TEST (2021) 30:935–959
https://doi.org/10.1007/s11749-021-00757-z

ORIGINAL PAPER

Robust multivariate estimation based on statistical depth filters

Giovanni Saraceno[1] · **Claudio Agostinelli**[1]

Received: 11 September 2019 / Accepted: 23 January 2021 / Published online: 22 February 2021

Abstract
In the classical contamination models, such as the gross-error (Huber and Tukey contamination model or case-wise contamination), observations are considered as the units to be identified as outliers or not. This model is very useful when the number of considered variables is moderately small. Alqallaf et al. (Ann Stat 37(1):311–331, 2009) show the limits of this approach for a larger number of variables and introduced the independent contamination model (cell-wise contamination) where now the cells are the units to be identified as outliers or not. 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. Here, we develop a general framework to build filters in any dimension based on statistical data depth functions. We show that previous approaches, e.g., Agostinelli et al. (TEST 24(3):441–461, 2015b) and Leung et al. (Comput Stat Data Anal 111:59–76, 2017), are special cases. We illustrate our method by using the half-space depth.

Keywords Case-wise contamination · Cell-wise contamination · Filters · Robust statistics · Statistical depth functions

Mathematics Subject Classification 62G35 · 62G05

1 Introduction

One of most common problems in real data is the presence of outliers, i.e., observations that are well separated from the bulk of data, that may be errors that affect the data

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11749-021-00757-z.

✉ Giovanni Saraceno
giovanni.saraceno@unitn.it

[1] Dipartimento di Matematica, Università degli studi di Trento, Sommarive 14, 38123 Povo, Trento, Italy

analysis or can suggest unexpected information. According to the classical Tukey–Huber contamination model (THCM), a small fraction of rows can be contaminated and these are the units considered as outliers. Since the 1960s, many methods have been developed in order to be less sensitive to such outlying observations. A complete introduction and explanation of the developments in robust statistics is given in the book by Maronna et al. (2006).

In some application, e.g., in modern high-dimensional data sets, the entries of an observation (or cells) can be independently contaminated. Alqallaf et al. (2009) first formulated the independent contamination model (ICM), taking into consideration this cell-wise contamination scheme. According to this paradigm, given a fraction ϵ of contaminated cells, the expected fraction of contaminated rows is

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

which exceeds the 50% breakdown point for increasing value of the contamination level ϵ and the dimension p. Traditional robust estimators may fail in this situation. Furthermore, Agostinelli et al. (2015a) showed that both type of outliers, case-wise and cell-wise, can occur simultaneously.

Gervini and Yohai (2002) introduced the idea of an adaptive univariate filter, identifying the proportion of outliers in the sample measuring the difference between the empirical distribution and a reference distribution. Then, it is used to compute an adaptive cutoff value, and finally, a robust and efficient weighted least squares estimator is defined. Starting from this concept of outlier detection, Agostinelli et al. (2015b) introduced a two-step procedure: in the first step, large cell-wise outliers are flagged by the univariate filter and replaced by NA's values (a technique called snipping in Farcomeni 2014); in the second step, a generalized S-estimator (Danilov et al. 2012) is applied to deal with case-wise outliers. The choice of using GSE is due to the fact that it has been specifically designed to cope with missing values in multivariate data. Leung et al. (2017) improved this procedure proposing the following modifications:

- they combined the univariate filter with a bivariate filter to take into account the correlations among variables;
- in order to handle also moderate cell-wise outliers, they proposed a filter as intersection between the univariate-bivariate filter and *detect deviating cells* (DDC), a filter procedure introduced by Rousseeuw and Van Den Bossche (2018);
- finally, they constructed a generalized Rocke S-estimator (GRE) replacing the GSE, to face the loss of robustness in case of high-dimensional case-wise outliers.

Here, we introduce a general idea of constructing filters in general dimension d, with $1 \leq d \leq p$, based on the statistical data depth functions, namely depth filters. In particular, we show that the previously mentioned univariate–bivariate filter is a special case, if an appropriate statistical depth function is used.

We develop one of these depth filters using the half-space depth, HS-filter. Thus, we repropose the two steps procedure. In the first step, we apply the HS-filter taking $d = 1$, $d = 2$ and $d = p$, in sequence. As in Leung et al. (2017), the univariate and bivariate filters are combined in order to identify outlying cells which are replaced by NA's values. Note that, if $d = 1$, we filter the cell-wise outliers considering the variables as

independent. Finally, the HS-filter with $d = p$ is performed on observations, so that we can find undetected case-wise outliers. In the second step, the generalized S-estimator is used. Therefore, we also took into account the improvements suggested by Leung et al. (2017). Indeed, we improved our procedure following such modifications.

The rest of the work is organized as follows. Section 2 introduces the main idea on how to construct filters based on statistical depth functions. In Sect. 3, we show that the filters used in Agostinelli et al. (2015b) and Leung et al. (2017), namely GY-filters, are special cases of our proposed depth-filter approach, that is, they can be written in terms of depth functions. In order to prove that, we introduce a statistical data depth function called Gervini–Yohai depth function and we prove that the filter based on this depth coincides with the GY-filter. In Sect. 4, as an important example, we consider the filter obtained by using the half-space depth function and in Sects. 4.1, we introduce the proposed strategy to mark observations/cells as outliers. Section 5 reports the results of a Monte Carlo experiment, while Sect. 6 illustrates the features of our approach using a simulation example and a real data set. Concluding remarks are given in Sect. 7. In the Supplementary Material, Section SM-1 discusses the general properties that a statistical data depth function should satisfy. The derivation of the claim in Remark 1 is provided in Section SM-2. In Section SM-3, we prove that the general properties introduced in SM-1 hold for the Gervini–Yohai depth. Section SM-4 illustrates the univariate HS-filter with two-tails control, and Section SM-5 contains full results of the Monte Carlo experiment. Finally, Section SM-6 reports the codes used for the simulation example and for the real data set.

2 Filters based on statistical data depth function

Let $\boldsymbol{X}$ be a $\mathbb{R}^d$-valued random variable and F a continuous distribution function. For a point $\boldsymbol{x} \in \mathbb{R}^d$, we consider the statistical data depth of $\boldsymbol{x}$ with respect to F be $d(\boldsymbol{x}; F)$, where $d(\cdot, F)$ satisfies the four properties given in Liu (1990) and Zuo and Serfling (2000a) and reported in Section SM-1 of the Supplementary Material. 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 $d(\boldsymbol{x}; \hat{F}_n)$ the sample depth. We assume that $d(\boldsymbol{x}; \hat{F}_n)$ is a uniform consistent estimator of $d(\boldsymbol{x}; F)$, that is,

$$\sup_{\boldsymbol{x}} |d(\boldsymbol{x}; \hat{F}_n) - d(\boldsymbol{x}; F)| \xrightarrow{a.s.} 0 \qquad n \to \infty,$$

a property enjoined by many statistical data depth functions, e.g., among others simplicial depth (Liu 1990) and half-space depth (Donoho and Gasko 1992). One important feature of the depth functions is the α-depth trimmed region given by

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

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. Throughout, subscripts and superscripts for depth regions are used for depth levels and probability contents, respectively. Let

$C^{\beta}(F)$ be the complement in $\mathbb{R}^d$ of the set $R^{\beta}(F)$. Let $m = \max_{\boldsymbol{x}} d(\boldsymbol{x}; F)$ be the maximum value of the depth (for simplicial depth $m \leq 2^{-p}$, for half-space depth $m \leq 1/2$).

Given a high-order probability β, we define a filter of dimension d based on

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

where $\{a\}^+$ represents the positive part of a. Then, we mark as outliers all the $n_0 = \left\lfloor \frac{nd_n}{2m} \right\rfloor$ observations with the smallest population depth (where $\lfloor a \rfloor$ is the largest integer less than or equal to a). Given a depth function $d(\cdot, F)$, a desired property is that $\frac{n_0}{n} \to 0$ as $n \to \infty$. We recall the definition of consistent filter.

Definition 1 Consider a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, where $\boldsymbol{X}_i$ are generated by the distribution F_0 and some cells can be independently contaminated. Let $\mathcal{F}$ be a filter, a procedure that flags some cells as cell-wise outliers replacing them by NA's, and let d_n be the proportion of cells flagged by the filter. A filter is said *consistent* for a given distribution F_0 if asymptotically it will not flag any cell if the data come from the true distribution F_0. That is,

$$\lim_{n \to \infty} d_n \to 0 \quad a.s.\,[F_0].$$

Note that a statistical depth function can assume values in $\mathbb{R}^+ \cup \{0\}$. Hence, in order to be sure that the value d_n is a proportion, we need to normalize this value dividing by the maximum m of the depth. Intuitively, we can understand that the proportion of contaminated observations cannot exceed the 50% since, in this case, it would not be possible to distinguish between the underlying distribution of data and the contaminating distribution. So, in addition, we divide also by 2 so that the final proportion of flagged observations as outliers lies between 0 and $1/2$.

Remark 1 We verified that the filter proposed by Leung et al. (2017) has a similar property. In particular, the probability that $d_n \geq \frac{1}{2}$ goes to 0 as $n \to \infty$. The derivation of this result is shown in Section SM-2 of the Supplementary Material.

3 Gervini–Yohai d-variate filter

In this section, we are going to show that the filters introduced in Agostinelli et al. (2015b) and Leung et al. (2017) are a special case of our general approach to construct filters, that is, they can be expressed in terms of a depth function. For this reason, we are going to define a new depth, namely Gervini–Yohai depth, as follows:

$$d_{\mathrm{GY}}(\boldsymbol{t}, F, G) = 1 - G(\Delta(\boldsymbol{t}, \boldsymbol{\mu}(F), \boldsymbol{\Sigma}(F))),$$

where G is a continuous distribution function, $\boldsymbol{\mu}(F)$ and $\boldsymbol{\Sigma}(F)$ are the location and scatter matrix functionals and $\Delta(t, F) = \Delta(\boldsymbol{t}, \boldsymbol{\mu}(F), \boldsymbol{\Sigma}(F)) = (\boldsymbol{t} -$

$\boldsymbol{\mu}(F))^\top \boldsymbol{\Sigma}(F)^{-1}(\boldsymbol{t} - \boldsymbol{\mu}(F))$ indicates the squared Mahalanobis distance. In the Supplementary Material, Section SM-3 shows that this is a proper statistical data depth function since it satisfies the four properties that characterize a depth function.

