Stat Methods Appl
https://doi.org/10.1007/s10260-017-0409-8

ORIGINAL PAPER

The power of monitoring: how to make the most of a contaminated multivariate sample

Andrea Cerioli[1] · **Marco Riani**[1] · **Anthony C. Atkinson**[2] · **Aldo Corbellini**[1]

Accepted: 10 November 2017

Abstract Diagnostic tools must rely on robust high-breakdown methodologies to avoid distortion in the presence of contamination by outliers. However, a disadvantage of having a single, even if robust, summary of the data is that important choices concerning parameters of the robust method, such as breakdown point, have to be made prior to the analysis. The effect of such choices may be difficult to evaluate. We argue that an effective solution is to look at several pictures, and possibly to a whole movie, of the available data. This can be achieved by monitoring, over a range of parameter values, the results computed through the robust methodology of choice. We show the information gain that monitoring provides in the study of complex data structures through the analysis of multivariate datasets using different high-breakdown techniques. Our findings support the claim that the principle of monitoring is very flexible and that it can lead to robust estimators that are as efficient as possible. We also address through simulation some of the tricky inferential issues that arise from monitoring.

✉ Andrea Cerioli
andrea.cerioli@unipr.it

Marco Riani
mriani@unipr.it

Anthony C. Atkinson
a.c.atkinson@lse.ac.uk

Aldo Corbellini
aldo.corbellini@unipr.it

[1] Department of Economics and Management, University of Parma, Via Kennedy 6, 43125 Parma, Italy

[2] Department of Statistics, The London School of Economics, London WC2A 2AE, UK

Published online: 23 November 2017

Keywords Data movie · Forward search · Outlier detection · MM-estimation · S-estimation · Trimming · Reweighting

1 Introduction

Assessing the effect of each individual observation on the result of a statistical analysis should be an essential ingredient of any applied statistical work. This goal is typically out of reach for classical diagnostic techniques, either from a model-based or a geometric perspective, since they can be grossly distorted in the presence of contamination by outliers, or under systematic deviation from the postulated data generating mechanism (Maronna et al. 2006; Huber and Ronchetti 2009; Avella-Medina and Ronchetti 2015; Farcomeni and Greco 2015). As a consequence, frequent examples can be found that use numerical and graphical inspection of robust residuals in regression and of robust Mahalanobis distances with multivariate data; see, e.g., Hubert et al. (2008) for an overview. Cerioli et al. (2009), Cerioli (2010) and Salini et al. (2016) show how to calibrate the robust diagnostics in order to obtain valid inferential conclusions in the case of small and moderate sample sizes, when asymptotic results are not reliable, thus enhancing their practical usefulness. Modern developments include the bagdistance map of Hubert et al. (2015) for the identification of multivariate functional outliers, regularized versions of the robust diagnostics to be used when the number of variables is large with respect to the sample size (Alfons et al. 2013; Boudt et al. 2017; Atkinson et al. 2017a) and extensions to non-normal models (Agostinelli et al. 2014; Amiguet et al. 2017).

However, the use of diagnostic tools derived from robust methods may not be entirely satisfactory. These tools, like those from the classical approach, typically end up with a single picture of the data, even if it may be uncorrupted. A persistent disadvantage of having a single summary is that several important choices have to be made prior to the analysis and their effect on the results is then difficult to evaluate. Among these choices, one crucial aspect is the selection of the precise value of the breakdown point, i.e. the fraction of contamination that the robust method is expected to tolerate, with large values (close to 50%) leading to high robustness, but also to low statistical efficiency. Other important features that can potentially affect the outcome include the selection of a specific robust technique among several well-established alternatives, and—as an extra level of complexity—further specific choices within the selected method, such as the downweighting function in soft trimming methods like S-estimation. In addition, each method requires a series of tuning constants for the numerical procedure (such as the number of subsets to extract, the number of refining steps, the number of best solutions to bring to full convergence, etc.) which have to be decided.

Some of the shortcomings described above can be overcome if we look at several pictures, and possibly to a whole movie, of the available data. We obtain this by monitoring the results computed from the selected estimator, by repeating the estimation process for different choices of the tuning parameters. We find that the simplest and most informative "data movies" are those achieved by varying the breakdown point, or the efficiency, of the estimator and we extensively explore this possibility in subsequent sections.

We became acquainted with the idea of monitoring more than twenty years ago through the forward search and, until recently, our monitoring experience has been limited to the development of this approach in a variety of contexts; see, e.g., Cerioli and Riani (1999); Atkinson and Riani (2000); Riani and Atkinson (2001); Atkinson et al. (2004); Riani et al. (2009); Atkinson et al. (2010); Riani et al. (2014c, 2015). However, it is now clear to us that the potential for monitoring is much wider, as the underlying idea can be extended to many other techniques. Riani et al. (2014a) and Cerioli et al. (2016) consider the case of robust estimation in regression, while in this work we focus on multivariate problems. A related multivariate methodology is the generalized radius process of García-Escudero and Gordaliza (2005), which uses ellipsoids of decreasing radius to define increasing levels of trimming. However, all these radii are computed from the same robust estimate and thus do not share the adaptive, i.e. data-driven, choice of the breakdown point of our monitoring approach. Adaptive trimming has also been advocated by Clarke and Schubert (2006), but not in connection with monitoring. Furthermore, Dotto et al. (2017) have recently proposed a data-driven approach to fix an appropriate trimming level in a clustering framework. An important bonus of monitoring, in our opinion, is that it conveys the idea of a (visual or numerical) comparison between the subsequent estimates and the related diagnostic measures. This paper is intended to show that such a comparison can have beneficial consequences in most practical implementations of the methodology, thus providing a positive step toward the hoped-for assessment of the effect of each individual observation.

Our primary goal is to support the “philosophy” of monitoring by showing the information gain that it provides in the analysis of complex data structures. This goal is reached by first reviewing, in a monitoring framework, some key ideas related to the forward search. The information provided by monitoring is there enhanced by the graphical tools for brushing and linking plots that are included in the FSDA Matlab toolbox (http://www.riani.it/MATLAB). Then, we extend the idea of monitoring to two popular classes of robust multivariate estimators. With all these techniques our approach is shown in action in four examples, for which the data can be found at the web site http://www.riani.it/smap17/, together with a Matlab file that allows the user to reproduce all the figures given in the paper. Although we are mainly oriented to the development of effective diagnostic tools to be used in real-world applications of robust statistical methods, we also address through simulation some of the tricky inferential issues that arise as a consequence of monitoring. Indeed, we see the development of a unified inferential framework for our monitored estimators, along the lines of Cerioli et al. (2014) and of Johansen and Nielsen (2016a, b), to be a challenging and compelling research goal for the future.

2 The forward search

2.1 Key ideas and Mahalanobis distances

The forward search (FS) provides an automatic form of monitoring. In this approach we start by fitting a small and supposedly homogenous subset of observations, often chosen through some robust criterion. The fitting subset is then repeatedly augmented in such a way that outliers and other influential observations enter toward the end of the algorithm. Even more importantly, their inclusion is typically signaled by a sharp increase in suitably selected diagnostic measures. It is thus very natural to monitor the values of such measures as the search progresses from the small starting subset to the final fit that corresponds to the classical statistical summary of the data.

The method of Riani et al. (2009) provides outlier tests for the FS with specified simultaneous size and good power when the sample comes from a single multivariate normal population, potentially contaminated by outliers. Atkinson et al. (2017b) exemplify this method and also illustrate its extension to the clustering of multivariate data. In the latter instance additional problems arise, such as the requirement of several random starting points and the use of trimming levels much larger than the usual bound of 0.5 (Cerioli et al. 2017). Therefore, we do not here address the problem of clustering in detail, but keep the basic assumption of a single multivariate normal population for the uncontaminated part of the data. Nevertheless, as the examples in Sects. 5–7 show, multi-population problems can still be solved through our outlier detection approach, provided that at least half of the observations come from the same population.

The search for a single population starts from a subset of m_0 observations, say $S^*(m_0)$, robustly chosen. The size of the fitting subset is increased from m to $m+1$ by forming the new subset $S^*(m+1)$ from those observations in the whole sample with the $m+1$ smallest squared Mahalanobis distances when the parameters are estimated from $S^*(m)$. Thus, some observations in $S^*(m)$ may not be included in $S^*(m+1)$. For each m $(m_0 \leq m \leq n-1)$, the test for the presence of outliers is based on the observation outside the subset with the smallest squared Mahalanobis distance.

The parameters μ and Σ of the v-dimensional multivariate normal distribution of y are estimated in the FS by the standard unbiased estimators from a subset of m observations, providing estimates $\hat{\mu}(m)$ and $\hat{\Sigma}(m)$. Using these estimates we calculate n squared Mahalanobis distances

$$d_i^2(m) = \{y_i - \hat{\mu}(m)\}' \hat{\Sigma}^{-1}(m)\{y_i - \hat{\mu}(m)\}, \qquad i = 1, \ldots, n. \tag{1}$$

To detect outliers we use the minimum Mahalanobis distance amongst observations not in the subset

$$d_{\min}(m) = \min d_i(m) \quad i \notin S^*(m). \tag{2}$$

Testing for outliers requires a reference distribution for $d_i^2(m)$ in (1) and hence for $d_{\min}(m)$ in (2). When Σ is estimated from all n observations, the squared statistics have a scaled beta distribution. However, the estimate $\hat{\Sigma}(m)$ in the search uses the central m out of n observations, so that the variability is underestimated. Results of Tallis (1963) on truncated distributions provide a scaling factor

$$c(m, n) = \frac{n}{m} C_{v+2}\{\chi^2_{v,m/n}\}, \tag{3}$$

where $C_r(y)$ is the c.d.f. of the χ^2 distribution on r degrees of freedom evaluated at y and $\chi^2_{r,\zeta} = C_r^{-1}(\zeta)$, for $0 < \zeta < 1$, is the ζth quantile of the same distribution. Then the scaled and asymptotically unbiased estimate of Σ is

$$\hat{\Sigma}^{\mathrm{sc}}(m) = \frac{1}{c(m,n)}\hat{\Sigma}(m).$$

The scaled minimum Mahalanobis distance $d^{\mathrm{sc}}_{\min}(m)$ follows from (2) when $\hat{\Sigma}(m)$ in (1) is replaced by $\hat{\Sigma}^{\mathrm{sc}}(m)$. In Eqs. (4) and (5) below, we show how the consistency factor $c(m,n)$ is taken into account in order to obtain distributional results for the unscaled minumum Mahalanobis distance $d_{\min}(m)$, which is the standard diagnostic tool in the FS.

