

Discussion of “The power of monitoring: how to make the most of a contaminated multivariate sample” by Andrea Cerioli, Marco Riani, Anthony C. Atkinson and Aldo Corbellini

Claudio Agostinelli¹ · Luca Greco² 

Accepted: 4 December 2017 / Published online: 22 December 2017
© Springer-Verlag GmbH Germany, part of Springer Nature 2017

Abstract Andrea Cerioli, Marco Riani, Anthony Atkinson, Aldo Corbellini (CRAC hereafter) have presented a powerful methodology aimed at improving robust fitting and related diagnostic tools. Monitoring is a very flexible approach that allows to tune the selected robust technique by looking at a whole movie of the available data. We contribute to the discussion of CRAC’s paper by applying the principle of monitoring to multivariate weighted likelihood estimation. The reliability of the method is illustrated through the analysis of the datasets taken from CRAC’s paper.

Keywords Monitoring · Outliers · Pearson residuals · Robust distances · Weighted likelihood

1 Introduction

We would like to sincerely congratulate the authors for this interesting and stimulating work, in which they pursue a powerful approach for robust estimation and the detection of anomalous values. Robust fitting and outlier detection are strictly connected tasks. The main goal of a robust analysis is to lead to reliable inferences that are not badly affected by the occurrence of outliers, at the cost of a negligible efficiency loss under the postulated model. On the other hand, the detection of such data inadequacies may be of interest itself, since they may unveil unexpected features in the sample at hand.

✉ Luca Greco
luca.greco@unisannio.it

Claudio Agostinelli
claudio.agostinelli@unitn.it

¹ Department of Mathematics, University of Trento, Trento, Italy

² DEMM Department, University of Sannio, Benevento, Italy

Actually, the data may have been contaminated by gross errors in the data collection process, the overlapping of several unexpected random mechanisms, sampling from a non homogenous population composed by some unknown and eventually rare sub-groups.

One obstacle to the diffusion of robust techniques in the statistical practice is that they depend on several choices to be made, whose effects on the results may be not really easy to evaluate or understand. A dangerous practice consists in using *default* settings, such as 50% breakdown point or 95% efficiency, whose employ may be not consistent with the sample at hand, leading to a worthless efficiency loss or to non robust solutions, respectively. The CRAC's paper has the unquestionable merit to warn researchers against an automatic use of robust methods with default settings and to lead them toward a conscious use of robust tools, instead. The authors suggest that the main features of a robust technique should be inferred by the data, in an adaptive fashion. Their monitoring approach may really give a new impulse to the use of robust methods in data analysis. The advantages of monitoring in multivariate estimation have been also addressed in Farcomeni and Greco (2016).

Here, we contribute to the discussion about the power of monitoring by applying this technique to weighted likelihood estimation of multivariate location and covariance. Weighted likelihood is a robust method falling in the category of soft-trimming, since robustness is achieved by attaching a weight in $[0, 1]$ to each observation, aiming at down-weighting outliers, as well as in M-type estimation. The monitoring of weighted likelihood analyses lead to two main achievements. From the one hand it is proved that monitoring is a valuable tool in order to tune the method, from the other hand monitoring shows that weighted likelihood highlights features that are not shared by other soft-trimming procedures such as S- and MM-estimation, whose behavior has been investigated in CRAC's paper. On the contrary, the monitoring unveils that weighted likelihood is able to deliver robust solutions that are in close agreement with those stemming from hard-trimming methods based on crispy weights $\{0, 1\}$, such as the Forward Search (FS) and the MCD.