Let $\{G_n\}_{n=1}^{\infty}$ be a sequence of discrete distribution functions that might depends on $\hat{F}_n$ and such that

$$\sup_t |G_n(t) - G(t)| \xrightarrow{a.s.} 0. \tag{2}$$

We might define the finite sample version of the Gervini–Yohai depth as

$$d_{\mathrm{GY}}(\boldsymbol{t}, \hat{F}_n, G_n) = 1 - G_n(\Delta(\boldsymbol{t}, \boldsymbol{\mu}(\hat{F}_n), \boldsymbol{\Sigma}(\hat{F}_n))).$$

However, for filtering purpose we will use two alternative definitions later on. The use of G_n, that might depend on the data, instead of G, makes this sample depth semiparametric.

Let $j_1, \ldots, j_d$, $1 \le d \le p$, be an d-tuple of the integer numbers in $\{1, \ldots, p\}$ and, for easy of presentation, let $\boldsymbol{Y}_i = (X_{ij_1}, \ldots, X_{ij_d})$ be a sub-vector of dimension d of $\boldsymbol{X}_i$. Consider a pair of initial location and scatter estimators

$$\boldsymbol{T}_{0n}^{(d)} = \begin{pmatrix} T_{0n,j_1} \\ \ldots \\ T_{0n,j_d} \end{pmatrix} \quad \text{and} \quad \boldsymbol{C}_{0n}^{(d)} = \begin{pmatrix} C_{0n,j_1j_1} & \ldots & C_{0n,j_1j_d} \\ \ldots & \ldots & \ldots \\ C_{0n,j_dj_1} & \ldots & C_{0n,j_dj_d} \end{pmatrix}.$$

Now, define the squared Mahalanobis distance for a data point $\boldsymbol{Y}_i$ by $\Delta_i = \Delta(\boldsymbol{Y}_i, \hat{F}_n) = \Delta(\boldsymbol{Y}_i, \boldsymbol{T}_{0n}^{(d)}, \boldsymbol{C}_{0n}^{(d)})$. Consider G the distribution function of a χ_d^2, H the distribution function of $\Delta = \Delta(\cdot, F)$ and let $\hat{H}_n$ be the empirical distribution function of Δ_i $(1 \le i \le n)$. We consider two finite sample version of the Gervini–Yohai depth, i.e.,

$$d_{\mathrm{GY}}(\boldsymbol{t}, \hat{F}_n, G) = 1 - G(\Delta(\boldsymbol{t}, \hat{F}_n)),$$

and

$$d_{\mathrm{GY}}(\boldsymbol{t}, \hat{F}_n, \hat{H}_n) = 1 - \hat{H}_n(\Delta(\boldsymbol{t}, \hat{F}_n)).$$

The proportion of flagged d-variate outliers is defined by

$$d_n = \sup_{\boldsymbol{t} \in A} \{d_{\mathrm{GY}}(\boldsymbol{t}, \hat{F}_n, \hat{H}_n) - d_{\mathrm{GY}}(\boldsymbol{t}, \hat{F}_n, G)\}^+.$$

Here, $A = \{\boldsymbol{t} \in \mathbb{R}^d : d_{\mathrm{GY}}(\boldsymbol{t}, F, G) \le d_{\mathrm{GY}}(\boldsymbol{\zeta}, F, G)\}$, where $\boldsymbol{\zeta}$ is any point in $\mathbb{R}^d$ such that $\Delta(\boldsymbol{\zeta}, F) = \eta$ and $\eta = G^{-1}(\alpha)$ is a large quantile of G. Then, we flag $\lfloor nd_n \rfloor$ observations. It is easy to see that

$$\begin{aligned} d_n &= \sup_{\boldsymbol{t} \in A} \{[1 - \hat{H}_n(\Delta(\boldsymbol{t}, \hat{F}_n))] - [1 - G(\Delta(\boldsymbol{t}, \hat{F}_n))]\}^+ \\ &= \sup_{\boldsymbol{t} \in A} \{G(\Delta(\boldsymbol{t}, \hat{F}_n)) - \hat{H}_n(\Delta(\boldsymbol{t}, \hat{F}_n))\}^+ \\ &= \sup_{\Delta \ge \eta} \{G(\Delta) - \hat{H}_n(\Delta)\}^+ \end{aligned}$$

since d_{GY} is a non-increasing function of the squared Mahalanobis distance of the point $\boldsymbol{t}$.

Remark 2 In principle, G_n could be any sequence of discrete distributions and for this reason we require that it satisfies condition (2). If G_n coincides with the empirical distribution of G, indicated as $\hat{G}_n$, such condition holds for the Glivenko–Cantelli lemma.

Remark 3 The Mahalanobis depth is defined as (Zuo and Serfling 2000a):

$$MHD(\boldsymbol{x}, F) = (1 + \Delta(\boldsymbol{x}, \boldsymbol{\mu}(F), \boldsymbol{\Sigma}(F))^{-1}, \ x \in \mathbb{R}^d.$$

Note that for a continuous distribution F, MHD is equivalent to the GY-depth. But the Mahalanobis depth, which is completely parametric, cannot be used in our approach to define filters.

We can rephrase Proposition 2 in Leung et al. (2017) that states the consistency property of the filter, as follows:

Proposition 1 *Consider a random vector $\boldsymbol{Y} = (X_1, \ldots, X_d) \sim F_0$ and a pair of location and scatter estimators $\boldsymbol{T}_{0n}$ and $\boldsymbol{C}_{0n}$ such that $\boldsymbol{T}_{0n} \to \boldsymbol{\mu}_0 = \boldsymbol{\mu}(F_0) \in \mathbb{R}^d$ and $\boldsymbol{C}_{0n} \to \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(F_0)$ a.s.. Consider any continuous distribution function G, and let $\hat{H}_n$ be the empirical distribution function of Δ_i and $H_0(t) = \Pr((\boldsymbol{Y} - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}_0) \leq t)$. If the distribution G satisfies:*

$$\max_{\boldsymbol{t} \in A} \{d_{\mathrm{GY}}(\boldsymbol{t}, F_0, H_0) - d_{\mathrm{GY}}(\boldsymbol{t}, F_0, G)\} \leq 0, \tag{3}$$

where $A = \{\boldsymbol{t} \in \mathbb{R}^d : d_{\mathrm{GY}}(\boldsymbol{t}, F_0, G) \leq d_{\mathrm{GY}}(\boldsymbol{\zeta}, F_0, G)\}$, where $\boldsymbol{\zeta}$ is any point in $\mathbb{R}^d$ such that $\Delta(\boldsymbol{\zeta}, F_0) = \eta$ and $\eta = G^{-1}(\alpha)$ is a large quantile of G, then

$$\frac{n_0}{n} \to 0 \quad a.s.$$

where

$$n_0 = \lfloor n d_n \rfloor.$$

Proof . Note that

$$d_{\mathrm{GY}}(\boldsymbol{t}, \hat{F}_n, \hat{H}_n) - d_{\mathrm{GY}}(\boldsymbol{t}, \hat{F}_n, G) = G(\Delta(\boldsymbol{t}, \boldsymbol{T}_{0n}, \boldsymbol{C}_{0n})) - \hat{H}_n(\Delta(\boldsymbol{t}, \boldsymbol{T}_{0n}, \boldsymbol{C}_{0n}))$$

and condition in equation (3) is equivalent to

$$\max_{\Delta \geq \eta} \{G(\Delta) - H_0(\Delta)\} \leq 0,$$

The rest of the proof is the same as in Proposition 2 of Leung et al. (2017). □

4 Filters based on half-space depth

In this section, we are going to give an example of depth filter considering the half-space depth $d_{\mathrm{HS}}(\cdot, F)$. In particular, we will prove the consistency property for this case.

Definition 2 (Half-space depth) Let $\boldsymbol{X}$ be a $\mathbb{R}^d$-valued random variable, and let F be a distribution function. 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 $\boldsymbol{x}$:

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

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 $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, we define the filter in general dimension d introduced previously, where here we use the half-space depth, as

$$d_n = \sup_{\boldsymbol{x} \in C^{\beta}(F)} \{d_{\mathrm{HS}}(\boldsymbol{x}; \hat{F}_n) - d_{\mathrm{HS}}(\boldsymbol{x}; F)\}^+, \tag{4}$$

where β is a high-order probability, $\hat{F}_n(\cdot)$ is the empirical distribution function and F is a chosen reference distribution which might depends, according to the assumed models, on unknown parameters, as in the case of location and dispersion models. In this last case, initial location and dispersion estimators, $\boldsymbol{T}_{0n}$ and $\boldsymbol{C}_{0n}$, are needed. As usual, $n_0 = \lfloor nd_n/2m \rfloor = \lfloor nd_n \rfloor$ observations with the smallest population depth are marked as outliers. Let F_0 be the true distribution of X, i.e., $\boldsymbol{X} \sim F_0$. Note that, so far we have no conditions on F_0. Here, we will prove the consistency property of the HS-filter when $\boldsymbol{X}$ is elliptically symmetric distributed.

Definition 3 A random vector $\boldsymbol{X} \in \mathbb{R}^d$ is said elliptically symmetric distributed, denoted by $\boldsymbol{X} \sim E_d(h_0, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, if it has a density function given by

$$f_0(\boldsymbol{x}) \propto |\boldsymbol{\Sigma}^{-1/2}| h_0((\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})).$$

where the density generating function h_0 is a non-negative scalar function, $\boldsymbol{\mu}$ is the location parameter vector and $\boldsymbol{\Sigma}$ is a $d \times d$ positive definite matrix.

