Statistical Papers (2023) 64:395–420
https://doi.org/10.1007/s00362-022-01319-7

REGULAR ARTICLE

A further study comparing forward search multivariate outlier methods including ATLA with an application to clustering

Brenton R. Clarke[1] · Andrew Grose[2]

Received: 23 September 2021 / Revised: 25 February 2022 / Accepted: 28 April 2022 /
Published online: 1 June 2022

Abstract

This paper makes comparisons of automated procedures for robust multivariate outlier detection through discussion and simulation. In particular, automated procedures that use the forward search along with Mahalanobis distances to identify and classify multivariate outliers subject to predefined criteria are examined. Procedures utilizing a parametric model criterion based on a χ^2-distribution are among these, whereas the multivariate Adaptive Trimmed Likelihood Algorithm (ATLA) identifies outliers based on an objective function that is derived from the asymptotics of the location estimator assuming a multivariate normal distribution. Several criterion including size (false positive rate), sensitivity, and relative efficiency are canvassed. To illustrate relative efficiency in a multivariate setting in a new way, measures of variability of the multivariate location parameter when the underlying distribution is chosen from a multivariate generalization of the Tukey–Huber ϵ-contamination model are used. Mean slippage models are also entertained. The simulation results here are illuminating and demonstrate there is no broadly accepted procedure that outperforms in all situations, albeit one may ascertain circumstances for which a particular method may be best if implemented. Finally the paper explores graphical monitoring for existence of clusters and the potential of classification through occurrence of multiple minima in the objective function using ATLA.

Keywords Efficiency · Forward search · Mahalanobis distance · Minimum covariance determinant estimator · Monte Carlo simulation · Multivariate normal distribution

✉ Brenton R. Clarke
B.Clarke@murdoch.edu.au

[1] Mathematics, Statistics, Chemistry, and Physics, College of Science, Health, Engineering, and Education, Murdoch University, Murdoch, WA 6150, Australia

[2] Statistics for the Australian Grains Industry (SAGI West), Curtin University, Bentley, WA 6102, Australia

Mathematics subject classification 62F35 · 62H12

1 Introduction

Outlier detection methods in multivariate data analysis that use the forward search begin with Hadi (1992, 1994) and have gained much publicity with a book and subsequent articles of Atkinson et al (2003), Riani et al (2009), Cerioli et al (2014, 2018, 2019). These articles show that multivariate outlier detection methods exist in a wide variety of settings effectively using different adopted techniques and methodology, with varying intended applications albeit with the same ultimate objective to detect and classify outliers. Riani et al (2009) and Cerioli et al (2019) cite an adaptive trimmed likelihood algorithm (ATLA) published in Clarke and Schubert (2006) which is the multivariate culmination of adaptive methods of trimming discussed in earlier settings including Clarke (1994, 2000), and Bednarski and Clarke (2002). This is a natural extension of the trimmed likelihood estimator countenanced in the univariate and multivariate discussion in Bednarski and Clarke (1993), Butler (1982), Butler et al (1993), Hadi and Luceno (1997), and Clarke et al (2017). See chapters 7 and 8 of Clarke (2018) for a panoramic discussion linking the estimators to the minimum covariance determinant (MCD) estimator of Rousseeuw (1983). The performance of the multivariate ATLA algorithm of Clarke and Schubert (2006) was not previously considered in comparisons even though it was cited. The aim of this paper is to highlight the performance of the original methods of Hadi (1992, 1994) and also the methods of Riani et al (2009) which are all based on the forward search, albeit in different ways, along with ATLA. Other algorithms are briefly considered such as the Blocked Adaptive Computationally Efficient Outlier Nominators method (BACON) (Billor et al, 2000) but only on an intermittent/ad-hoc basis.

Measures of performance indicated by earlier authors vary. This combined with the wide variety of classification techniques made under various assumptions can make comparisons difficult. The importance of this paper is to show empirically at least that there is no universally superior method in outlier detection and subsequent multivariate estimation. There is no single all-encompassing measure or statistic for the performance of an outlier method for any given situation or simulation. For example, a single univariate observation that is known to be outlying may not explicitly imply that of a multivariate outlier with the addition of new variate(s) and the opposite may apply for a single multivariate outlier not necessarily implying that of a univariate one when considering one of its components. Therefore it is important that one defines what constitutes an outlier in the context of the above methods and, in-turn derive and outline a motivation for such methods. Briefly summarizing, an outlier in this context, constitutes an observation that lies at a sufficient distance away from the [centroid of the] majority of the data. By using the word *majority* it is acknowledged that the methods rely on an initial robust calculation to derive a subset of size $h = \lfloor (n + p + 1)/2 \rfloor$ that is assumed to be outlier free, in order to maximize what is termed the finite sample breakdown point (see Rousseeuw (1983) and Clarke (2018)). Here n is the sample size and p is the dimension of the multivariate data in question. Furthermore, regarding the magnitude distance, this may be dependent on the underlying algorithm and its inherent methodology involving a chosen metric. Unlike the Euclidean distance which does not account for correlated variates, the scale-invariant Mahalanobis distance has been shown to be a useful measure of a multivariate observation's outlyingness and thus has led to its continued use in outlier detection, cluster analysis and classification techniques (Mahalanobis 1936; Wilks 1963). However, it is common knowledge that even a single outlier is able to distort measures of multivariate location and scale which can cause an obvious disconnect between the value of Mahalanobis distance that a given observation takes on and whether or not it is outlying. Such perturbations can often result in instances of *masking* and *swamping*. See Barnett and Lewis (1994). While it may be possible to overcome these issues through consideration of all possible subsets by way of exhaustive enumeration, this would be precluded due to the combinatorial explosion for large n and p in sorting out all cases. The forward search procedure is a commonly adopted technique that aims to rectify such issues through iterative exploration of subsets by way of a computationally simple algorithm. The method partitions the observations to form what is called a basic subset. This basic subset is assumed to be outlier free and is based on an initial robust calculation. It is then iteratively redefined by inflating this subset based on Mahalanobis distances calculated with respect to this subset. Subsequently one arrives at a final classification subject to a predefined criterion.

It could be said the above should only be treated as a simplified explanation of this procedure as there exists a number of variations which employ different methodology.

In an alternative development Cabana et al (2021) and Leys et al (2018) use robust Mahalanobis distances. While these have assisted in overcoming issues of masking and swamping, they may not be entirely appropriate when used iteratively in the forward search due to computational burden. Should the data size and dimension permit it, the opportunity to utilise such developments are possible yet these will not be exercised in the context of this paper. See also Filzmoser et al (2014).

A principal motivation for this study is the focus on an adaptive procedure known as the multivariate *adaptive trimmed likelihood algorithm* (ATLA); described in Clarke and Schubert (2006) and Clarke (2018). This numerical routine serves as a multivariate outlier detection method based on the use of the forward search to locate a subset which minimizes a measure of the asymptotic variance of the multivariate location estimator. This method utilizes the minimum covariance determinant (MCD) (Rousseeuw, 1983) based on the Fast-MCD algorithm (Rousseeuw and Driessen, 1999) to obtain a robust initial starting subset preceding the forward search. Hubert et al (2012) proposed an improved MCD algorithm which came as a later development to the initial realization of the ATLA algorithm in Clarke and Schubert (2006). Hubert et al (2012) and Garciga and Verbrugge (2021) demonstrate the improved performance of this deterministic algorithm which serves as motivation for its inclusion in the ATLA algorithm used in this paper.

The use of sample covariance correction factors in the calculation of Mahalanobis Distances, as in Hadi (1992, 1994), for example, that are designed to achieve consistency under the multivariate normal distribution in order to combat bias, do prevent application to certain combinations of n and p. This also brings into question the possible over-reliance on an assumed distribution such as the χ^2 shown by Krazanowski (1988).

2 Simulation study

2.1 Size

In order to demonstrate the adaptive nature of the multivariate ATLA, estimates of size are given based on Monte Carlo Simulation of $N = 1,000$ samples of size n and variables p generated from the multivariate distribution $\mathcal{N}_p(\mathbf{0}, I_p)$. Here $\mathbf{0}$ is the $p \times 1$ mean vector of zeroes and I_p is the assumed $p \times p$ identity covariance matrix. The size is the proportion of samples where at least one outlier is detected. There may be more than one outlier detected in any one sample for example. In addition we will also incorporate size estimates of three similar multivariate outlier detection methods that utilise the forward search in some capacity. These include the Blocked Adaptive Computationally Efficient Outlier Nominators (**BACON**) method by Billor et al (2000), an automatic multivariate outlier detection procedure by Riani et al (2009) (we will refer to as **FSM**) and the original forward search procedure (referred to here as **FS**) proposed by Hadi (1992, 1994). The method **BACON** is available in the R package *robustX* (Stahel and Maechler, 2019) and the method **FSM** is in the package *fsdaR* (Todorov and Sordini, 2020). The results from these simulations are presented in Table 1.