2 Preliminaries

Let $y = (y_1, \dots, y_n)^\top$ be a random sample from a random variable Y with unknown distribution function $M(y; \theta)$ and corresponding probability (density) function $m(y; \theta)$, $\theta \in \Theta \subset \mathbb{R}^p$, with $p \geq 1$ and let \hat{M}_n be the empirical distribution function. A weighted likelihood estimate (WLE) is defined as the root of the weighted likelihood estimating equation (WLEE) (Markatou et al. 1998)

$$\sum_{i=1}^n w(y_i; \theta, \hat{M}_n) s(y_i; \theta) = 0,$$

where $s(y_i; \theta)$ denotes the i th contribution to the score function and the weight function is defined as

$$w(y; \theta, \hat{M}_n) = \frac{[A(\delta(y; \theta, \hat{M}_n)) + 1]^+}{\delta(y; \theta, \hat{M}_n) + 1}, \quad (1)$$

where $[\cdot]^+$ denotes the positive part. The function $\delta(y; \theta, \hat{M}_n)$ is the Pearson residual function

$$\delta(y; \theta, \hat{M}_n) = \frac{\hat{m}_n(y)}{m^*(y; \theta)} - 1$$

where

$$\hat{m}_n(y) = \int k(y; t, h) d\hat{M}_n(t)$$

is a kernel density estimate of $m(y; \theta)$ with kernel $k(\cdot; t, h)$ and smoothing parameter h , and

$$m^*(y; \theta) = \int k(y; t, h) dM(t; \theta)$$

is a smoothed model density (Basu and Lindsay 1994), with $\delta \in [-1, +\infty)$. The Pearson residual function measures the agreement between the data and the assumed model. Hence, outliers are expected to show large residuals. Pearson residuals are made large when $\hat{m}_n(y)$ presents *bumps* not shared by the model but also when the expected density $m^*(y; \theta)$ is close to zero. This feature enhances the meaning of outliers as values that are *unlikely* to occur under the assumed model rather than values that are *distant* from the bulk of the data. The methodology has been extended to the regression framework, as well (Agostinelli and Markatou 1998; Agostinelli 2002; Alqallaf and Agostinelli 2016). In this case the construction of Pearson residuals is different even if conceptually equivalent, i.e. the kernel density estimate is evaluated over the (standardized) residuals that depend on the coefficients vector θ , whereas the assumed model might depend on nuisance parameters only. A similar argument has been also applied in the multivariate framework in Agostinelli and Greco (2017). More insights on this approach will be given in the next subsection.

The function $A(\cdot)$ is the Residual Adjustment Function (RAF) that plays the role to bound the effect of large Pearson residuals. This function is related to minimum disparity estimation problems (Lindsay 1994; Park et al. 2002).

Weighted likelihood estimation leads to consistent and asymptotically fully efficient estimators at the postulated model but also characterized by a high breakdown point under contamination. Full efficiency at the assumed model stems from the fact that the weight function converges uniformly to one for non contaminated samples (Agostinelli and Greco 2013).

The smoothing parameter h , indexing the kernel function $k(\cdot; \cdot, h)$, does not play any role in controlling asymptotic efficiency and it has a limited impact on the asymptotic robust behavior of the method. On the contrary, in finite samples it controls the robustness/efficiency trade-off of the weighted likelihood methodology. Small values of h are a reasonable choice in order to obtain a non parametric model density estimate that is sensitive to outlying observations. On the contrary, large values of h lead to smooth density estimates that are expected to be close to the assumed model when the data are not prone to contamination. The selection of the smoothing parameter h may be troublesome. Actually, it is not an easy task to relate its choice to efficiency or break-

down arguments, in a fashion similar to M-type estimation. In Markatou et al. (1998), the authors suggested that: “*the sum of the final fitted weights is a useful diagnostic statistic for the comparison of solutions to the WLEE*”. In other words the authors’ advice is to monitor the sum of the fitted weights by varying h . Then, the quantity $1 - \bar{w}$, where \bar{w} denotes the average of the weights evaluated at the WLE, can be used as a rough measure of the rate of contamination in the sample at hand. We refer to $1 - \bar{w}$ as the empirical downweighting level. The same approach has been pursued in Greco (2016). As it is shown in Markatou et al. (1998), the mean downweighting under the correct model for finite samples can be approximated by

$$\Lambda = -\frac{w''(0)}{2} \left[\int \frac{\int k^2(y; t, h) dM(t; \theta)}{(m^*(t; \theta))^2} dM(t; \theta) - 1 \right],$$

where $w''(0)$ is the second derivative of the weight function with respect to δ at 0 and it is equal to $2 - \tau$ for the Generalized Kullback–Leibler (GKL) residual adjustment function. The mean downweighting parameter Λ gives on average the rate of contamination at the assumed model and represents “*a simple measure of the interplay of the various parameters in the degree of downweighting that will occur when the model is correct*”.

