Statistics and Computing (2019) 29:707–723
https://doi.org/10.1007/s11222-018-9831-5

Outlyingness: Which variables contribute most?

Michiel Debruyne[1] · **Sebastiaan Höppner**[2] · **Sven Serneels**[3] · **Tim Verdonck**[2]

Received: 10 October 2017 / Accepted: 11 September 2018 / Published online: 20 September 2018

Abstract
Outlier detection is an inevitable step to most statistical data analyses. However, the mere *detection* of an outlying case does not always answer all scientific questions associated with that data point. Outlier detection techniques, classical and robust alike, will typically flag the entire case as outlying, or attribute a specific case weight to the entire case. In practice, particularly in high dimensional data, the outlier will most likely not be outlying along all of its variables, but just along a subset of them. If so, the scientific question why the case has been flagged as an outlier becomes of interest. In this article, a fast and efficient method is proposed to detect variables that contribute most to an outlier's outlyingness. Thereby, it helps the analyst understand in which way an outlier lies out. The approach pursued in this work is to estimate the univariate direction of maximal outlyingness. It is shown that the problem of estimating that direction can be rewritten as the normed solution of a classical least squares regression problem. Identifying the subset of variables contributing most to outlyingness, can thus be achieved by estimating the associated least squares problem in a sparse manner. From a practical perspective, sparse partial least squares (SPLS) regression, preferably by the fast sparse NIPALS (SNIPLS) algorithm, is suggested to tackle that problem. The performed method is demonstrated to perform well both on simulated data and real life examples.

Keywords Partial least squares · Robust statistics · Sparsity · Variable selection

1 Introduction

Statistical analysis usually encompasses a step in which outliers need to be processed. It depends on the application what happens to them. Potentially, one is only interested in fitting a model for the bulk of the data, in which case outlier removal fits the purpose, given the outliers have correctly been detected. However, often one would like to know more about these outliers: are they manual errors or measurement errors, or are they just extreme values occurring naturally? Possibly even the outliers belong to separate clusters in the data, previously unassumed? As data dimensions increase, it becomes more likely that outliers of any of these natures will be predominantly outlying only with respect to a subset of the variables they consist of. Ample methodology exists to detect outliers. In this article, methodology will be developed to analyze in which way outliers lie out, given they have been detected by an appropriate statistic. Consider detection of transfer fraud as an example where the methodology proposed in this article, can have a great practical advantage. Fraud detection is all about outlier detection: typically only few transactions out of a vast number are fraudulent. Therefore, the outliers are the cases of highest interest. Once fraudulent transactions have been detected, one wants to investigate in which way these transactions are suspicious. A method that explains a fraudulent transaction's outlyingness, can speed up that analysis significantly or even automate it.

The aim of nonrobust or *classical* statistical methods, such as maximum likelihood or least squares techniques, is to optimally fit an assumed model to all observations in the data. However, real data often contain outliers, i.e. observations that deviate from the assumed model. In their presence,

This work was supported by the BNP Paribas Fortis Chair in Fraud Analytics and Internal Funds KU Leuven under Grant C16/15/068.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11222-018-9831-5) contains supplementary material, which is available to authorized users.

✉ Tim Verdonck
tim.verdonck@kuleuven.be

[1] Credit Risk Modelling, Dexia, Marsveldplein 5, 1050 Brussels, Belgium

[2] Department of Mathematics, KU Leuven, Celestijnenlaan 200B, 3001 Leuven, Belgium

[3] BASF Corporation, 540 White Plains Road, Tarrytown, NY 10591, USA

classical methods may become unreliable. Therefore, robust high-breakdown methods have been developed that are not heavily influenced by outliers. These robust alternatives can still reliably estimate the parameters of the postulated model, while a minority (i.e. less than 50%) of the data are allowed to deviate arbitrarily far from this model. As an additional benefit, one can detect the outliers as the observations that deviate substantially from the robust fit (Rousseeuw and Leroy 1987). Note that the outliers are often not detected using the classical fit, since this fit itself is also influenced by these atypical observations, an effect known as *masking*. Moreover, the effect of the outliers on a nonrobust fit can be so large that some regular observations may appear to be outlying, which is called *swamping* (Davies and Gather 1993).

Nowadays, many robust statistical methods are available that are able to detect outliers in multivariate data, both in high and low dimensions (Maronna et al. 2006). Popular robust mean and covariance estimators are, for example, the MCD estimator (Rousseeuw 1984; Rousseeuw and Van Driessen 1999), S-estimators
(Rousseeuw and Leroy 1987) and τ-estimators (Lopuhaä 1991). When the dimension exceeds the sample size, one can use the OGK estimator (Maronna and Zamar 2002) or the MRCD estimator (Boudt et al. 2017). Alternatively, a robust PCA method (e.g. Hubert et al. (2005), Croux and Ruiz-Gazen (2005)) can be applied to detect outliers. A good overview of robust dimension reduction methods is presented in Farcomeni and Greco (2015).

It is important to note that the detected outliers are not necessarily errors in the data. However, examining the structure of outliers found by robust estimators is a diagnostic effort that is often neglected. Willems et al. (2009) proposed several diagnostics which can help to obtain a better understanding of the data. The presence of outliers may reveal that the data are more heterogeneous than previously assumed and also more heterogeneous than what could be handled by the original statistical model. Outliers can be isolated or may come in clusters, indicating that there are subgroups in the population that behave differently. Sometimes outliers can even be the most interesting cases in the entire sample. Robust analysis can provide a better insight in the structure of the data and reveal structures in the data that would remain hidden in a classical analysis.

The robust estimation methods described above, as well as outlier detection techniques based on classical statistics, typically flag entire cases as outliers. In reality, outliers may only be outlying with respect to a small subset of the variables they consist of. A question that, up to our knowledge, remains unanswered in the robust statistical literature, is the following: once an outlier has been detected in a multivariate data set, how can the subset of variables that contribute most to its outlyingness be identified? Some outliers may be deviating along all of the variables, whereas other outliers may only deviate along just a few of them. Robust statistics treat such outliers in exactly the same way: they down-weight the entire observation. However, if an outlier is only deviating along some variables, it might be more useful to only adjust the atypical values in these variables. In this way, the non-contaminated and potentially valuable information in the other variables is retained. This has become more important in recent years since technical advances have led to the availability of (very) high-dimensional data sets. For instance in genetics, it is perfectly reasonable that an observation deviates from the majority of data points only for a few genes, not for all of them. Obviously, finding this subset of genes would be of high practical interest. Note that in practice, this would imply finding a subset of a few out of several hundreds of thousands of genes. Similar examples can be found in climatology, geology, neurology, process and analytical chemistry, economics and finance, among others. This extra information may be very interesting and useful for gaining insight in the data. By studying the selected variables, one can explain in which direction the outlier is deviating from the pattern of the majority of the observations.

This paper focuses on developing methods to analyze which variables contribute most to outlyingness. In what follows, it will be assumed that the outlier(s) in the corresponding data can be detected correctly by an appropriate method from the literature cited above. That method detects outliers and, as a by-product, yields a set of case weights. These weights will be used as input to the method developed here and will not change throughout the procedure.

In order to investigate an outlier's outlyingness, consider the following problem: given a multivariate data set $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ and the fact that observation $\boldsymbol{x}_i \in \mathbb{R}^p$ is an outlier with large outlyingness, find the subset of variables contributing most to the outlyingness of $\boldsymbol{x}_i$. This problem is akin to variable selection, with the objective of determining those variables contributing most to outlyingness instead of to predictive power. A simple idea to find relevant variables is to check the univariate direction in which the observation is most outlying. In Sect. 2, it is shown that the problem of estimating this direction of maximal outlyingness can be rewritten as the normed solution of a classical least squares regression problem. The proofs of the propositions found there are given in "Appendix A". Thanks to this result, identifying the subset of variables that contribute most to outlyingness becomes a variable selection problem: investigating the direction of maximal outlyingness is equivalent to investigating the vector of regression coefficients of the associated regression problem. Therefore, any widely accepted method for variable selection can be applied to this associated regression problem, ranging from visual inspection of normalized regression coefficients to application of sparse estimation procedures. The latter may be preferable in

automated analyses, yet one should keep in mind that these methods still depend on the selection of a sparsity parameter. Various methods exist that allow to estimate the vector of regression coefficients in a sparse way: the lasso (Tibshirani 1996) or the elastic net (Zou and Hastie 2005), for instance, would be suitable to accomplish this task. In this article, however, it is suggested to apply sparse partial least squares (SPLS) regression (Chun and Keleş 2010) for these purposes. It has the advantage that a single model component suffices to detect the relevant variables specifically for the outlyingness problem, which is illustrated in the simulation study reported in Sect. 5. Moreover, thanks to the univariate nature of the regression problem, SPLS can be calculated by the sparse NIPALS (SNIPLS) algorithm (as first described by Hoffmann et al. (2016)), which has the advantage of using exact PLS solutions instead of the numerical optimization applied in Chun and Keleş (2010). Both of these advantages make the SNIPLS algorithm up to our knowledge the computationally most elegant and efficient way to search for variables contributing to outlyingness in individual cases. Computational efficiency is an important positive property, since this method will realistically be applied to every single outlying case of the data.

The article is organized as follows. In Sect. 2, the direction of maximal outlyingness is defined, and its alternative formulation as a least squares regression problem is introduced. Section 3 outlines how to search for variables contributing to outlyingness by estimating the regression coefficients of the corresponding regression problem through SNIPLS. Section 4 describes several graphical tools as well as an automatic approach for selecting the optimal value for the sparsity parameter. In Sect. 5, the validity of the approach is illustrated in an extensive simulation study. In Sect. 6, the method is applied to real life data. Finally, Sect. 7 concludes.

2 Outlyingness as a regression problem

Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ be an $n \times p$ data matrix with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ and $X_j = (x_{1j}, \ldots, x_{nj})^T$ respectively the ith row and the jth column of $\boldsymbol{X}$ (note that both are column vectors). Denote by $\hat{\boldsymbol{\mu}}_r$ and $\hat{\boldsymbol{\Sigma}}_r$ robust estimates of location and scatter for $\boldsymbol{X}$. One can then compute the squared robust Mahalanobis distance for every point $\boldsymbol{x} \in \mathbb{R}^p$ as

$$m(\boldsymbol{x}; \hat{\boldsymbol{\mu}}_r, \hat{\boldsymbol{\Sigma}}_r)^2 = \left(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_r\right)^T \hat{\boldsymbol{\Sigma}}_r^{-1} \left(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_r\right). \tag{1}$$

These robust Mahalanobis distances measure the distance between point $\boldsymbol{x}$ and the robust center taking into account the covariance structure of the data.