Let $\boldsymbol{X} \sim E_d(h_0, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Denote by F_0 its distribution function and by $\Delta_{\boldsymbol{x}} = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ the squared Mahalanobis distance of a d-dimensional point $\boldsymbol{x}$. By Theorem 3.3 of Zuo and Serfling (2000b), if a depth $d(\cdot, \cdot)$ is affine equivariant (P1) and has maximum at $\boldsymbol{\mu}$ (P2) (see the Supplementary Material—Section SM-1), then the depth is of the form $d(\boldsymbol{x}; F_0) = g(\Delta_{\boldsymbol{x}})$ for some non-increasing function g. In this case, we can restrict ourselves, without loss of generality, to the case $\boldsymbol{\mu} = \boldsymbol{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix of dimension d. Under this setting, it is easy to see that the half-space depth of a given point $\boldsymbol{x}$ is given by

$$d_{\mathrm{HS}}(\boldsymbol{x}; F_0) = 1 - F_{0,1}(\sqrt{\Delta_{\boldsymbol{x}}}), \tag{5}$$

where $F_{0,1}$ is a marginal distribution of $\boldsymbol{X}$. Denoting the reference distribution by F, let $f \propto h(\Delta_x)$ be the corresponding density function. Note that if the function h is such that

$$\frac{h_0(\Delta_{\boldsymbol{x}})}{h(\Delta_{\boldsymbol{x}})} \to 0 \qquad \Delta_{\boldsymbol{x}} \to \infty, \tag{6}$$

then there exists a Δ^* such that, for all $\boldsymbol{x}$ with $\Delta_{\boldsymbol{x}} > \Delta^*$

$$d_{\mathrm{HS}}(\boldsymbol{x}; F) \geq d_{\mathrm{HS}}(\boldsymbol{x}; F_0).$$

Hence,

$$\sup_{\{\boldsymbol{x}:\Delta_{\boldsymbol{x}} > \Delta^*\}} [d_{\mathrm{HS}}(\boldsymbol{x}; F_0) - d_{\mathrm{HS}}(\boldsymbol{x}; F)] \leq 0$$

and therefore, for all $\beta > 1 - 2F_{0,1}(-\sqrt{\Delta^*})$,

$$\sup_{C^\beta(F)} [d_{\mathrm{HS}}(\boldsymbol{x}; F_0) - d_{\mathrm{HS}}(\boldsymbol{x}; F)] \leq 0.$$

In order to compute the value d_n, we have to identify the set $C^\beta(F) = \{\boldsymbol{x} \in \mathbb{R}^p : d_{\mathrm{HS}}(\boldsymbol{x}, F) \leq d_{\mathrm{HS}}(\eta_\beta, F)\}$ where η_β is such that the probability of $C^\beta(F)$ is equal to $1 - \beta$. In case we use the normal distribution as reference distribution, that is $F = N(\boldsymbol{T}_{0n}, \boldsymbol{C}_{0n})$, then, by Corollary 4.3 in Zuo and Serfling (2000b), the computation of $C^\beta(F)$ is particularly simple. In fact, denoting with $\Delta_x = (\boldsymbol{x} - \boldsymbol{T}_{0n})^\top \boldsymbol{C}_{0n}^{-1} (\boldsymbol{x} - \boldsymbol{T}_{0n})$ the squared Mahalanobis distance of $\boldsymbol{x}$ using the initial location and dispersion estimates, the set $C^\beta(F)$ can be rewritten as:

$$C^\beta(F) = \{\boldsymbol{x} \in \mathbb{R}^p : \Delta_x \geq (\chi^2_d)^{-1}(\beta)\}, \tag{7}$$

where $(\chi^2_d)^{-1}(\beta)$ is a large quantile of a Chi-squared distribution with d degrees of freedom. Now, we can state the consistency property for the HS-filter.

Proposition 2 *Consider a random vector* $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \sim F_0(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ *and suppose that* F_0 *is an elliptically symmetric distribution. Also consider a pair of location and dispersion estimators* $\boldsymbol{T}_{0n}$ *and* $\boldsymbol{C}_{0n}$ *such that* $\boldsymbol{T}_{0n} \to \boldsymbol{\mu}_0$ *and* $\boldsymbol{C}_{0n} \to \boldsymbol{\Sigma}_0$ *a.s.. Let* F *be a chosen reference distribution and* $\hat{F}_n$ *the empirical distribution function. Assume that* $F(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *is continuous with respect to* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$*. If the reference distribution satisfies*

$$\sup_{\boldsymbol{x} \in C^\beta(F)} [d_{\mathrm{HS}}(\boldsymbol{x}; F_0) - d_{\mathrm{HS}}(\boldsymbol{x}; F)] \leq 0 \tag{8}$$

where β *is some large probability, then*

$$\frac{n_0}{n} \to 0 \text{ as } n \to \infty$$

where $n_0 = \lfloor n d_n \rfloor$.

Proof In Donoho and Gasko (1992), it is proved that for $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ i.i.d. with distribution F_0, as $n \to \infty$

$$\sup_{\boldsymbol{t} \in \mathbb{R}^d} |d_{\mathrm{HS}}(\boldsymbol{t}, F_0) - d_{\mathrm{HS}}(\boldsymbol{t}, \hat{F}_n)| \to 0 \text{ a.s.}$$

Note that, by the continuity of F, $F(\boldsymbol{T}_{0n}, \boldsymbol{C}_{0n}) \to F(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ a.s.. Hence, for each $\varepsilon > 0$ there exists n^* such that for all $n > n^*$, we have

$$\begin{aligned}
&\sup_{\boldsymbol{x} \in C^{\beta}(F)} \{d_{\mathrm{HS}}(\boldsymbol{x}; \hat{F}_n) - d_{\mathrm{HS}}(\boldsymbol{x}; F(\boldsymbol{T}_{0n}, \boldsymbol{C}_{0n}))\} \\
&\leq \sup_{\boldsymbol{x} \in C^{\beta}(F)} \{d_{\mathrm{HS}}(\boldsymbol{x}; \hat{F}_n) - d_{\mathrm{HS}}(\boldsymbol{x}; F_0(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0))\} \\
&\quad + \sup_{\boldsymbol{x} \in C^{\beta}(F)} \{d_{\mathrm{HS}}(\boldsymbol{x}; F_0(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)) - d_{\mathrm{HS}}(\boldsymbol{x}; F(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0))\} \\
&\quad + \sup_{\boldsymbol{x} \in C^{\beta}(F)} \{d_{\mathrm{HS}}(\boldsymbol{x}; F(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)) - d_{\mathrm{HS}}(\boldsymbol{x}; F(\boldsymbol{T}_{0n}, \boldsymbol{C}_{0n}))\} \\
&\leq \frac{\varepsilon}{2} + 0 + \frac{\varepsilon}{2} = \varepsilon
\end{aligned}$$

which implies that $d_n = \sup_{\boldsymbol{x} \in C^{\beta}(F)} \{d_{\mathrm{HS}}(\boldsymbol{x}; \hat{F}_n) - d_{\mathrm{HS}}(\boldsymbol{x}; F(\boldsymbol{T}_{0n}, \boldsymbol{C}_{0n}))\}^+$ goes to zero as $n \to \infty$. Hence, $\frac{n_0}{n} \to 0$ as $n \to \infty$. □

Remark 4 We showed that if condition (6) holds, then assumption (8) of Proposition 2 is satisfied. In other words, even if the actual distribution is unknown, asymptotically, the filter will not wrongly flag any outlier when the tail of the chosen reference distribution is heavier than that of the actual distribution. In case F coincides to F_0, assumption 8 is clearly satisfied. We suggest to use for F the same distribution assumed for the model of the data.

Remark 5 When the underlying F_0 distribution is elliptical, a natural choice for T_{0n} and C_{0n} is as follows. For an univariate filter, $d = 1$, T_{0n} and C_{0n} might be, for example, the median and the MAD. In our study, when $d > 1$, as $\boldsymbol{T}_{0n}$ and $\boldsymbol{C}_{0n}$ we adopted the observation with maximum half-space depth, since the half-space depth corresponds to a generalization of the median in multivariate space, and the estimate given by a generalized S-estimator, respectively. Notice that these initial estimates satisfy the almost sure convergence assumption, under the nominal model F_0.

In Section SM-4 of the Supplementary Material, we added an example which illustrates the filter based on half-space depth for $d = 1$. In this case, it is possible to control independently the left and the right tail of the distribution and equation (4) assumes a simpler form. However, in our implementation, we always use the general formulation that does not make this distinction.

On the other hand, the computation of the sample half-space depth is demanding for $d > 1$, even in low dimensions, since it is based on all possible one-dimensional projections. Here, we decided to use the random Tukey depth introduced by Cuesta-Albertos and Nieto-Reyes (2008), a random approximation of the exact sample half-space depth, implemented in the `R` (R Core Team 2019) package `ddalpha` (Lange

et al. 2012). The reason is that approximate algorithms seem to be promising and, as pointed out in Cuesta-Albertos and Nieto-Reyes (2008), they may outperform exact algorithms in terms of computational time. Note that, the random Tukey depth is able to handle also the case $d = 50$, even if the computational time slightly increases. More information about exact algorithms can be found in Dyckerhoff and Mozharovskyi (2016). These algorithms allow the exact computation of half-space depth for moderate dimensions and sample sizes.

4.1 A consistent univariate, bivariate and *p*-variate filter

Consider a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ where $\boldsymbol{X}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$. In this subsection, we describe a filtering procedure which consists in applying the d-dimensional HS-filter given in equation (4) three times in sequence, using $d = 1$, $d = 2$ and $d = p$.