2.2 Monitoring plots, envelopes and multiple testing for outlier detection

Atkinson et al. (2004, pp. 43–44) give results on the distribution of deletion Mahalanobis distances from a sample of size n. These results yield that the required quantile of order γ (say d_γ) of the distribution of the minimum Mahalanobis distance (2) is given by

$$d_\gamma = \sqrt{\frac{v(m^2-1)}{m(m-v)c(m,n)}F^{-1}_{v,m-v}\left(\frac{m+1}{m+1+(n-m)F^{-1}_{2(n-m),2(m+1)}(1-\gamma)}\right)}, \tag{4}$$

where $F_{a,b}(y)$ is the c.d.f. of the F distribution with a and b degrees of freedom evaluated at y, and $c(m,n)$ is the scaling factor given in (3). Correspondingly, for $d^* > 0$,

$$P\left[\{d_{\min}(m)\}^2 \le d^*\right] = 1 - F_{2(n-m),2(m+1)}\left(\left[\frac{1}{F_{v,m-v}\left\{\frac{m(m-v)}{v(m^2-1)}c(m,n)d^*\right\}} - 1\right]\frac{m+1}{n-m}\right). \tag{5}$$

As we shall see, it is extremely helpful to look at forward plots of quantities of interest such as $d_{\min}(m)$ during the search and to compare them with the envelopes from several values of γ. Such monitoring plots, drawn for series of values of m, are exceptionally rich in information about departures of the data from the assumed structure.

For precise outlier identification we perform a series of tests, one for each $m \ge m_0$. To allow for multiple testing, we use a rule which depends on the sample size to determine the relationship between the envelopes calculated for the distribution of $d_{\min}(m)$ and the significance of the observed values. If at some point $m^\dagger$ in the search the nearest observation to those already in the subset appears to be an outlier, as judged by an appropriate envelope of the distribution of the test statistic, we call this a “signal”. Appearance of a signal indicates that observation $m^\dagger$, and the remaining observations not in the subset, may be outliers. But, we need to judge the values of the statistics against envelopes from appropriately smaller population sizes that exclude potential outliers. The second stage of the analysis consists of superimposing envelopes for a series of smaller sample sizes $n^\dagger$, starting from $m^\dagger - 1$ onwards, until the first introduction of an observation recognised as an outlier. The details of the procedure are described in Riani et al. (2009).

In this paper we also look at monitoring plots of all n squared Mahalanobis distances as m increases. As we show, these plots can be combined with brushing to relate Mahalanobis distances to data points exhibited in scatterplot matrices, thus making a closer connection between statistical results and individual observations.

2.3 Regression

The structure of the FS for regression is similar to that for multivariate data. Although we are not here concerned with regression, there are some computational advances in regression which are incorporated in our paper. Instead of scaled Mahalanobis distances, the test statistic for outlyingness in regression is the deletion residual (Atkinson and Riani 2000, Chapter 2) with the estimate of the error variance σ^2 scaled by a consistency factor similar to (3), but for a sample of univariate normal observations. Now the distribution of the test statistic is Student's t (Riani and Atkinson 2007), to give the analogue of (5). Allowing for these changes, the procedure for outlier detection again involves a signal and resuperimposition of envelopes (see, e.g., Riani et al. 2014c).

3 Other robust methods for multivariate data

The FS is one of several methods for detecting outliers in multivariate data. We compare our analyses with results from monitoring high-breakdown techniques in which extreme observations are either downweighted by a function ρ or trimmed. Extended discussion of these methods is given in Maronna et al. (2006). We again use brushing and linking, in conjunction with monitoring, to highlight the effect of each individual observation on inference.

3.1 S-estimation

In estimation of Mahalanobis distances, as in (1), the estimate of the mean μ does not depend on the estimate of Σ. However, this is not the case in such a robust method as S-estimation. This is derived from M-estimation (see Huber and Ronchetti 2009), in which the downweighting function ρ is used with the variance assumed known. To make possible an estimate of scale, the covariance matrix Σ is rewritten as $\Sigma = \sigma^2 \Gamma$, with $|\Gamma| = 1$. For given σ^2, the estimates of μ and Γ minimize the objective function

$$\sum_{i=1}^{n} \rho \left\{ d_i^2(\mu, \Gamma)/\sigma^2 \right\}, \tag{6}$$

where ρ is a function that reduces the importance of observations with large Mahalanobis distances. The robust estimate of the squared scale, say $\tilde{\sigma}^2$, is found by solution of the equation

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{d_i^2(\mu,\Gamma)}{\sigma^2}\right)=K, \tag{7}$$

where $0 < K < \sup\rho$. Taking the minimum value of $\tilde{\sigma}^2$ which satisfies equation (7) yields the S-estimate of squared scale ($\tilde{\sigma}_S^2$) and the associated estimates of μ and Γ ($\tilde{\mu}_S$ and $\tilde{\Sigma}_S$).

Some properties of the class of functions ρ are important for the robustness of the estimator. Specifically, we focus on the replacement version of the breakdown point, which is defined as the smallest fraction of outliers that can take the estimate over all bounds; see, e.g., Rousseeuw and Leroy (1987, §2) and Farcomeni and Greco (2015, p. 10). Rousseeuw and Leroy (1987, p. 139) show that if ρ satisfies the following conditions:

1. It is symmetric and continuously differentiable, and $\rho(0)=0$;
2. There exists a $c > 0$ such that ρ is strictly increasing on $[0, c]$ and constant on $[c, \infty)$;
3. It is such that

$$K/\rho(c) = \text{bdp, with } 0 < \text{bdp} \leq 0.5, \tag{8}$$

the breakdown point of the S-estimator tends to bdp when $n \to \infty$. As c increases, fewer observations are downweighted, so that the estimate of σ^2 approaches that for maximum likelihood estimation and bdp → 0. For consistency when the errors are normally distributed, we require

$$K = E_{\Phi_{0,1}}\left[\rho\left(\frac{d_i^2}{\tilde{\sigma}^2}\right)\right], \tag{9}$$

where $\Phi_{0,1}$ is the c.d.f. of the standard normal distribution. It is also possible to rescale ρ (see, e.g., Maronna et al. 2006, p. 31). If $\rho(x)$ is normalized in such a way that $\rho(c)=1$, the constant K gives the asymptotic value of the breakdown point of the S-estimator. If we fix bdp it follows from (8) and (9) that c and K are determined. The exact relationship will depend upon the function ρ, but Riani et al. (2014b, §3.1) show how to obtain computationally efficient calculations for finding the value of c once the value of bdp is specified. Although the ρ functions in (6) and (7) may be different, in our calculations we use the same ρ for both equations and, specifically, we take ρ as Tukey's biweight.

3.2 MM-estimation

The results of Riani et al. (2014b) show an asymptotic relationship between the breakdown point and efficiency of S-estimators; as one increases, the other decreases. In an attempt to break out of this relationship, Yohai (1987) introduced MM-estimation, which can be seen as a two-step extension of S-estimation. In the first stage the breakdown point of the scale estimate is set at 0.5, thus ensuring high robustness. This fixed estimate is then used to obtain new estimates of μ and Γ, for which K can be chosen to provide high efficiency. We start the empirical analyses that follow by taking a value of 0.99 for this efficiency. However, we of course look over a range of values when we monitor MM-estimates. In this paper we always refer to location efficiency, but our ideas could also be applied to scale efficiency.

It is worth noting that at present there is not a universal recipe for which level of efficiency must be used. Maronna et al. (2006, §5.9) recommend an efficiency of 0.85 as a generally safe choice in regression problems, even if MM-estimators are often advocated with an efficiency of 0.95 or 0.99. The aim of our approach is to reach a data-driven balance between robustness and efficiency in MM-estimation. Automatic computation of this balance clearly enhances the practical advantages of estimation using several trial values, as envisaged by Maronna et al. (2006, p. 144).

3.3 Methods using hard trimmed estimates of the covariance matrix

The contours of constant squared Mahalanobis distances form ellipsoids in v-dimensional space. This simple geometric interpretation suggests two further estimators of μ and Σ found by "hard" trimming. That is, the number of observations h to be used in fitting is decided before the data are analysed, although, of course, which $n - h$ observations are to be trimmed is a matter of calculation. The estimators that we consider are the minimum volume ellipsoid (MVE) and the minimum covariance determinant (MCD) (Rousseeuw and Leroy 1987). In both methods the value of h is often taken as just greater than $n/2$ (see formula (10) below, with bdp $= 0.5$), yielding the highest possible finite-sample value of the breakdown point for an affine equivariant estimator (Davies 1987; Lopuhaä and Rousseeuw 1991). Larger values of h give more efficient estimates of the parameters but with lower breakdown point.

In accordance with our established approach we use monitoring to provide an adaptive estimate of the highest value of h which provides a robust fit; see Farcomeni and Greco (2015, §2.5 and §3.7) and Boudt et al. (2017, §5) for recent applications in the same direction, also envisaged by Croux and Haesbroeck (1999, p. 170). The MVE has the undesirable property that its consistency rate is only $n^{-1/3}$ and we see the effects of the resulting instability in monitoring plots such as Fig. 15. Finally, we also include in our comparisons the more efficient reweighted MCD estimate that is computed on a second subset of $h^* > h$ observations for which the squared robust distances computed from the raw MCD estimate are below a fixed threshold, often taken from the χ^2_v distribution.

In all instances of hard trimmed estimates of the covariance matrix a scaling factor similar to (3) must be used to ensure consistency in the absence of contamination. However, for simplicity we omit explicit reference to this scaling in the examples that follow.

4 Summary of empirical analysis

We use four examples to explore the properties and advantages of monitoring robust analyses. The first data example, in Sect. 5, is of 272 observations on successive eruptions of the 'Old Faithful' geyser in Yellowstone National Park, Wyoming. This example shows that monitoring MM analyses provides useful information, but that interpretation of the analysis with S-estimation is less straightforward. We accordingly, in Sect. 6, look at two simulated data sets, the first with comparatively few outliers, all remote, to help understand conditions under which monitoring is helpful in establishing the value of bdp for S-estimation. The second simulated example, in Sect. 6.2, has a higher proportion of outliers, none of which are particularly remote. This provides a very different assessment of the various methods of robust analysis, although useful information is still gathered by monitoring. This second simulated example has a structure related to that of our second data example in Sect. 7 which contains 488 four dimensional observations on cows with bovine dermatitis.

5 A first data example: eruptions of old faithful

The data are taken from the MASS library (Venables and Ripley 2002). There are 272 observations with y_{1i} the duration of the ith eruption and y_{2i} the waiting time to the start of that eruption from the start of eruption $i-1$. There are several similar data sets in the literature and we may thus take this example as a specimen for a much wider class of statistical applications. The related literature and the physics of the problem are discussed by Azzalini and Bowman (1990) who employ a time series analysis. Here we use multivariate analysis of the two-dimensional observations, so ignoring any time series structure. We assume that the bulk of the data come from a single bivariate (normal) population, for which we robustly estimate μ and Σ.

Figure 1 shows results from the analysis using S-estimation with (asymptotic) breakdown point of 50% and Tukey's biweight function. The left-hand panel shows that the estimator has found 82 outliers when these are identified at a pointwise level of 99%. The right-hand panel of the figure shows that two groups have been found. The larger group, taken as the main population, with higher values of the two variables is in general well separated from the smaller group of outliers. If this were a clustering problem it might be argued that a few of the observations between the two groups could be assigned to the smaller group and this argument could be explored using clustering methods. However, the very robust S analysis has revealed the salient features of the data.