For high-dimensional data (i.e. $p > n$), the sample covariance matrix is not existing (since some of its eigenvalues will be zero) and many robust alternatives (such as the MCD estimator) cannot be computed. Note that the OGK estimator (Maronna and Zamar 2002), the MRCD estimator (Boudt et al. 2017) and the robust precision matrix (i.e. inverse of the scatter matrix) of Öllerer and Croux (2015) can still be used. As an alternative, one can also use a robust principal component analysis (PCA) method (e.g. Hubert et al. (2005), Croux and Ruiz-Gazen (2005)), to obtain a spectral decomposition of the covariance matrix as $\boldsymbol{P}\boldsymbol{L}\boldsymbol{P}^T$ where $\boldsymbol{P}$ and $\boldsymbol{L}$ respectively contain the eigenvectors and eigenvalues of the covariance matrix.

Based on these distances, a weight w_i can then be assigned to each observation $\boldsymbol{x}_i$, indicating whether the observation is outlying or not. Under the assumption of multivariate normal data, squared Mahalanobis distances are asymptotically χ^2_p distributed (Bibby et al. 1979). Therefore, weights for each observation are often obtained as follows:

$$w_i = \begin{cases} 1 \text{ if } m(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_r, \hat{\boldsymbol{\Sigma}}_r)^2 \leq \chi^2_{p,0.975} \\ 0 \qquad\qquad \text{otherwise} \end{cases}. \tag{2}$$

Note that other weight functions can of course be used and let $\sum_{i=1}^n w_i = n_w$. These weights can then be used to compute a weighted mean and weighted covariance matrix:

$$\hat{\boldsymbol{\mu}}_w = \frac{1}{n_w} \sum_{i=1}^n w_i \boldsymbol{x}_i, \tag{3}$$

$$\hat{\boldsymbol{\Sigma}}_w = \frac{1}{n_w - 1} \sum_{i=1}^n w_i (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_w)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_w)^T. \tag{4}$$

The outlyingness of a point $\boldsymbol{x} \in \mathbb{R}^p$ is defined as the robust Mahalanobis distance using the weighted mean and weighted covariance matrix of sample $\boldsymbol{X}$:

$$\begin{aligned} r(\boldsymbol{x}; \boldsymbol{X})^2 &= m\left(\boldsymbol{x}; \hat{\boldsymbol{\mu}}_w, \hat{\boldsymbol{\Sigma}}_w\right)^2 \\ &= \left(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_w\right)^T \hat{\boldsymbol{\Sigma}}_w^{-1} \left(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_w\right). \end{aligned} \tag{5}$$

Note that this robust Mahalanobis distance is reminiscent of the robust Mahalanobis distances from a re-weighted MCD estimator as proposed in Cerioli (2010) and of the robust distance measure based on the forward search (Riani et al. 2009).

The following proposition is well known in the unweighted case. In "Appendix A.1", it is proven that it also holds in the weighted case.

Proposition 1 *The outlyingness of any point $\boldsymbol{x} \in \mathbb{R}^p$ can be expressed as the solution of a maximization problem as follows:*

$$r(\boldsymbol{x}; \boldsymbol{X}) = \max_{\boldsymbol{a} \in \mathbb{R}^p, \|\boldsymbol{a}\|=1} \frac{|\boldsymbol{x}^T \boldsymbol{a} - \hat{\boldsymbol{\mu}}_w^T \boldsymbol{a}|}{\sqrt{\boldsymbol{a}^T \hat{\boldsymbol{\Sigma}}_w \boldsymbol{a}}}. \tag{6}$$

and the direction $\boldsymbol{a}$ *that maximizes the right-hand side of the above equation, is equal to*

$$\boldsymbol{a} = \frac{\hat{\boldsymbol{\Sigma}}_w^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_w)}{\|\hat{\boldsymbol{\Sigma}}_w^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_w)\|}.$$

This can be interpreted as searching for the direction $\boldsymbol{a}$ such that the distance between the projected point $\boldsymbol{x}^T\boldsymbol{a}$ and the projected weighted mean $\hat{\boldsymbol{\mu}}_w^T\boldsymbol{a}$, standardized by a measure of spread of the projected observations, is maximal. The direction for which the maximum in proposition 1 is attained, is the *direction of maximal outlyingness* for point $\boldsymbol{x}$ and will be denoted by $\boldsymbol{a}(\boldsymbol{x})$. This direction of maximal outlyingness is potentially interesting, because its coefficients reflect how individual variables contribute to the outlyingness of a point.

The direction of maximal outlyingness can alternatively be expressed as a normalized least squares problem.

Theorem 1 *Let* $\boldsymbol{x}$ *be an arbitrary point in* $\mathbb{R}^p$ *and* $\varepsilon \in \mathbb{R}$ *with* $\varepsilon > 0$. *Denote* $\boldsymbol{y}_{w,\varepsilon}^{n+1} = \boldsymbol{e}_{n+1}$ *with* $\boldsymbol{e}_{n+1}$ *the* $(n+1)$*th basis vector in* $\mathbb{R}^{n+1}$ *containing* 1 *at component* $(n+1)$ *and* 0 *elsewhere. Let* $n_{w,\varepsilon} = n_w + \varepsilon$ *and*

$$\hat{\boldsymbol{\mu}}_{w,\varepsilon} = \frac{1}{n_{w,\varepsilon}}\left(\sum_{i=1}^{n} w_i\boldsymbol{x}_i + \varepsilon\boldsymbol{x}\right).$$

Let $\boldsymbol{X}_{w,\varepsilon} = (\sqrt{w_1}(\boldsymbol{x}_1 - \hat{\boldsymbol{\mu}}_{w,\varepsilon})^T, \ldots, \sqrt{w_n}(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_{w,\varepsilon})^T, \sqrt{\varepsilon}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{w,\varepsilon})^T)^T$, *the weighted data to which the row* $\sqrt{\varepsilon}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{w,\varepsilon})^T$ *is added, centred around the robust location estimate. Suppose that* $n > p$. *Then*

$$\boldsymbol{a}(\boldsymbol{x}) = \lim_{\varepsilon \to 0} \frac{\boldsymbol{\theta}_\varepsilon}{\|\boldsymbol{\theta}_\varepsilon\|},$$
$$\text{with } \boldsymbol{\theta}_\varepsilon = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{argmin} \|\boldsymbol{y}_{w,\varepsilon}^{n+1} - \boldsymbol{X}_{w,\varepsilon}\boldsymbol{\beta}\|^2 \tag{7}$$

where $\boldsymbol{a}(\boldsymbol{x})$ *is the direction which maximizes* $r(\boldsymbol{x}; \boldsymbol{X})$ (Eq. 6). *When* $\boldsymbol{x} \in \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, *the definition of* $\boldsymbol{a}(\boldsymbol{x})$ *as a limit for* $\varepsilon \to 0$ *also holds, in which case*

$$\boldsymbol{a}(\boldsymbol{x}_i) = \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|}, \text{ with } \boldsymbol{\theta} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{argmin} \|\boldsymbol{y}_w^i - \boldsymbol{X}_w\boldsymbol{\beta}\|^2 \tag{8}$$

where $\boldsymbol{y}_w^i$ *is the* i*th basis vector in* $\mathbb{R}^n$ *and* $\boldsymbol{X}_w = (\sqrt{w_1}(\boldsymbol{x}_1 - \hat{\boldsymbol{\mu}}_w)^T, \ldots, \sqrt{w_n}(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_w)^T)^T$.

For proof of Theorem 1, the reader is referred to "Appendix A.2". From now onwards, we will focus on the particular case when $\boldsymbol{x}$ belongs to the sample $\boldsymbol{X}$.

Many robust estimators will assign exact zero case weights to outliers far away from the data centre. Note that if our case of interest is assigned a zero weight ($w_i = 0$), then this can be circumvented by replacing the zero weight by a very small weight (e.g. 0.0001). This is equivalent with adding observation i to the data matrix $\boldsymbol{X}$ and assigning it a small weight ε. The weight needs to be small such that the outlier does not affect the estimates $\hat{\boldsymbol{\mu}}_w$ and $\hat{\boldsymbol{\Sigma}}_w$. The overall effect of choosing a small weight is limited since the vector of regression coefficients is normalized:

$$\lim_{\varepsilon \to 0} \frac{\boldsymbol{\theta}_\varepsilon}{\|\boldsymbol{\theta}_\varepsilon\|} = \frac{\hat{\boldsymbol{\Sigma}}_w^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_w)}{\|\hat{\boldsymbol{\Sigma}}_w^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_w)\|} = \boldsymbol{a}(\boldsymbol{x})$$

Owing to Theorem 1, outlyingness can be estimated by calculating the vector of regression coefficients $\boldsymbol{\beta}$ in this problem. Different regression estimators can now be plugged in to estimate $\boldsymbol{\beta}$, and as such, outlyingness. The most straightforward choice for a plug in regression estimate is least squares regression. Interpreting which variables contribute most to outlyingness, can then be done by examining the absolute magnitude of the standardized least squares regression coefficients. In practice, however, this can be a tedious process.

Moreover, least squares regression has several important drawbacks. At first, when the number of variables exceeds the sample size, the least squares fit is not well defined and cannot be calculated. Another problem frequently encountered in practice is multicollinearity. Even when some regressors are nearly collinear, it is well known that the results obtained from least squares become unstable. Moreover, least squares regression is not sparse, which implies that it typically yields a set of regression coefficients with very few non-zero elements, or none at all. As dimensions increase, this complicates interpretation and is challenging to automate. It will be discussed in the next Section how to go on about these issues.

3 Sparse direction of maximal outlyingness

In order to obtain an estimate of the direction of maximal outlyingness that can (i) easily be interpreted and (ii) from which automatically the non-zero elements can be selected, a regression plug-in estimate should be applied to Eq. (7), which has the capability to produce a sparse vector of regression coefficients. Plenty sparse regression estimators have been described in the literature. These estimators all have in common that they can yield sparse regression coefficients by including a term in their respective objective functions that puts a penalty on the norm of these regression coefficients. The idea of such a penalization goes back to ridge regression (Hoerl and Kennard 1970), where an L_2-penalty term on the Euclidean norm of the parameter vector is imposed. This effectively solves ill-posed problems in least squares regres-

sion, such as the ones discussed at the end of the previous Section.