2.1 Multivariate weighted likelihood

In a multivariate setting, the weighting scheme based on the computation of a multivariate density estimate becomes troublesome for large dimensions, because of the curse of dimensionality (Huber 1985; Scott and Wand 1991). With growing dimensions the data are more sparse and kernel density estimation may become unfeasible. One can get round this hindrance by following the approach proposed in Agostinelli and Greco (2017), in the standard framework of robust fitting of a multivariate normal distribution $N_p(\mu, \Sigma)$. The authors suggested to compute Pearson residuals based on Mahalanobis distances rather than on multivariate data. Then, by paralleling what happens in regression problems, Pearson residuals are obtained by comparing a univariate (unbiased at the boundary) kernel density estimate evaluated over the squared distances $d^2(y_i; \mu, \Sigma)$ and their underlying χ_p^2 distribution.

3 Monitoring the smoothing parameter h

In this section, we apply the monitoring to obtain some guidance in the selection of the smoothing parameter h in weighted likelihood estimation of multivariate location and scatter. The same four example discussed in CRAC’s paper are considered. The examples all show that monitoring the smoothing parameter h in the weighted likelihood analyses provides useful information. First of all, monitoring is always able to detect a value beyond whom the analysis becomes non robust. Furthermore, monitoring allows us to highlight the peculiar and satisfactory features of weighted likelihood estimation with respect to the other robust methods that have been compared in CRAC’s paper. In all the example that follows, we used the Generalized Kullback–Leibler RAF

$$A_{GKL}(\delta) = \frac{\log(\delta\tau + 1)}{\tau}, \quad 0 \leq \tau \leq 1$$

and a folded normal kernel. The results all depend on our personal choices concerning $\tau = 0.9$, the kernel and the grid of h values to monitor the WLE analyses. Another setting may lead to different solutions, but the general features of the procedure will be very similar. Pointwise thresholds to detect outliers have been set by approximating the distribution of squared robust distances by a scaled Beta distribution, as conjectured in Agostinelli and Greco (2017), but the asymptotic χ_p^2 could have been used as well.

3.1 Eruptions of Old Faithful

This data give the duration of the eruption and the waiting time to the start of that eruption from the previous for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. The sample size is $n = 272$. The robust analyses discussed in CRAC’s paper all highlight the presence of two groups. Monitoring leads to select the tuning parameters characterizing S-, MM- and MCD estimation before the fit abruptly changes and becomes non robust and no more able to distinguish the two sub-groups.

Here, in a similar fashion we use monitoring to determine a value of h leading to an efficient but reasonably robust estimate. The results from monitoring are displayed in Fig. 1. Both panels clearly show that beyond a certain value $h = 0.01175$, given by the horizontal dotted line, the analysis becomes non robust, since the two sub-groups structure is no more detected. The left-hand panel shows the trajectories of individual robust distances as h varies on the considered grid. For small values of h several robust distances exceed the 0.99-level pointwise threshold, hence leading to detect the smallest sub-group as outliers. On the contrary, in the right-hand part of the plot, it is evident that no more outliers are detected and the underlying structure of the data is lost. In a fashion similar to CRAC’s paper, a color map has been used that goes

