Case-wise and cell-wise outliers detection based on statistical depth filters

Giovanni Saraceno and Claudio Agostinelli

Abstract According to the classical case-wise contamination model, observations are considered as the units to be identified as outliers. Alqallaf et al. (2009) showed the limits of this approach, especially for a larger number of variables, and introduced the Independent contamination model, or cell-wise contamination, where the cells are the units to be identified as outliers. For the estimation problem, one approach to deal, at the same time, with both type of contamination is filter out the contaminated cells from the data set and then apply a robust procedure able to handle case-wise outliers and missing values. In this work we deal with the outliers detection task, taking into account both types of contamination. We propose to use the depth filters introduced by Saraceno and Agostinelli (2021) as detection procedure which is able to identify both case-wise and cell-wise outliers. We investigated the finite sample performance by a small simulation study, comparing the depth filters with the detection rules available in literature.

1 Introduction

It is well-known that one of the common problem in real data is the possible presence of outliers. According to the situation, they may be errors that affect the data analysis or can suggest unexpected information. In both cases, it can be crucial to detect such observations. Indeed, the investigation of their source could reveal hidden random mechanisms. Furthermore, the outliers are

Giovanni Saraceno (✉)
Department of Mathematics, University of Trento, Italy, e-mail: giovanni.saraceno@unitn.it

Claudio Agostinelli
Department of Mathematics, University of Trento, Italy, e-mail: claudio.agostinelli@unitn.it

model dependent then an effective detection rule is a successful strategy to improve the model estimation. In statistics, the outliers are typically referred to the observations, the rows of a data matrix. All the methods developed since 1960s in the field of robust statistics had the objective to be less sensitive to the presence of these row-wise outliers. For example, outlying observations can be identified by their large residuals from the fit or by large Mahalanobis distances. See Maronna et al. (2006) for a complete description of the developments in robust statistics. Among others, it is worth to cite the Forward Search (FS), a powerful general model for detecting multiple masked outliers in continuous multivariate data (Cerioli et al. , 2018). The idea is that the search starts by fitting a multivariate normal model starting from a subset of size d. Then, all the n observations are ordered by their Mahalanobis distance and the subset size is updated to $d+1$ by taking the $d+1$ observations with the smallest Mahalanobis distance until all the observations are included. This approach not only identifies the outlying observations but also provide a monitoring tool of the main statistics during the search.

The case-wise contamination paradigm can be not sufficient in modern applications with high-dimensional data, where only some of the entries, or cells, of a row can be contaminated. Alqallaf et al. (2009) firstly formulated this cell-wise contamination scheme and they noticed how they propagate, i.e. given a proportion ε of contaminated cells, the expected proportion of rows that contain at least one outlying cell is

$$1 - (1 - \varepsilon)^p$$

where p denotes the number of variables. This proportion easily exceed the 50% breakdown point for increasing contamination level ε and dimension p. For this reason, the existing methods may fail under the cell-wise contamination scheme. Finally, Agostinelli et al. (2015) showed that case-wise and cell-wise outliers can occur at the same time. One of the first and successful method proposed to cope with the cell-wise contamination is called *Detecting Deviating Cells* (DDC), introduced by Rousseeuw and Van den Bossche (2018), which takes into account the correlations between variables and provides predicted values of the outlying cells.

2 Half-space depth-filter

Let $\boldsymbol{X}$ be a $\mathbb{R}^d$-valued random variable and F a continuous distribution function.

Definition 1 (Half-space depth) For a point $\boldsymbol{x} \in \mathbb{R}^d$, the half-space depth of $\boldsymbol{x}$ with respect to F is defined as the minimum probability of all closed half-spaces including $\boldsymbol{x}$:

$$d_{HS}(\boldsymbol{x}; F) = \min_{H \in \mathcal{H}(\boldsymbol{x})} P_F(\boldsymbol{X} \in H),$$

where $\mathcal{H}(\boldsymbol{x})$ indicates the set of all half-spaces in $\mathbb{R}^d$ containing $\boldsymbol{x} \in \mathbb{R}^d$.