Applying ridge regression to estimate the vector of regression coefficients $\boldsymbol{\beta}$ in (7), actually yields an entire path of regularized directions of maximal outlyingness as follows:

Definition 1 A path of regularized directions of maximal outlyingness $\boldsymbol{a}(\lambda, \boldsymbol{x}_i)$ is defined by

$$\boldsymbol{a}(\lambda, \boldsymbol{x}_i) = \frac{\boldsymbol{\theta}(\lambda)}{\|\boldsymbol{\theta}(\lambda)\|}, \text{ with}$$
$$\boldsymbol{\theta}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \|\boldsymbol{y}_w^i - \boldsymbol{X}_w\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} \boldsymbol{\beta}_j^2 \right\}.$$

Once the path $\boldsymbol{a}(\lambda, \boldsymbol{x}_i)$ is obtained, the objective is to select a subset of k variables contributing most to the outlyingness.

Ridge regression, however, does not yield a set of regression coefficients with a subset of elements exactly equal to zero. Since it cannot produce parsimonious models, alternative, sparse plug-in regression estimators have to be considered. Tibshirani (1996) has proposed the LASSO which uses L_1-norm regularization to effectively shrink many parameter estimates to zero and hence perform an intrinsic variable selection. Other penalty methods that yield sparse models can be applied as well, e.g. the SCAD penalty (Fan and Li 2001), the minimax concave penalty (Zhang 2010), the adaptive lasso (Zou 2006) or the Dantzig selector (Candès and Tao 2007). The elastic net (Zou and Hastie 2005) combines the LASSO and ridge penalties to obtain a method that can provide sparse model estimates in the presence of multicollinearity. Among these methods, the LASSO is one of the most frequently applied techniques.

Using the LASSO as a plug-in estimate into Eq. (7), actually corresponds to a path of sparse (and still regularized) directions of maximal outlyingness:

Definition 2 A path of sparse directions of maximal outlyingness $\boldsymbol{a}(\lambda, \boldsymbol{x}_i)$ is defined by

$$\boldsymbol{a}(\lambda, \boldsymbol{x}_i) = \frac{\boldsymbol{\theta}(\lambda)}{\|\boldsymbol{\theta}(\lambda)\|}, \text{ with}$$
$$\boldsymbol{\theta}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \|\boldsymbol{y}_w^i - \boldsymbol{X}_w\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\boldsymbol{\beta}_j| \right\}. \quad (9)$$

This path of sparse directions of maximal outlyingness tackles the issue with interpretability of the direction of maximal outlyingness. Yet from a computational perspective, it can still be burthensome. Recall that the procedure should be applied to every single outlying case in a data set. In order to have both the benefits of interpretability (many zero elements) and computational elegance, the approach pursued in this work is to combine dimension reduction and a penalty term. This is accomplished by applying sparse partial least squares (SPLS) regression (Chun and Keleş 2010) as the plug-in regression estimate into Eq. (7). Should one plug in SPLS with a maximal number of latent variables, then this approach becomes very similar to scanning a LASSO based path such as defined in Eq. (9). Yet the elegance SPLS offers over the other methods, is that it can actually be applied with fewer, or just one, latent variable, without losing interpretative power. Application of SPLS with few latent variables is computationally very efficient and yields good and reliable results for high-dimensional data in practice.

The reason why SPLS performs well in this context, can be interpreted reflecting on how PLS and its sparse counterpart have been conceived. Partial least squares regression (PLS) is a regression method developed in the 1960s (Wold 1966) that is particularly suited to model data where the number of variables exceeds the number of cases, as well as multicollinear data. PLS thanks these properties to its implicit dimension reduction step, wherein it typically decomposes the original data $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ onto a subset of $h << p$ latent variables $\boldsymbol{T}$. The latent variables are defined according to a criterion that maximizes covariance with the predictand, which ensures that the latent variables capture a maximal amount of information in the data relevant for prediction.

The regression problem to estimate outlyingness is particular in the sense that the dependent variable is the unit vector in the n dimensional space, which only has one nonzero entry in the cell that corresponds to the outlying case. Therefore, one can assume that few partial least squares components should be able to capture all variance in the data relevant for predicting this atypical $\boldsymbol{y}$ vector. Based on preliminary studies, we found that often a single PLS component captures sufficient amount of information for the particular task of outlyingness estimation. This assumption has in fact been corroborated in the course of establishing the results presented in the simulation study (Sect. 5).

Partial least squares has the drawback, however, that it is non-sparse, which implies that the vector of regression coefficients will only seldomly have zero entries. One could put a threshold on the absolute magnitude of the individual coefficients to determine which variables contribute most to outlyingness. However, sparse partial least squares offers a more elegant alternative, yielding a model consistent estimate for the vector of regression coefficients that is based on a PLS-alike dimensionality reduction on the one hand, but also consistently has a subset of zero entries thanks to a sparsity penalty η being imposed to the weighting vector (as long as $\eta > 0$). Here, $\eta \in [0, 1)$ plays the role of the sparsity parameter and was introduced by Chun and Keleş (2010) to facilitate the parameter selection since the range of η is known.

When using sparse PLS regression instead of least squares regression, the resulting direction $\boldsymbol{a}(\eta, \boldsymbol{x})$ is not the one which

maximizes (6) anymore since it differs from the direction $\boldsymbol{a}(\boldsymbol{x})$ given in Proposition 1. The goal of introducing a sparse direction of maximal outlyingness, however, is not to make a better estimate of the outlyingness of an observation.The outlyingness measure based on $\boldsymbol{a}(\eta, \boldsymbol{x})$ will converge to the outlyingness $r(\boldsymbol{x}, \boldsymbol{X})$ based on $\boldsymbol{a}(\boldsymbol{x})$ when $\eta \to 0$, provided that the number of latent variables is equal to p. When using a single latent variable and $\eta > 0$, the estimated direction of maximal outlyingness is sparse and the corresponding outlyingness measure can be far from $r(\boldsymbol{x}, \boldsymbol{X})$, especially for values of η close to 1.

Sparse partial least squares regression has two drawbacks: on the one hand, its intrinsic minimization may be time consuming, and secondly, it depends on two parameters to be optimized: the sparsity parameter and the number of latent variables. Owing to the univariate nature of the predictand, the former drawback can be avoided by applying the sparse NIPALS (SNIPLS) algorithm instead of the algorithm described in the original paper by Chun and Keleş (2010). For a univariate predictand, the SNIPLS algorithm is equivalent to the SPLS algorithm, but it is significantly more efficient from a computational perspective. The SNIPLS algorithm was published as an internal subroutine used in the construction of the SPRM-DA classifier (Hoffmann et al. 2016), and is also used as an internal step for computing SPRM regression (Hoffmann et al. 2015).

In what follows, it will be described how to select the optimal sparsity parameter.

4 Determining the optimal SPLS sparsity parameter

The optimal sparsity parameter η is the value for which the minimal number of variables is selected, such that the reduced case (i.e. the observation after removing those selected columns from the data set) is no longer outlying (in the lower dimensional feature space). This optimal SPLS parameter combination has to be determined from the data. For $\eta = 0$ the model is estimated including all variables and for η close to 1, almost all variables are equal to zero. Therefore, typically a grid of values for $\eta \in [0, 1)$ is searched. In Sect. 4.1, an automatic approach to determine the optimal η is described, which leads up to the SPADIMO algorithm (SPArse DIrections of Maximal Outlyingness). In Sect. 4.2, two graphical tools are presented that give more insight about the selection of the parameter η.

4.1 SPADIMO procedure

The approach followed by the authors is the following. From the perspective of the number of variables retained, SPLS converges from none (or some) to all variables in two dimensions: when keeping η constant, increasing the number of latent variables (h) will eventually yield a model with nonzero entries for all variables. Likewise, a model with constant h will eventually use all predictors available as η approaches zero. Based on the conjecture that an SPLS model with one latent variable should be able to capture all variance in the data relevant to predict a unit vector, it is plausible to fix the SPLS number of latent variables to one, and then screen η in a given range from high to low. This order of proceeding will make sure that in the first iteration the most sparse estimate is constructed, based on none to just a few of the original set of variables.

It then becomes a good question at which value of η to terminate the algorithm. For that purpose, consider the following remark. If the case consisted entirely of cells that follow the pattern of the majority of the data, it would not be flagged as an outlier. Therefore, it makes sense to proceed as follows: for each η, compute an SPLS regression estimate (with h=1) determining a subset of variable(s) contributing to outlyingness. Then, omit these entire column(s) from the data set, and estimate the case's outlyingness with respect to this reduced data set: after applying a robust estimator for location and scale on this data one calculates the weights as in formula (2) and then obtain a weighted mean and weighted covariance matrix as in Eqs. (3) and (4). The outlyingness can then be calculated using formula (5).

If it is still flagged as an outlier, proceed to the next value of η and re-estimate the SNIPLS model on the original data set, determine which variables contribute to outlyingness, and check if the observation is still an outlier. Repeat this procedure until the case is no longer outlying. The procedure, called SPADIMO, is outlined in Algorithm 1.

The algorithm is sensitive to the initial choice of the grid values $\mathcal{L}$ for η. Reasonable values for $\mathcal{L}$ will be provided from the simulation study. It is strongly advised not to scan the entire range [.01, .99], which may cause the algorithm to break off either too early or too late. Furthermore, the algorithm depends on the stopping criterion applied to $r(\boldsymbol{z}_i^{(\eta)}; \boldsymbol{Z}^{(\eta)})^2$. The approach adopted in Algorithm 1, follows Rousseeuw and Van Zomeren (1990). It provides the most generally applicable stopping criterion and can justifiably be adopted in Algorithm 1 for weights w_i deriving from various classes of robust estimators that yield continuous weights in [0, 1], e.g. MM or S estimators. However, note that in the case of reweighted MCD, in Cerioli (2010), more accurate distributional properties have been derived for distances. It has been shown there that distances for cases that have not been flagged as an outlier follow a specific β distribution, whereas distances for cases corresponding to $w_i = 0$ are F distributed. Given case weights provided as input to SPADIMO have been obtained from RMCD, these distributional results should be applied as a more accurate stopping criterion to Algorithm

1. Likewise, distributional results based on more specific assumptions than those in Rousseeuw and Van Zomeren (1990) could be plugged into the SPADIMO stopping criterion, in accordance with the outlier detection algorithm used to calculate the case weights.