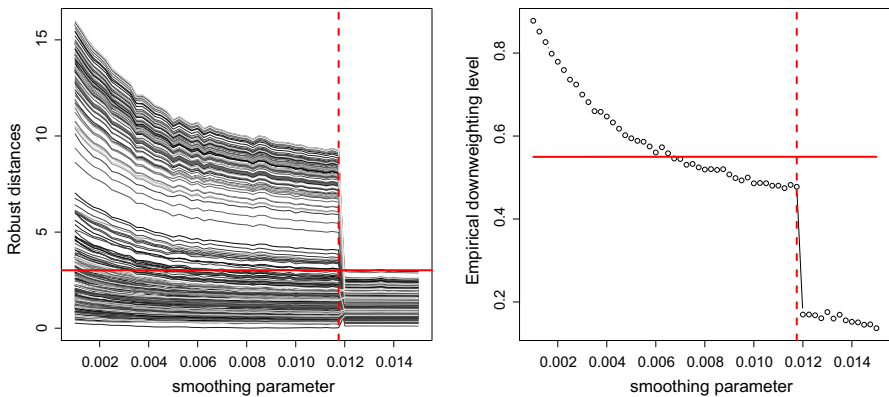


Fig. 1 Eruptions of Old Faithful. Left-hand panel, robust distances from monitoring WLE, the solid red line corresponds to 0.99 level threshold. Right-hand panel, empirical downweighting level from monitoring WLE, the solid red line corresponds to λ (color figure online)

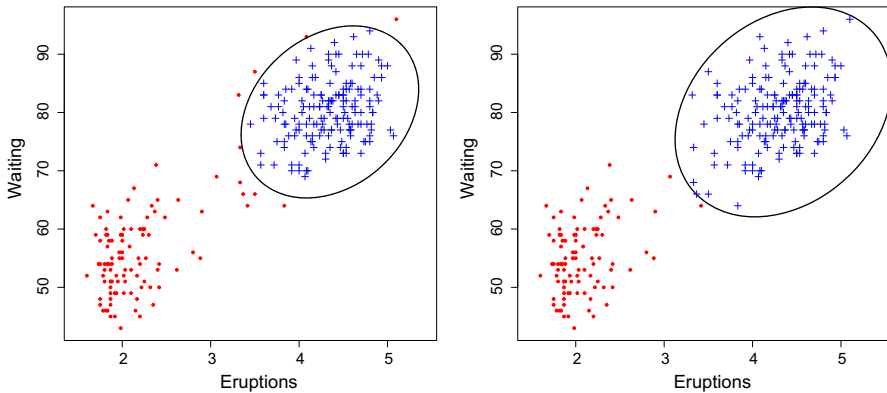


Fig. 2 Eruptions of Old Faithful. Left-hand panel, 0.99 tolerance ellipse. Right-hand panel, 0.999 tolerance ellipse. Outliers are plotted as filled red circles, clean data as blue + (color figure online)

from light gray to dark gray in order to highlight those trajectories corresponding to observations that are flagged as outlying for most of the monitoring. The right-hand plot monitors the empirical downweighting level $1 - \bar{w}$. Beyond $h = 0.01175$, $1 - \bar{w}$ abruptly decreases, meaning that the fitted model has drastically changed. It is worth mention that one could have monitored the trajectory of each single weight by varying h . Figure 2 displays the fitted model in the form of tolerance ellipses and shows the bulk of the data and those data points flagged as outliers with different colors and symbols. The WLE is able to recover the underlying structure of the data formed by two sub-groups. By varying the pointwise threshold, a slightly different number of outliers is detected: 108 in the left-hand panel at a 0.99-level, 99 in the right-hand panel at a 0.999 level.

3.2 Lightly contaminated data

This is a simulated data set with $n = 200$, $p = 5$. The first 30 rows are outliers. Outliers are grouped with no evident overlap with the bulk of the data. Figure 3 displays the monitoring for robust distances (left-hand panel) and the empirical downweighting level (right-hand panel). For a value of h up to 0.095, weighted likelihood provides a robust solution. Figure 4 shows the results from WLE with $h = 0.095$. Robust distances are shown in the left-hand panel, whereas weights are plotted in the right-hand panel. By using a 0.99-level pointwise threshold, 32 outliers are found, three of which are false positives. They all receive a final weight less than 0.2. The diagnostic analysis also leads to one false negative.