We first apply the univariate filter to each variable separately. Let $\boldsymbol{X}^{(j)} = \{X_{1j}, \ldots, X_{nj}\}$, $j = 1, \ldots, p$, be a single variable. The univariate filter will flag $\lfloor nd_{nj} \rfloor$ observations as outliers, where d_{nj} is as in equation (4), and these values are replaced by NA's values. Note that the initial location and variance estimators used here are the median and the MAD of $\boldsymbol{X}^{(j)}$. Filtered data are indicated through an auxiliary matrix $\boldsymbol{U}$ of zeros and ones, with zero corresponding to a NA value.

Next, we identify the bivariate outliers by iterating the filter over all possible pairs of variables. Consider a pair of variables $\boldsymbol{X}^{(jk)} = \{(X_{ij}, X_{ik}), i = 1, \ldots, n\}$. The initial location and dispersion estimators are, respectively, the observation with maximum depth and the 2×2 covariance matrix estimate S computed by the generalized S-estimator on non-filtered data $\boldsymbol{X}^{(jk)}$. 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 $d_n^{(jk)}$ given in equation (4). In particular, to compute the sample depth $d_{\mathrm{HS}}(\cdot, \hat{F}_n)$, we use the random Tukey depth, as mentioned before, through the function `depth.halfspace` implemented in the `R` package `ddalpha` (Lange et al. 2012).

Then, $n_0^{(jk)}$ couples will be identified as bivariate outliers. But, at the end, we want to identify the cells (i, j) which have to be flagged as cell-wise outliers. The procedure used for this purpose is described in Leung et al. (2017) and reported here. 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 $i = 1, \ldots, n$ indicates the row. For each cell (i, j) in the data, we count the number of flagged pairs in the ith row in which the considered cell is involved:

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

In the 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, we perform the p-variate filter to the full data matrix. Detected observations (rows) are directly flagged as p-variate (case-wise) outliers. We denote the procedure based on univariate, bivariate and p-variate filters as HS-UBPF.

4.2 A sequencing filtering procedure

Suppose we would like to apply a sequence of k filters with different dimension $1 \leq d_1 < d_2 < \ldots < d_k \leq p$. For each $d_i, i = 1, \ldots, k$, the filter updates the data matrix adding NA values to the d_i-tuples identified as d_i-variate outliers. In this way, each filter applies only those d_i-tuples that have not been flagged as outliers by the filters with lower dimension.

Initial values for each procedures rather than d_1 would be obtained by using the observation with the maximum half-space depth for location and the estimate given by GSE for the scatter matrix.

This procedure aims to be a valid alternative to that used in the presented HS-UBPF filter to perform a sequence of filters with different dimensions. However, this is a preliminary idea; indeed, it has not been implemented yet.

5 Monte Carlo results

We performed a Monte Carlo simulation to assess the performance of the proposed filter based on half-space depth. After the filter flags the outlying observations, the generalized S-estimator is applied to the data with added missing values. Part of our simulation study is based on the same setup described in Leung et al. (2017) since it seems a good choice to test our filter in the presence of contamination and the comparison with previous methods is easier. In particular, we compare the filter introduced in Agostinelli et al. (2015b) (indicated as GY-UF in case of univariate filter and GY-UBF for univariate and bivariate filter) and the same filter with the improvements proposed in Leung et al. (2017) (indicated here as GY-UBF-DDC-C) to the presented filter based on statistical data depth functions obtained using the half-space depth (HS-UF for the univariate filter, HS-UBF for the univariate-bivariate filter, HS-UBPF for the univariate-bivariate-p-variate filter and HS-UBPF-DDC-C for the combination of the HS-UBPF with the modifications in Leung et al. (2017)). The already existing filters are implemented in the `R` (R Core Team 2019) package `GSE` (Leung et al. 2015), whereas the `R` code for the proposed filter based on half-space depth is available in the `R` package `GSEdepth` provided as supplementary material.

We considered samples from a $N_p(\mathbf{0}, \boldsymbol{\Sigma}_0)$, where all values in $diag(\boldsymbol{\Sigma}_0)$ are equal to 1, $p = 10, 20, 30, 40, 50$ and the sample size is $n = 10p$. Since our model is the normal distribution, we choose the normal distribution as reference distribution. 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 $k = 1, \ldots, 10$.

– 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 $c = \sqrt{k(\chi^2_p)^{-1}(0.99)}$, $k = 2, 4, \ldots, 100$ and $\boldsymbol{v}$ is the eigenvector corresponding to the smallest eigenvalue of $\boldsymbol{\Sigma}_0$ with length such that $(\boldsymbol{v} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{v} - \boldsymbol{\mu}_0) = 1$.
– Mixed contamination: case-wise and cell-wise contaminations are introduced at the same time (after replacing a proportion of cases, a proportion of the remaining cells is contaminated).

The proportions of contaminated rows chosen for case-wise contamination are $\epsilon = 0.1, 0.2$, and $\epsilon = 0.02, 0.05, 0.1$ for cell-wise contamination. For the mixed contamination, we combined the proportions $\epsilon = 0.05, 0.1$ and $\epsilon = 0.02, 0.05$ for case-wise and cell-wise contamination, respectively. Finally, we tested the behavior of the procedure for increasing n. We considered $p = 5$ variables and $n = (10p, 50p, 100p)$ observations. Case-wise contamination and cell-wise contamination scenarios, as explained above, were performed on this setting. The number of replicates in our simulation study is $N = 200$.

We measure the performance of a given pair of location and scatter estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ using the mean-squared error (MSE) and the likelihood ratio test (LRT) distance:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_0)^\top(\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_0)$$

$$LRT(\hat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}_0) = \frac{1}{N}\sum_{i=1}^{N} D(\hat{\boldsymbol{\Sigma}}_i, \boldsymbol{\Sigma}_0)$$

where $\hat{\boldsymbol{\Sigma}}_i$ is the estimate of the ith replication and $D(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0)$ is the Kullback–Leibler divergence between two Gaussian distributions with the same mean and variances $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_0$. Finally, we computed the maximum average LRT distances and maximum average MSE considering all contamination values k.

Table 1 shows the maximum average LRT distances under cell-wise contamination. The univariate and univariate-bivariate filters have a similar behavior, while HS-UBPF has a lightly better performance. GY-UBF-DDC-C and HS-UBPF-DDC-C have lower maximum average LRT distances if the number of variables is not large, but their LRT distances are higher with respect to the other filters for large k. This behavior is shown in Fig. 1 (left) where the average LRT distances versus different contamination values are displayed, with 5% of cell-wise contamination level and $p = 30$.

Table 2 shows the maximum average LRT distances under case-wise contamination. Overall, the GY-UBP-DDC-C and HS-UBPF-DDC-C outperform all the other filters obtaining better results. An illustration of their behavior is given in Fig. 2 (top) which shows the average LRT distances for different values of k, with 10% of case-wise contamination level and $p = 30$.

Tables 3 and 4 show the maximum average MSE under cell-wise and case-wise contamination, respectively. The values in the tables are the MSE values multiplied by 100 for a better visualization and model comparison. Under case-wise contamination, the GY-UBF-DDC-C and HS-UBPF-DDC-C outperform the other filters, and have

Table 1 Maximum average LRT distance under cell-wise contamination

p	ϵ	UF		UBF		HS-UBPF	DDC-C		MLE
		GY	HS	GY	HS		GY-UBF	HS-UBPF	
10	0	0.8	0.7	0.9	0.7	0.8	1.0	1.0	0.6
	0.02	1.2	1.1	1.3	1.1	1.1	1.1	1.1	113.0
	0.05	4.6	4.8	4.6	4.9	4.8	2.4	2.5	290.5
	0.1	16.4	16.7	16.4	16.9	16.8	13.3	13.2	555.3
20	0	1.3	1.2	1.4	1.3	1.3	1.8	1.8	1.1
	0.02	3.9	3.8	4.2	4.0	3.8	2.5	2.5	146.4
	0.05	11.0	11.3	11.3	11.6	11.4	8.2	8.3	380.8
	0.1	24.4	24.6	24.5	25.1	24.7	21.6	21.8	742.7
30	0	1.9	1.8	2.0	1.9	1.9	3.4	3.4	1.6
	0.02	6.0	5.8	6.5	6.1	5.8	5.0	5.1	179.5
	0.05	14.5	14.7	15.1	15.3	14.9	13.4	13.4	470.5
	0.1	30.5	30.6	30.5	31.4	31.0	31.1	31.5	930.4
40	0	2.4	2.3	2.6	2.4	2.5	5.8	5.8	2.1
	0.02	7.5	7.4	8.2	7.8	7.4	9.2	9.2	213.2
	0.05	17.4	17.7	18.1	18.3	17.9	20.0	20.1	565.0
	0.1	35.6	35.7	35.6	36.5	36.1	41.4	42.4	1117.5
50	0	2.9	2.8	3.1	3.0	3.0	5.1	5.0	2.6
	0.02	8.8	8.6	9.7	9.1	8.8	12.2	12.3	245.7
	0.05	19.9	20.1	20.8	21.0	20.7	24.5	24.5	653.0
	0.1	40.0	40.1	40.0	41.0	40.6	44.7	44.3	1291.1

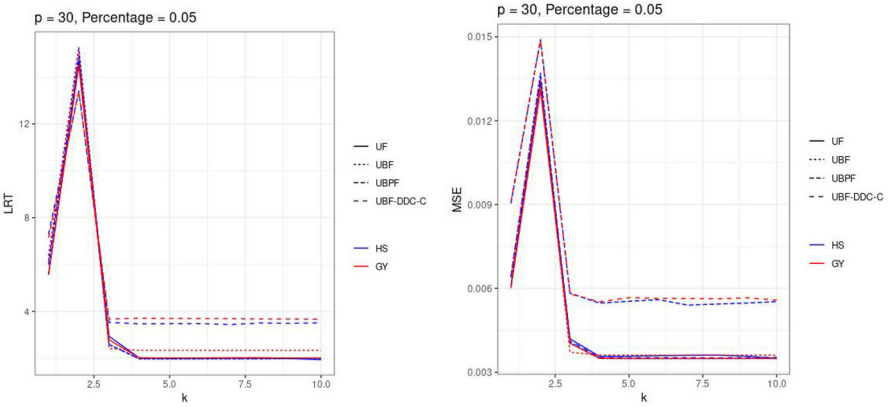

Fig. 1 Average LRT (left) and average MSE (right) versus the contamination value k, for 5% cell-wise contamination level and $p = 30$