On top of more specific distributional results, it is also possible to account for the size of the subset containing outliers. This can be achieved by implementing a stopping criterion based on a Bonferroni-inspired family-wise error rate correction. In practice, one would replace the stopping criterion in Algorithm 1 by Wilks' rule, i.e. stop the algorithm if

Algorithm 1: SPADIMO
Input: Data matrix $\boldsymbol{X}$ (dimension $n \times p$), vector of case weights $\boldsymbol{w}$ obtained from a given robust outlier detection procedure, index $i \in \{1, \ldots, n\}$ of the observation on which to apply SPADIMO, and grid of values $\mathcal{L} = [\ell_1, \ell_2]$ within $[0, 1]$
Output: Sparse direction of maximal outlyingness $\boldsymbol{a}(\eta, \boldsymbol{x}_i)$ for each $\eta \in \mathcal{L}$ and the corresponding subset of variable(s) contributing to outlyingness

Step 1: Standardize $\boldsymbol{X}$ to $\boldsymbol{Z}$ by subtracting a robust estimate for location (e.g. weighted mean or columnwise median) and dividing by robust scale estimate (e.g. columnwise Q_n scale estimator of Rousseeuw and Croux (1993))
Step 2: If the weight w_i of the observation to which we want to apply our method, is equal to zero, then replace that weight by a very small weight (e.g. 0.0001)
Step 3: Construct $\boldsymbol{Z}_w = (\sqrt{w_1}\boldsymbol{z}_1^T, \ldots, \sqrt{w_n}\boldsymbol{z}_n^T)^T$ and $\boldsymbol{y}_w^i$ as outlined in Sect. 2
Step 4: Set $\boldsymbol{Z}^{(\eta)} = \boldsymbol{Z}_w$ to start the algorithm
Step 5: Decreasing from ℓ_2 to ℓ_1, **for** each $\eta \in \mathcal{L}$ **do**:

1. Estimate $\boldsymbol{\theta}(\eta)$, the sparse PLS vector of regression coefficients regressing $\boldsymbol{y}_w^i$ on $\boldsymbol{Z}^{(\eta)}$ at $h = 1$
2. Calculate $\boldsymbol{a}(\eta, \boldsymbol{x}_i) = \boldsymbol{\theta}(\eta)/\|\boldsymbol{\theta}(\eta)\|$
3. Determine $v = \{j : \boldsymbol{\theta}_j(\eta) \neq 0\}$, the subset of variable(s) contributing to outlyingness
4. Update $\boldsymbol{Z}^{(\eta)} = \boldsymbol{Z}^{(\eta)} \backslash \{Z_j | j \in v\}$, with Z_j denoting the jth column of $\boldsymbol{Z}$
5. Compute $r(\boldsymbol{z}_i^{(\eta)}; \boldsymbol{Z}^{(\eta)})$, where $\boldsymbol{z}_i^{(\eta)}$ denotes the ith row of $\boldsymbol{Z}^{(\eta)}$

Stop the algorithm if $r(\boldsymbol{z}_i^{(\eta)}; \boldsymbol{Z}^{(\eta)})^2 < \chi^2_{\alpha,q}$, where α (e.g. 0.975) denotes the required χ^2 significance level and q denotes the number of remaining columns of $\boldsymbol{Z}^{(\eta)}$

$r^2_{(\tilde{n})} < b_{1-\alpha/\tilde{n}} \cdot (\tilde{n} - 1)^2/\tilde{n}$ where $b_{1-\alpha/\tilde{n}}$ is the $1 - \alpha/\tilde{n}$ quantile of the $Beta(\frac{\tilde{n}-p-1}{2}, \frac{p}{2})$ distribution and $\tilde{n}$ denotes the size of the subsample of flagged outliers. However, applying family-wise error rate corrections to significance tests has been reported to increase the amount of false negatives, which leads to more variables being flagged as contributing to outlyingness. Therefore, the more general chi-squared stopping rule has been implemented as a default into the SPADIMO algorithm.

Besides, note that the approach suggested in the SPADIMO algorithm could be applied similarly using another sparse regression estimate such as the LASSO, at the cost of increased computational complexity.

4.2 Graphical tools

Since the SNIPLS algorithm is computationally very efficient, the sparse PLS vector of regression coefficients can easily be obtained for a whole grid of values for η. The optimal η can then be selected by analyzing figures which show the number of flagged variables for each grid value, and studying how the sparsity of the direction of maximal outlyingness changes depending on the sparsity parameter. A simple example is presented to illustrate these graphical tools.

Figure 1 considers 50 points generated from a bivariate normal distribution with correlated standard normal components and correlation value 0.85. One outlier (with case number 51) is put at position (10,0). By construction, the first variable contributes most to the large outlyingness of case 51. Next 28 independent standard normal noise variables are added to this data set. By construction the first variable is still the only variable for which case 51 is outlying.

The MCD estimator is used to obtain weights for each observation and then SPADIMO is applied on case 51 with η belonging to the grid $\{0.1, 0.15, \ldots, 0.9\}$. For small values of η the direction of maximal outlyingness becomes less sparse, whereas it contains more zero components when η is close to 1. Figure 2 is akin to a screeplot which shows the number of variables that, according to SPADIMO, contribute most to outlyingness for different values of η. For $\eta \in \{0.3, \ldots, 0.9\}$,

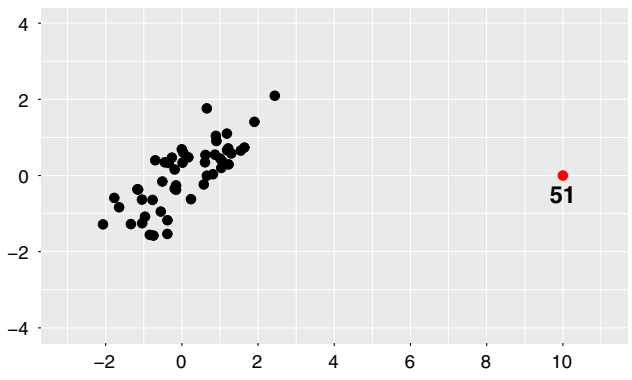

Fig. 1 Normal distributed data with one outlier

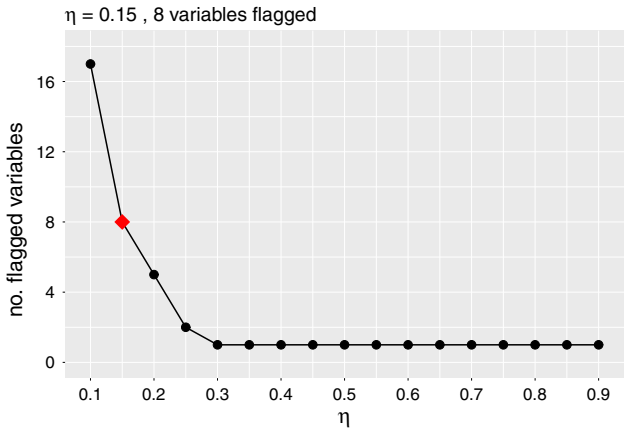

Fig. 2 Number of flagged variables versus sparsity parameter. The result of the automatic approach to determine an optimal η, as described in Sect. 4.1, is indicated by the red triangle ($\eta = 0.15$ and 8 variables are flagged)

SPADIMO identifies only one variable which contributes to outlyingness while several variables are flagged for $\eta < 0.3$. The screeplot can be used to select the number of variables that contribute most to the outlyingness of the outlier. This can be achieved by identifying an interval for η for which the number of flagged variables remains more or less constant. Based on Fig. 2, we would indeed select only one variable.

Figure 3 shows how the sparse direction of maximal outlyingness changes depending on η. The flagged variables correspond with the nonzero components of this direction. Figure 3 is a heatmap wherein positive components (i.e. SPLS regression coefficients) are coloured red and negative components are in blue. This reflects whether the variable is respectively outlying upwards or downwards. Recall from (8) that direction $\boldsymbol{a}(\boldsymbol{x}_i)$ is standardized to have an L_2-norm equal to one. The first variable is clearly identified as causing the outlyingness of observation 51 since it is the only nonzero component for $\eta \in \{0.3, \ldots, 0.9\}$. For very small values of η, e.g. 0.15, we see that there are other nonzero components, however they are very small (in absolute value) compared to the first one and thus their contribution to the outlyingness is clearly negligible. Both the screeplot and the heatmap are graphical tools to enhance the interpretation when analyzing outliers.

5 Simulation study

In this Section, the performance of SPADIMO using artificial data, is investigated. Since the results of various settings considered led to similar results and conclusions, only a part of our extensive simulation study is reported here.

Data generation setup In the simulation experiment, 1000 data sets of size n were generated from a p-variate gaussian distribution, with $n \times p$ taken as either 500×50, 200×200, 50×500 or 50×5000.