For this simulation two variations of ATLA have been presented. Each pertain to different initial robust estimates based on the aforementioned MCD algorithms; FASTMCD and Deterministic MCD respectively. The FASTMCD algorithm was originally utilized in the ATLA in order to achieve a robust initial starting subset of size $h = \lfloor (n + p + 1)/2 \rfloor$ out of $n_h = \log(0.05)/\log(1 - \binom{(n+p)/2}{p+1}/\binom{n}{p+1})$ possible subsets. Unlike the FASTMCD algorithm which initially considers random subsets, the proposal of Hubert et al (2012) arrives at a subset with the smallest MCD based on six estimators computed in a deterministic way. For the simulations presented in this paper use is made of the 'Deterministic MCD', that is ATLA^b, algorithm due to reasons that will become known.

It is worth mentioning here that unlike the ATLA, these procedures utilize a fixed simultaneous significance level for an assumed parametric distribution which we have chosen to set at $\alpha = 0.01$. The BACON algorithm also requires the user to define the initial subset size, m. As per recommendation in the literature we have chosen to use the default of $m = 4 \cdot p$ for these simulations. It is noted that while BACON and FS cannot be used when $n = 25$ and $p = 10$, see footnotes c and d, this does not preclude use of ATLA and FSM. Here ATLA^b has a smaller average size of 1% while FSM has a size of 10.6%. Also note the decrease in average size between ATLA^a and ATLA^b for these particular parameters.

The adaptive nature of the ATLA method in this analysis demonstrates that if the data are multivariate normal the method soon with sample size of $n \geq 100$ shows almost no unnecessary rejection of outliers under the assumed model.

While other methods than ATLA use a nominal fixed size α with $\alpha = 0.01$ recommended, the maximum average size for ATLA^b is 1.9% in Table 1 and often is much lower. Yet the ATLA^b technique works powerfully to identify outliers in the

Table 1 Estimated size and total number of false positives for simulations of $N = 1,000$ randomly generated samples of size n from $\mathcal{N}_p(\mathbf{0}, I_p)$

n	p	Average Size/Proportion					Total Number of FP				
		ATLA[a]	*ATLA*[b]	*FSM*†	*BACON*†	*FS*†	*ATLA*[a]	*ATLA*[b]	*FSM*†	*BACON*†	*FS*†
25	2	0.026	0.019	0.007	0.008	0.009	211	135	22	22	29
	5	0.021	0.004	0.040	0.001	0.013	183	23	197	1	91
	10	0.107	0.010	0.106	NA[c]	NA[d]	679	51	534	NA[c]	NA[d]
50	2	0.009	0.006	0.005	0.012	0.012	119	49	5	12	12
	5	0.003	0.003	0.014	0.010	0.010	39	39	85	10	10
	10	0.010	0.000	0.105	0.006	0.009	187	0	1346	6	45
100	2	0.001	0.001	0.004	0.010	0.010	1	1	33	10	10
	5	0.000	0.000	0.014	0.007	0.007	0	0	178	8	8
	10	0.000	0.000	0.023	0.006	0.007	0	0	545	6	7
200	2	0.000	0.000	0.007	0.011	0.011	0	0	10	11	11
	5	0.000	0.000	0.008	0.006	0.006	0	0	10	6	6
	10	0.000	0.000	0.009	0.002	0.005	0	0	9	2	5
500	2	0.000	0.000	0.007	0.010	0.010	0	0	9	10	10
	5	0.000	0.000	0.015	0.011	0.011	0	0	17	11	11
	10	0.000	0.000	0.010	0.008	0.009	0	0	13	8	9
1000	2	0.000	0.000	0.013	0.014	0.014	0	0	13	14	14
	5	0.000	0.000	0.017	0.012	0.012	0	0	21	13	13
	10	0.000	0.000	0.011	0.008	0.009	0	0	13	8	9

Note a simultaneous significance level of $\alpha = 0.01$ is used for methods marked †

[a] Initial robust MCD estimates based on the Fast MCD, and;

[b] Deterministic MCD algorithms respectively.

[c] Not applicable due to $n \geq m$.

[d] Not applicable due to $n > 3p + 1$ required for correction factor

tabulations below. This is so much so that it is thought not to be a disadvantage to not be able to set the size specifically.

2.2 Average power and performance measures

In order to assess and compare the performance of ATLA in situations involving contaminated data we will be generating $N = 1,000$ samples of size n for which $n - k$ observations are generated from $\mathcal{N}_p(\mathbf{0}, I_p)$ and k observations from the mean shifted distribution $\mathcal{N}_p(4 \cdot \mathbf{J}, I_p)$ representing the contaminated distribution. Here $\mathbf{J}$ is a $p \times 1$ vector of ones. The parameter k will be chosen in accordance with varying levels of contamination $\epsilon = k/n$ up to a maximum of say, $\epsilon = 0.4$. The following measures of performance are considered:

– Average power = For each sample one calculates the proportion of the k planted outliers that are identified as such and then takes the average over all samples of all such proportions.
– p_1 = Proportion with which exactly the k outlying observations are identified as outliers.
– p_2 = Proportion with which at least one planted outlier is identified.
– p_3 = Proportion with which there is false identification.
– p_4 = Proportion with which at least all the k outliers are trimmed.
– p_5 = Proportion with which observations are identified.

The results from these simulations are presented in Tables 2 and 3 respectively. Note for simplicity we have chosen to omit the performance measures p_2, p_4 and p_5, however, these are available in the Supplementary Materials.

In response to referees' request we have included in Tables 2 and 3 smaller proportions of contamination $\epsilon = 0.02$ and 0.04 where possible, to illustrate what can happen with just one or a few outliers for $n = 25,\ 50$ or 100 as this may be typical in practice.

In those instances that have resulted in a violation in the breakdown point for the respective algorithm, data have been omitted. Due to such restrictions, values for $n = 25$, $p = 10$ and $\epsilon = 0.4$ have been omitted. It is as a result of achieving inconsistent results and in order to ensure brevity that we have chosen not to include BACON in this comparison . One can argue that for the parameters used in this investigation this would not be a fair comparison due to BACON's comparatively low breakdown point. It can also be noted that BACON may be preferred in the case of very large data sets including those of higher dimension because of its computational efficiency.

In terms of the average power it appears that ATLA consistently achieves the highest probability out of the four tested methods with FSM equalling or falling closely behind in most circumstances. It is only in some simulations where the average power of FSM exceeds that of ATLA albeit only by a slight margin. In particular, simulations involving lower contamination ($\epsilon = 0.04$) FSM achieves a higher average power. Unlike the average power, the results for p_1 are not as clearly defined, with all four methods demonstrating differing optimal situations. In some instances of particularly

Table 2 Performance measures of respective outlier methods through simulations of $N = 1,000$ spherically symmetric data [$n \leq 100$] generated from the mean slippage model