In our initial MM implementation we have used an efficiency of 99%. The resulting analysis shows no outliers at all. We now use monitoring to determine, for example, whether the choice of 99% efficiency is too optimistic.

To monitor the parameters of the procedures, bdp for S-estimation and efficiency for MM, we see how plots of the squared Mahalanobis distances for all n observations vary with the parameters. The left-hand panel of Fig. 2 shows a zoom of the monitoring plot for S-estimation. As we have already seen, for a bdp of 50% a robust analysis is obtained. However, the plot shows that, even for a bdp of 0.49, the analysis becomes non-robust. In order to emphasize the behaviour at the beginning of our monitoring we have zoomed the full plot which runs from 0.5 to 0.01. Over the range 0.49–0.01 the plot remains the same as it is for the greatest part of the left-hand panel of Fig. 2. Only for a bdp of 0.5 is a robust analysis obtained, which provides parameter estimates with poor efficiency.

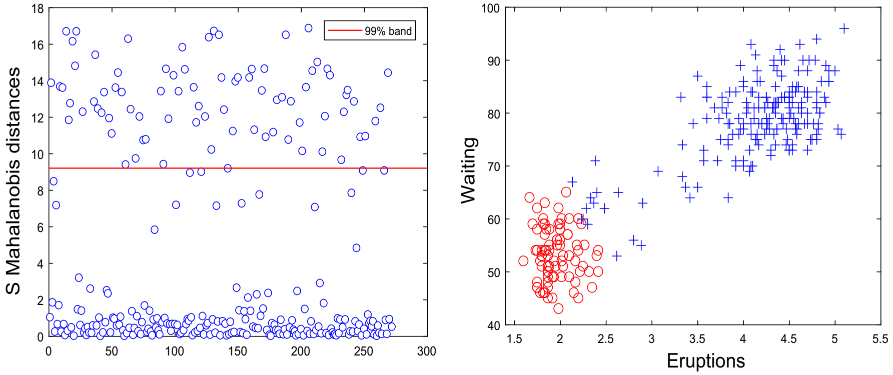

Fig. 1 Eruptions of Old Faithful, S-estimation with a bdp of 50%. Left-hand panel, index plot of squared Mahalanobis distances with threshold $\chi^2_{2,0.99}$; 82 outliers are identified. Right-hand panel, scatterplot matrix showing the identified outliers plotted as (red) circles (colour figure online)

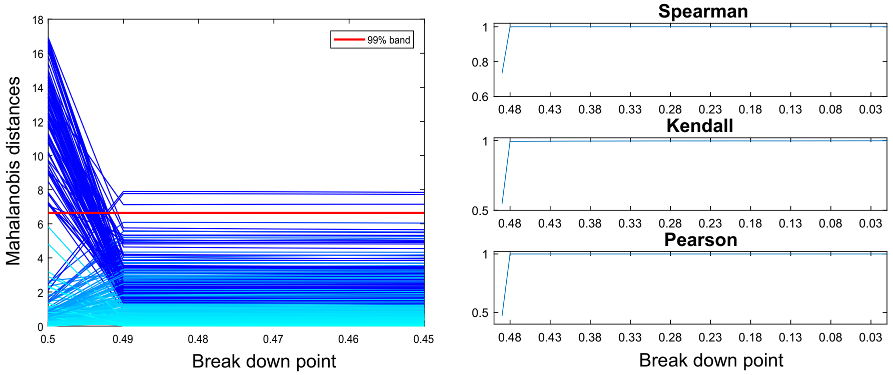

Fig. 2 Eruptions of Old Faithful. Left-hand panel, squared Mahalanobis distances from monitoring S-estimation—note the small range of bdp. Right-hand panel, monitoring correlation between consecutive distances

In all our plots monitoring Mahalanobis distances we use a colour map which goes from light blue (light grey in the black and white version) to dark blue (dark grey). The colour becomes darker as the maxima of the individual trajectories increase. Consequently, the eye is drawn to the behaviour of the most outlying units.

For simple structures, as here, there is a clear division of the solutions into a robust fit and a non-robust one, with a sharp break between them. For more complicated examples the point of transition is not so clearly visible. But in all cases we find that the structure of the plot is well summarized by the correlation of the squared Mahalanobis distances, or their ranks, at adjacent monitoring values. The right-hand panel of Fig. 2 shows the monitoring plot of three standard measures of correlation:

1. Spearman. Correlation between the ranks of the two sets of observations.
2. Kendall. Concordance of the pairs of ranks.
3. Pearson. Product-moment correlation coefficient.

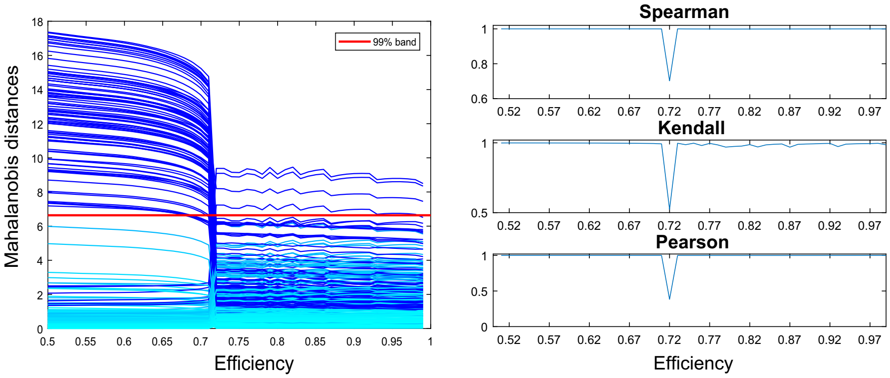

Fig. 3 Eruptions of Old Faithful. Left-hand panel, squared Mahalanobis distances from monitoring MM-estimation. Right-hand panel, monitoring correlation between consecutive distances

All three panels clearly indicate the failure of the robust procedure even for a bdp of 0.49.

Figure 3 is, on the other hand, a fine example of the power of monitoring. The left-hand panel shows the Mahalanobis distances for a series of robust MM fits. These are stable until an efficiency of 0.71, when the fit changes abruptly to one corresponding to the maximum likelihood estimate which remains stable as the efficiency increases to one. The right-hand panel shows the monitoring plots of the three correlation measures, all of which very sharply confirm 0.71 as the highest possible efficiency.

To extract further information from this plot, we show in Fig. 4 the effect of brushing the more extreme squared Mahalanobis distances in the stable left-hand part of the monitoring plot. The right-hand panel gives the scatterplot matrix for the units in the brush and those outside. The brushed units indeed correspond to the outliers. In comparison with the scatterplot matrix from S-estimation in Fig. 1, slightly more outliers are indicated (97 instead of 82) and the two groups are more similar in shape. All of this is of course a detail. What we have shown is how to find a robust MM-estimate with an adaptively established efficiency which is the greatest possible for these data. The choice of the greatest possible efficiency yields a clearer separation between the "main population" and the "outliers" than the robust but inefficient S fit with bdp $= 0.5$.

In both monitoring plots of the squared Mahalanobis distances we have included a horizontal (red) line corresponding to $\chi^2_{2,0.99}$. In the left-hand part of the left-hand panels of the figures this indicates many outliers. It is however a pointwise bound and its properties are not obvious (they are investigated by simulation in Sect. 8). It is interesting that in the right-hand part of both plots, where we have a non-robust fit, three outliers are indicated, in excellent agrement with the 1% band for outliers and a sample size of 272.

We now turn to hard trimming methods. The left-hand panel of Fig. 5 is an index plot of the squared Mahalanobis distances from the raw MCD with (asymptotic) breakdown point of 50%; 97 outliers are found. The right-hand panel shows the 93 outliers found after reweighting using (0,1) weights determined by comparison with $\chi^2_{2,0.99}$. More

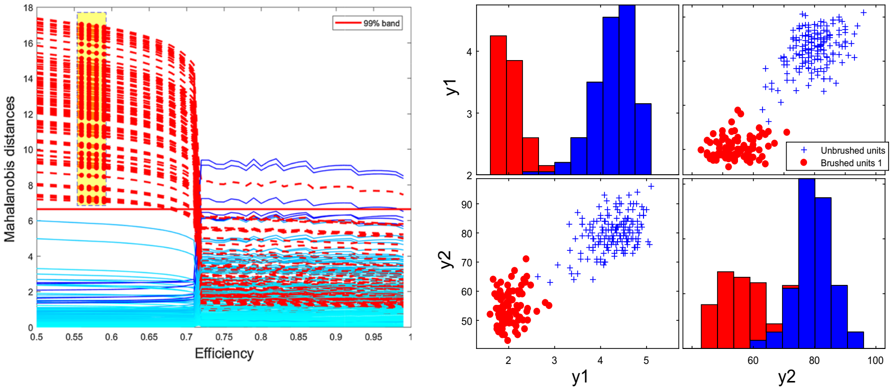

Fig. 4 Eruptions of Old Faithful. Left-hand panel, brushing the monitoring plot for MM-estimation (Fig. 3). Right-hand panel, scatterplot matrix of the units with the 97 most extreme squared Mahalanobis distances shown as filled red circles (colour figure online)

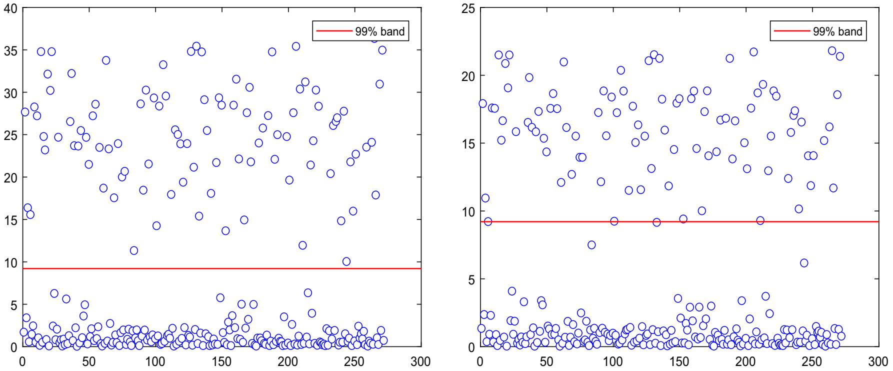

Fig. 5 Eruptions of Old Faithful, MCD analysis. Left-hand panel, index plot of squared Mahalanobis distances from raw MCD with 50% bdp; 97 outliers lie above the 99% band from χ^2_2. Right-hand panel, reweighted MCD; now 93 outliers are identified

accurate thresholds could be adopted for precise outlier identification (see Hardin and Rocke 2005; Cerioli 2010; Cerioli and Farcomeni 2011; Farcomeni and Greco 2015, §2), but we keep the simple (asymptotic) χ^2_2 approximation not to distract from the main goal of our work, which is the study of monitoring. We again refer to Sect. 8 for discussion of this issue. The scatterplot of the division into two groups for the raw MCD is identical to that for the brushed MM-estimator. However, the parameter estimates from raw MCD have a higher variance than those from MM-estimation.