Given an independent and identically distributed sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of size n, we denote by $\hat{F}_n(\cdot)$ its empirical distribution function and by $d_{HS}(\boldsymbol{x}; \hat{F}_n)$ the sample half-space depth. We have that $d_{HS}(\boldsymbol{x}; \hat{F}_n)$ is a uniform consistent estimator of $d_{HS}(\boldsymbol{x}; F)$ (Donoho and Gasko, 1992), that is,

$$\sup_{\boldsymbol{x}} |d_{HS}(\boldsymbol{x}; \hat{F}_n) - d_{HS}(\boldsymbol{x}; F)| \overset{a.s.}{\rightarrow} 0 \qquad n \rightarrow \infty.$$

Given a statistical depth function, it is possible to define the α-depth trimmed region $R_\alpha(F)$, with $\alpha \in [0, m]$ and m denotes the maximum value obtained by the chosen depth. For the half-space depth it is given by

$$R_\alpha(F) = \{\boldsymbol{x} \in \mathbb{R}^d : d_{HS}(\boldsymbol{x}; F) \geq \alpha\},$$

and $\alpha \in [0, \frac{1}{2}]$. For any $\beta \in [0, 1]$, $R^\beta(F)$ will denote the smallest region $R_\alpha(F)$ that has probability larger than or equal to β according to F. Let $C^\beta(F)$ be the complement in $\mathbb{R}^d$ of the set $R^\beta(F)$.

Given a high order probability β, the half-space depth-filter of general dimension d is defined by (Saraceno and Agostinelli , 2021)

$$d_n = \sup_{\boldsymbol{x} \in C^\beta(F)} \{d_{HS}(\boldsymbol{x}; \hat{F}_n) - d_{HS}(\boldsymbol{x}; F)\}^+, \tag{1}$$

where $\{a\}^+$ represents the positive part of a. Then, the $n_0 = \lfloor n d_n \rfloor$ d-variate observations with the smallest population half-space depth are marked as outliers, where $\lfloor a \rfloor$ is the largest integer less then or equal to a. The half-space depth-filter satisfies the desired property of a consistent filter, that is $\frac{n_0}{n} \rightarrow 0$ as $n \rightarrow \infty$.

3 Outlier detection based on half-space depth-filters

Consider a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ where $\boldsymbol{X}_i \in \mathbb{R}^p, i = 1, \ldots, n$. The definition of the d-dimensional half-space depth-filter given in equation (1) allows to use the filtering procedure in a versatile way. In principle, it is possible to apply the d-variate depth-filter for all the values in the set $\{1, \ldots, p\}$. Each filter identifies the d-dimensional outliers that can be used to study possible substructures in the data. In practical situations, the choice of d can be dictated by previous knowledge about the phenomenon under investigation.

An efficient way to detect the outlying cells is to combine the output of the univariate and bivariate depth-filter, for $d = 1$ and $d = 2$, respectively.

The procedure used for this purpose is described in Leung et al. (2017) and summarized here.

We first apply the univariate filter to each variable separately. Let $\boldsymbol{X}^{(j)} = \{X_{1j}, \dots, X_{nj}\}$, $j = 1, \dots, p$, be a single variable. The univariate filter will flag $\lfloor n d_{nj} \rfloor$ observations as outliers, where d_{nj} is as in equation (1). Filtered data are indicated through an auxiliary matrix $\boldsymbol{U}$ of zeros and ones, with zero corresponding to an outlying cell. Next, we identify the bivariate outliers by iterating the filter over all possible pairs of variables. Consider a pair of variables $\boldsymbol{X}^{(jk)} = \{(X_{ij}, X_{ik}), i = 1, \dots, n\}$. For bivariate points with no flagged components by the univariate filter, we apply the bivariate filter. Given the pair of variables $\boldsymbol{X}^{(jk)}$, $1 \leq j < k \leq p$, we compute the value $d_n^{(jk)}$. Then, $n_0^{(jk)}$ couples will be identified as bivariate outliers. Finally, in order to identify the cells (i, j) which have to be flagged as cell-wise outliers, let

$$J = \{(i, j, k) : (X_{ij}, X_{ik}) \text{ is flagged as bivariate outlier}\}$$

be the set of triplets which identifies the pairs of cells flagged by the bivariate filter where $i = 1, \dots, n$ indicates the row. For each cell (i, j) in the data, we count the number of flagged pairs in the i-th row in which the considered cell is involved:

$$m_{ij} = \#\{k : (i, j, k) \in J\}.$$

In absence of contamination, m_{ij} follows approximately a binomial distribution $Bin(\sum_{k \neq j} \boldsymbol{U}_{jk}, \delta)$ where δ represents the overall proportion of cell-wise outliers undetected by the univariate filter. Hence, we flag the cell (i, j) if $m_{ij} > c_{ij}$, where c_{ij} is the 0.99-quantile of $Bin(\sum_{k \neq j} \boldsymbol{U}_{jk}, 0.1)$.