3.3 Heavily contaminated data

In this second simulated dataset, there are $n = 400$ observations with $p = 4$ features. The last 100 rows correspond to outlying observations. There is some overlapping between genuine and contaminated data that makes robust fitting and outlier detection

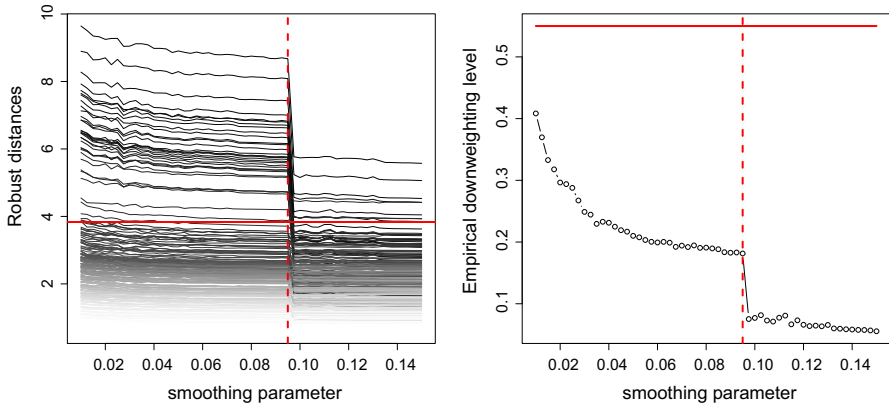


Fig. 3 Lightly contaminated data. Left-hand panel, robust distances from monitoring WLE, the solid red line corresponds to 0.99 level threshold. Right-hand panel, empirical downweighting level from monitoring WLE, the solid red line corresponds to λ (color figure online)

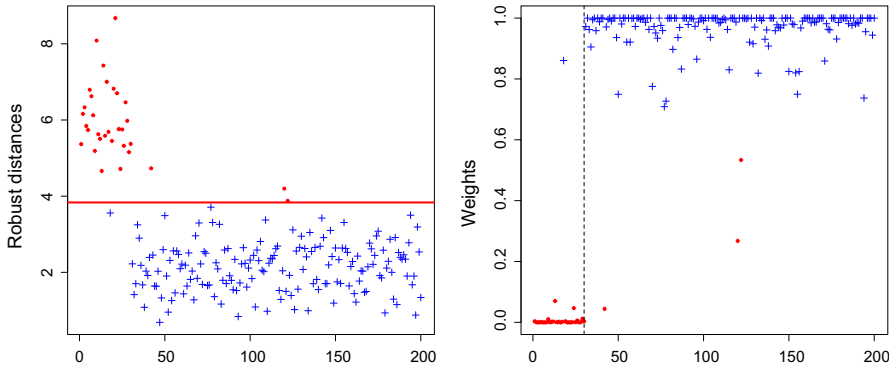


Fig. 4 Lightly contaminated data. Left-hand panel, robust distances, the solid red line corresponds to 0.99 level threshold. Right-hand panel, final weights, the vertical dashed line separates true outliers from true genuine data. Outliers are plotted as filled red circles, clean data as blue + (color figure online)

more challenging tasks. Monitoring of S-, MM- and MCD along with FS leads to remark that soft-trimming fails in this example whereas hard trimming is necessary in order to recover a reliable robust solution. In the following, the monitoring of WLE will show that soft-trimming based on the weighted likelihood carries on a robust solution that is well comparable with those stemming from FS and MCD. Both panels in Fig. 5 show the signal corresponding to the abrupt change in the monitored distances or empirical downweighting level, at $h = 0.00135$. A similar structure is not present in the monitoring plot driven by S- and MM-estimation. Outlier detection based on the 0.99-level pointwise threshold leads to flag 115 data points as outliers. The level of swamping is 19/300, whereas masking is 4/100. Figure 6 displays a scatterplot of the data in which genuine observations and outliers are plotted by using different colors and symbols. The reader may appreciate the close agreement with the solution stemming from the FS.