n	p	ϵ	k	Average Power			p_1			p_3		
				ATLA	FSM	FS	ATLA	FSM	FS	ATLA	FSM	FS
25	5	0.04	1	0.833	0.967	0.968	0.826	0.930	0.946	0.007	0.037	0.022
		0.08	2	0.914	0.950	0.939	0.904	0.879	0.912	0.009	0.047	0.027
		0.16	4	0.923	0.914	0.869	0.901	0.747	0.838	0.022	0.067	0.031
		0.24	6	0.921	0.904	0.802	0.866	0.681	0.752	0.055	0.036	0.051
		0.32	8	0.906	0.850	0.527	0.851	0.509	0.482	0.055	0.008	0.048
		0.40	10	0.617	0.002	0.184	0.616	0.000	0.180	0.001	0.004	0.009
	10	0.04	1	0.796	0.970	NA[b]	0.786	0.849	NA[b]	0.010	0.124	NA[b]
		0.08	2	0.921	0.960	NA[b]	0.903	0.791	NA[b]	0.019	0.152	NA[b]
		0.16	4	0.927	0.921	NA[b]	0.885	0.664	NA[b]	0.042	0.162	NA[b]
		0.24	6	0.930	0.908	NA[b]	0.861	0.688	NA[b]	0.068	0.078	NA[b]
		0.32	8	NA[a]	0.824	NA[b]	NA[a]	0.655	NA[b]	NA[a]	0.027	NA[b]
50	5	0.02	1	0.998	1.000	1.000	0.979	0.981	0.986	0.009	0.019	0.014
		0.04	2	0.996	0.996	0.996	0.977	0.966	0.983	0.017	0.026	0.011
		0.08	4	0.998	0.997	0.996	0.960	0.943	0.984	0.037	0.049	0.010
		0.16	8	0.998	0.997	0.991	0.884	0.939	0.971	0.113	0.041	0.017
		0.24	12	0.999	0.992	0.978	0.843	0.933	0.960	0.156	0.030	0.016
		0.32	16	0.998	0.972	0.877	0.746	0.877	0.855	0.252	0.024	0.022
		0.40	20	0.846	0.232	0.465	0.635	0.000	0.449	0.209	0.001	0.018
	10	0.02	1	0.999	1.000	1.000	0.998	0.906	0.996	0.001	0.094	0.004
		0.04	2	1.000	0.999	1.000	0.987	0.877	0.993	0.013	0.122	0.007
		0.08	4	0.999	0.999	0.998	0.974	0.858	0.991	0.025	0.140	0.007
		0.16	8	1.000	0.998	0.978	0.916	0.862	0.969	0.084	0.136	0.009
		0.24	12	1.000	0.992	0.644	0.877	0.923	0.632	0.123	0.068	0.013
		0.32	16	0.997	0.983	0.146	0.800	0.953	0.135	0.199	0.030	0.013
		0.40	20	0.674	0.028	0.017	0.671	0.000	0.017	0.000	0.003	0.001
100	5	0.02	2	1.000	1.000	1.000	0.995	0.984	0.991	0.005	0.016	0.009
		0.04	4	0.999	1.000	1.000	0.982	0.973	0.992	0.017	0.026	0.008
		0.08	8	1.000	0.999	0.998	0.933	0.952	0.983	0.065	0.041	0.010
		0.16	16	1.000	1.000	1.000	0.855	0.953	0.984	0.145	0.043	0.013
		0.24	24	1.000	0.996	0.996	0.776	0.963	0.985	0.224	0.026	0.008
		0.32	32	1.000	0.990	0.982	0.719	0.943	0.950	0.281	0.029	0.028
		0.40	40	0.974	0.946	0.770	0.641	0.000	0.749	0.333	0.000	0.014
	10	0.02	2	1.000	1.000	1.000	0.998	0.963	0.994	0.002	0.037	0.006
		0.04	4	1.000	1.000	1.000	0.991	0.942	0.986	0.009	0.058	0.014
		0.08	8	1.000	1.000	1.000	0.974	0.920	0.988	0.026	0.080	0.012
		0.16	16	1.000	1.000	1.000	0.959	0.946	0.988	0.041	0.054	0.012

Table 2 continued

n	p	ϵ	k	Average Power			p_1			p_3		
				ATLA	FSM	FS	ATLA	FSM	FS	ATLA	FSM	FS
		0.24	24	1.000	0.998	0.936	0.899	0.950	0.918	0.101	0.048	0.018
		0.32	32	1.000	0.994	0.416	0.845	0.963	0.402	0.155	0.031	0.017
		0.40	40	0.901	0.950	0.062	0.750	0.000	0.058	0.151	0.000	0.006

[a] Not applicable due to breakdown point

[b] Not applicable due to $n > 3p + 1$ required for correction factor

high contamination FSM and also FS perform extremely poorly. This can be explained since the default breakdown point of FSM, for example, is chosen to be 0.4, whereas the nominal percent of contamination is set at $\epsilon = 0.4$. This may be "corrected" by resetting the default breakdown point for large amounts of contamination to 0.5 in FSM, whereupon a better result for FSM ensues, for example the "Average power" $n = 50,\ p = 10,\ \epsilon = 0.4,\ k = 20$ is 0.964, which compares with the reported value of Table 2 of 0.028. However, it is the default value of 0.4 that is given in the use of the algorithm, and one is not to know in the case of $p = 10$ dimensions whether or not there will be large amounts of contamination in order to adjust the FSM algorithm. ATLA does not have this problem.

On the other hand ATLA remains stable in this event, given that ATLA is the multivariate extension of its univariate estimation algorithm developed in Clarke (1994) which was shown to have breakdown point of near one half.

One will also note that ATLA has a propensity to over-trim for larger n and small p through false identification demonstrated in p_3; a trait which is consistent with the influence of swamping.

2.2.1 Correlated data

The previous section dealt with spherically symmetric ($\Sigma = I_p$) multivariate distributions generated out of the mean slippage model. While algorithms presented here are based on affine equivariant statistics which are accounted for by linear transformations, it is interesting to examine empirical performance under a correlated error structure. Briefly here we consider examples where both distributions are correlated and this is done for $p = 2$.

For each distribution we have chosen to use a simple first order autoregressive covariance structure to generate bivariate ($p = 2$) samples with correlation coefficient $\rho = 0.5$ and 0.9 respectively. That is,

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad \rho = 0.5, 0.9. \tag{1}$$

Due to the proximity of the distributions for $\rho = 0.9$, as seen in Fig. 1b), we have chosen to let the contaminating distribution have shifted mean $5 \cdot \mathbf{J}$. Simulations

Table 3 Performance measures of outlier detection methods through simulations of $N = 1,000$ spherically symmetric data [$n > 100$] generated from the mean slippage model

n	p	ϵ	k	Average Power			p_1			p_3		
				ATLA	FSM	FS	ATLA	FSM	FS	ATLA	FSM	FS
200	5	0.04	8	1.000	1.000	1.000	0.963	0.971	0.984	0.035	0.026	0.014
		0.08	16	1.000	1.000	1.000	0.895	0.964	0.990	0.105	0.034	0.008
		0.16	32	1.000	1.000	1.000	0.775	0.965	0.981	0.225	0.031	0.016
		0.24	48	1.000	1.000	0.999	0.690	0.959	0.981	0.310	0.032	0.011
		0.32	64	1.000	0.998	0.996	0.609	0.966	0.978	0.391	0.020	0.010
		0.40	80	0.996	0.972	0.935	0.530	0.000	0.914	0.466	0.000	0.012
	10	0.04	8	1.000	1.000	1.000	0.997	0.964	0.994	0.003	0.036	0.006
		0.08	16	1.000	1.000	1.000	0.978	0.940	0.987	0.022	0.060	0.013
		0.16	32	1.000	1.000	1.000	0.938	0.953	0.988	0.062	0.047	0.012
		0.24	48	1.000	1.000	0.997	0.888	0.953	0.983	0.112	0.047	0.014
		0.32	64	1.000	0.997	0.700	0.837	0.976	0.695	0.163	0.021	0.006
		0.40	80	0.994	0.985	0.148	0.764	0.000	0.146	0.230	0.000	0.005
500	5	0.04	20	1.000	1.000	1.000	0.924	0.964	0.989	0.075	0.034	0.009
		0.08	40	1.000	1.000	1.000	0.821	0.957	0.987	0.177	0.038	0.008
		0.16	80	1.000	1.000	1.000	0.582	0.961	0.982	0.418	0.032	0.009
		0.24	120	1.000	1.000	1.000	0.400	0.959	0.970	0.600	0.025	0.010
		0.32	160	1.000	1.000	1.000	0.315	0.951	0.972	0.685	0.032	0.008
		0.40	200	0.999	0.994	0.993	0.269	0.000	0.966	0.730	0.001	0.006

Table 3 continued

n	p	ϵ	k	Average Power			p_1			p_3		
				ATLA	FSM	FS	ATLA	FSM	FS	ATLA	FSM	FS
	10	0.04	20	1.000	1.000	1.000	0.989	0.953	0.989	0.011	0.047	0.011
		0.08	40	1.000	1.000	1.000	0.970	0.961	0.990	0.030	0.039	0.010
		0.16	80	1.000	1.000	1.000	0.900	0.952	0.990	0.100	0.048	0.010
		0.24	120	1.000	1.000	1.000	0.793	0.972	0.991	0.207	0.028	0.009
		0.32	160	1.000	1.000	0.960	0.737	0.965	0.949	0.263	0.035	0.011
		0.40	200	1.000	0.990	0.378	0.683	0.000	0.374	0.317	0.000	0.005
1000	5	0.04	40	1.000	1.000	1.000	0.889	0.966	0.987	0.11	0.030	0.009
		0.08	80	1.000	1.000	1.000	0.682	0.956	0.979	0.317	0.035	0.012
		0.16	160	0.999	1.000	1.000	0.352	0.954	0.972	0.647	0.032	0.011
		0.24	240	1.000	1.000	1.000	0.178	0.937	0.953	0.822	0.035	0.009
		0.32	320	1.000	1.000	1.000	0.134	0.941	0.945	0.866	0.025	0.004
		0.40	400	1.000	0.995	0.999	0.081	0.000	0.930	0.919	0.000	0.001
	10	0.04	40	1.000	1.000	1.000	0.982	0.967	0.987	0.018	0.033	0.013
		0.08	80	1.000	1.000	1.000	0.944	0.972	0.993	0.056	0.028	0.007
		0.16	160	1.000	1.000	1.000	0.804	0.957	0.988	0.196	0.043	0.012
		0.24	240	1.000	1.000	1.000	0.681	0.968	0.990	0.319	0.032	0.010
		0.32	320	1.000	1.000	0.997	0.586	0.969	0.988	0.414	0.031	0.009
		0.40	400	1.000	0.998	0.575	0.481	0.000	0.569	0.519	0.000	0.008

have been limited to bivariate samples for illustration purposes. Results from these simulations are presented in Tables 4 and 5 respectively.