Finally, in order to detect both case-wise and cell-wise outliers simultaneously we can consider the filtering procedure which consists in applying the d-dimensional depth-filter three times in sequence, using $d = 1$, $d = 2$ and $d = p$. In practice, after applying the univariate and bivariate filters as described above, the p-variate filter is performed to the full data matrix. Detected observations (rows) are directly flagged as p-variate (case-wise) outliers. This procedure based on univariate, bivariate and p-variate filters has been denoted as HS-UBPF.

4 Simulation study

In order to illustrate the behaviour of the depth-filter HS-UBPF as detection rule, we consider a small simulation study where their performance is compared with the *Detecting Deviating Cells* algorithm, as implemented in the R (R Core Team, 2019) package `cellWise`, and the detection rule based on the Minimum Covariance Determinant (MCD).

We considered samples from a $N_p(\mathbf{0}, \boldsymbol{\Sigma}_0)$, where all values in $diag(\boldsymbol{\Sigma}_0)$ are equal to 1, $p = 20$ and the sample size is $n = 200$. We consider the following scenarios:

- Clean data: data without changes.
- Cell-Wise contamination: a proportion ε of cells in the data is replaced by $X_{ij} \sim N(k, 0.1^2)$, where $k = 1, \dots, 10$.
- Case-Wise contamination: a proportion ε of cases in the data is replaced by $\boldsymbol{X}_i \sim 0.5N(c\boldsymbol{v}, 0.1^2\boldsymbol{I}) \ + \ 0.5N(-c\boldsymbol{v}, 0.1^2\boldsymbol{I})$, where $c = \sqrt{k(\chi^2_p)^{-1}(0.99)}$, $k = 2, 4, \dots, 20$ and $\boldsymbol{v}$ is the eigenvector corresponding to the smallest eigenvalue of $\boldsymbol{\Sigma}_0$ with length such that $(\boldsymbol{v} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{v} - \boldsymbol{\mu}_0) = 1$.

The proportions of contaminated rows chosen for case-wise contamination are $\varepsilon = 0.1, 0.2$, and $\varepsilon = 0.02, 0.05, 0.1$ for cell-wise contamination. The number of replicates in our simulation study is $N = 50$.

We measure the performance of the considered rules by computing the values of accuracy and precision across the simulation parameters. Figure 1 and Figure 2 show the average accuracy and precision for different contamination values k and in case of case-wise and cell-wise contamination, respectively. The half-space depth-filter shows competitive performances with respect to DDC in terms of both accuracy and precision in case of cell-wise contamination. When the case-wise contamination is considered, the depth-filter has poorer results, while DDC confirms its high-quality results.

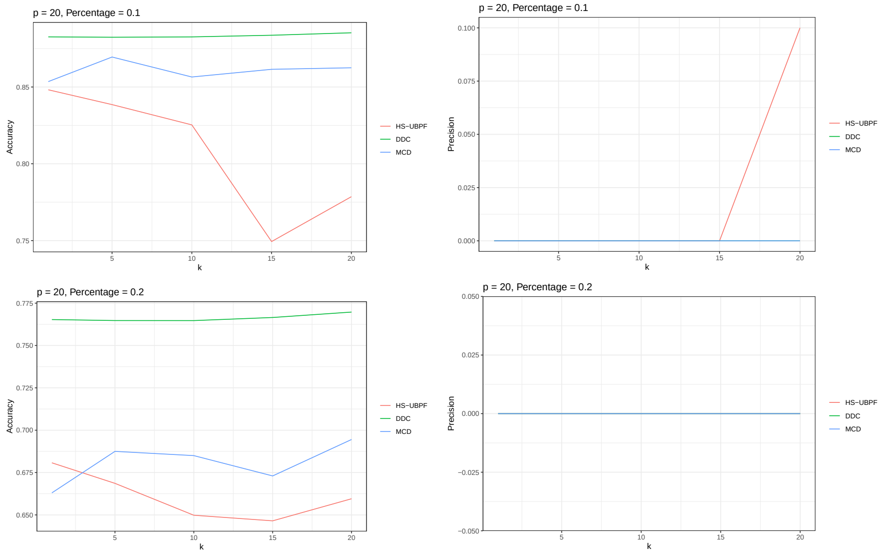

Fig. 1 Average accuracy (left) and precision (right) versus the contamination value k, for different case-wise contamination level and $p = 20$.