In this example, the trimmed estimators can both be improved through monitoring. The left-hand panel of Fig. 6 shows the monitoring of the squared Mahalanobis distances for the MCD. As in the previous plots for S-estimation, the asymptotic value of the breakdown point (bdp) is reported on the horizontal axis. For each bdp ≤ 0.5 the estimates are computed using a subset of

$$h = \lfloor (1 - \text{bdp})(n + v + 1) \rfloor \tag{10}$$

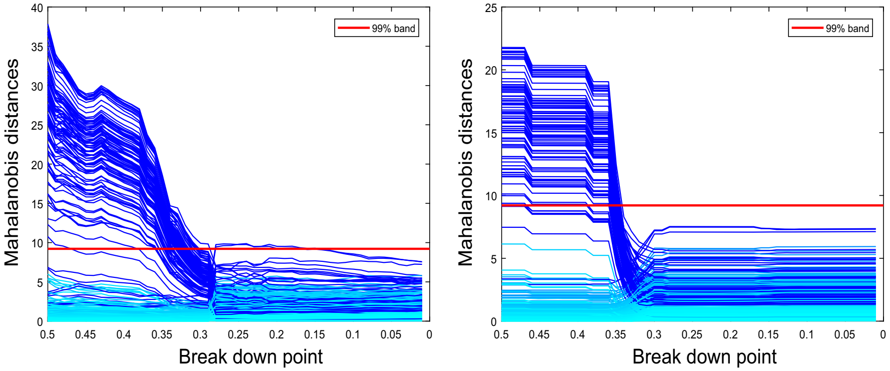

Fig. 6 Eruptions of Old Faithful. Left-hand panel, squared Mahalanobis distances from monitoring raw MCD estimation. Right-hand panel, monitoring the reweighted MCD (reweighting threshold $\chi^2_{2,0.99}$)

observations, where $\lfloor\ \rfloor$ denotes the floor function. The squared robust distances initially decrease steadily, while highlighting the same set of units. Then, at a bdp around 0.29 there is an abrupt change to a fit displaying two or three large distances which remains sensibly constant until the MLE is reached at bdp $= 0$.

A similar plot for the reweighted MCD is displayed in the right-hand panel of the figure. This is much more stable than that for the crude MCD until 0.37, at which bdp there is a collapse to the MLE. The right-hand parts of both panels are similar. However, the left-hand panel shows the distances for the crude MCD decreasing as successive observations are added to the subset used in fitting. On the other hand, the reweighted MCD shows three regions during which the distances are constant. In these regions the effect of changing the bdp in the (raw) first stage does not cause any change in the units chosen by the reweighting procedure.

Our monitoring approach also helps to appreciate the effect of the threshold used in the reweighting step. Figure 7 repeats monitoring of the squared Mahalanobis distances from the reweighted MCD when weights are determined by $\chi^2_{2,0.95}$ (left) and $\chi^2_{2,0.999}$ (right). Although the message conveyed by the two plots is broadly the same, the less efficient 95% threshold produces a few more outliers and a neater separation between the two populations, while increasing efficiency in the reweighting step causes the inclusion of some contaminated units at a slightly larger bdp than 0.37. We reach the same conclusions by looking at Fig. 8, which shows the 0.99 tolerance ellipses obtained through the two alternative reweighted estimates with bdp $= 0.5$. It is apparent that the less biased estimate of Σ computed on the main population of the left-hand panel is based on a smaller number of observations and correspondingly has lower statistical efficiency. We thus argue that monitoring can also help to select this additional tuning parameter, leading to the best data-specific balance between robustness and efficiency for the reweighted estimator.

The latter claim suggests the possibility of monitoring the robust Mahalanobis distances as a function of the reweighting probability itself, for a given value of bdp, in much the same way as we have seen for MM-estimation. Such a monitoring plot, not reported here, is fairly stable but indeed shows that the squared robust distances tend

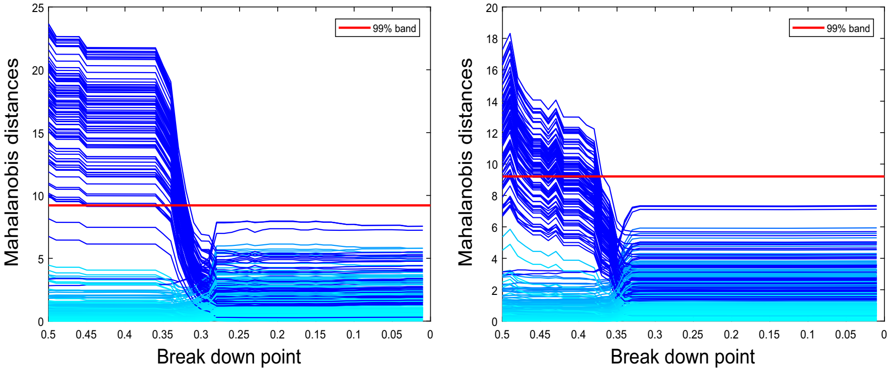

Fig. 7 Eruptions of Old Faithful. Squared Mahalanobis distances from reweighted MCD. Left-hand panel, reweighting threshold $\chi^2_{2,0.95}$. Right-hand panel, reweighting threshold $\chi^2_{2,0.999}$

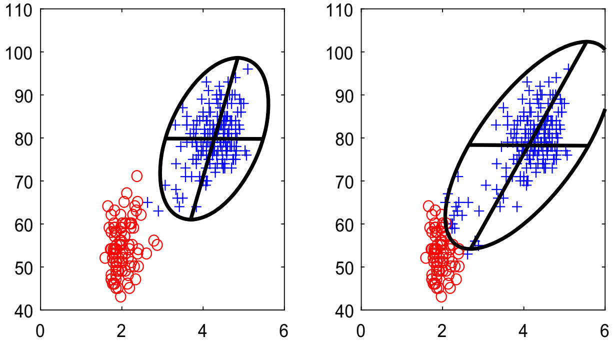

Fig. 8 Eruptions of Old Faithful. 0.99 tolerance ellipse for reweighted MCD based on $\chi^2_{2,0.95}$ (left-hand panel) and on $\chi^2_{2,0.999}$ (right-hand panel), with bdp $= 0.5$

to be smaller when the reweighting probability increases, as implied by the ellipses in Fig. 8. Further evidence of this behaviour is provided in the simulation study of Sect. 8.

We now turn to the totally adaptive and parameter free analysis from the FS. The left panel of Fig. 9 shows the forward plot of $d_{\min}(m)$. A signal is found at $m = 160$ and superimposition leads to the identification of 95 outliers. The scatterplot matrix, not shown, is virtually identical to that for monitored MM-estimation in Fig. 4. The shape of this trajectory is typical of that obtained from data with two clusters. The peak arises because the next unit to enter the search is remote from those in the cluster providing the subset of the search. After several units from the other cluster have entered this subset, the parameter estimates change and units in the second cluster no longer appear remote.

These ideas are cogently illustrated by the forward plot of squared Mahalanobis distances in the right-hand panel of Fig. 9. In the central part of the search, almost for m in the whole range 50–200, the two clusters are apparent. The lower set of distances are from the observations forming the first cluster. There is then a gap in the

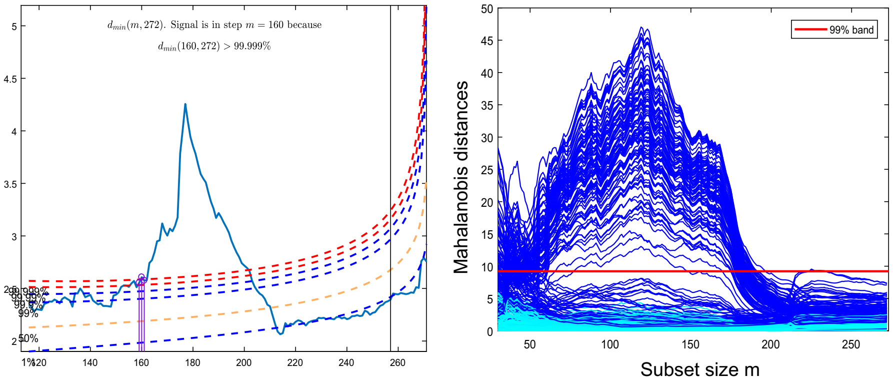

Fig. 9 Eruptions of Old Faithful; analyses with the FS. Left-hand panel, forward plot of minimum Mahalanobis distance illustrating outlier detection. Right-hand panel, forward plot of scaled squared Mahalanobis distances showing evidence of a cluster of outliers

plot, because observations in the second cluster are clearly separated from those in the first cluster. Although we found a signal at $m = 160$ the final number of members in the first group was found by resuperimposition to be 177. The figure shows that intense masking only occurs a little later, around $m = 201$. There are three relatively large Mahalanobis distances at the end of the search, which are from observations lying between the two groups. However, the present paper is about monitoring, not clustering. A fuller discussion of the clustering of these data is given by Atkinson and Riani (2007) and Cerioli et al. (2017).

6 Two examples with simulated data sets

6.1 Lightly contaminated data

In our first simulated example there are 200 five-dimensional observations, all simulated with standard normal co-ordinates. Thirty of the observations had a displacement of 2.4 added to each co-ordinate. As a result the outliers are grouped, with virtually no overlap with the central 170 observations.

Figure 10 shows the result of S-estimation using Tukey's biweight with a bdp of 0.5; 30 observations (all the contaminated units excluding unit 18 and plus unit 42) are declared as outliers when $\chi^2_{5,0.99}$ is used to perform the test. The right-hand panel of the figure shows the group of outliers found, virtually all of which have higher values of all co-ordinates than the observations not declared as outlying. A question to be answered by monitoring is whether the same structure can be revealed by a lower value of bdp, leading to estimates of μ and Σ with higher efficiency that do not require preliminary outlier removal.