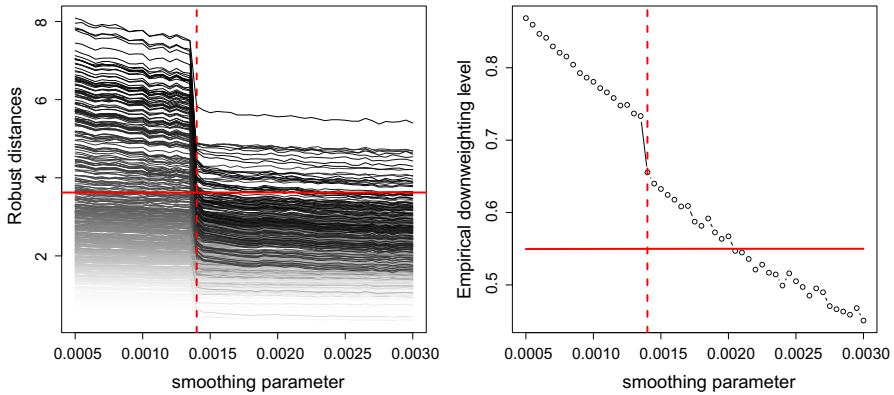


Fig. 5 Heavily contaminated data. Left-hand panel, robust distances from monitoring WLE, the solid red line corresponds to 0.99 level threshold. Right-hand panel, empirical downweighting level from monitoring WLE, the solid red line corresponds to λ (color figure online)

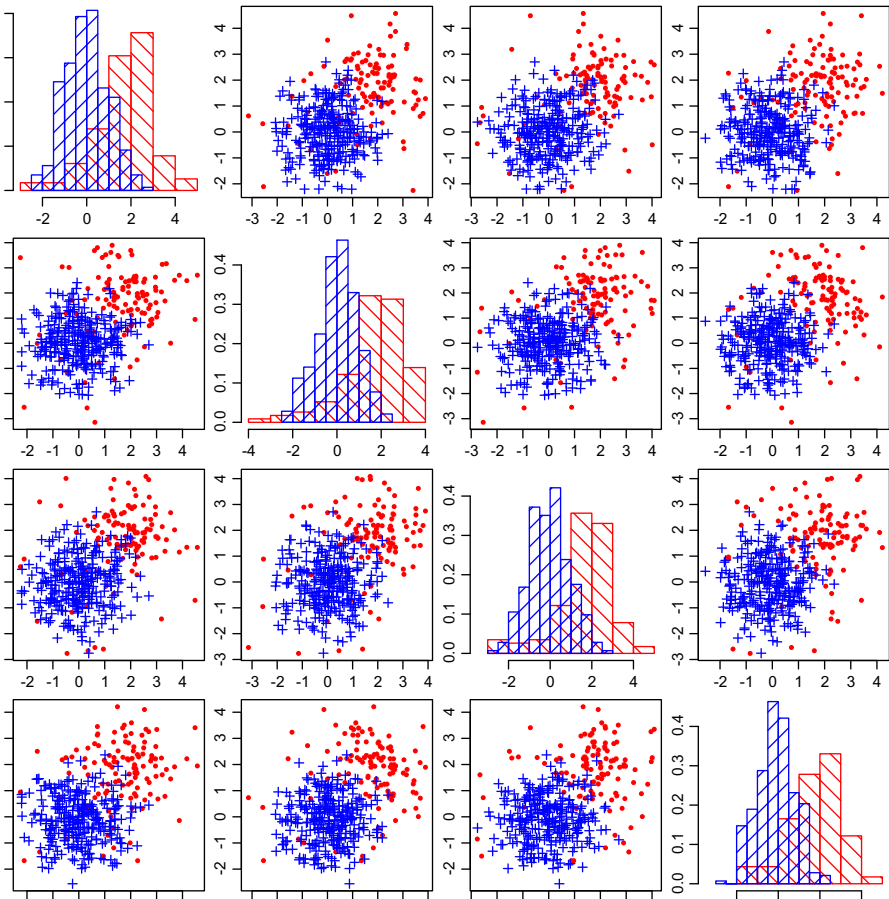


Fig. 6 Heavily contaminated data. Scatterplot of the data. Outliers are plotted as filled red circles, clean data as blue + (color figure online)