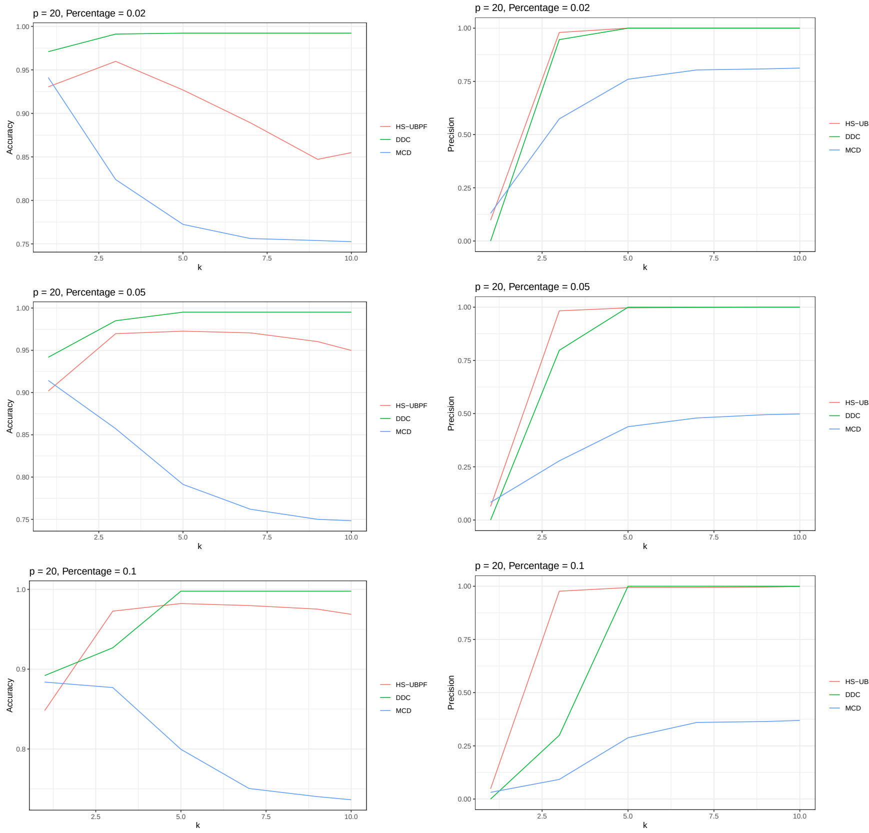

Fig. 2 Average accuracy (left) and precision (right) versus the contamination value k, for different cell-wise contamination level and $p = 20$.

References

Agostinelli, C., Leung, A., Yohai, V.J., Zamar, R.H. (2015) Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. TEST **24**(3), 441–461

Alqallaf, F., Van Aelst, S., Zamar, R.H., Yohai V.J. (2009) Propagation of outliers in multivariate data. The Annals of Statistics **37**(1), 311–331

Cerioli, A., Riani, M., Atkinson, A.C. and Corbellini, A. (2018), The power of monitoring: how to make the most of a contaminated multivariate sample, Statistical Methods & Applications, **27**(4), 559–587

R Core Team (2019). R: A Language and Environment for Statistical Computing. url: https://www.R-project.org/.

Donoho, D.L. and Gasko, M. (1992) Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness. The Annals of Statistics, **20**(4). doi: 10.1214/aos/1176348890.

Leung, A., Yohai, V.J., Zamar, R.H. (2017) Multivariate location and scatter matrix estimation under cellwise and casewise contamination. Computational Statistics and Data Analysis **111**, 59–76

Maronna, R.A., Martin, R.D., Yohai, V.J. (2006) Robust statistic: theory and methods. Wiley, Chichister

Rousseeuw, P.J. and Van den Bossche, W. (2018) Detecting Deviating Data Cells. Technometrics, **60**(2):135–145. doi: 10.1080/00401706.2017.1340909.

Saraceno, G., Agostinelli, C. (2021) Robust multivariate estimation based on statistical depth filters. TEST **30**, 935-–959. doi: 10.1007/s11749-021-00757-z