The correlation matrices were generated randomly following Agostinelli et al. (2015), henceforth ALYZ, to ensure that the performance is not tied to a particular choice of correlation matrix. The ALYZ random correlation matrices

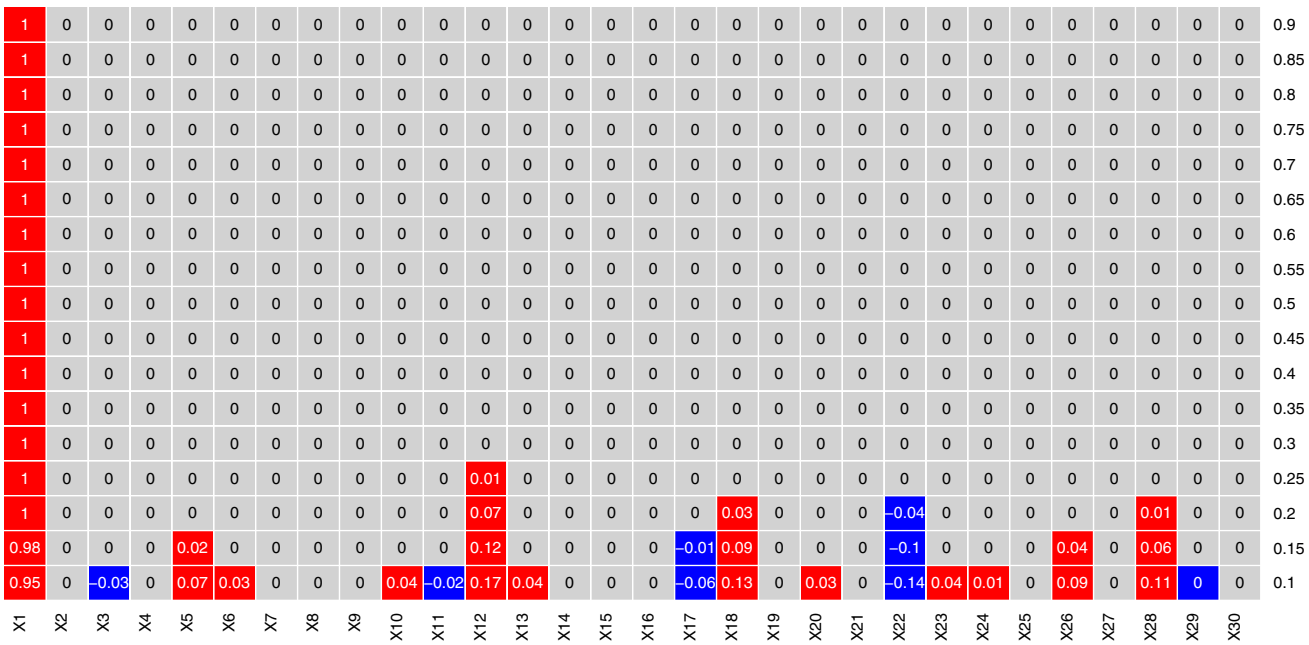

Fig. 3 Estimated sparse direction of maximal outlyingness for different values of η

Table 1 Results with A09 data

	5% of variables replaced by γ			10%			25%		
	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$
500×50, A09									
# flagged	3.585	3.381	3.402	5.558	5.367	5.377	13.455	13.289	13.292
detected (%)	99.800	100.000	100.000	99.760	100.000	100.000	99.808	100.000	100.000
swamped (%)	1.257	0.811	0.855	1.267	0.816	0.838	1.297	0.781	0.789
η	0.872	0.863	0.855	0.864	0.856	0.850	0.842	0.839	0.835
200×200, A09									
# flagged	19.177	12.468	11.309	27.959	22.154	21.145	56.019	51.601	50.894
detected (%)	99.950	100.000	100.000	99.990	100.000	100.000	100.000	100.000	100.000
swamped (%)	4.833	1.299	0.689	4.423	1.197	0.636	4.013	1.067	0.596
η	0.599	0.592	0.582	0.599	0.591	0.581	0.598	0.590	0.581
50×500, A09									
# flagged	38.195	32.118	31.866	59.717	55.920	56.284	130.329	129.172	130.129
detected (%)	94.964	97.360	98.140	93.826	97.208	98.176	95.230	98.131	99.006
swamped (%)	3.043	1.637	1.543	2.845	1.626	1.599	3.011	1.735	1.699
η	0.584	0.552	0.521	0.574	0.538	0.506	0.543	0.506	0.477
50×5000, A09									
# flagged	265.507	254.148	254.747	503.546	504.376	504.480	1275.086	1258.361	1256.832
detected (%)	80.701	92.761	95.890	86.320	95.810	97.999	94.981	98.594	99.360
swamped (%)	1.342	0.468	0.316	1.599	0.563	0.322	5.019	1.406	0.640
η	0.576	0.544	0.516	0.544	0.508	0.482	0.488	0.462	0.438

yield relatively low correlations between the variables and therefore Rousseeuw and Van den Bossche (2017) proposed the A09 correlation matrices, given by $\rho_{jh} = (-0.9)^{|h-j|}$. These matrices yield both high and low correlations.

Without loss of generality, only a single outlier is added to each data set (since SPADIMO needs to be applied to each of the outliers separately and other detected outliers do not affect this analysis). In order to test the efficiency of the SPADIMO algorithm at detecting individual variables that contribute to the outlier's outlyingness, it is of course imperative that the outliers generated are only outlying along a subset of the variables. In order to create outliers along a few of their variables, the strategy from Rousseeuw and Van den Bossche (2017) is adopted and we randomly replace $\lceil \varepsilon p \rceil$ of its variables by a value γ, which was varied to study its effect. In this study, the fraction of ε is set to either 5%, 10% or 25%.

Evaluation setup In each of the generated datasets, one observation is contaminated by replacing a certain number of its variables (given by ε) by a constant γ and then SPADIMO was applied on this single contaminated observation. Based on extended experiments like the ones in the simulation study, we found that a good starting value for η is 0.9 when $n >> p$ while it is better to start at 0.6 otherwise. Using these values, the number of variables that are flagged by SPADIMO were always very close to the real number of cells that have been contaminated.

In a simulation study one needs performance measures in order to evaluate the performance of the method. Therefore, the measures itemized below are calculated for the contaminated case and the average values over 1000 simulation runs are reported in Tables 1 and 2. Note that all results reported correspond to SNIPLS estimates using a single component, since using more components did not improve the results significantly (see Section 1 of the Supplementary Material for the simulation results obtained using two components instead of a single one).
For each contaminated case:

- # flagged: How many of its variables are flagged as outlying
- detected: How many of its contaminated variables are detected (in %; optimal value is 100)
- swamped: How many of its good variables are flagged as outlier (in %, optimal value is 0)
- η: optimal value of η that is used.

The detection and swamping rate can be understood as follows. When SPADIMO correctly identifies a contaminated variable as contributing to the outlier's outlyingness, we call that flagged variable a true positive (TP). On the other hand,

Table 2 Results with ALYZ data

	5% of variables replaced by γ			10%			25%		
	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$
500×50, ALYZ									
# flagged	3.389	3.238	3.250	5.296	5.194	5.215	13.023	13.067	13.135
detected (%)	97.400	99.567	99.967	97.240	99.400	99.920	97.685	99.385	99.908
swamped (%)	0.994	0.534	0.534	0.964	0.498	0.487	0.876	0.397	0.397
η	0.879	0.868	0.858	0.873	0.862	0.852	0.856	0.846	0.837
200×200, ALYZ									
# flagged	18.589	11.940	10.610	27.452	21.586	20.515	55.714	51.256	50.52
detected (%)	99.970	100.000	100.000	99.980	100.000	100.000	99.984	100.000	100.000
swamped (%)	4.522	1.021	0.321	4.142	0.881	0.286	3.815	0.837	0.347
η	0.600	0.598	0.589	0.600	0.598	0.588	0.600	0.595	0.584
50×500, ALYZ									
# flagged	34.858	27.442	26.260	55.970	51.432	51.051	126.920	126.042	125.941
detected (%)	93.880	97.668	98.756	92.794	97.316	98.534	94.658	98.523	99.328
swamped (%)	2.397	0.637	0.331	2.127	0.616	0.396	2.293	0.770	0.475
η	0.594	0.578	0.556	0.586	0.564	0.537	0.552	0.519	0.492
50×5000, ALYZ									
# flagged	256.918	248.166	247.305	503.867	501.135	498.420	1273.590	1256.568	1250.408
detected (%)	79.974	93.925	97.304	87.204	96.915	98.749	95.161	98.905	99.600
swamped (%)	1.200	0.281	0.085	1.508	0.368	0.104	2.242	0.540	0.144
η	0.575	0.549	0.524	0.541	0.511	0.489	0.486	0.463	0.446

when SPADIMO incorrectly flags an uncontaminated variable of the outlier, we call that flagged variable a false positive (FP). Since the number of contaminated variables is $\lceil \varepsilon p \rceil$ with ε either 5%, 10% or 25%, we get

$$\text{detection rate} = \frac{\text{number of TP's}}{\lceil \varepsilon p \rceil}$$

$$\text{swamping rate} = \frac{\text{number of FP's}}{p - \lceil \varepsilon p \rceil}$$

Discussion of results For $p \in \{50, 200, 500\}$, it can be seen that the average detection rate lies between 92.794% (ALYZ, 50×500, 10% contaminated with $\gamma = 3$) and 100% (A09, 500×50, 5% contaminated with $\gamma = 4$). The results are slightly less overwhelming when the number of variables far exceeds the number of cases ($p >> n$), but even for data of dimension 50×5000, still at least 80% of the outlying cells are detected. This implies that SPADIMO is able to detect almost all contaminated variables. This does not come at the expense of wrongly flagging clean variables, since the swamping rate remains between 0.286% and 5.019%, illustrating the very good performance of the proposed methodology and the automatic SPLS parameter selection. Note that there is of course a trade-off between these performances and the value of η. Depending on the application, it might be more important to detect at least all the variables contributing to the outlyingness (i.e. small η), while accepting a few too many; or perhaps one only wants to flag the most important variables and certainly not too much (i.e. large η). The graphical tools described in Sect. 4.2 may certainly be helpful to make a decision. This effect can also be tuned by using different starting values for η. Deriving from the extended simulation study that led up to the results reported here, we recommend to start at 0.9 when $n >> p$ and at 0.6 otherwise. Note that the optimal value of η is very close to the suggested starting value. Furthermore, the performance is similar when using ALYZ or A09 for generating the correlations, which indicates that the methodology works for both highly and moderately correlated data. For a data set of dimension 50×5000, it takes less than 3 seconds to run SPADIMO for a single case, while it takes around a second or even less for data sets with few dimensions. The required resources were measured on an Intel Core i5 with 2.7 GHz and 8 GB RAM.

Remark Note that it has always been assumed that the outliers in the data can be detected correctly by an appropriate robust methodology. While it is beyond the scope of this article to comment on the advantages and disadvantages of each individual outlier detection technique, it is interesting to know what is to be expected by SPADIMO if different choices are made at the outlier detection stage. Therefore,

the behaviour of the method has been tested (1) if applied on regular observations that have been wrongly flagged as outliers, and (2) if applied on observations that are contaminated in all the variables.