Ostensibly, comparing performances with previous simulations one is able to gauge an apparent decrease in performance across all methods. Although this may not be an indication of their inadequacy but rather a by-product of the selected simulation framework. As the proximity of the two distributions grow closer for increasing ρ, the distinction between which distribution a particular observation comes from, becomes difficult (See Fig. 1). Again, the supplementary performance measures can be found in the Supplementary Materials. In terms of the average power, in the majority of cases ATLA performs better in comparison to other methods. A similar but inherently different approach based on regression can be found in Riani et al (2014).

2.3 Further discussion of supplementary performance measures

In the Supplementary Materials from the mean slippage model there are some notable advantages of ATLA in that it performs well in identifying at least one outlier, p_2, and at least the k planted outliers, p_4 in comparison to FSM and FS for such cases involving $\epsilon \geq 0.08$. The latter two methods appear to falter with large 40% proportions of contamination for sample sizes n greater than or equal 50. This is as explained in previous discussion of the breakdown point of FSM.

For the mean slippage model with correlation $\rho = 0.5$ or $\rho = 0.9$, again in the Supplementary Materials show that ATLA consistently trims all k outliers, at a greater rate than FSM or FS, considering the reported values for p_4.

Values for p_2 and p_5 are included for completeness.

2.4 Relative efficiency of outlier trimmed location estimates

Another way we might assess the performance of an outlier detection algorithm is through the relative efficiency of the outlier-trimmed location estimates.

By results of the Cramér Rao Lower Bound a measure of efficiency of a unbiased multivariate estimator say $\mathbf{T}$ is as follows,

$$\mathrm{Eff}(\mathbf{T}) = \frac{|\mathcal{I}(\boldsymbol{\theta})^{-1}|}{|\mathrm{Var}(\mathbf{T})|} \tag{2}$$

Noting here that $|\mathrm{Var}(\mathbf{T})|$ corresponds to the generalized variance which in this case is the determinant of the covariance matrix of the estimator $\mathbf{T}$, call it $|\boldsymbol{\Sigma}_{\boldsymbol{T}}|$. The matrix $\mathcal{I}(\boldsymbol{\theta})^{-1}$ is the inverse of Fisher Information of the unknown parameter vector $\boldsymbol{\theta}$. See Rao (1973) for further information on these arguments.

Hence if one were to compare the efficiencies of two multivariate location estimates, $\mathbf{T}_2$ relative to $\mathbf{T}_1$ for example, then this would involve the following calculation,

$$\mathrm{RelEff}(\mathbf{T}_1, \mathbf{T}_2) = \frac{\mathrm{Eff}(\mathbf{T}_2)}{\mathrm{Eff}(\mathbf{T}_1)} = \frac{|\boldsymbol{\Sigma}_{\mathbf{T}_1}|}{|\boldsymbol{\Sigma}_{\mathbf{T}_2}|}, \tag{3}$$

Table 4 Performance measures of outlier detection methods through simulations of $N = 1,000$ bivariate data generated through the mean slippage model with correlation $[\rho = 0.5]$

ρ	n	ϵ	k	Average Power			p_1			p_3		
				ATLA	FSM	FS	ATLA	FSM	FS	ATLA	FSM	FS
0.5	25	0.04	1	0.651	0.686	0.798	0.625	0.677	0.779	0.026	0.010	0.020
		0.08	2	0.747	0.611	0.707	0.707	0.512	0.655	0.029	0.012	0.016
		0.16	4	0.814	0.549	0.588	0.720	0.322	0.544	0.073	0.023	0.025
		0.24	6	0.791	0.578	0.460	0.677	0.229	0.431	0.092	0.017	0.018
		0.32	8	0.755	0.556	0.418	0.628	0.169	0.380	0.112	0.002	0.033
		0.40	10	0.622	0.002	0.371	0.526	0.000	0.345	0.083	0.001	0.021
	50	0.02	1	0.813	0.859	0.903	0.788	0.854	0.895	0.025	0.005	0.009
		0.04	2	0.906	0.837	0.873	0.822	0.727	0.797	0.054	0.014	0.013
		0.08	4	0.942	0.817	0.762	0.791	0.553	0.642	0.101	0.024	0.011
		0.16	8	0.961	0.859	0.520	0.680	0.347	0.446	0.205	0.034	0.006
		0.24	12	0.965	0.867	0.401	0.636	0.235	0.364	0.269	0.035	0.010
		0.32	16	0.965	0.856	0.334	0.568	0.186	0.308	0.341	0.036	0.008
		0.40	20	0.908	0.280	0.306	0.531	0.000	0.283	0.336	0.009	0.017
	100	0.02	2	0.907	0.891	0.912	0.843	0.793	0.838	0.026	0.011	0.007
		0.04	4	0.941	0.889	0.866	0.789	0.620	0.673	0.067	0.021	0.005
		0.08	8	0.965	0.905	0.733	0.725	0.424	0.499	0.148	0.030	0.007
		0.16	16	0.982	0.924	0.425	0.593	0.240	0.299	0.296	0.039	0.005
		0.24	24	0.990	0.933	0.246	0.514	0.160	0.188	0.383	0.041	0.002
		0.32	32	0.989	0.931	0.171	0.447	0.105	0.143	0.480	0.067	0.007
		0.40	40	0.968	0.864	0.132	0.408	0.000	0.115	0.488	0.067	0.006

Table 4 continued

ρ	n	ϵ	k	Average Power			p_1			p_3		
				ATLA	FSM	FS	ATLA	FSM	FS	ATLA	FSM	FS
	200	0.04	8	0.964	0.916	0.847	0.687	0.444	0.450	0.117	0.029	0.007
		0.08	16	0.982	0.929	0.571	0.558	0.214	0.223	0.272	0.053	0.004
		0.16	32	0.990	0.943	0.169	0.391	0.083	0.078	0.488	0.077	0.005
		0.24	48	0.992	0.951	0.060	0.293	0.035	0.027	0.604	0.084	0.003
		0.32	64	0.995	0.956	0.024	0.241	0.030	0.019	0.681	0.112	0.002
		0.40	80	0.993	0.923	0.009	0.201	0.000	0.006	0.725	0.134	0.004
	500	0.04	20	0.973	0.925	0.764	0.465	0.118	0.099	0.221	0.061	0.004
		0.08	40	0.986	0.943	0.184	0.305	0.026	0.009	0.493	0.097	0.004
		0.16	80	0.995	0.959	0.004	0.122	0.007	0.000	0.794	0.147	0.002
		0.24	120	0.996	0.964	0.000	0.056	0.008	0.000	0.895	0.179	0.003
		0.32	160	0.997	0.969	0.000	0.036	0.001	0.000	0.943	0.230	0.004
		0.40	200	0.997	0.956	0.000	0.033	0.000	0.000	0.943	0.250	0.000
	1000	0.04	40	0.975	0.935	0.587	0.234	0.012	0.002	0.437	0.130	0.004
		0.08	80	0.988	0.952	0.009	0.086	0.005	0.000	0.745	0.193	0.002
		0.16	160	0.994	0.963	0.000	0.020	0.001	0.000	0.945	0.302	0.001
		0.24	240	0.996	0.970	0.000	0.003	0.000	0.000	0.983	0.384	0.001
		0.32	320	0.997	0.975	0.000	0.002	0.000	0.000	0.994	0.465	0.002
		0.40	400	0.998	0.971	0.000	0.003	0.000	0.000	0.995	0.441	0.004