Table 2 Maximum average LRT distance under case-wise contamination

p	ϵ	UF		UBF		HS-UBPF	DDC-C		MLE
		GY	HS	GY	HS		GY-UBF	HS-UBPF	
10	0	0.8	0.7	0.9	0.7	0.8	1.0	1.0	0.6
	0.1	9.8	7.6	14.9	8.5	6.2	3.5	3.4	893.9
	0.2	93.0	79.6	161.1	120.1	77.1	18.7	17.5	1593.6
20	0	1.3	1.2	1.4	1.3	1.3	1.8	1.8	1.1
	0.1	25.7	21.2	38.1	27.2	26.0	6.8	6.9	894.1
	0.2	368.0	322.3	428.9	441.0	373.8	19.6	20.1	1593.8
30	0	1.9	1.8	2.0	1.9	1.9	3.4	3.4	1.6
	0.1	50.8	44.9	64.0	70.3	68.6	9.0	8.7	895.0
	0.2	745.8	708.7	620.0	744.2	751.3	17.1	17.6	1595.1
40	0	2.4	2.3	2.6	2.4	2.5	5.8	5.8	2.1
	0.1	64.2	89.8	97.0	70.7	67.7	16.2	16.3	898.0
	0.2	1156.9	1112.1	852.0	1078.4	1088.0	22.7	21.4	1600.2
50	0	2.9	2.8	3.1	3.0	3.0	5.1	4.8	2.6
	0.1	175.2	215.6	123.3	156.6	163.9	30.5	29.9	898.0
	0.2	1528.8	1468.0	1081.6	1354.5	1364.5	21.2	20.1	1599.9

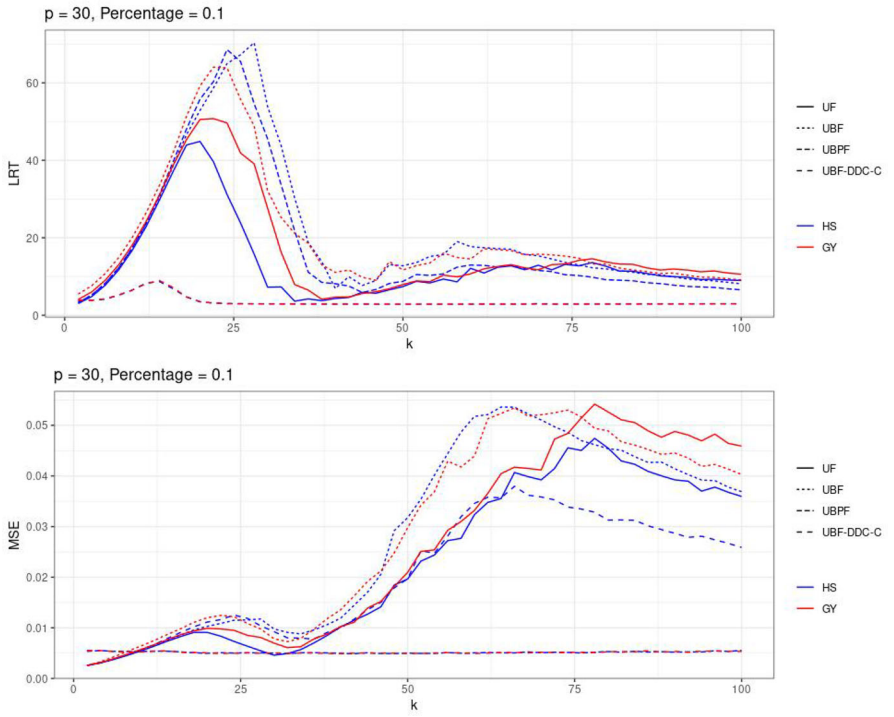

Fig. 2 Average LRT (top) and average MSE (bottom) versus the contamination value k, for 10% case-wise contamination level and $p = 30$

Table 3 Maximum average MSE distance under cell-wise contamination

p	ϵ	UF		UBF		HS-UBPF	DDC-C		MLE
		GY	HS	GY	HS		GY-UBF	HS-UBPF	
10	0	1.1	1.1	1.1	1.1	1.1	1.3	1.3	1.0
	0.02	1.3	1.3	1.3	1.3	1.3	1.5	1.5	6.8
	0.05	1.9	2.0	2.0	2.0	2.0	2.0	2.0	30.2
	0.1	4.8	4.9	4.8	4.9	4.9	5.0	5.0	109.2
20	0	0.5	0.5	0.5	0.5	0.5	0.7	0.7	0.5
	0.02	0.7	0.7	0.7	0.7	0.7	0.8	0.8	5.4
	0.05	1.5	1.5	1.5	1.5	1.5	1.6	1.6	27.8
	0.1	4.4	4.5	4.5	4.6	4.6	4.6	4.7	104.7
30	0	0.3	0.3	0.4	0.3	0.4	0.6	0.6	0.3
	0.02	0.5	0.5	0.5	0.5	0.5	0.7	0.7	4.9
	0.05	1.3	1.3	1.3	1.4	1.4	1.5	1.5	26.8
	0.1	4.3	4.3	4.3	4.4	4.4	4.5	4.7	103.2
40	0	0.3	0.3	0.3	0.3	0.3	0.6	0.6	0.2
	0.02	0.4	0.4	0.5	0.4	0.4	0.7	0.7	4.7
	0.05	1.3	1.3	1.3	1.3	1.3	1.5	1.6	26.4
	0.1	4.3	4.3	4.3	4.4	4.4	4.5	4.6	102.5
50	0	0.2	0.2	0.2	0.2	0.2	0.4	0.3	0.2
	0.02	0.4	0.4	0.4	0.4	0.4	0.6	0.6	4.6
	0.05	1.2	1.2	1.2	1.3	1.3	1.4	1.4	26.1
	0.1	4.2	4.2	4.2	4.4	4.3	4.3	4.5	101.9

also competitive results for cell-wise contamination. In Fig. 1 (right) and Fig. 2 (bottom), the average MSE versus different contamination values k are displayed, with $p = 30$ and 0.05 of cell-wise contamination and 0.1 of case-wise contamination, respectively.

The results given by the mixed contamination scenario do not show any additional information and they are not reported.

Finally, Figs. 3 and 4 show the average LRT and average MSE with respect to different value of k, for 10% of case-wise contamination and 5% of cell-wise contamination, respectively, for $p = 5$ and different number of observations n. For increasing n, the filters perform better showing smaller average LRT and average MSE values. In particular, depth filters present better improvements in case of case-wise contamination and they seem to perform better than those in combination with DDC.

In a second Monte Carlo experiment, we use the location-scale family of multivariate Student's t-distribution with 5 degrees of freedom as reference distribution F. We consider two data generation processes: In the first case, data are simulated from the multivariate normal distribution and in the second case, data are simulated from a t_5 distribution with 5 degrees of freedom. Apart from this, the setup of the experiment is the same of the previous one. The construction of the half-space-filter for this case follows directly from the definition given in equation (4), with just one change.

Table 4 Maximum average MSE distance under case-wise contamination

p	ϵ	UF		UBF		HS-UBPF	DDC-C		MLE
		GY	HS	GY	HS		GY-UBF	HS-UBPF	
10	0	1.1	1.1	1.1	1.1	1.1	1.3	1.3	1.0
	0.1	2.8	2.5	3.2	2.9	1.9	1.9	1.9	21.8
	0.2	15.1	14.2	20.1	16.1	9.7	2.5	2.8	84.4
20	0	0.5	0.5	0.5	0.5	0.5	0.7	0.7	0.5
	0.1	3.5	2.9	4.2	4.0	2.7	0.8	0.8	10.8
	0.2	28.6	25.8	34.1	25.9	21.3	1.3	1.2	41.9
30	0	0.3	0.3	0.4	0.3	0.4	0.6	0.6	0.3
	0.1	5.4	4.7	5.3	5.4	3.8	0.6	0.6	7.1
	0.2	50.6	46.7	37.2	46.5	48.0	0.8	0.8	27.6
40	0	0.3	0.3	0.3	0.3	0.3	0.6	0.6	0.2
	0.1	7.1	6.6	6.1	6.4	4.7	0.5	0.5	5.3
	0.2	41.6	38.1	34.7	38.9	39.7	0.7	0.7	20.6
50	0	0.2	0.2	0.2	0.2	0.2	0.4	0.3	0.2
	0.1	7.9	7.6	6.3	6.2	5.0	0.5	0.5	4.3
	0.2	32.4	30.0	30.6	31.9	32.5	0.5	0.5	16.5

In particular, since the t distribution belongs to the family of elliptically symmetric distribution, equation (5) holds and it is used to compute the theoretical depth. On the other hand, the sample depth is again computed using the random Tukey depth. Complete results are not reported. In this new setup, the HS-filters are still competitive for case-wise contamination, while they outperform the GY-filters in case of cell-wise contamination. This performance does not change if observations are sampled from a normal distribution or a t-distribution.

6 Examples

In Sect. 6.1, we illustrate how depth-filters approach can be used in models different from the location and scatter model with elliptical contours. In particular, we provide details of applying such filters to multivariate Skew-Normal distributions. A real-data application is reported in Sect. 6.2. The `R` package `GSEdepth`, available as supplementary material, implements the new procedures and contains the used data set.