Figure 11 shows the result of MM-estimation with an efficiency of 99%. This method starts from the successful S-estimates in Fig. 10, but the high efficiency requirement has the consequence that the estimates change sufficiently so that many of the

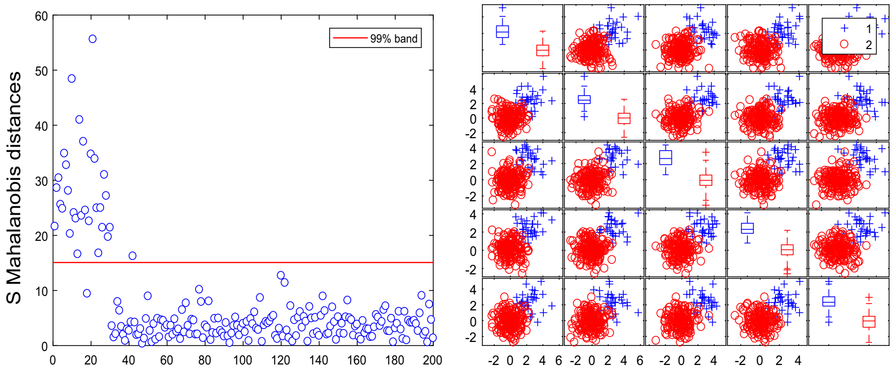

Fig. 10 Lightly contaminated data; S-estimation with bdp 0.5. Left-hand panel, index plot of squared Mahalanobis distances; 30 outliers are identified by the χ^2_5 band. Right-hand panel, scatterplot showing outliers as (blue) crosses (colour figure online)

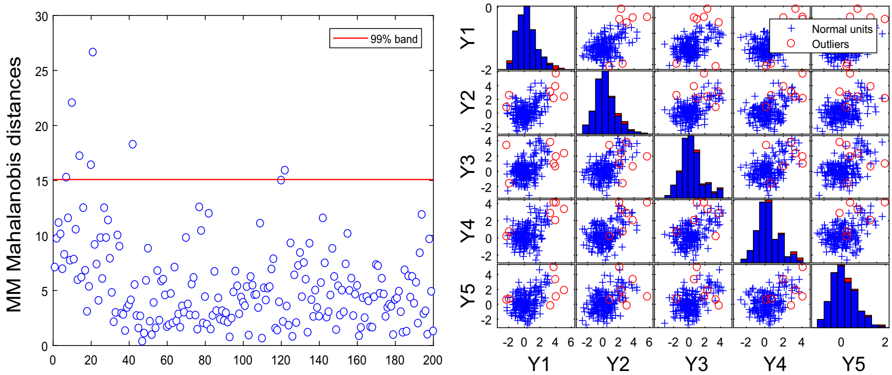

Fig. 11 Lightly contaminated data; MM-estimation with efficiency 99%. Left-hand panel, index plot of squared Mahalanobis distances; now only seven outliers are identified by the χ^2_5 band. Right-hand panel, scatterplot showing outliers as (red) circles and the distorted group of central observations (colour figure online)

outliers have high weights. As the left-hand panel of the figure shows, only seven observations are declared as outliers. The right-hand panel of the plot shows how the clustered nature of the outliers has caused the original spherical group of "good" observations to become ellipsoidal due to the incorporation of outliers. A series of panels detailing this process for a similar example is given in Figure 10 of Riani et al. (2014b).

Before turning to the monitoring of these two estimators, we look at the two hard trimming methods. The left-hand panel of Fig. 12 shows the 31 outliers found by the MCD and the right-hand panel shows the effect of reweighting using the χ^2_5 band. In this case the effect is slight; one wrongly declared outlier is reclassified. A similar set of outliers is found by the MVE.

We now consider the effect of monitoring these procedures. The left-hand panel of Fig. 13 shows the plot of robust squared Mahalanobis distances of the 200 units using S-estimation with bdp decreasing from 0.5 to 0.01. Despite there only being 15% of

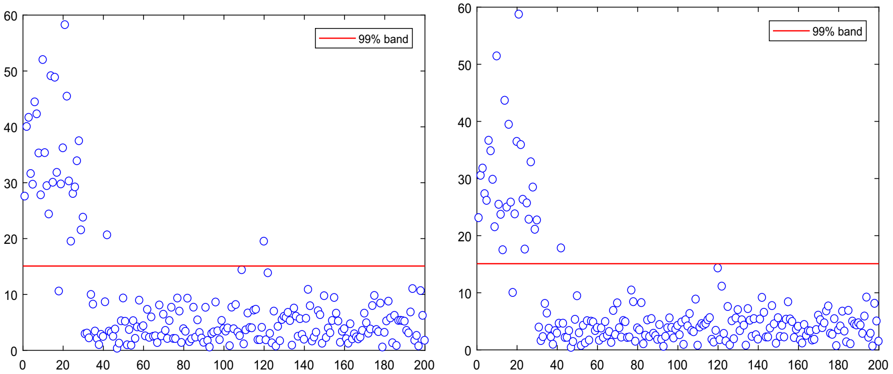

Fig. 12 Lightly contaminated data, index plots of squared Mahalanobis distances. Left-hand panel, raw MCD; 31 outliers are identified by the χ^2_5 band. Right-hand panel, reweighted MCD; 30 outliers

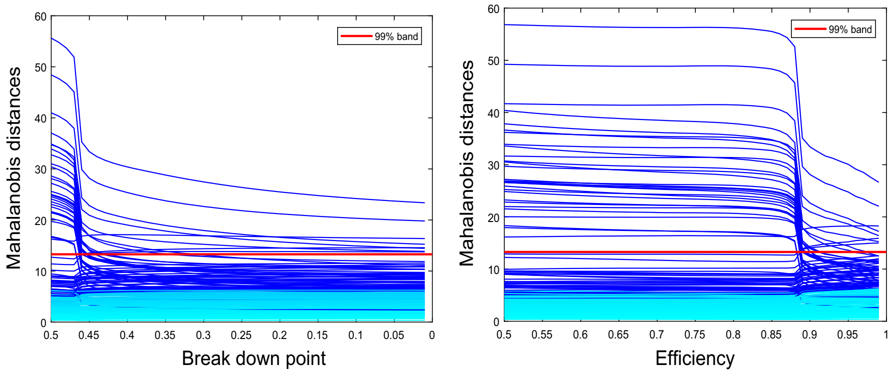

Fig. 13 Lightly contaminated data, monitoring plots of squared Mahalanobis distances. Left-hand panel, S-estimation. Right-hand panel MM-estimation

outliers, the robust solution is only found for breakdown points of 0.46 or higher. It is not possible to obtain an efficient robust estimate for these data using S-estimation.

The right-hand panel of Fig. 13 shows the monitoring plot for MM-estimation. It is clear from this stable plot that the set of outliers found by the very robust versions of S and MCD is also found here for an efficiency of up to 88%. Thus, despite the indication of Fig. 11, MM works well here. The failure arises because the common advice of an efficiency of 95% or 99% does not hold for these data. To confirm that these are virtually the same outliers as those found by S and MCD estimation, Fig. 14 shows the effect of brushing the most outlying observations for efficiencies between 0.5 and 0.88. The scatterplot matrix in the right-hand panel of the plot shows the strong similarity with the scatterplot from S estimation in Fig. 10. Through the use of monitoring we are able to adaptively choose the highest efficiency that gives a robust analysis. This result is an example justifying the use of monitoring.

More briefly, we show in Fig. 15 the plots of squared Mahalanobis distances from monitoring the MCD and the MVE. The plot for the MCD is more jagged than those

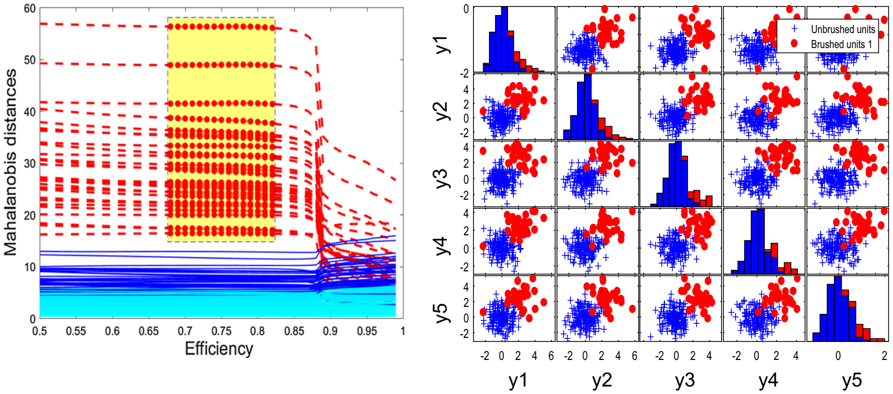

Fig. 14 Lightly contaminated data. Left-hand panel, brushing the monitoring plot for MM-estimation (Fig. 13). Right-hand panel, scatterplot matrix of the units with the most extreme squared Mahalanobis distances, plotted as (red) dots (colour figure online)

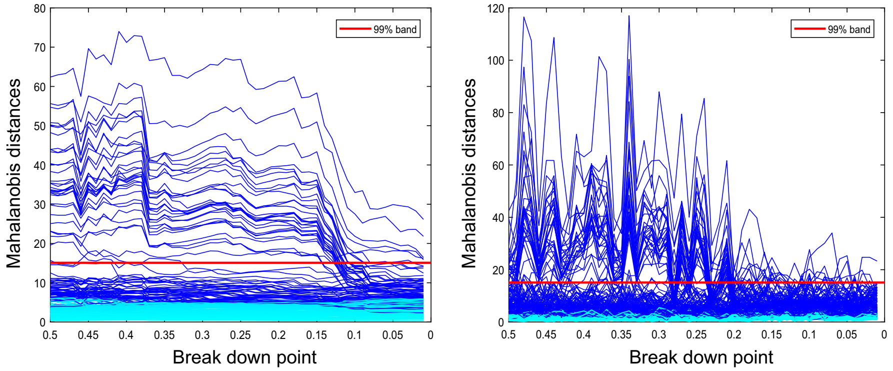

Fig. 15 Lightly contaminated data, monitoring plots of squared Mahalanobis distances. Left-hand panel, raw MCD. Right-hand panel MVE

we have seen before, but does show the change in the pattern of distances around a bdp of 0.14. The right-hand panel of the plot shows the monitoring plot for the MVE. This also finds the outliers for a high breakdown point but is so jagged as to be of little diagnostic use. As we stated in Sect. 3.3, such a plot is a reflection of the poor rate of convergence of this estimator. We now exclude the MVE from further study.

The monitoring plot for the reweighted MCD with a pointwise threshold of 0.99 is much the same as that for the original MCD, including a dramatic change at a bdp of 0.14. We do not give it here. The FS provides a clear indication of the outliers and a forward plot of scaled squared Mahalanobis distances which, like Fig. 9, for a large part of the search exhibits a clear gap between central units and the outliers with large distances. This plot is likewise not given here.

The conclusion of this example is that most of the methods work well with a light amount of contamination well separated from the main body of the data. In general monitoring allows us to choose values of efficiency or breakdown point that give esti-

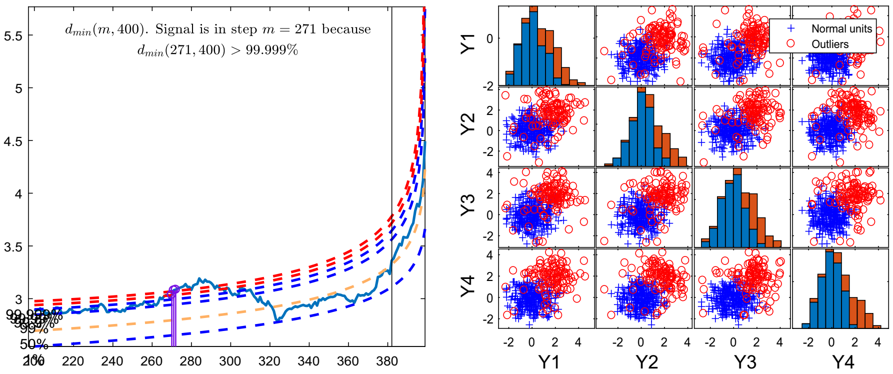

Fig. 16 Heavily contaminated data; analysis with the FS. Left-hand panel, forward plot of minimum Mahalanobis distance showing the signal for the presence of outliers. Right-hand panel, scatterplot of the 116 outliers indicated by the search, plotted as red circles (colour figure online)

mators that are as efficient as possible: that is, they exclude the outliers while fitting the "good" observations. Our monitoring has also provided important information about S-estimation; even with this advantageous data configuration, the method does not yield an efficient robust estimator. Equally importantly, we have shown the excellent performance of MM-estimation, provided unrealistic requirements are not made for efficiency.