(1) If a case is wrongly detected as an outlier, SPADIMO cuts off very fast and typically stops after one iteration indicating that the oberservation contains only one or just a few exceptional features that might be examined. This behaviour is illustrated on the Top Gear dataset and tested in simulations whose results can be consulted in Section 2 of the Supplementary Material.
(2) In the setup of the simulation study in Sect. 5, the generated outliers are only outlying along a subset of the variables. If an outlier is contaminated in all variables, then SPADIMO typically suggests that almost all variables are contributing to the outlyingness. A detailed description and the results are described in Section 3 of the Supplementary Material. A general finding on the stability of SPADIMO is that "approximately equivalent" outlier detection rules give comparable SPADIMO results.

6 Examples

6.1 Top gear data

The data for the first example were taken from the website of the popular British television show Top Gear by Alfons (2012). It consists of 297 cars quantified in 32 variables, but as in Rousseeuw and Van den Bossche (2017), we will only focus on the 11 objectively measured numerical variables. Five of these variables (such as top speed) were logarithmically transformed first, since they were skewed. Moreover, 52 cars with missing values are omitted, resulting in a data set with 245 observations and 11 variables.

Since $n > p$, the MCD estimator is applied as a robust estimator for location and scale and the case weights are calculated as in formula (2). As a result, 59 out of 245 observations are flagged as outliers and these will be studied using SPADIMO so as to detect which variables contribute most to their outlyingness. The results are plotted as a heatmap in Fig. 4, where the 59 outliers are represented as rows. For each outlier detected by MCD, the variables that contribute most to the outlyingness, according to the SPADIMO algo-

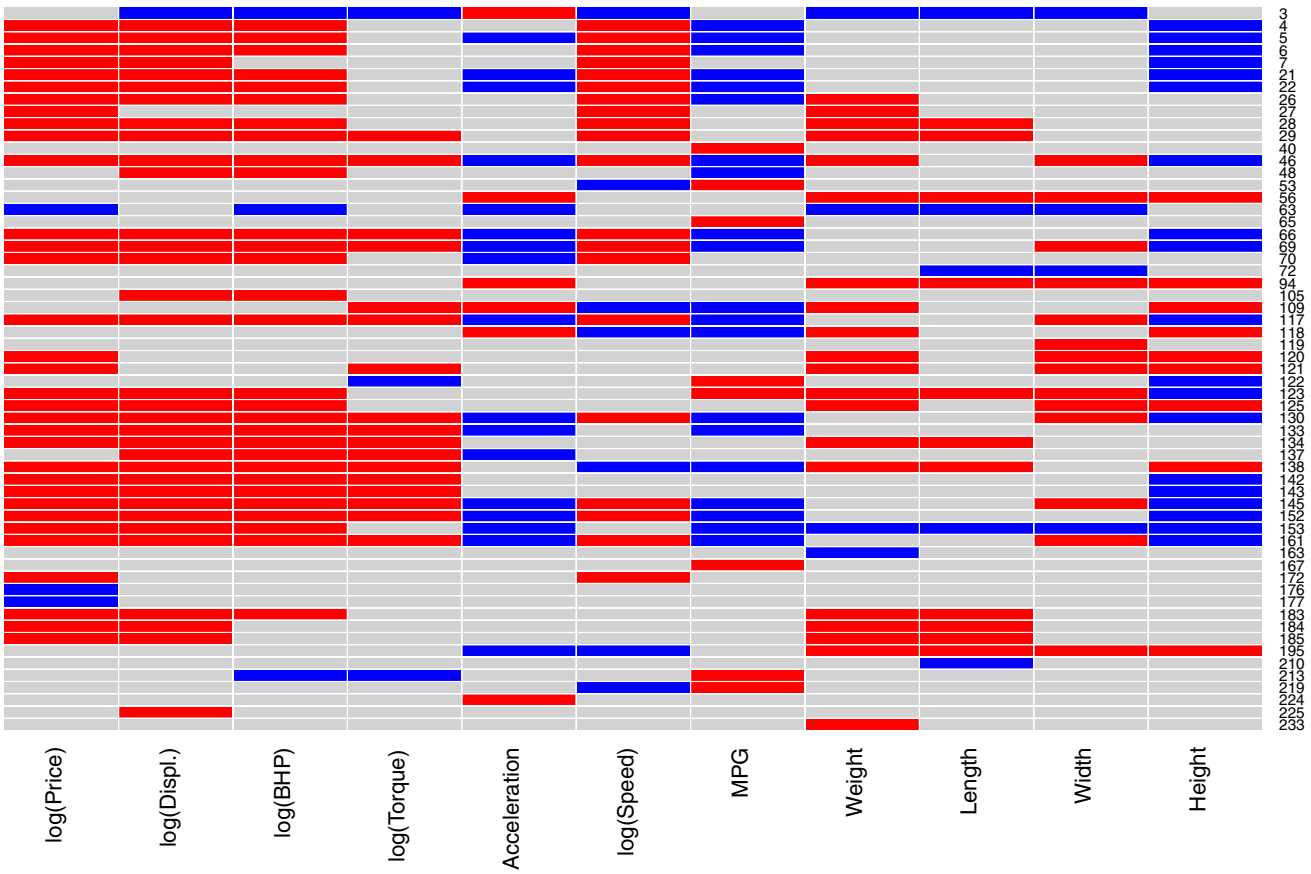

Fig. 4 Heatmap of outliers (detected by MCD) in the Top Gear data set. The red and blue boxes indicate the variables that are flagged as contributing most to outlyingness according to SPADIMO

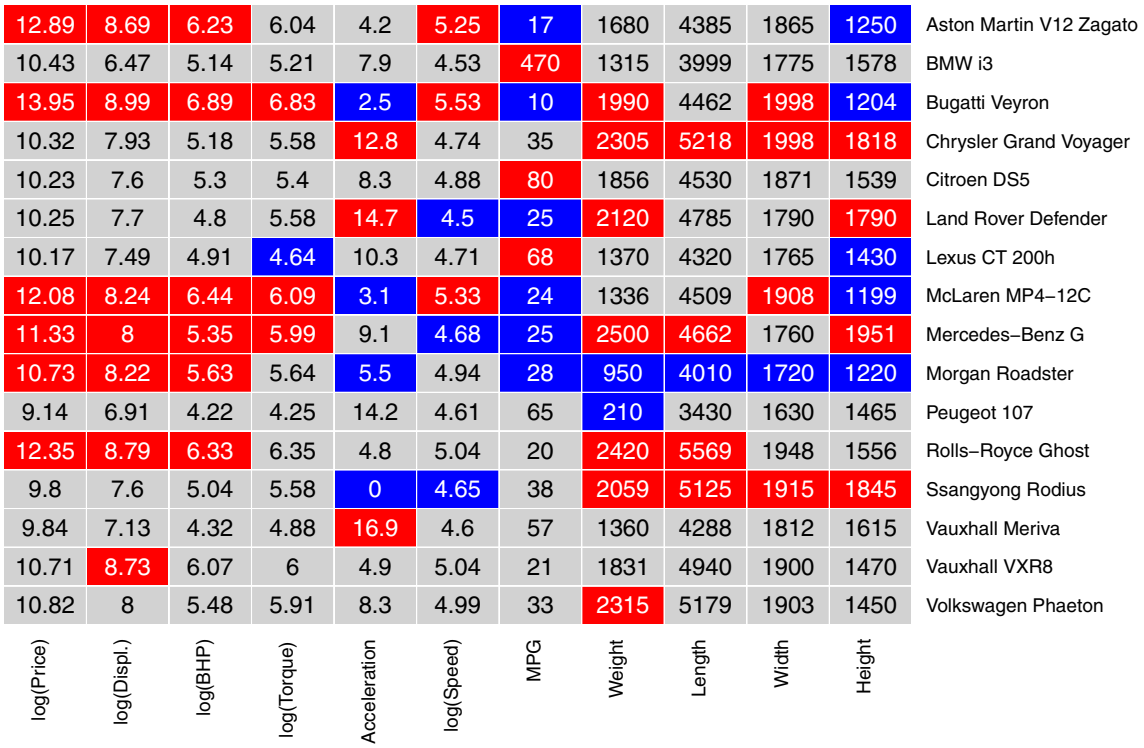

Fig. 5 Heatmap of some outliers (detected by MCD) in the Top Gear data set. The red and blue boxes indicate the variables that are flagged as contributing most to outlyingness according to SPADIMO

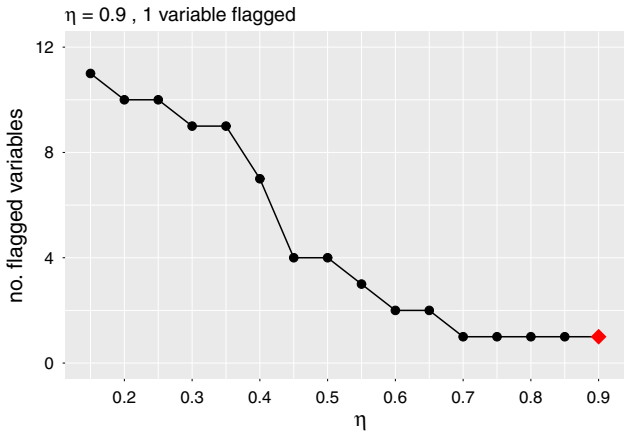

Fig. 6 Number of flagged variables versus sparsity parameter for the Peugeot 107

rithm, are shown as a colored box. The anomalous variables whose corresponding component in the sparse direction of maximal outlyingness is positive are colored red and those with a negative component are in blue. It can immediately be seen that some outliers are deviating in a lot of cells, whereas others only have an atypical value for a few cells.

Let us focus on some examples (see Fig. 5). From this analysis, it can be seen that some outliers only have atypical values in a single column, such as the BMW i3 (MPG of 470), Citroën DS5 (MPG of 80), Lexus CT 200h (log torque of 4.64), Peugeot 107 (weight of 210), Vauxhall Meriva (acceleration of 16.9), Vauxhall VXR8 (log displacement of 8.73) and Volkswagen Phaeton (weight of 2315). The weight of the Peugeot 107 is clearly an error, but not all of these atypical values are errors, since for example the BMWI i3 is an electrical vehicle with a small additional gas engine which explains its very high MPG.

Moreover, a fair amount of cars are multivariate outliers. None of their properties are unusual in the univariate sense, but in combination with other characteristics, the corresponding cars are flagged as outliers. The SPADIMO analysis can definitely distinguish between both types of outliers; moreover the color of the anomalous variables reflects whether the observed value is outlying upwards or downwards, which helps interpreting this complexly structured data set.