6.1 Multivariate skew-normal distributions

In this example, we consider a p-multivariate skew-normal random variable $\boldsymbol{X} \sim SN_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha})$, with a location parameter $\boldsymbol{\xi}$, a positive definite scatter matrix $\boldsymbol{\Omega}$, and a skewness vector parameter $\boldsymbol{\alpha}$. We point out the reader to Azzalini (2014) for the details

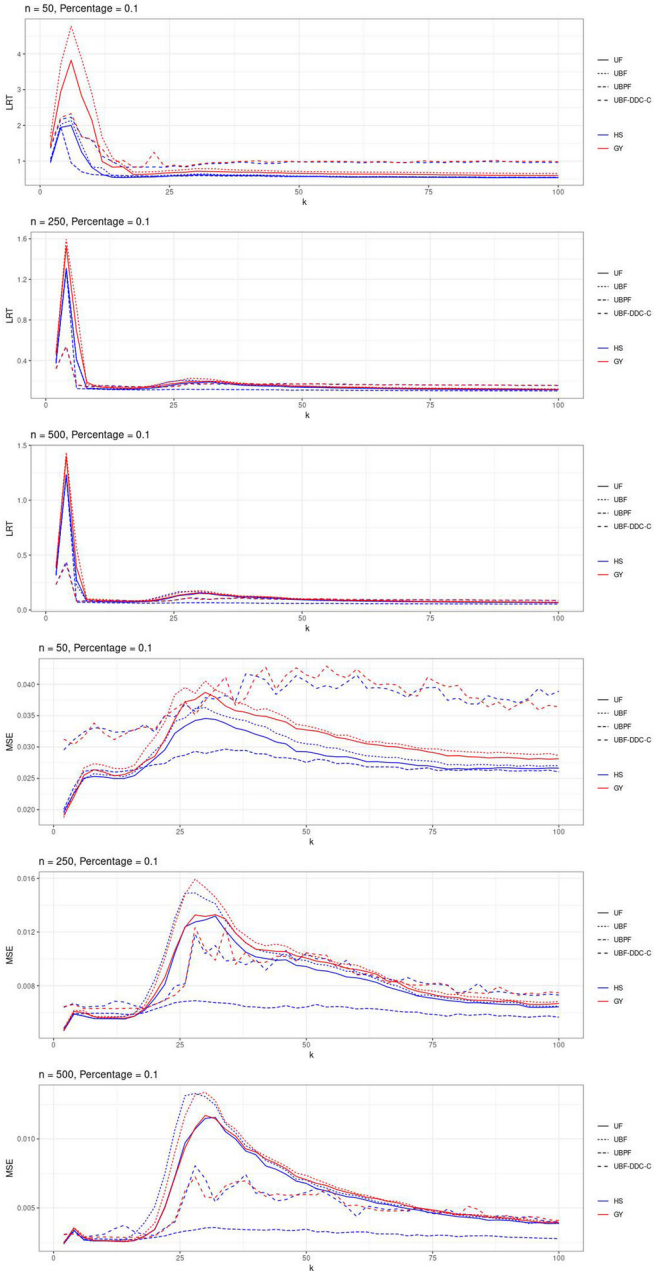

Fig. 3 Average LRT (top) and average MSE (bottom) in 0.1 case-wise contamination level versus the contamination value k, for $p = 5$ and $n = 50, 250, 500$

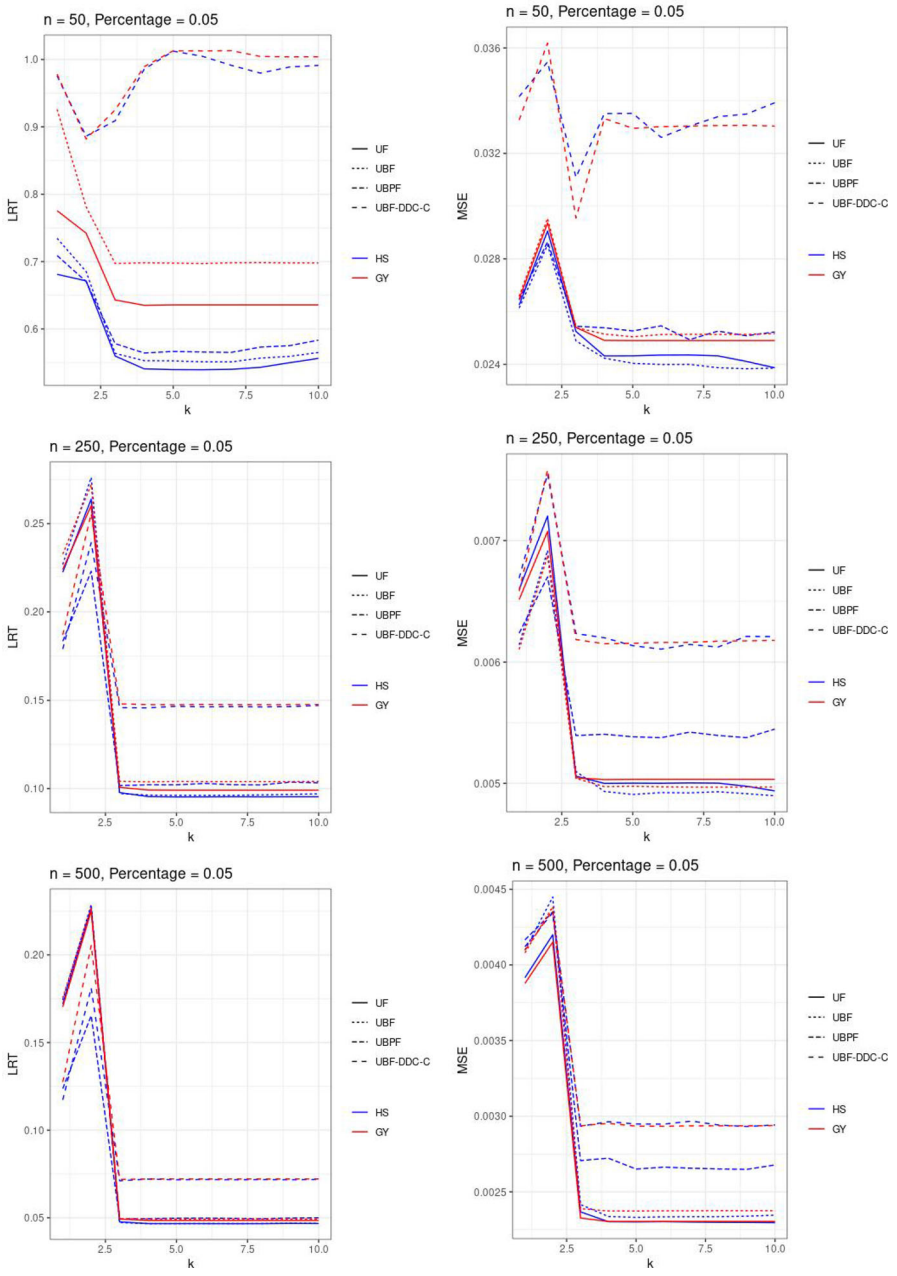

Fig. 4 Average LRT (left) and average MSE (right) in 0.05 cell-wise contamination level versus the contamination value k, for $p = 5$ and $n = 50, 250, 500$

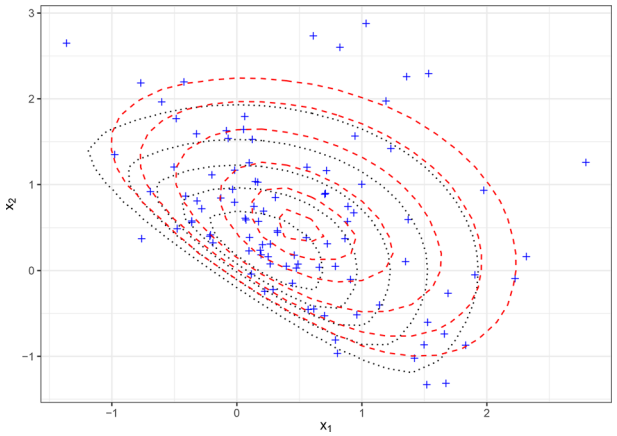

Fig. 5 Contour plot of the density of the skew-normal (black dotted lines) and of the half-space depth (red dashed lines). Sample observations are blue crosses (color figure online)

on multivariate skew-normal distributions. The mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ do not coincide with the distribution parameters; however, they are easily evaluated as (Azzalini 2014, formulas 2.27, 5.31 and 5.32)

$$\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{X}) = \boldsymbol{\xi} + \omega\boldsymbol{\nu}, \quad \Sigma = \mathbb{C}\mathrm{ov}(\boldsymbol{X}) = \boldsymbol{\Omega} - \omega\nu\nu^\top\omega\,,$$

where $\nu = \sqrt{\frac{2}{\pi}}(1+\boldsymbol{\alpha}^\top\bar{\Omega}\boldsymbol{\alpha})^{-1/2}\bar{\Omega}\boldsymbol{\alpha}$ while $\bar{\Omega}$ and ω are, respectively, the correlation matrix obtained from Ω and a diagonal matrix with the square-root of the diagonal elements of Ω. We are going to apply the GY-filter and the HS-filter in this framework, using as reference distribution the skew-normal model, evaluated at the true parameters value. Subsection SM-6-1 of the Supplementary Material provides all the necessary code to replicate the results and the figures.

A sample of size $n = 100$ is obtained, and it is represented in Fig. 5 (blue crosses) together with the density contours (black dotted lines) and the half-space depth contours (red dashed lines).

The GY-filters that are based on Mahalanobis distances need the mean vector and the variance–covariance matrix to be computed. The half-space-depth filters work directly with the actual parametrization of the reference distribution. While the set $C^\beta(F)$ is always an ellipse for GY-filters, this is not the case for half-space-depth filters, which, instead, depends on the shape of the reference distribution, and in this case, it is able to take into account the asymmetry of the skew-normal distribution.