Table 5 Performance measures of outlier detection methods through simulations of $N = 1,000$ bivariate data generated through the mean slippage model with correlation $[\rho = 0.9]$

ρ	n	ϵ	k	Average Power			p_1			p_3		
				ATLA	FSM	FS	ATLA	FSM	FS	ATLA	FSM	FS
0.9	25	0.04	1	0.457	0.494	0.641	0.428	0.487	0.626	0.030	0.008	0.017
		0.08	2	0.564	0.418	0.510	0.516	0.307	0.450	0.032	0.011	0.015
		0.16	4	0.602	0.309	0.336	0.498	0.148	0.294	0.073	0.022	0.021
		0.24	6	0.555	0.345	0.244	0.425	0.101	0.222	0.096	0.019	0.016
		0.32	8	0.517	0.339	0.205	0.389	0.079	0.174	0.111	0.004	0.028
		0.40	10	0.407	0.002	0.185	0.323	0.000	0.162	0.069	0.001	0.021
	50	0.02	1	0.604	0.667	0.751	0.579	0.663	0.743	0.025	0.005	0.009
		0.04	2	0.751	0.642	0.694	0.646	0.477	0.566	0.054	0.015	0.012
		0.08	4	0.795	0.584	0.513	0.585	0.288	0.385	0.095	0.031	0.009
		0.16	8	0.817	0.664	0.228	0.499	0.137	0.186	0.188	0.051	0.004
		0.24	12	0.822	0.668	0.123	0.455	0.062	0.102	0.240	0.061	0.004
		0.32	16	0.805	0.635	0.098	0.396	0.058	0.088	0.303	0.079	0.004
		0.40	20	0.707	0.138	0.079	0.363	0.000	0.072	0.272	0.004	0.008
	100	0.02	2	0.700	0.699	0.729	0.601	0.523	0.578	0.024	0.012	0.006
		0.04	4	0.785	0.728	0.634	0.534	0.308	0.342	0.059	0.023	0.006
		0.08	8	0.837	0.795	0.344	0.481	0.136	0.160	0.120	0.049	0.005
		0.16	16	0.858	0.820	0.080	0.351	0.049	0.053	0.252	0.070	0.004
		0.24	24	0.850	0.832	0.030	0.304	0.030	0.023	0.343	0.132	0.001
		0.32	32	0.837	0.853	0.018	0.253	0.022	0.015	0.417	0.197	0.004
		0.40	40	0.771	0.739	0.013	0.230	0.000	0.013	0.387	0.162	0.002

Table 5 continued

ρ	n	ϵ	k	Average Power			p_1			p_3		
				ATLA	FSM	FS	ATLA	FSM	FS	ATLA	FSM	FS
	200	0.04	8	0.811	0.804	0.538	0.376	0.113	0.108	0.086	0.063	0.003
		0.08	16	0.858	0.840	0.114	0.268	0.032	0.022	0.221	0.103	0.002
		0.16	32	0.893	0.865	0.004	0.167	0.014	0.001	0.427	0.172	0.003
		0.24	48	0.880	0.881	0.000	0.127	0.003	0.000	0.519	0.228	0.003
		0.32	64	0.855	0.896	0.000	0.096	0.002	0.000	0.579	0.343	0.002
		0.40	80	0.774	0.882	0.000	0.070	0.000	0.000	0.566	0.419	0.004
	500	0.04	20	0.855	0.840	0.296	0.133	0.009	0.002	0.176	0.103	0.004
		0.08	40	0.906	0.870	0.005	0.059	0.000	0.000	0.416	0.197	0.004
		0.16	80	0.936	0.900	0.000	0.013	0.000	0.000	0.715	0.371	0.002
		0.24	120	0.927	0.916	0.000	0.005	0.000	0.000	0.820	0.512	0.003
		0.32	160	0.891	0.929	0.000	0.003	0.000	0.000	0.836	0.625	0.004
		0.40	200	0.871	0.928	0.000	0.000	0.000	0.000	0.812	0.708	0.000
	1000	0.04	40	0.874	0.856	0.111	0.003	0.000	0.000	0.355	0.250	0.002
		0.08	80	0.927	0.890	0.001	0.003	0.000	0.000	0.656	0.424	0.002
		0.16	160	0.956	0.920	0.000	0.000	0.000	0.000	0.915	0.647	0.001
		0.24	240	0.954	0.934	0.000	0.000	0.000	0.000	0.956	0.786	0.001
		0.32	320	0.928	0.948	0.000	0.000	0.000	0.000	0.936	0.873	0.002
		0.40	400	0.891	0.956	0.000	0.000	0.000	0.000	0.899	0.917	0.004

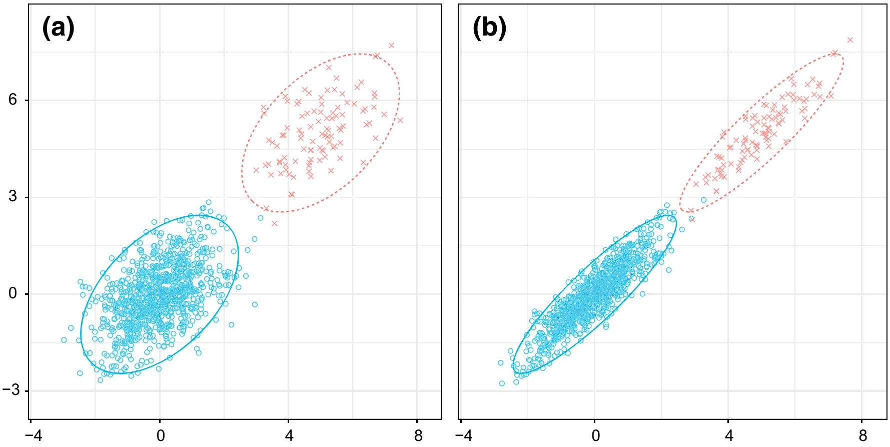

Fig. 1 Scatterplot matrix of a simulation of bivariate data with correlation of $\rho = 0.5$ [a)] and $\rho = 0.9$ [b)] respectively. The ellipses here represent the 95% bivariate normal density contours

where $| \cdot |$ denotes the determinant of the associated covariance matrix of location estimates. For example, the efficiency of ATLA relative to FSM would show ATLA as the better estimate for relative efficiencies greater than one.

Now for simulations of data generated through the Tukey–Huber ϵ-contamination model,

$$f(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \epsilon) = (1-\epsilon) \cdot \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \epsilon \cdot \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \cdot \sigma^2) \tag{4}$$

we choose to fix $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma} = I_p$ and $\sigma^2 = 9$ for simplicity.

Here we will produce estimates of the relative efficiencies of the respective algorithms with respect to ATLA. That is we identify $\mathbf{T}_2$ in equation (3) as the ATLA estimate for location and a scaled estimate of the denominator in the second part of the equation is the determinant of the variance covariance matrix of the ATLA trimmed location estimates. For example, if we calculate the relative efficiency of FSM, then the numerator in equation (3) would be the the determinant of the variance covariance matrix of the estimates of location achieved using FSM. Note with the underlying model (4) the estimates of location are unbiased and consistent estimates of $\boldsymbol{\mu}$. Since according to Cator et al (2012) the estimators are asymptotically normal even at the distribution (4), it follows the ratio of the generalized variances can be estimated as we have done here.

The results from simulations of $N = 1,000$ for various levels of contamination ϵ, sample sizes n and variables p are presented in Table 6 where covMCD and DetMCD are explained below.