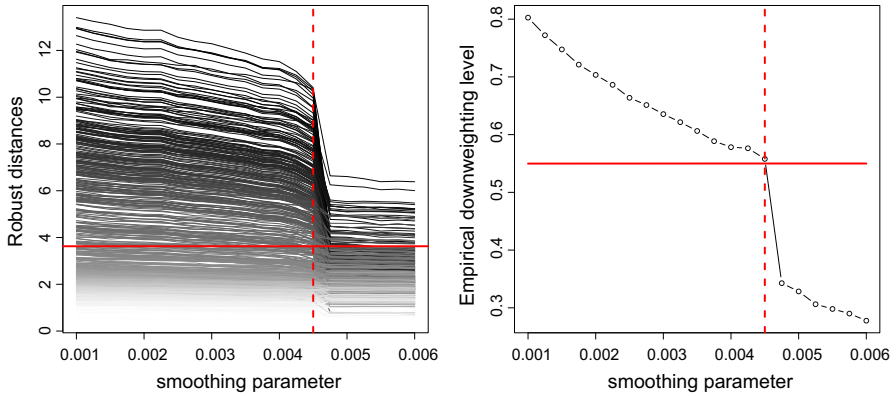


Fig. 7 Cows with Bovine Dermatitis. Left-hand panel, robust distances from monitoring WLE, the solid red line corresponds to 0.99 level threshold. Right-hand panel, empirical downweighting level from monitoring WLE, the solid red line corresponds to λ (color figure online)

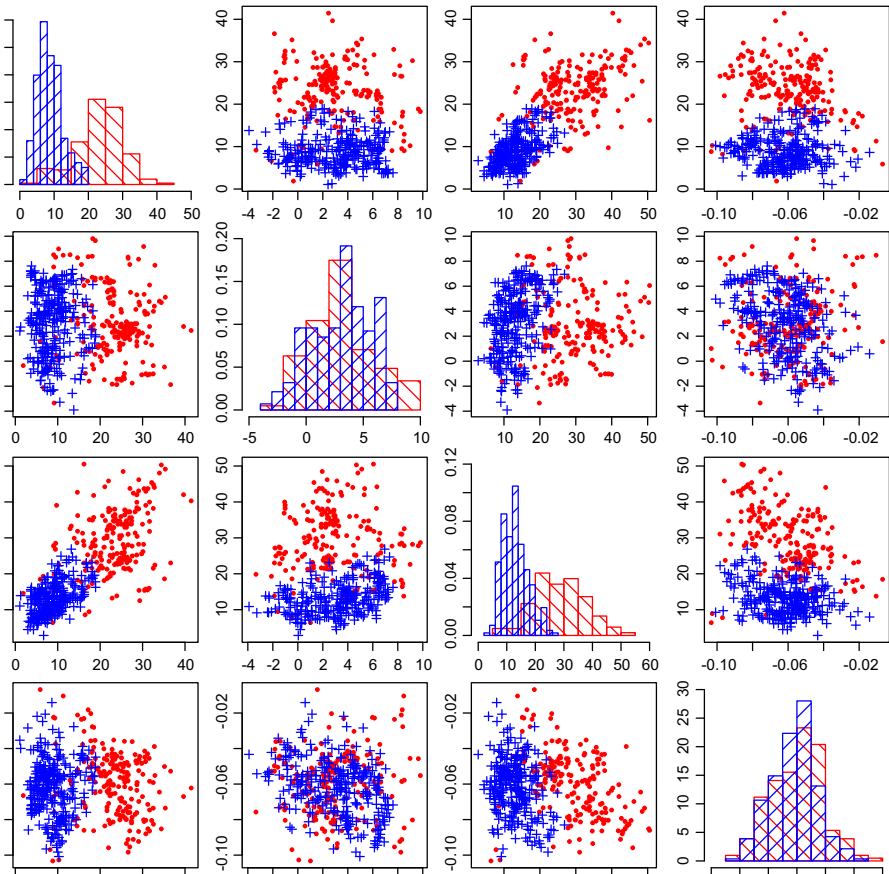


Fig. 8 Cows with Bovine Dermatitis. Scatterplot of the data. Outliers are plotted as filled red circles, clean data as blue + (color figure online)