We are going to add artificially 20 outlying observations sampled from a $N_2((-0.2,-0.25), 0.01\boldsymbol{I}_2)$ in an iterative procedure. Note that these points, with high probability, lie inside the boundary set given by the Mahalanobis distance but outside the boundary set computed using the half-space depth. This position is clearly crucial; however, it is a region of low density according to the true model. In each iteration, an outlier is added to the data set and the number of flagged observations n_0 is computed and reported in Table 5. The GY-filter is insensitive to this kind of

Table 5 Number of flagged observations by the GY-filter and the HS-filter for increasing number of added outliers placed at $N_2((-0.2, -0.25), 0.01\boldsymbol{I}_2)$

n^o of outliers	1	2	3	4	5	6	7	8	9	10
GY-filter	4	4	4	4	4	4	4	4	4	4
HS-filter	3	3	3	3	4	5	6	7	8	9
n^o of outliers	11	12	13	14	15	16	17	18	19	20
GY-filter	4	4	4	3	3	3	3	3	3	3
HS-filter	10	11	12	13	14	15	15	17	18	19

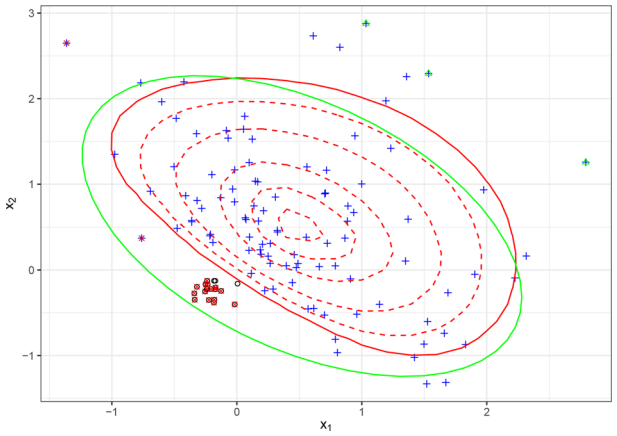

Fig. 6 $C^{\beta}(F)$ based on GY-filter is in solid green, while for HS-filter is in solid red. Half-space depth contours are red dashed lines, and sample observations are blue crosses. The 20 added outliers are black circles. Observations flagged by the GY-filter are green triangles, while those flagged by the HS-filter are red crosses. Outliers are placed at $N_2((-0.2, -0.25), 0.01\boldsymbol{I}_2)$ (color figure online)

outliers; indeed, the number of detected cells is stable or decreases as the number of added outliers increases. Vice versa, the number of detected cells by the HS-filter is almost always equal to the amount of added outliers.

In this simulation, we are also interested in identifying such flagged points. Figure 6 shows the added outliers at the final iteration (as black circles). Observations flagged by the HS-filter are red crosses, while those flagged by the GY-filter are green triangles. The HS-filter correctly identifies the majority of the added cells, while these are never detected by the GY-filter. Indeed, GY-filter flags regular observations which lead to a more symmetric empirical distribution.

In a second experiment, we sampled the added outliers from $N_2((-0.5, -0.6), 0.01\boldsymbol{I}_2)$ so that with high probability, the outliers lie in a region outside the boundary set given by the Mahalanobis distance. While in this case the GY-filter flags the right amount of observations (see Table 6), most of them do not belong to the set of added outliers. The only effect is, again, to reduce the asymmetry of the observed empirical distribution. Figure 7 shows the flagged observations after 10 added outliers (left panel) and at the final step (right panel).

Table 6 Number of flagged observations by the GY-filter and the HS-filter for increasing number of added outliers placed at $N_2((-0.5, -0.6), 0.01\boldsymbol{I}_2)$

n^o of outliers	1	2	3	4	5	6	7	8	9	10
GY-filter	4	5	5	6	6	7	8	9	10	10
HS-filter	3	3	3	4	5	6	6	7	8	8
n^o of outliers	11	12	13	14	15	16	17	18	19	20
GY-filter	11	12	13	14	15	16	17	18	19	20
HS-filter	9	10	11	12	13	13	14	15	16	17

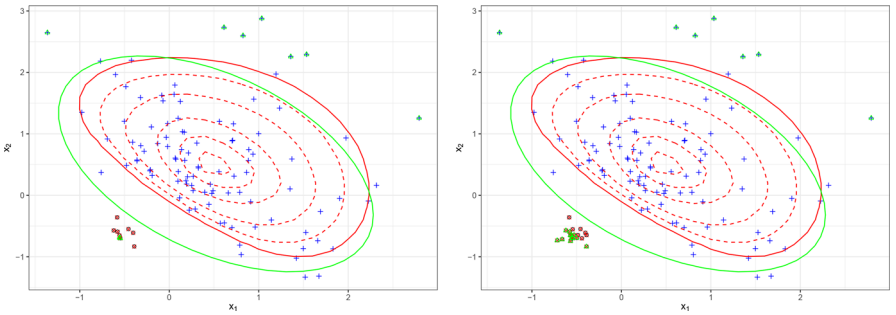

Fig. 7 $C^{\beta}(F)$ based on GY-filter is in solid green, while for HS-filter is in solid red. Half-space depth contours are red dashed lines, and sample observations are blue crosses. The added outliers are black circles. Observations flagged by the GY-filter are green triangles, while those flagged by the HS-filter are red crosses. Outliers are placed at $N_2((-0.5, -0.6), 0.01\boldsymbol{I}_2)$. Left panel: 10 added outliers, right panel: 20 added outliers (color figure online)

6.2 Small-cap stock returns

We consider the weekly returns from 01/01/2008 to 12/28/2010 for a portfolio of 20 small-cap stocks from Martin (2013). The data set is publicly available at the link "http://www.bearcave.com/finance/smallcapweekly.csv" and can be found in the `R` package `GSEdepth`. Subsection SM-6-2 of the Supplementary Material provides the necessary code to replicate the results and the figures.

With this example, we want to compare the filter introduced in Agostinelli et al. (2015b) and the same filter with the improvements proposed in Leung et al. (2017) to the presented filter based on statistical data depth functions obtained using the half-space depth.

Figure 8 shows the normal QQ-plots of the 20 variables. The returns in all stocks seem to roughly follow a normal distribution, but with the presence of large outliers. The returns in each stock that lie 3 MAD's away from the coordinate-wise median are displayed in green in the figure. These indicated cells, which are considered cell-wise outliers, correspond to the 4.4% of the total cells, and they propagate to 37.6% of the cases.

Figure 9 shows the squared Mahalanobis distances (MDs) of the weekly returns based on the estimates given by the MLE, the GY-UF, the GY-UBF, the HS-UF, the HS-UBF and the HS-UBPF. Observations with one or more cells flagged as outliers

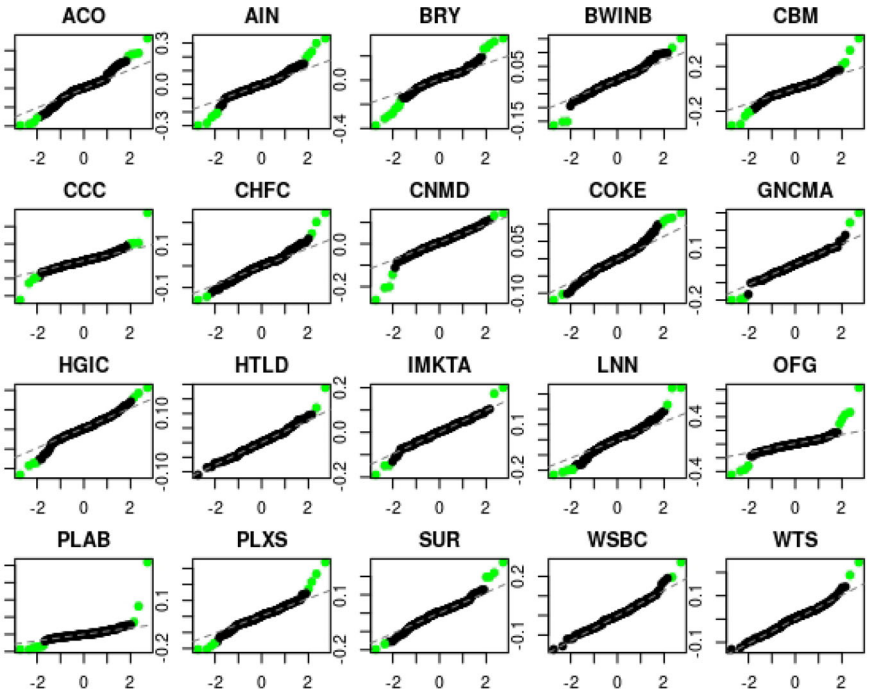

Fig. 8 Small-cap stock returns. QQ-plots of the variables, green: observations marked as outliers (color figure online)

are displayed in green. We say that the estimate identifies an outlier correctly if the MD exceeds the 99.99% quantile of a Chi-squared distribution with 20 degrees of freedom. We see that the MLE estimate does a very poor job recognizing only 8 of the 59 cases. The GY-UF, HS-UF, HS-UBF and HS-UBPF show a quite similar behavior, doing better than the MLE, but they miss about one-third of the cases. The GY-UBF identifies all but seven of the cases.

Figure 10 shows the Mahalanobis distances produced by GY-UBF-DDC-C and HS-UBPF-DDC-C. Here, we can see that the GY-UBF-DDC-C misses 13 of 59 cases, while the HS-UBPF-DDC-C has missed 12 cases. Although they seem not to do a better job, these two filters are able to flag some other observations, not identified before, as case-wise outliers.