In discussing efficiency we allude to two well known methods of estimating multivariate location. For these simulations we have included for comparison the R function *covMcd* available from the *robustbase* package (Maechler et al, 2019) which is a robust location estimation method that utilizes the MCD. This function contains arguments that allow use of two possible MCD proposals; the Fast MCD algorithm spawned out

Table 6 Relative Efficiencies of ATLA to the respective outlier algorithms for the trimmed multivariate location estimates through simulations of $N = 1,000$ samples generated from the Tukey–Huber Model

n	ϵ	p=5				p=10			
		FSM	FS	covMCD[a,b]	DetMCD[a]	FSM	FS	covMCD[a,b]	DetMCD[a]
25	0.1	0.802	0.879	2.620	2.284	0.754	NA[c]	1.926	0.881
	0.2	0.615	0.784	0.650	0.536	0.570	NA[c]	0.209	0.082
	0.3	0.555	0.908	0.223	0.167	0.516	NA[c]	0.094	0.039
	0.4	0.684	0.930	0.207	0.126	0.561	NA[c]	0.222	0.089
50	0.1	1.059	1.187	3.380	2.802	1.556	1.254	50.563	42.410
	0.2	0.946	1.832	1.366	1.174	1.602	2.553	10.880	12.820
	0.3	0.847	3.434	0.583	0.522	2.554	4.863	2.428	2.469
	0.4	0.611	2.780	0.328	0.268	3.273	4.255	0.963	0.713
100	0.1	0.989	1.200	2.013	1.706	1.169	1.083	11.738	8.657
	0.2	0.957	2.401	1.131	1.059	1.129	1.420	4.648	3.847
	0.3	0.674	4.707	0.552	0.500	1.229	1.907	2.056	1.845
	0.4	0.401	5.448	0.246	0.243	3.267	3.638	1.009	1.052
200	0.1	0.969	1.269	1.530	1.356	1.040	1.077	3.658	2.998
	0.2	0.980	2.659	0.869	0.803	1.061	1.435	1.637	1.570
	0.3	0.764	9.876	0.624	0.596	1.103	1.953	1.365	1.198
	0.4	0.406	8.444	0.296	0.323	1.876	4.638	0.879	0.896
500	0.1	0.964	1.439	1.182	1.084	0.990	1.116	1.947	1.883
	0.2	0.920	4.131	0.931	0.899	1.072	1.538	1.490	1.355
	0.3	0.847	21.638	0.716	0.749	1.051	2.601	1.065	0.998
	0.4	0.231	8.715	0.214	0.254	1.429	19.347	0.897	0.941

Table 6 continued

n	ϵ	p=5				p=10			
		FSM	FS	covMCD[a,b]	DetMCD[a]	FSM	FS	covMCD[a,b]	DetMCD[a]
1000	0.1	0.946	1.466	1.127	1.124	0.985	1.126	1.600	1.549
	0.2	0.890	5.886	0.919	0.894	0.990	1.493	1.220	1.138
	0.3	0.811	25.738	0.727	0.746	1.005	2.977	0.931	0.915
	0.4	0.246	9.850	0.247	0.323	1.119	479.266	0.913	1.005

[a] $\alpha = \lfloor (n+1)/2 \rfloor / n$

[b] $n_h = \log(0.05)/\log(1 - \binom{(n+p)/2}{p+1}/\binom{n}{p+1})$

[c] Not applicable due to $n > 3p + 1$ required for correction factor

of Rousseeuw and Driessen (1999) as well as the "Deterministic MCD" algorithm proposed by Hubert et al (2012). In the case of FASTMCD this algorithm searches through n_h possible subsets of size $h = \alpha \cdot n$ for some predefined $0.5 \leq \alpha < 1$, whose covariance matrix yields the lowest determinant. Both of these procedures are known to have a particularly high breakdown point and have been included in this simulation to highlight what to expect in performance given data that consists of a high percentage of contamination. Here the name "DetMCD" is used to denote the deterministic MCD algorithm and "covMCD" is used to denote the initial Fast MCD algorithm. The relative efficiencies of these algorithms are compared to ATLA where just the Det MCD is employed in both cases for the initial start of the forward search algorithm.

In consideration of the two MCD methods there is an apparent lack of efficiency of the high breakdown point estimators of location, ATLA performing remarkably well in simulations with lower proportions of contamination. One will also notice the Deterministic MCD algorithm consistently produces a higher efficiency in comparison with the FASTMCD; one of the motivating factors for its substitution in ATLA. It is only in a few such cases where *covMCD* is the more efficient estimator but mostly by a small margin. ATLA beats FS generally for $n \geq 50$ and is not as efficient for $n = 25$. There are mixed results for relative efficiencies of FSM and ATLA. There is no clear winner. ATLA appears to be better in cases when the dimension, p, is higher.

3 Cluster monitoring and detection

Due to the ability to detect outliers, the use-case of the forward search has shown that it can be further extended into the area of cluster analysis. It has been explored in a number of circumstances including Atkinson and Riani (2004), and Cerioli et al (2019).

However most notably graphical techniques can be employed to aid in monitoring successive iterations of the forward search in an attempt to divulge the structure of the data and assist in possible outlier or cluster identification. This includes *forward-plots* to enable monitoring of the subset inflations by displaying the minimum Mahalanobis distance among units in the non-basic subset. Examples of this for parametric cases are in Atkinson et al (2003), Atkinson and Riani (2004), Atkinson et al (2018) and Riani et al (2009).

Utilizing multiple minima for a chosen objective function is not necessarily a new finding. Examples are given in say Rocke and Woodruff (1999). It is intuitively reasonable to justify that the majority of works in a solution to classification/clustering problems are simply those of optimization; with techniques such as k-means, Gaussian mixture models, Mean-shift or perhaps support-vector machines that utilize this in some capacity. In the case of ATLA, the employed objective function, which is optimally chosen based on the sample size, is evaluated based on the occurrence (or lack there-of) of minima. That is, observations are deemed as outlying based on the minimum of any minima occuring for an $\alpha > 0$ corresponding to the proportion of trimming. Otherwise if no such minima exist then the data set may be considered outlier free. Here we argue the benefit of incorporating plots of the objective function to assist in divulging the structure of the data, which may be clustered (based on occurrence of multiple minima), and by monitoring the search.

3.1 Objective functions of ATLA

Now one of the by-products of not utilizing a particular stopping criteria, as in ATLA, is that the entire search must be conducted. While this results in a slight increase in computational cost, it on the other hand opens the possibility for graphical monitoring and also mitigates the possibility for erroneous miss-classifications that may be a result of type II error; an inevitable characteristic for large samples. For instance, algorithms BACON, and FS when including a good observation, leave no path for that observation to leave once included. The FSM procedure which is contained within the FSDA Toolbox for MATLAB also contains similar routines which allow for the entire search to be conducted and thus monitored.

Moreover the procedure utilized in ATLA allows observations to leave the basic subset which enables spurious subset inflations to be "corrected" in subsequent iterations. This also minimizes over-reliance on the initial robust location/scale estimate facilitating the path of dividing the basic subset and the non basic subset. It can be remarked here that an observation may leave the basic subset at any point in the FSM algorithm.

Through existence of multiple minima (for an $\alpha > 0$) in the objective function, one is able to discern the possible existence of multiple contaminating distributions/clusters. It is possible to demonstrate this with a simulation of clustered data composed of a total of five clustered samples with $n = 500$ generated as follows:

- Cluster 1: $\mathcal{N}_3(\mathbf{0}, I_3)$ of size $n_1 = 275$ representing the majority population.
- Cluster 2: $\mathcal{N}_3(\mathbf{1} \cdot \sqrt{\chi^2_{0.975,3}}, 0.1 \cdot I_3)$ of size $n_2 = 50$.
- Cluster 3: $\mathcal{N}_3([0, 0, -2.5 \cdot \sqrt{\chi^2_{0.975,3}}], I_3)$ of size $n_3 = 75$.
- Cluster 4: $\mathcal{N}_3([0, 0, 5 \cdot \sqrt{\chi^2_{0.975,3}}], I_3)$ of size $n_4 = 75$.
- Cluster 5: $\mathcal{N}_3([0, 4 \cdot \sqrt{\chi^2_{0.975,3}}, 1.5 \cdot \sqrt{\chi^2_{0.975,3}}], \boldsymbol{\Sigma})$ of size $n_5 = 25$ where $\boldsymbol{\Sigma} = \text{diag}(1, 0.1, 0.1)$.

Figure 2 shows the pairwise scatterplot matrix for simulated data generated from these distributions. As one can gauge from such plots, Cluster 2 with points shown as yellow △-symbols are representative of what is referred to as a *point mass cluster* while points in Cluster 5 shown as purple ⊠-symbols follow a *line mass cluster*.

Now performing the multivariate ATLA procedure on this simulated data it is possible to plot the objective function over successive basic subsets. Looking at Fig. 3 one is able to discern the existence of five minima at relative basic subset sizes of 275, 325, 400, 475 and 500 respectively. The latter four minima correspond to subsets classifications that contain the exact cumulative cluster distributions whence they were originally generated from.