6.2 Heavily contaminated data with appreciable overlap

Now we consider a simulated example which is similar to the structure we detect in our second data example in Sect. 7. There are 400 four-dimensional standard normal random variables, one hundred of them being displaced by an amount 2 in each dimension. There is thus some overlap between the 25% of outliers and the uncontaminated data.

We start by describing the results of the FS, which conveniently also allows us to display the structure of the data. The left-hand panel of Fig. 16 shows the monitoring of the minimum Mahalanobis distance during the search. There is a signal at $m = 271$. Resuperimposition leads to the detection of 116 outliers (97 of the 116 belong to the group of contaminated units), that is 29% of the data. The right-hand panel of the figure shows the scatterplot with the 116 outliers plotted as (red) circles. The result of brushing the scaled squared Mahalanobis distances which are above the $\chi^4_{0.99}$ threshold in the central part of the search (left-hand panel of Fig. 17) shows the spherical structure of the uncontaminated data (right-hand panel of Fig. 17).

There is much less structure evident in some of the other analyses. Both S-estimation with the highest breakdown point and MM-estimation with 99% efficiency fail to detect most of the outliers. Using $\chi^2_{4,0.99}$ for outlier detection, the S-estimator identifies 7 outliers and the MM-estimator 6. In these cases, monitoring does not help. The left-hand panel of Fig. 18 shows the smooth forward plot of the squared Mahalanobis distances from S-estimation as a function of bdp. There is slightly more structure in the likewise smooth plot for MM-estimation in the right-hand panel of the figure, but nothing that indicates an appreciable number of outliers.

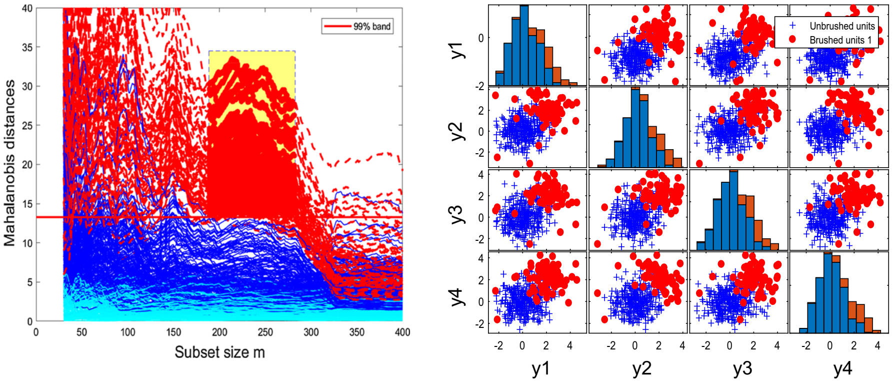

Fig. 17 Heavily contaminated data; analysis with the FS. Left-hand panel, forward plot of squared scaled Mahalanobis distance after brushing the units above the $\chi^4_{0.99}$ threshold in the central part of the search. Right-hand panel, scatterplot of the brushed units plotted as filled red circles (colour figure online)

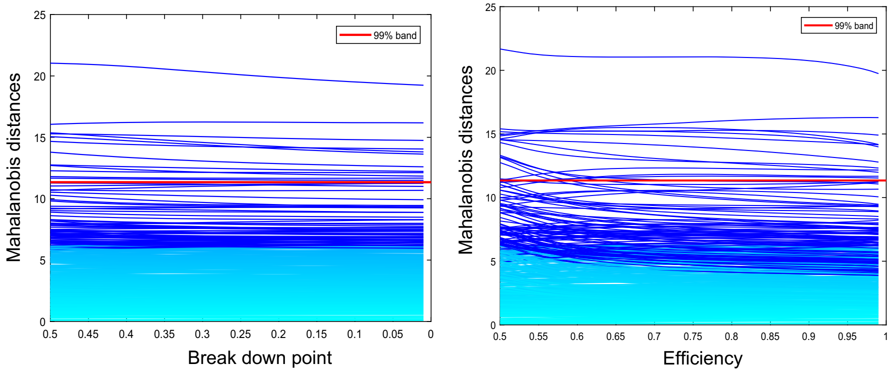

Fig. 18 Heavily contaminated data, monitoring plots of squared Mahalanobis distances. Left-hand panel, S-estimation. Right-hand panel MM-estimation

We now consider the MCD. When bdp is 50%, 64 outliers are found using $\chi^2_{4,0.99}$ (60 belong to the group of contaminated units). These are shown in the left-hand panel of Fig. 19. Reweighting the output of this analysis leads to the detection of only 9 outliers (7 belong to the group of contaminated units) as is shown in the right-hand panel of the figure. The panel shows how the distribution of distances for the 100 contaminated units is changed by the parameter estimates from reweighting. However, the distribution of these distances is even so quite distinct from those from the uncontaminated units. This effect of reweighting, which is not substantially affected by the choice of the reweighting threshold, is quite different from that shown in Fig. 12, where weighted and unweighted analyses were comparable. However, monitoring the MCD is still very informative. The left-hand panel of Fig. 20 shows a striking change around a bdp of 0.27 as the outliers start to be included in the central part of the data. The right-hand panel of the figure shows the monitoring plot for the reweighted MCD. There is a change around a bdp of 0.28 when some Mahalanobis distances slightly increase in

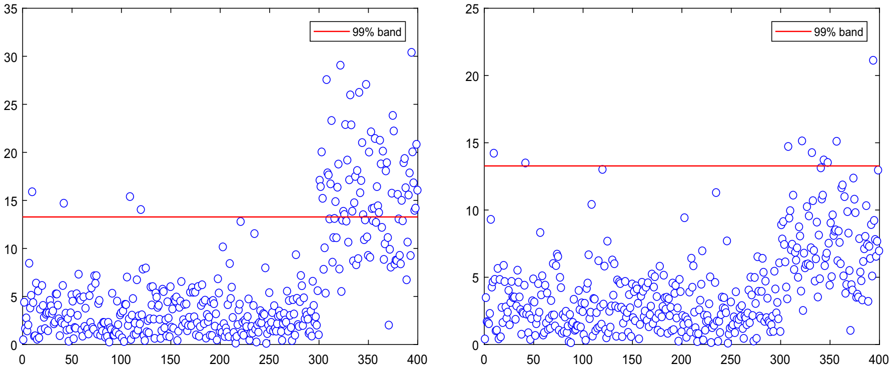

Fig. 19 Heavily contaminated data, MCD analysis. Left-hand panel, index plot of squared Mahalanobis distances from raw MCD with 50% bdp. Right-hand panel, reweighted MCD

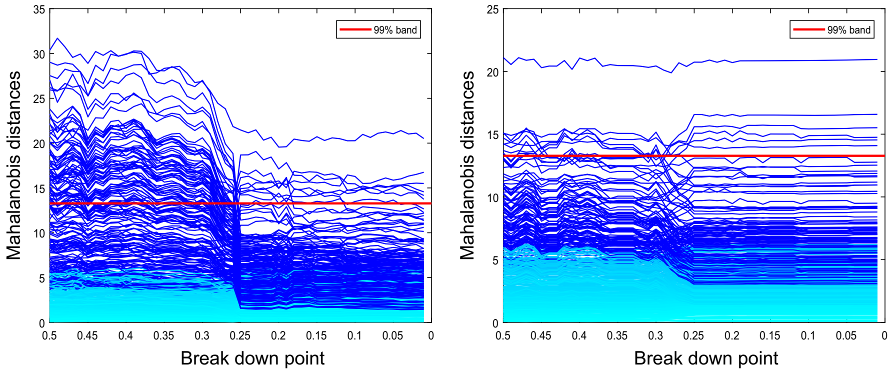

Fig. 20 Heavily Contaminated Data, monitoring plots of squared Mahalanobis distances. Left-hand panel, raw MCD, right-hand panel, reweighted MCD

magnitude. For lower values of the bdp the plots in the two panels are similar; just 9 observations are identified as outlying, those shown in the right-hand panel of Fig. 19.

Figure 21, showing index plots of the squared Mahalanobis distances computed from the raw MCD with different values of the breakdown point, provides further insight into our data-driven choice of bdp. The left-hand panel is obtained with bdp $=$ 0.29, immediately before the sudden change pointed out by the monitoring plot. The structure of the data implied by the new index plot is similar to that already given in the left-hand panel of Fig. 19, with only a slight reduction in the number of detected outliers (now 46, with just 1 of them belonging to the group of uncontaminated observations) and essentially the same number of uncontaminated observations taken as non-outlying (now 299 instead of 296). On the other hand, it is clear that the choice of a smaller breakdown point, such as bdp $=$ 0.23 (right-hand panel), does not not guarantee against masking and provides a completely different (non-robust) fit, with only few contaminated observations identified as outliers.

The conclusion of this potentially problematic example is that smooth downweighting, as in M-estimation and its derivatives, is not enough with such a high level of

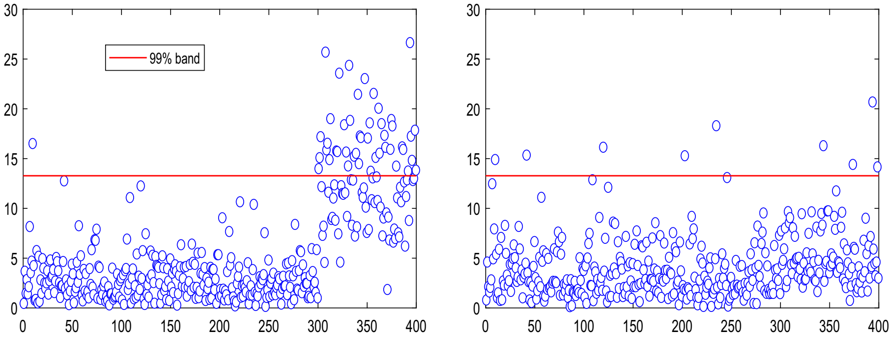

Fig. 21 Heavily contaminated data, index plots of squared Mahalanobis distances from the raw MCD with different values bdp: 0.29 (left) and 0.23 (right)

Table 1 Heavily contaminated data, number of uncontaminated (n_u) and contaminated (n_c) observations in the fitting subset of the raw MCD, for different values of bdp

	bdp = 0.5	bdp = 0.45	bdp = 0.35	bdp = 0.29	bdp = 0.23
n_u	201	220	258	281	245
n_c	1	1	3	4	63

contamination. Hard trimming, either from the FS or from the MCD, is necessary. Efficient estimates are automatically computed from the adaptive choice of subset size in the FS. Such estimates can also be obtained by monitoring the MCD which again introduces adaptive trimming into the fitting algorithm. The advantage of selecting a degree of robustness which is specifically tailored to the data at hand can be appreciated by noting the number (n_u) of uncontaminated observations that are used for parameter estimation. Table 1 reports these numbers for the raw MCD and for different values of the breakdown point. The same table also gives the corresponding numbers (n_c) of contaminated observations that are included in the fitting subset. It is seen that the maximally robust MCD estimate obtained with bdp $= 0.5$ is computed on considerably fewer uncontaminated observations than the still robust estimate with bdp $= 0.29$, although the values of n_c are comparable in the two cases. While introducing a very modest amount of bias (three more mildly contaminated units in the fitting subset), our data-driven choice of bdp leads to an increase in efficiency of the order of $\sqrt{281/201} - 1 \approx 0.18$ for the estimate of μ and to an even larger gain for the estimate of Σ (see, e.g., Croux and Haesbroeck 1999). In contrast, Table 1 shows that the effect of masking is paramount when bdp $= 0.23$.