Figure 11 shows the bivariate scatter plot of WTS versus HTLD, HTLD versus WSBC and WSBC versus SUR where the GY-UBF and HS-UBF filters are applied, respectively. The bivariate observations with at least one component flagged as outlier are in blue, while outliers detected by the bivariate filter, but excluded by the univariate filter, are in orange. We see that the HS-UBF identifies less outliers with respect to the GY-UBF.

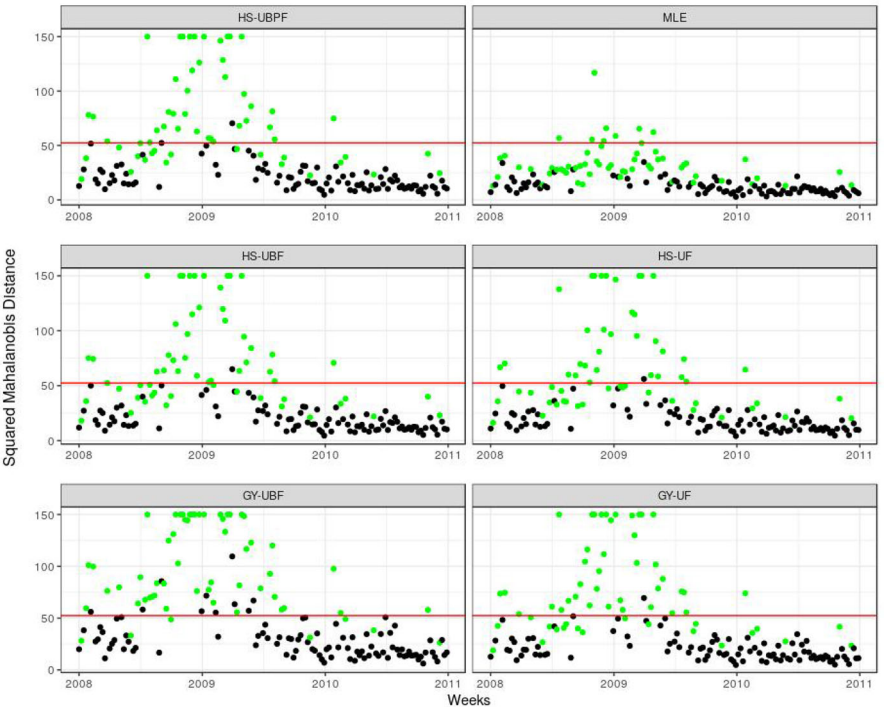

Fig. 9 Squared Mahalanobis distances of the weekly returns based on the MLE, the GY filters (GY-UF, GY-UBF) and the filters based on half-space depth (HS-UF, HS-UBF, HS-UBPF). Observations with one or more cells flagged as outliers are displayed in green. Large Mahalanobis distance are truncated for a better visualization (color figure online)

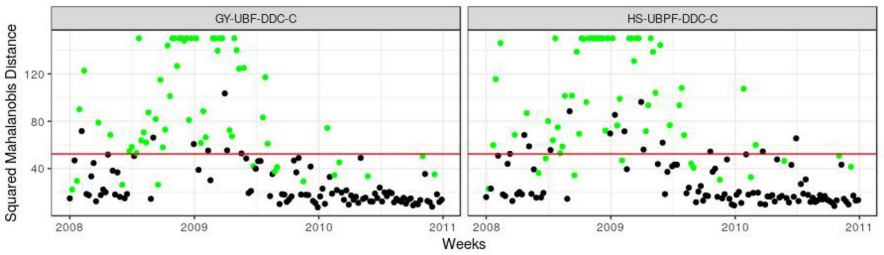

Fig. 10 Squared Mahalanobis distances of the weekly returns based on the GY-UBF-DDC-C and the corresponding filter based on half-space depth, HS-UBPF-DDC-C. Observations with one or more cells flagged as outliers are displayed in green (color figure online)

7 Conclusions

We presented a general idea to construct filters based on statistical data depth functions, called depth filters. We also showed that previously defined filters can be derived from our general method. We developed one filter, belonging to the family of depth filters, using the half-space depth, namely HS-filter. Furthermore, our filter is very versatile

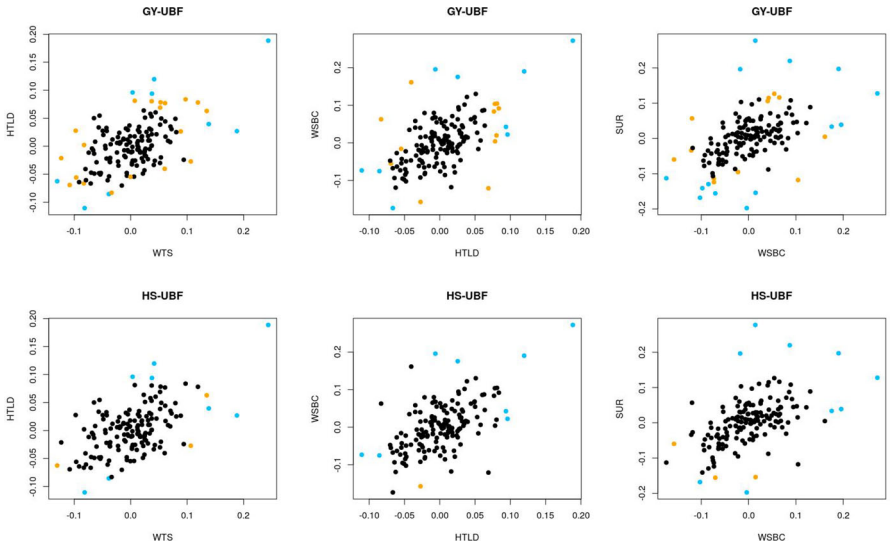

Fig. 11 Bivariate scatter plot of small-cap stock returns. In the first row, the Gervini–Yohai depth is used. Blue: outliers detected by the GY-UF univariate filter; orange: outliers detected by the bivariate step of GY-UBF but not in the univariate step. In the second row, the half-space depth is used. Blue: outliers detected by the HS-UF univariate filter; orange: outliers detected by the bivariate step of HS-UBF but not in the univariate step

since it is defined in general dimension d, $1 \leq d \leq p$. Indeed, considering the idea of an univariate and univariate-bivariate filter, we applied our HS-filter using both $d = 1$ and $d = 2$, and we proposed a new filtering procedure adding the case $d = p$, in sequence. Finally, we combined the depth-filter HS-UBPF and DDC, as suggested by Leung et al. (2017). After the filtering process, the generalized S-estimator was applied, following the two-step procedure introduced in Agostinelli et al. (2015b).

The results of the simulation study show that GY-UBF and HS-UBPF, combined with DDC, outperform the other filters in the case-wise contamination scenario. However, for small p, HS-UBPF outdoes the other filters, even if its computational time could slightly increase, in both case-wise and cell-wise contamination, and improves for increasing n. Finally, it is not suggested to combine any filter with DDC if cell-wise outliers are present, indeed, even if GY-UBF-DDC-C and HS-UBPF-DDC-C may show lower maximum average LRT and average MSE values, they do not have the best behavior with respect different contamination values k.

Further research on this filter could be needed to explore the performance of the estimator in different types of data, for example in flat data sets (e.g., $n \approx 2p$). In addition, different statistical data depth functions could be used in place of the half-space depth to construct new filters. The choice of the appropriate statistical data depth function could be helpful to analyze different types of data.

Funding Open access funding provided by Universitá degli Studi di Trento within the CRUI-CARE Agreement.

References

Agostinelli C, Leung A, Yohai V, Zamar R (2015a) Rejoinder on: robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. TEST 24(3):484–488

Agostinelli C, Leung A, Yohai V, Zamar R (2015b) 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 RH, Yohai VJ (2009) Propagation of outliers in multivariate data. Ann Stat 37(1):311–331

Azzalini A (2014) The skew-normal and related families. Institute of mathematical statistics monographs. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139248891

Cuesta-Albertos J, Nieto-Reyes A (2008) The random Tukey depth. Comput Stat Data Anal 52(11):4979–4988. https://doi.org/10.1016/j.csda.2008.04.021

Danilov M, Yohai V, Zamar R (2012) Robust estimation of multivariate location and scatter in the presence of missing data. J Am Stat Assoc 107:1178–1186

Donoho D, Gasko M (1992) Breakdown properties of location estimates based on halfspace depth and projected outlyingness. Ann Stat 20(4):1803–1827

Dyckerhoff R, Mozharovskyi P (2016) Exact computation of the halfspace depth. Comput Stat Data Anal 98:19–30. https://doi.org/10.1016/j.csda.2015.12.011

Farcomeni A (2014) Robust constrained clustering in presence of entry-wise outliers. Technometrics 56(1):102–111

Gervini D, Yohai V (2002) A class of robust and fully efficient regression estimators. Ann Stat 30(2):583–616

Lange T, Mosler K, Mozharovskyi P (2012) Fast nonparametric classification based on data depth. Stat Pap. https://doi.org/10.1007/s00362-012-0488-4

Leung A, Danilov M, Yohai V, Zamar R (2015) Gse: robust estimation in the presence of cellwise and casewise contamination and missing data p R package

Leung A, Yohai V, Zamar R (2017) Multivariate location and scatter matrix estimation under cellwise and casewise contamination. Comput Stat Data Anal 111:59–76

Liu R (1990) On a notion of data depth based on random simplices. Ann Stat 18(1):405–414

Maronna R, Martin R, Yohai VJ (2006) Robust statistic: theory and methods. Wiley, Chichister

Martin R (2013) Robust covariances: common risks versus specific risk outliers. In: Presented at the 2013 R-finance conference, Chicago, IL

R Core Team (2019) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/

Rousseeuw P, Van Den Bossche W (2018) Detecting deviating data cells. Technometrics 60(2):135–145

Zuo Y, Serfling R (2000a) General notions of statistical depth function. Ann Stat 28(2):461–482

Zuo Y, Serfling R (2000b) Structual properties and convergence results for contours of sample statistical depth functions. Ann Stat 28(2):483–499