Let us have a closer look at the Peugeot 107. Figure 6 shows the number of flagged variables for varying values of η. For $\eta > 0.4$, the number of identified variables ranges from 1 to 4. The method itself, as described in Algorithm 1, selects a single variable at $\eta = 0.9$. Figure 7 indicates that variable *weight* causes mostly the outlyingness of Peugeot 107 as it is the first variable to be flagged. Other variables that seem to contribute to the oulyingness are *length*, *width* and *log(torque)*. The Peugeot 107 is a small city car as can be seen

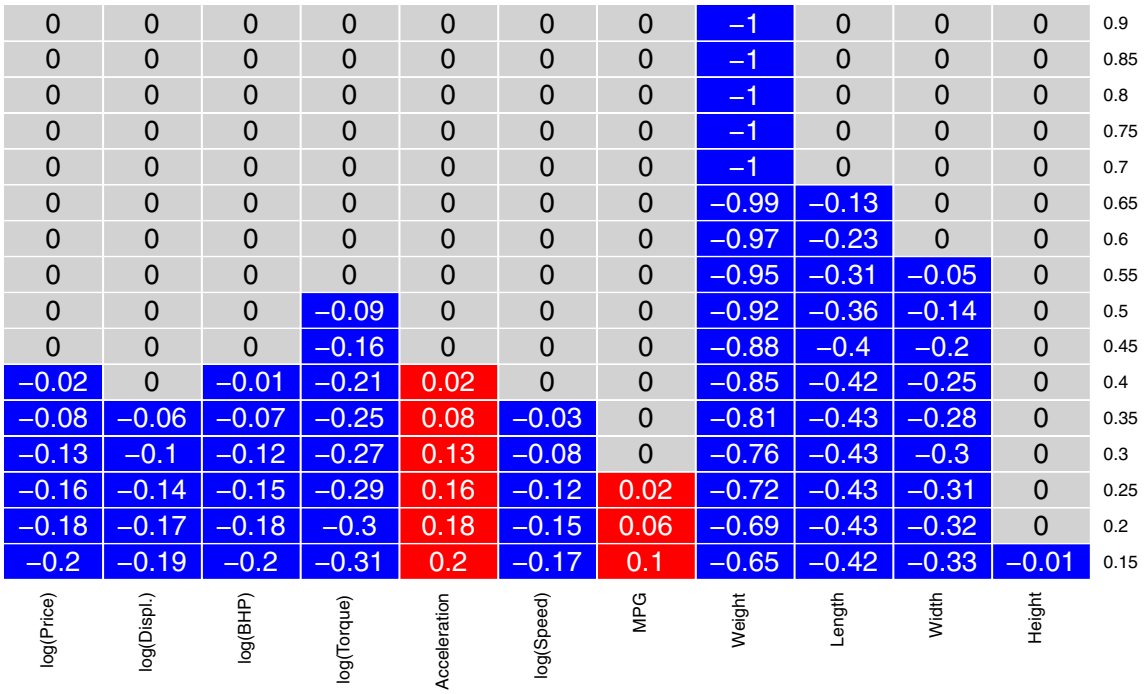

Fig. 7 The estimated sparse direction of maximal outlyingness of the Peugeot 107 for different values of η

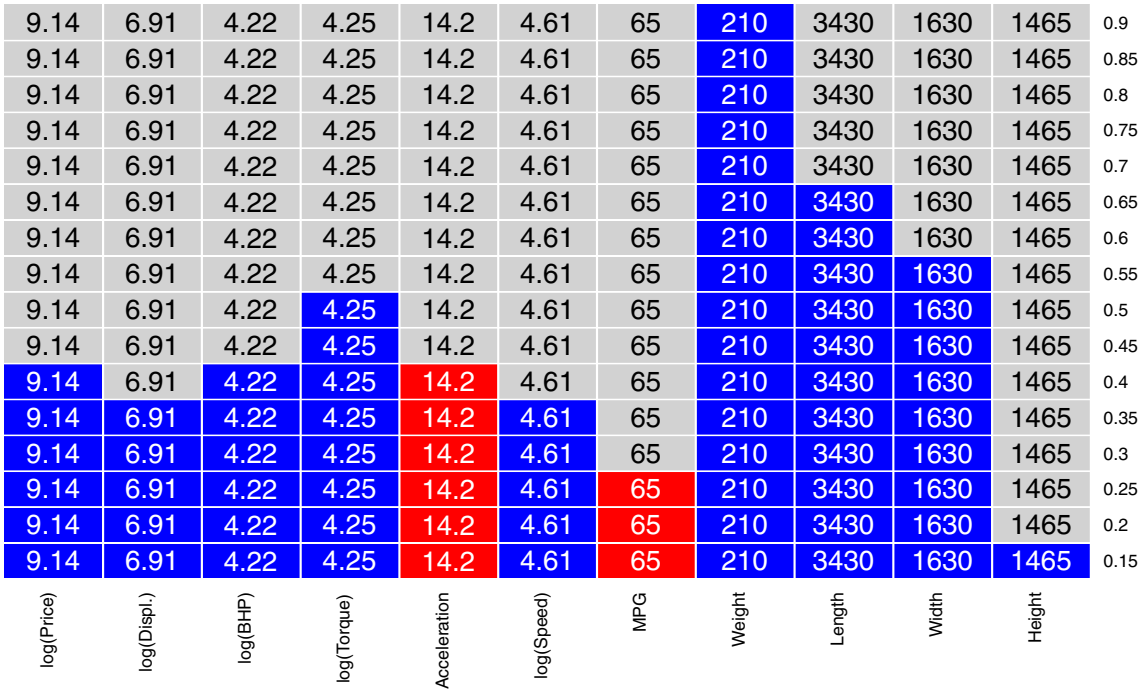

Fig. 8 The anomalous cells of the Peugeot 107 as flagged by SPADIMO for varying η

from the values listed in Fig. 8. More variables are flagged when $\eta < 0.4$, but their corresponding components remain rather small (in absolute value) so they clearly contribute less to the outlyingness of this car.

6.2 Glass data

The second example concerns a data set consisting of electron probe X-ray microanalysis (EPXMA) spectra over $p = 750$ energy channels measured on 180 archæological glass ves-

sels excavated in the Anwerp, Belgium area. The data set has been described extensively before, see e.g. Janssens et al. (1998), Lemberge et al. (2000), Hubert et al. (2005) and Serneels et al. (2005). It is known that cases 143–180 are outliers since they were measured with a different detector efficiency. On top of that, the data have been reported to consist of four clusters corresponding to four glass types: *sodic*, *potassic*, *calcic* and *potasso–calcic* glass (Janssens et al. 1998). The vast majority of the data correspond to sodic glass vessels.

Outlier detection was carried out by means of ROBPCA (Hubert et al. 2005) which detects 68 outliers. This set of outliers contains all 38 measurement outliers (i.e. cases 143-180), but it also contains the cases corresponding to the non-sodic glass vessels. In the top of Fig. 9, a heatmap is plotted showing the individual cells detected as contributing to outlyingness by SPADIMO in the cases detected as outliers by ROBPCA. The parameters used in the SPADIMO scan are $\mathcal{L} = [.1, .6]$ and $\alpha = .99$. The cells that correspond to variables that contribute most to outlyingness, are plotted in red and blue.

At first, Fig. 9 shows a clear difference between the top group of outliers, that have outlying cells more or less across the entire range of energies, whereas the bottom group of outliers contain outlying cells only in very specific areas. By analyzing the case numbers, it is clear that the former correspond to the non-sodic glass vessels, whereas the latter ones correspond to the measurement outliers.

The non-sodic glass vessels have a different chemical composition and contain different trace elements. Hence it is very plausible that deviations may be present across the entire spectrum when compared to sodic glass. Moreover, based on the SPADIMO analysis it is even possible to detect the individual types of glass composition which the outlying glass samples are made of. When looking at the cells that correspond to variables roughly in the range 300 through 375, one can see that the compositional outliers (case number < 143) can be subdivided into three groups: (1) For some samples, only roughly variables 300 through 330 are outlying. These cases correspond to the potassic samples. (2) For another group, only roughly variables 330 through 375 are outlying. These cases correspond to the calcic samples. (3) Cases where the entire range 300–375 is detected as outlying, correspond to the potasso–calcic samples. Note that a similar discrepancy can be observed in variable range 540–620. This range corresponds to iron and manganese peaks, elements of which the different glass types typically also contain different amounts.

Finally, the measurement outliers can also clearly be distinguished from the compositional outliers. It turned out that the window of the detector system had been cleaned before the last 38 spectra were measured. This resulted in a slight decrease in detector efficiency which only affects low energy X-rays. Therefore, as energy increases toward the right hand side of the plot, it is logical that all outlying cells are being detected at the left hand side. For all measurement outliers, the detector efficiency issue has generated an artifact at the beginning of the spectrum (very low energetic X-rays). This artifact is being detected by SPADIMO. Furthermore, due to individual element concentration within the glass samples, one can see from Fig. 9 that only for selected samples, outlying cells are also detected in a few of the peaks corresponding to low elements with low characteristic energies such as sodium and silicon.

7 Further applications and conclusions

In this paper, it has been proven that calculation of the direction of maximal outlyingness can be rewritten as a regression problem. It has been shown that applying variable selection methodology to this regression problem leads to detection of individual variables that contribute to a case's outlyingness. Therefore, it eventually helps understanding in which way an outlier lies out, albeit without analyzing causality. An algorithm, called SPADIMO, has been proposed to accomplish this topic in practice. Two graphical tools have been presented that are helpful to gain insight in the studied observation. SPADIMO is based on the estimation of the regression problem associated with outlyingness by sparse NIPALS regression. The latter is a multivariate regression technique for integrated variable selection and regression, that is both suitable for low and high dimensional data. By consequence, the proposed SPADIMO algorithm can be applied with equal convenience to both data configurations. An implementation of SPADIMO will be made publicly available as an R package. Note that other sparse regression techniques instead of NIPALS regression (such as LASSO or elastic net) can also be incorporated in the SPADIMO procedure.

In an extensive simulation study, the method has been proven to by and large detect exactly those outlying cells that had been set to outlying in simulated data. Even for very high dimensional data, SPADIMO does detect over 80% of cells truly contributing to simulated cellwise outliers, whereas it only yields less than 5% false positives. For swamped cases, i.e. cases that are not outlying in reality, but have wrongly been detected as outliers, SPADIMO typically breaks off after the first variable has been screened.

