimmed likelihood estimator. Comput Stat Data Anal 52(1):299–308

Neykov NM, Müller CH (2003) Breakdown point and computation of trimmed likelihood estimators in generalized linear models. In: Dutter R, Filzmoser P, Gather U, Rousseeuw PJ (eds) Developments in robust statistics. Physica-Verlag, Heidelberg, pp 277–286

Noirhomme-Fraiture M, Brito P (2011) Far beyond the classical data models: symbolic data analysis. Stat Anal Data Min 4(2):157–170

Pison G, Van Aelst S, Willems G (2002) Small sample corrections for LTS and MCD. Metrika 55(1–2):111–123

Ramos-Guajardo AB, Grzegorzewski P (2016) Distance-based linear discriminant analysis for interval-valued data. Inf Sci 372:591–607

Rousseeuw PJ (1984) Least median of squares regression. J Am Stat Assoc 79(388):871–880

Rousseeuw PJ (1985) Multivariate estimation with high breakdown point. Math Stat Appl 8:283–297

Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41(3):212–223

Rousseeuw PJ, Van Zomeren BC (1990) Unmasking multivariate outliers and leverage points. J Am Stat Assoc 85(411):633–639

Van Rijsbergen CJ (1979) Information retrieval, 2nd edn. Butterworth, London

Vandev DL, Neykov NM (1998) About regression estimators with high breakdown point. Statistics 32:111–129

Viattchenin D (2012) Detecting outliers in interval-valued data using heuristic possibilistic clustering. J Comput Sci Control Syst 5(2):39–44