By allowing the original ATLA procedure to iteratively classify and trim each of the subsets based on the objective function criterion (Fig. 4) than this results in a sim-

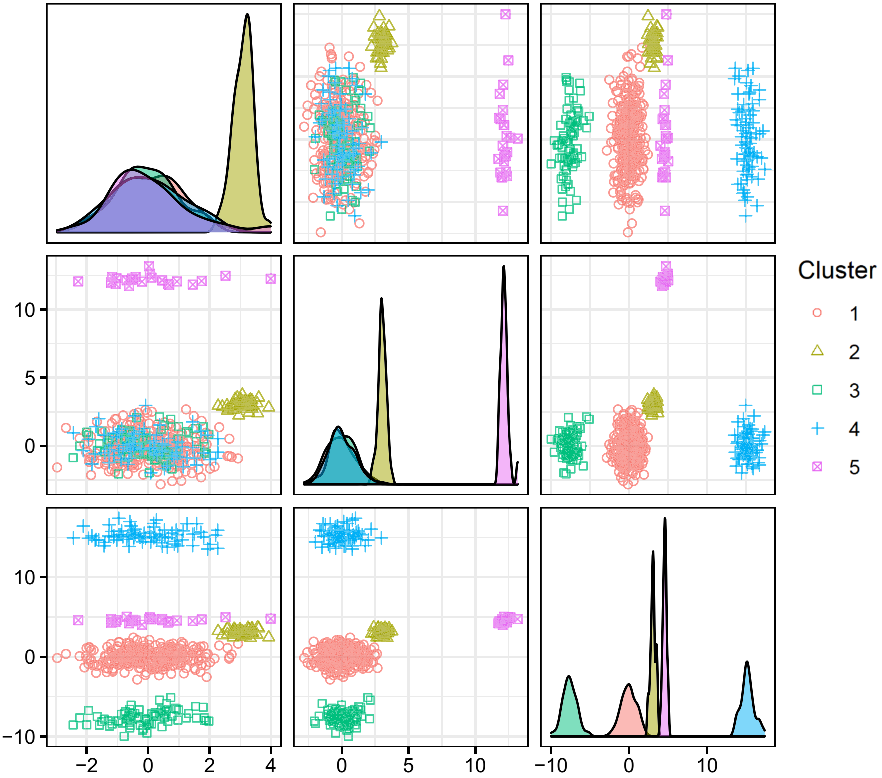

Fig. 2 Scatterplot matrix of simulated data containing 5 sample clusters generated out of different distributions with point coloring and symbols to highlight the cluster whence it belongs to

ilar final cluster classification. Doing so also ensures the optimal objective function proposal is used for a given application upholding the adaptive nature of the algorithm. Although this methodology would not be recommended for large data sets due to computational expense, it does present the procedure and motivation behind the optimization criteria for the objective function proposals.

In addition to the classification of these clusters, either through multiple application trimmings or minima subset comparisons, one is able to identify intra-cluster outliers. This can be observed in the simulated data shown in Fig. 5 which presents four identified intra-cluster outliers and the exact classification of clusters whence they originally belong to.

Here it is important to emphasize that the T1 and T2 proposals of Clarke and Schubert (2006) assume samples are taken from populations which follow a multivariate normal distribution, hence departures from such distribution will not guarantee results. It can be noted that this is an ad-hoc feature of the algorithm and not the sole intended purpose. Due to its construction of finding an initial subset of size $h = \lfloor (n+p+1)/2 \rfloor$, successful cluster discrimination may only be possible for clusters which are smaller than this value. Nevertheless, relaxation of these restrictions are elementary yet come

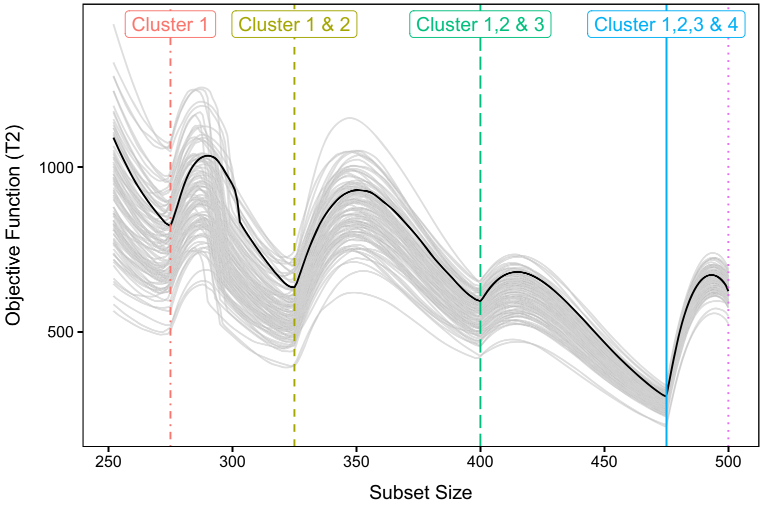

Fig. 3 Objective Functions of ATLA based on $N = 100$ simulated datasets generated from the clustered distributions highlighted above. The darker shaded line corresponds to the dataset shown in Fig. 2 with labels to demonstrate the occurrence of minima at basic subsets containing the exact associated cluster(s)

at the cost of increased probability of breakdown. Further research will benefit in the context of clustering and use of the forward search. Perhaps utilization of soft-trimming through weights in the ATLA procedure may prove useful for better cluster discrimination and monitoring.

4 Conclusion

The comparison and discussion of outliers here is limited to the case of the multivariate normal distribution. We do not entertain here the multivariate t-distribution, for example, which was also adequately explained in Clarke and Schubert (2006). Outliers are usually modelled at the multivariate normal distribution because of the central limit theorem. Further improvements in the ATLA algorithm were implemented based on developments in multivariate estimation in particular MCD and subsequently demonstrated in simulation results. We have explained the power of the three methods that use the forward search and there are varying terms of performance, with no outright winner. ATLA has a good all-round performance, vindicating its introduction in Clarke and Schubert (2006). To illustrate further the ATLA approach we apply it to clustering, albeit in an elementary dataset simulated out of predefined cluster distributions.

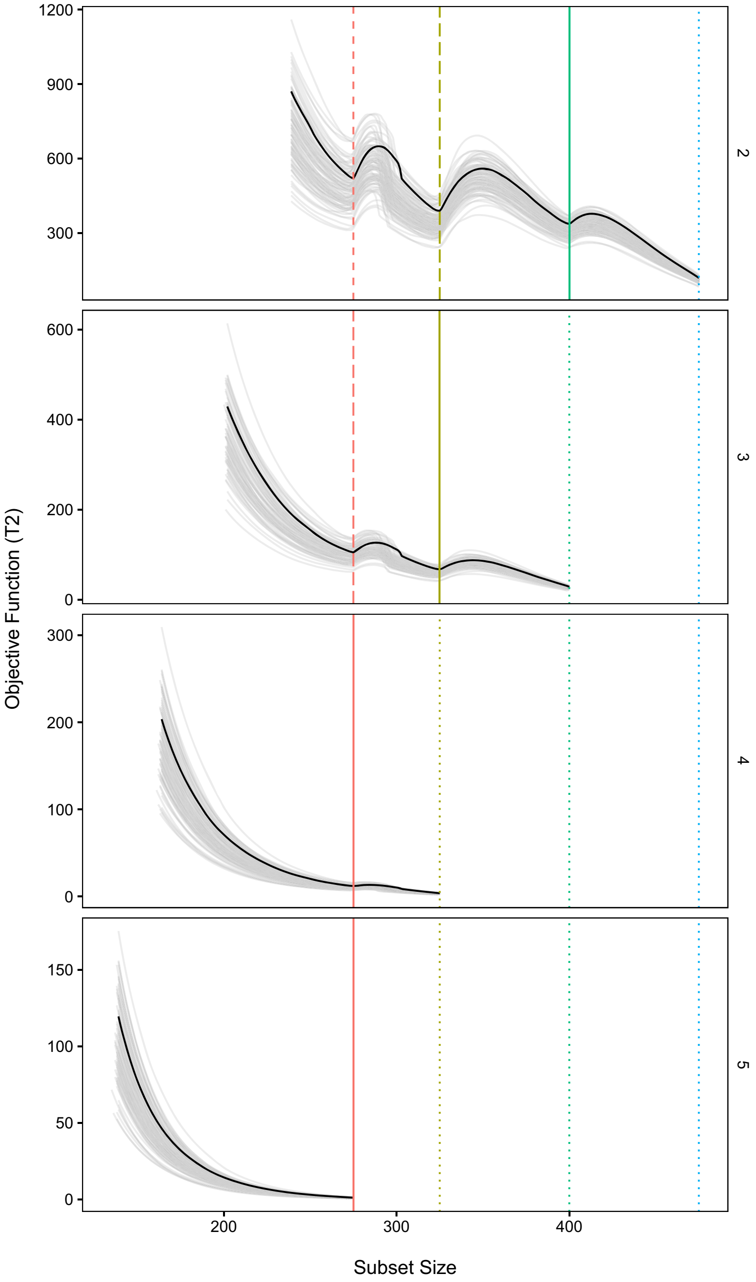

Fig. 4 Objective functions for successive applications of ATLA based on a continuation of the trimming of $N = 100$ simulated datasets as shown in Fig. 3. The solid lines denote the approximate minima which yields the final outlier classification/trimming (if any). Note the positioning of the vertical lines are consistent with cluster locations shown previously with each application imposing trimmings of size approximately 50, 75, 75 and 25 respectively