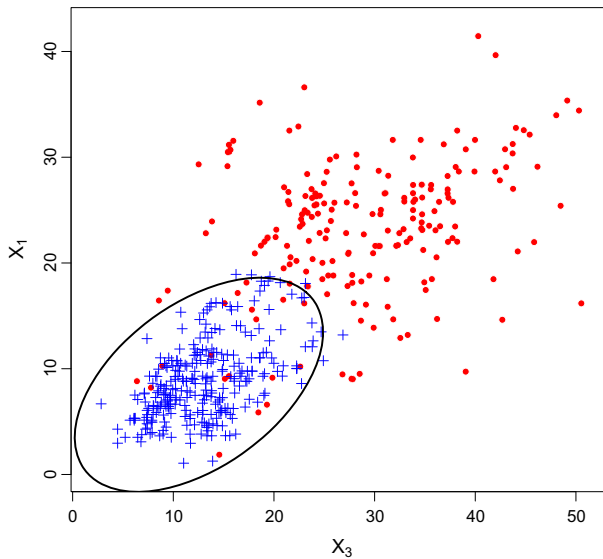


Fig. 9 Cows with Bovine Dermatitis. Scatterplot of the data with respect to X_1 and X_3 with 0.99 tolerance ellipse over-imposed. Outliers are plotted as filled red circles, clean data as blue + (color figure online)

3.4 Cows with Bovine Dermatitis

The last dataset concerns 488 cows with bovine dermatitis and four measurements per cow. As well as in the previous example, the monitoring plots for S- and MM-estimation are essentially smooth. Hence, they do not lead to a reasonable robust solution. On the contrary, the FS and the MCD are able to catch the underlying structure of the data, that is characterized by two groups of similar size. The inspection of Fig. 7 leads to state that up to $h = 0.0045$ the WLE is also able to provide a robust solution. Outlier detection based on 0.99-level pointwise threshold identifies 235 outliers. Figures 8 and 9 display the groups that have been identified, that are relatively well separated in some dimensions such as X_1 and X_3 , but overlapping in the other two. As well as before, the reader may appreciate the close agreement with the solution stemming from the FS.

References

- Agostinelli C (2002) Robust model selection in regression via weighted likelihood methodology. *Stat Probab Lett* 56:289–300
- Agostinelli C, Greco L (2013) A weighted strategy to handle likelihood uncertainty in Bayesian inference. *Comput Stat* 28(1):319–339
- Agostinelli C, Greco L (2017) Weighted likelihood estimation of multivariate location and scatter. arXiv preprint [arXiv:1706.05876](https://arxiv.org/abs/1706.05876)
- Agostinelli C, Markatou M (1998) A one-step robust estimator for regression based on the weighted likelihood reweighting scheme. *Stat Probab Lett* 37(4):341–350
- Alqallaf F, Agostinelli C (2016) Robust inference in generalized linear models. *Commun Stat Simul Comput* 45(9):3053–3073

- Basu A, Lindsay BG (1994) Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Ann Inst Stat Math* 46(4):683–705
- Farcomeni A, Greco L (2016) Robust methods for data reduction. CRC press, Boca Raton, FL
- Greco L (2016) Weighted likelihood based inference for $p(x < y)$. *Commun Stat Simul Comput*. <https://doi.org/10.1080/03610918.2016.1252396>
- Huber P (1985) Projection pursuit. *Ann Stat* 13(2):435–475
- Lindsay B (1994) Efficiency versus robustness: the case for minimum hellinger distance and related methods. *Ann Stat* 22:1018–1114
- Markatou M, Basu A, Lindsay BG (1998) Weighted likelihood equations with bootstrap root search. *J Am Stat Assoc* 93(442):740–750
- Park C, Basu A, Lindsay B (2002) The residual adjustment function and weighted likelihood: a graphical interpretation of robustness of minimum disparity estimators. *Comput Stat Data Anal* 39(1):21–33
- Scott D, Wand M (1991) Feasibility of multivariate density estimates. *Biometrika* 78(1):197–205