7 A second data example: cows with bovine dermatitis

We now consider our second data example, that of 488 cows with bovine dermatitis (perhaps, more strictly, vaccine dermatitis). The disease, which causes lameness in

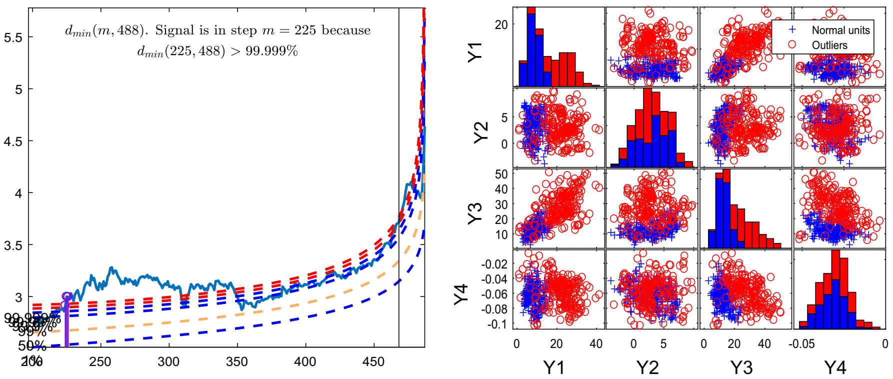

Fig. 22 Cows with Bovine Dermatitis; analysis with the FS. Left-hand panel, forward plot of minimum Mahalanobis distance showing the signal for the presence of outliers. Right-hand panel, scatterplot of the 230 outliers indicated by the FS, plotted as red circles (colour figure online)

cattle, was first discovered in Italy in 1974. It can reduce the yield of milk, but, we are assured, the quality of the milk is not affected.

There are four measurements per cow derived from photographic images. There is no reason to believe that they are approximately normally distributed, but we start with the FS assuming this to be true. The left-hand panel of Fig. 22 shows the shape in the middle of the search that is often associated with two clusters of similar size; after a signal, here at $m = 225$, the trace of minimum distances returns inside the envelopes as the observations from the second cluster cause masking. For these data, resuperimposition of envelopes leads to the identification of 230 outliers. The right-hand panel of Fig. 22 clearly shows the two groups that have been identified, relatively well separated in some dimensions, such as y_1 and y_3, but overlapping in the other two dimensions. The groups have plausibly normal distributions in all projections in the scatterplot matrix.

As might be expected from the previous examples, both the S-estimator and MM without monitoring fail to indicate any structure.

Figure 23 shows the monitoring of the Mahalanobis distances for these two estimators. The left-hand panel shows the S-estimator and is essentially smooth—under these conditions the estimator with bdp 0.5 is little different from the maximum likelihood estimator. As illustrated in the right-hand panel of Fig. 23 there is slightly more of interest in the plot for the MM-estimator, caused by the few outliers that enter as the efficiency approaches one.

The MCD, which uses hard trimming, is more informative about the structure of the data. The raw MCD indicates 186 outliers. Reweighting with $\chi^2_{4,0.95}$ and $\chi^2_{4,0.99}$ reduces this number respectively to 144 and 58. More insight is obtained by monitoring the MCD as in the left-hand panel of Fig. 24, even if this is not a particularly easy plot to interpret. There is a first region from a bdp of 0.5–0.4. Brushing this narrow range of values gives the set of 44 outliers shown in the scatterplot of the right-hand panel of the figure. These have a similar structure to the outliers exhibited for the FS in Fig. 22. Of course, with such a complex plot, the number of outliers selected will depend both

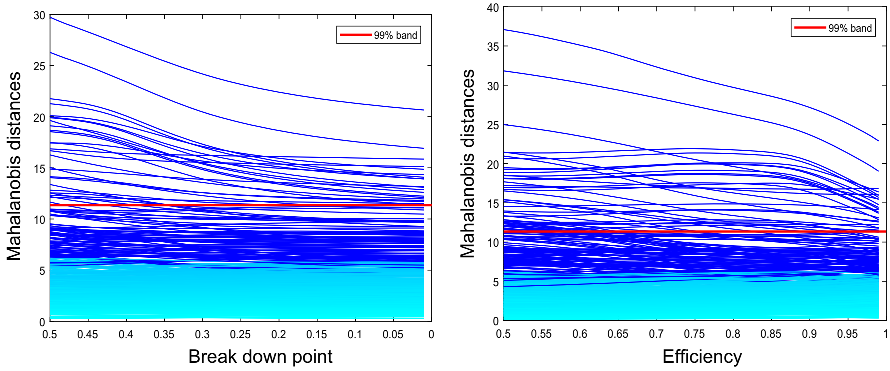

Fig. 23 Cows with Bovine Dermatitis; monitoring plots of squared Mahalanobis distances. Left-hand panel S-estimation and, right-hand panel, MM-estimation

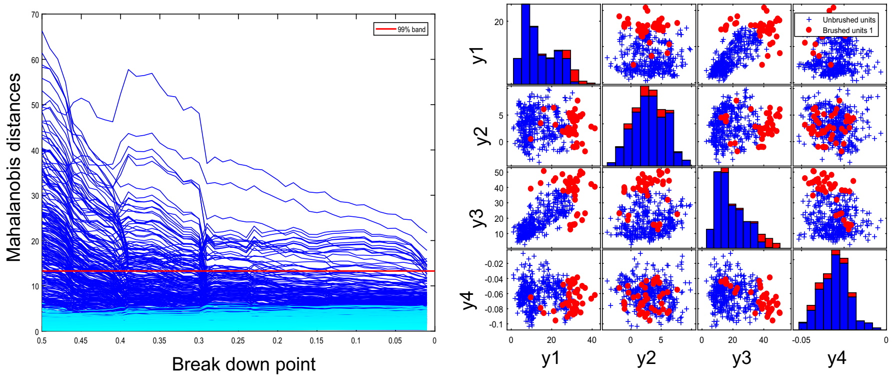

Fig. 24 Cows with Bovine Dermatitis. Left-hand panel, monitoring the raw MCD. Right-hand panel, scatterplot showing, as red dots, the 44 observations obtained by brushing large Mahalanobis distances (colour figure online)

on the range of bdp considered and the minimum value of distance included in the brush.

Although robust clustering methods might be preferable in this example, given the presence of two overlapping and almost equally-sized groups, we have seen that monitoring the value of the breakdown point of MCD does help to understand the structure of the data. It does not, as in the other examples, lead to the specification of the bdp for an efficient robust estimator; a reasonable separation between the two groups can only be obtained with values of bdp close to 0.5, even if hard trimming is used.

8 Assessing the pointwise bounds for the squared Mahalanobis distances from monitoring

In our examples we have used thresholds for outlier detection based on the chi-squared distribution. This is the limiting distribution to which, in the absence of contamination,

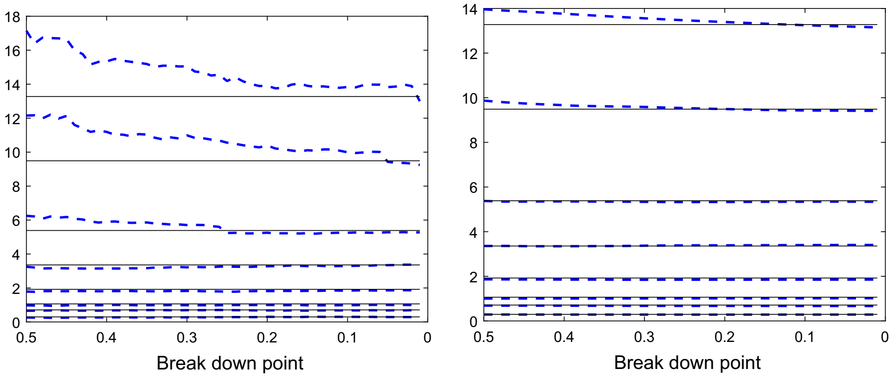

Fig. 25 Estimated (dashed lines) and asymptotic (solid lines) quantiles of squared robust Mahalanobis distances for samples of size $n = 200$ from the four-variate normal distribution, as a function of bdp. Left-hand panel: raw MCD. Righ-hand panel: S-estimation

all the squared robust Mahalanobis distances of this paper converge as $n \to \infty$. A relevant issue is then to understand the quality of this distributional approximation in monitoring plots for samples of moderate size. An additional problem, that we do not address here but leave for future research, is the use of pointwise bounds when performing a sequence of n outlier tests, as in the FS, with the same dataset being scrutinized for each value of breakdown point or efficiency.