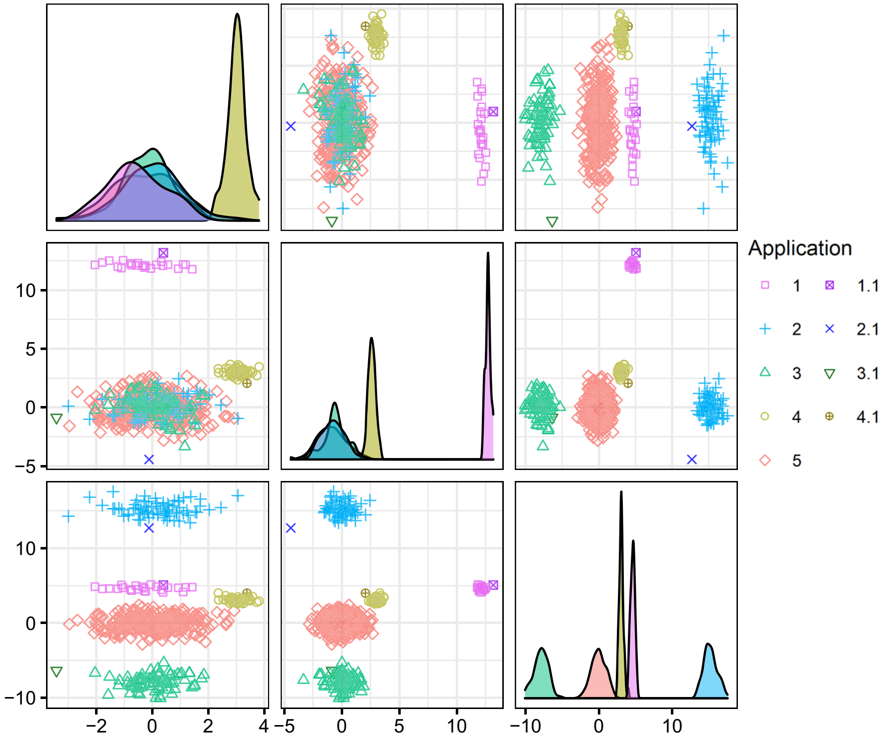

Fig. 5 Scatterplot matrix of a simulated dataset with point coloring to denote the cluster designation obtained through multiple applications of ATLA. Four points shown as ⊠, ×, ▽ and ⊕-respectively, correspond to outliers found when performing ATLA on trimming subsets discovered after each application

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00362-022-01319-7.

Acknowledgements The authors would like to acknowledge the use and explanation of the published work in the original ATLA algorithm discussed in Clarke and Schubert (2006). Daniel Schubert who received his doctorate for his thesis (2005) passed away in 2007. The authors are also grateful for insightful comments by referees which helped improve the paper.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

References

Atkinson AC, Riani M (2004) The forward search and data visualisation. Comput Stat 19(1):29–54

Atkinson AC, Riani M, Cerioli A (2003) Exploring multivariate data with the forward search. Springer, New York

Atkinson AC, Riani M, Cerioli A (2018) Cluster detection and clustering with random start forward searches. J Appl Stat 45(5):777–798

Barnett V, Lewis T (1994) Outliers in statistical data, 3rd edn. Wiley, New York

Bednarski T, Clarke BR (1993) Trimmed likelihood estimation of location and scale of the normal distribution. Austral J Stat 35:141–153

Bednarski T, Clarke BR (2002) Asymptotics for an adaptive trimmed likelihood estimator. Statistics 36:1–8

Billor N, Hadi AS, Vellamen PF (2000) BACON: blocked adaptive computationally efficient outlier nominators. Comput Stat Data Anal 34:27–298

Butler RW (1982) Nonparametric interval and point prediction using data trimmed by a Grubbs-type outlier rule. Ann Stat 10:197–204

Butler RW, Davies PL, Jhun M (1993) Asymptotics for the minimum covariance determinant estimator. Ann Stat 21:1385–1400

Cabana E, Lillo RE, Laniado H (2021) Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. Stat Pap 62:1583–1609

Cator EA, Lopuhaä HP et al (2012) Central limit theorem and influence function for the mcd estimators at general multivariate distributions. Bernoulli 18(2):520–551

Cerioli A, Farcomeni A, Riani M (2014) Strong consistency and robustness of the forward search estimator of multivariate location and scatter. J Multivar Anal 126:167–183

Cerioli A, Riani M, Atkinson AC, Corbellini A (2018) The power of monitoring: how to make the most of a contaminated multivariate sample. Stat Methods Appl 27:559–587

Cerioli A, Farcomeni A, Riani M (2019) Wild adaptive trimming for robust estimation and cluster analysis. Scand J Stat 46(1):235–256

Clarke BR (1994) Empirical evidence for adaptive confidence intervals and identification of outliers using methods of trimming. Austral J Stat 36:45–58

Clarke BR (2000) An adaptive method of estimation and outlier detection applicable for small to medium sample sizes. Discussiones Mathematicae 20:25–50

Clarke BR (2018) Robustness theory and application, 1st edn. Wiley, Hoboken, NJ

Clarke BR, Schubert DD (2006) An adaptive trimmed likelihood algorithm for identification of multivariate outliers. Austral New Zealand J Stat 48:353–371

Clarke BR, Höller A, Müller CH, Wamahiu K (2017) Investigation of the performance of trimmed estimators of the life time distributions with censoring. Austral New Zealand J Stat 59:513–525

Filzmoser P, Ruiz-Gazen A, Thomas-Agnan C (2014) Identification of local multivariate outliers. Stat Pap 55:29–47

Garciga C, Verbrugge R (2021) Robust covariance matrix estimation and identification of unusual data points: new tools. Res Econ 75(2):176–202

Hadi AS (1992) Identifying multiple outliers in multivariate data. J R Stat Soc Ser B (Methodol) 54:761–771

Hadi AS (1994) A modification of a method for the detection of outliers in multivariate samples. J R Stat Soc Ser B (Methodol) 56:393–396

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

Hubert M, Rousseeuw PJ, Verdonck T (2012) A deterministic algorithm for robust location and scatter. J Comput Gr Stat 21(3):618–637

Krazanowski WJ (1988) Principles of multivariate analysis. Clarendon Press, ISBN0 19:852,230

Leys C, Klein O, Dominicy Y, Ley C (2018) Detecting multivariate outliers: use a robust variant of the Mahalanobis distance. J Exp Soc Psychol 74:150–156

Maechler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller M, C ELT, Anna di Palma M (2019) robustbase: Basic Robust Statistics. http://robustbase.r-forge.r-project.org/, r package version 0.93-5

Mahalanobis PC (1936) On the generalized distance in statistics. Natl Inst Sci India 2:49–55

Rao CR (1973) Linear statistical inference and its applications, vol 2. Wiley, New York

Riani M, Atkinson AC, Cerioli A (2009) Finding an unknown number of multivariate outliers. J R Stat Soc: Ser B (Stat Methodol) 71(2):447–466

Riani M, Atkinson AC, Perrotta D (2014) A parametric framework for the comparison of methods of very robust regression. Stat Sci 29(1):128–143. https://doi.org/10.1214/13-STS437

Rocke DM, Woodruff DL (1999) A synthesis of outlier detection and cluster identification. Technical Report, University of California, Davis, CA

Rousseeuw PJ (1983) Multivariate estimation with high breakdown point. In: Grossman W, Pflug G, Vincze I, Wertz W (eds) Mathematical statistics and applications (1985), vol B. Reidel Publishing Co, Dordrecht, pp 283–297

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

Schubert DD (2005) A multivariate adaptive trimmed likelihood algorithm. PhD thesis, Murdoch University, Murdoch, Western Australia

Stahel W, Maechler M (2019) 'eXtra' / 'eXperimental' Functionality for Robust Statistics. https://CRAN.R-project.org/package=robustX , version 1.2-4

Todorov V, Sordini E (2020) fsdaR: robust data analysis through monitoring and dynamic visualization. https://CRAN.R-project.org/package=fsdaR, r package version 0.4-9

Wilks SS (1963) Multivariate statistical outliers. Sankhyā Indian J Stat Ser A (1961-2002) 25(4):407–426

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.