To answer our present question we simulate 1000 samples of size $n = 200$ from the four-variate standard normal distribution. For each sample we repeat the monitoring analyses described in Sects. 5–7 and compute Monte Carlo estimates of the quantiles of the squared robust distances for

$$\zeta \in \{0.01\ 0.05\ 0.10\ 0.25\ 0.5\ 0.75\ 0.95\ 0.99\}.$$

We start our comparison in Fig. 25, where we show the estimated quantiles and their asymptotic $\chi^2_{4,\zeta}$ counterparts, for the raw MCD and for S-estimation, as a function of breakdown point. With such a sample size the null asymptotic distribution already provides a reliable approximation for maximum likelihood estimation, since $\chi^2_{4,\zeta}$ is close to the ζ-quantile of the exact scaled beta distribution of squared Mahalanobis distances. For instance, $\chi^2_{4,0.99} = 13.28$, while the exact 0.99-quantile is 12.97. It is evident that, as the degree of trimming in MCD increases, the χ^2_4 approximation rapidly deteriorates, leading to liberal outlier tests. Consequently, the consistency factor (3) is not enough to accommodate even moderate levels of trimming in the right tail of the distance distribution which we require for precise outlier identification. This result confirms the need to adopt more accurate thresholds (Hardin and Rocke 2005; Cerioli 2010; Cerioli and Farcomeni 2011), perhaps by including further correction factors (Pison et al. 2002; Cerioli et al. 2009), when the goal is outlier detection (which is not, we recall, the main focus of this paper). The result also quantifies the change implied by alternative degrees of robustness on the accuracy of asymptotic distributional results for MCD-based squared distances, which may be relevant in several application fields

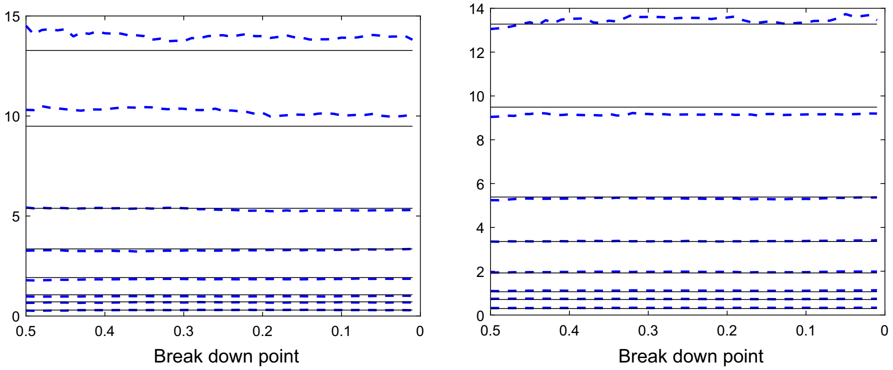

Fig. 26 As Fig. 25, but now for reweighted MCD. Left-hand panel: reweighting based on $\chi^2_{4,0.90}$. Righ-hand panel: reweighting based on $\chi^2_{4,0.99}$

(see, e.g., Green and Martin 2014). The quality of the χ^2_4 approximation is consistently better for the squared robust distances computed from S-estimates. This distributional advantage with uncontaminated data must of course be contrasted with the potential masking effects that we have seen in Sects. 5–7 when many outliers (or groups) are present.

Figure 26 repeats the comparison for the reweighted MCD, with two alternative thresholds for reweighting. Now the empirical quantiles are virtually constant for all values of bdp. However, the resulting outlier test is liberal, even when bdp is close to 0, if reweighting is based on $\chi^2_{4,0.90}$. This is another harmful consequence of using inaccurate distributional results for robust distances, both in the reweighting step—leading to discarding a proportion of units larger than the nominal one (10% in this case)—and in the final test for outlier nomination.

We conclude our assessment by looking in the left-hand panel of Fig. 27 at the estimated quantiles for MM-estimation as a function of efficiency. The right-hand panel shows, for the reweighted MCD, the estimated quantiles of the squared distances computed as a function of the probability used in reweighting. We see that both plots broadly repeat the pattern already depicted in Fig. 25 by their less efficient counterparts, but with a reduced agreement between empirical and asymptotic bands. This again is the price to be paid in order to have more powerful detection tools with contaminated data.

9 Discussion

The four examples analysed in this paper show how monitoring can be used adaptively to obtain robust estimators that are as efficient as possible. Depending on the estimator we are often able to choose the lowest bdp, the highest efficiency or the largest value of h consistent with downweighting or trimming outlying observations. Perhaps not surprisingly, the four examples also show how this is decreasingly easy as the number of outliers increases and as they become closer to the main body of the data. The limit

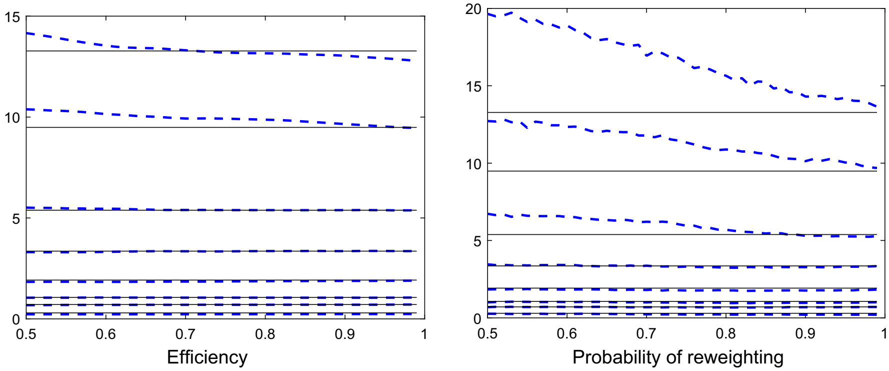

Fig. 27 Estimated (dashed lines) and asymptotic (solid lines) quantiles of squared robust Mahalanobis distances for samples of size $n = 200$ from the four-variate normal distribution. Left-hand panel: MM-estimation as a function of efficiency. Right-hand panel: reweighted MCD as a function of the probability of reweighting

is in the heavily contaminated examples, illustrated in Figs. 18 and 23 for S and MM estimation, where monitoring is not able to detect any parameter values that provide a robust fit.

The adaptive estimators we have obtained through monitoring, when they exist, show a strong relationship to the results obtained from the FS. Like the FS they avoid the awkward choices of efficiency and breakdown point which bedevil practical applications of robust statistics. Of course, such tiresome choices can also be avoided by using estimators such as the MCD with the highest possible breakdown point. However, our results show that efficiency can often be appreciably improved through the adaptive choice of the number of observations to be trimmed. In addition, as illustrated by Riani et al. (2014a), our approach could also be useful to assess the effect of other decisions required by robust techniques, such as the function ρ in S and MM-estimation, not specifically addressed in this work.

Our paper, following the assertion of our title, demonstrates that monitoring can provide the way of extracting the maximum information from a contaminated sample. We look forward to the confirmation of our claim through the evidence of other data analyses.

Acknowledgements We are very grateful to the Editor, Tommaso Proietti, for inviting this paper and for organizing its discussion. We also thank Alessio Farcomeni, Luca Greco, Domenico Perrotta and two anonymous reviewers for helpful comments on a previous draft. MR and ACA gratefully acknowledge support from the CRoNoS project, reference CRoNoS COST Action IC1408.

References

Agostinelli C, Marazzi A, Yohai V (2014) Robust estimators of the generalized log-gamma distribution. Technometrics 56:92–101

Alfons A, Croux C, Gelper S (2013) Sparse least trimmed squares regression for analyzing high-dimensional large data sets. Ann Appl Stat 7:226–248

Amiguet M, Marazzi A, Valdora M, Yohai V (2017) Robust estimators for generalized linear models with a dispersion parameter. Technical Report 1703.09626v1, arXiv

Atkinson AC, Corbellini A, Riani M (2017a) Robust Bayesian regression with the forward search: theory and data analysis. Test, in press, https://doi.org/10.1007/s11749-017-0542-6

Atkinson AC, Riani M (2000) Robust diagnostic regression analysis. Springer, New York

Atkinson AC, Riani M (2007) Exploratory tools for clustering multivariate data. Comput Stat Data Anal 52:272–285

Atkinson AC, Riani M, Cerioli A (2004) Exploring multivariate data with the forward search. Springer, New York

Atkinson AC, Riani M, Cerioli A (2010) The forward search: theory and data analysis (with discussion). J Korean Stat Soc 39:117–134

Atkinson AC, Riani M, Cerioli A (2017) Cluster detection and clustering with random start forward searches. J Appl Stat, in press, https://doi.org/10.1080/02664763.2017.1310806

Avella-Medina M, Ronchetti E (2015) Robust statistics: a selective overview and new directions. WIREs Comput Stat 7:372–393

Azzalini A, Bowman A (1990) A look at some data on the Old Faithful geyser. Appl Stat 39:357–365

Boudt K, Rousseeuw P, Vanduffel S, Verdonck T (2017) The minimum regularized covariance determinant estimator. Technical Report 1701.07086v1, arXiv

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

Cerioli A, Farcomeni A (2011) Error rates for multivariate outlier detection. Comput Stat Data Anal 55:544–553

Cerioli A, Riani M (1999) The ordering of spatial data and the detection of multiple outliers. J Comput Gr Stat 8:239–258

Cerioli A, Riani M, Atkinson AC (2009) Controlling the size of multivariate outlier tests with the MCD estimator of scatter. Stat Comput 19:341–353

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, Atkinson AC, Riani M (2016) How to marry robustness and applied statistics. In: Di Battista T, Moreno E, Racugno W (eds) Topics on methodological and applied statistical inference. Springer, Heidelberg, pp 51–64

Cerioli A, Farcomeni A, Riani M (2017) Wild adaptive trimming for robust estimation and cluster analysis. Submitted

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

Croux H, Haesbroeck G (1999) Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. J Multivar Anal 71:161–190

Davies PL (1987) Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices ellipsoid estimator. Ann Stat 15:1269–1292

Dotto F, Farcomeni A, García-Escudero LA, Mayo-Iscar A (2017) A reweighting approach to robust clustering. Stat Comput, in press, https://doi.org/10.1007/s11222-017-9742-x

Farcomeni A, Greco L (2015) Robust methods for data reduction. Chapman and Hall/CRC, Boca Raton

García-Escudero LA, Gordaliza A (2005) Generalized radius processes for elliptically contoured distributions. J Am Stat Assoc 100:1036–1045

Green CG, Martin D (2014) An extension of a method of Hardin and Rocke, with an application to multivariate outlier detection via the IRMCD method of Cerioli. Technical Report available at http://christopherggreen.github.io/papers, Department of Statistics, University of Washington

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

Huber PJ, Ronchetti EM (2009) Robust statistics, 2nd edn. Wiley, Hoboken

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

Hubert M, Rousseeuw PJ, Siegaert P (2015) Multivariate functional outlier detection (with discussion). Stat Methods Appl 24:177–202

Johansen S, Nielsen B (2016a) Analysis of the Forward Search using some new results for martingales and empirical processes. Bernoulli 22:1131–1183

Johansen S, Nielsen B (2016b) Asymptotic theory of outlier detection algorithms for linear time series regression models (with discussion). Scand J Stat 43:321–348

Lopuhaä HP, Rousseeuw PJ (1991) Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. Ann Stat 19:229–248

Maronna RA, Martin RD, Yohai VJ (2006) Robust statistics. Wiley, Chichester

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

Riani M, Atkinson AC (2001) Regression diagnostics for binomial data from the forward search. J R Stat Soc Ser D 50:63–78

Riani M, Atkinson AC (2007) Fast calibrations of the forward search for testing multiple outliers in regression. Adv Data Anal Classif 1:123–141

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

Riani M, Cerioli A, Atkinson AC, Perrotta D (2014a) Monitoring robust regression. Electron J Stat 8:646–677

Riani M, Cerioli A, Torti F (2014b) On consistency factors and efficiency of robust S-estimators. Test 23:356–387

Riani M, Atkinson AC, Perrotta D (2014c) A parametric framework for the comparison of methods of very robust regression. Stat Sci 29:128–143

Riani M, Perrotta D, Cerioli A (2015) The forward search for very large datasets. J Stat Softw 67:1

Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley, New York

Salini S, Cerioli A, Laurini F, Riani M (2016) Reliable robust regression diagnostics. Int Stat Rev 84:99–127

Tallis GM (1963) Elliptical and radial truncation in normal samples. Ann Math Stat 34:940–944

Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York

Yohai VJ (1987) High breakdown-point and high efficiency estimates for regression. Ann Stat 15:642–656