SPADIMO can turn out to be of great practical importance for various fields of science. One can think of the detection of transfer fraud, where fraudulent transactions are both outliers and the most interesting cases at the same time, and where it is of utmost importance to be able to analyze in which way the transaction is fraudulent. In bioinformatics, one often wants to distinguish cancer cells from regular ones. The cancer cells will have different expressions in just a sub-

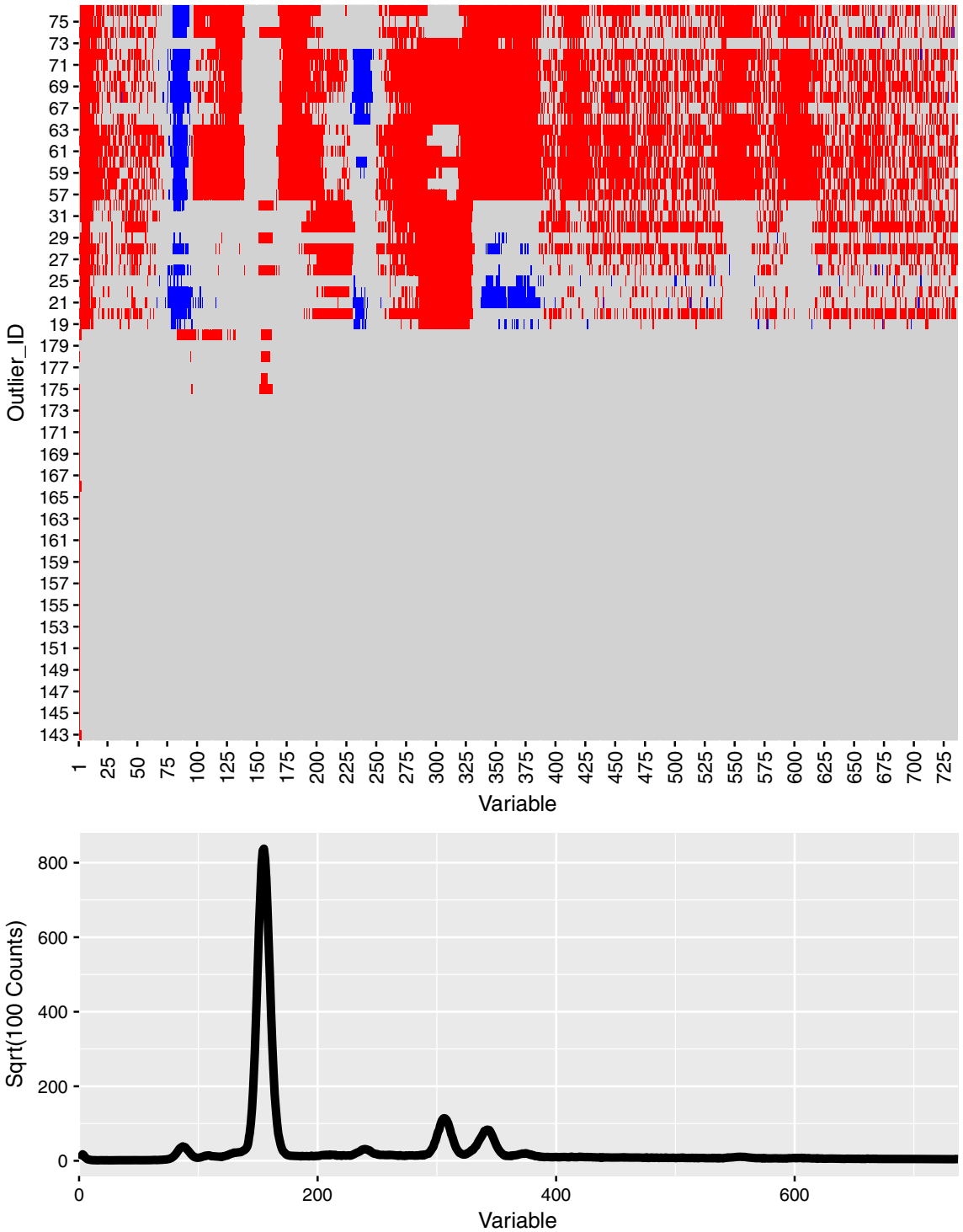

Fig. 9 Top: Heatmap of outlying cells in outliers in the glass data set. The red and blue boxes indicate the variables that are flagged as contributing most to outlyingness according to SPADIMO. Bottom: For illustrative purposes, EPXMA spectrum corresponding to outlier 145

set of the genes. Therefore, they will be outlying with respect to the bulk of the data, and it is even more interesting to know which are the outlying genes. In process chemistry, a plant may produce out-of-specification product without an *obvious* established cause according to experienced operators. In that case, the off-spec production streak will be outlying in a multivariate way, and SPADIMO can help select which combination of variables to tune. Even though the aforementioned examples are all very realistic, the examples shown in this article are yet of another nature. It has been shown

that SPADIMO allows to distinguish univariate from multivariate oultiers in the Top Gear data set, which helped to analyze the different types of cars in it more efficiently. In a second data set, SPADIMO allowed to detect outlying cells in a data set of archæological glass vessels. There, the outcome from SPADIMO can perfectly be traced back to the nature of the individual outliers: some outliers belong to a deviating glass type, whereas others can be traced back to measurement error. In these two examples of different natures, it has been illustrated that SPADIMO yields highly interpreta-ble information regarding individual outliers.

SPADIMO can be a great tool to enhance interpretation when analyzing outliers. It can even have more potential than being a standalone tool used in one-off analyses. It can, for instance, become integrated in software packages, showing variables contributing to outlyingness on a one click basis. It can also become an ancillary part of cellwise robust estimation procedures, particularly so in procedures based on cellwise re-weighting. How this approach compares to various algorithmic approaches to construct cellwise robust estimation procedures, such as in or Agostinelli et al. (2015), Öllerer et al. (2016) or Rousseeuw and Van den Bossche (2017), can be an exciting topic for future research.

A Appendix: Proofs

A.1 Proof of Proposition 1

Proof Note that our weighted covariance matrix $\hat{\boldsymbol{\Sigma}}_w$, like all covariance matrices, is a positive-semidefinite matrix. Since we also assume it is not singular and $\hat{\boldsymbol{\Sigma}}_w^{-1}$ exists, we know that $\hat{\boldsymbol{\Sigma}}_w$ is positive-definite. We now apply the Cauchy-Bunyakovskiy-Schwarz inequality to $\boldsymbol{x} = \hat{\boldsymbol{\Sigma}}_w^{-1/2}\boldsymbol{x}_1$ and $\boldsymbol{y} = \hat{\boldsymbol{\Sigma}}_w^{1/2}\boldsymbol{y}_1$, for arbitrary $\boldsymbol{x}_1, \boldsymbol{y}_1 \in \mathbb{R}^p$. This results in the following inequality

$$(\boldsymbol{x}_1^T\boldsymbol{y}_1)^2 \leq \boldsymbol{x}_1^T\hat{\boldsymbol{\Sigma}}_w^{-1}\boldsymbol{x}_1\boldsymbol{y}_1^T\hat{\boldsymbol{\Sigma}}_w\boldsymbol{y}_1$$

We have equality if $\boldsymbol{y} = c\boldsymbol{x}$ with $c \in \mathbb{R}$, which means $\hat{\boldsymbol{\Sigma}}_w^{1/2}\boldsymbol{y}_1 = c\hat{\boldsymbol{\Sigma}}_w^{-1/2}\boldsymbol{x}_1$ or $\boldsymbol{y}_1 = c\hat{\boldsymbol{\Sigma}}_w^{-1}\boldsymbol{x}_1$. So summarized, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$ we have the inequality

$$(\boldsymbol{x}^T\boldsymbol{y})^2 \leq \boldsymbol{x}^T\hat{\boldsymbol{\Sigma}}_w^{-1}\boldsymbol{x}\boldsymbol{y}^T\hat{\boldsymbol{\Sigma}}_w\boldsymbol{y},$$

where there is equality if and only if $\boldsymbol{y} = c\hat{\boldsymbol{\Sigma}}_w^{-1}\boldsymbol{x}$.

We now look at

$$\frac{(\boldsymbol{x}^T\boldsymbol{a} - \hat{\boldsymbol{\mu}}_w^T\boldsymbol{a})^2}{\boldsymbol{a}^T\hat{\boldsymbol{\Sigma}}_w\boldsymbol{a}}$$

and apply this inequality:

$$\begin{aligned}\frac{((\boldsymbol{x} - \hat{\boldsymbol{\mu}}_w)^T\boldsymbol{a})^2}{\boldsymbol{a}^T\hat{\boldsymbol{\Sigma}}_w\boldsymbol{a}} &\leq \frac{(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_w)^T\hat{\boldsymbol{\Sigma}}_w^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_w)\boldsymbol{a}^T\hat{\boldsymbol{\Sigma}}_w\boldsymbol{a}}{\boldsymbol{a}^T\hat{\boldsymbol{\Sigma}}_w\boldsymbol{a}} \\ &= (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_w)^T\hat{\boldsymbol{\Sigma}}_w^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_w).\end{aligned}$$

We have equality in the above inequality if $\boldsymbol{a} = c\hat{\boldsymbol{\Sigma}}_w^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_w)$. So

$$\boldsymbol{a} = \frac{\hat{\boldsymbol{\Sigma}}_w^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_w)}{\|\hat{\boldsymbol{\Sigma}}_w^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_w)\|}$$

is the direction $\boldsymbol{a}$ that maximizes

$$\frac{|\boldsymbol{x}^T\boldsymbol{a} - \hat{\boldsymbol{\mu}}_w^T\boldsymbol{a}|}{\sqrt{\boldsymbol{a}^T\hat{\boldsymbol{\Sigma}}_w\boldsymbol{a}}}$$

and for this $\boldsymbol{a}$ we have

$$\left(\frac{|\boldsymbol{x}^T\boldsymbol{a} - \hat{\boldsymbol{\mu}}_w^T\boldsymbol{a}|}{\sqrt{\boldsymbol{a}^T\hat{\boldsymbol{\Sigma}}_w\boldsymbol{a}}}\right)^2 = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_w)^T\hat{\boldsymbol{\Sigma}}_w^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_w) = r(\boldsymbol{x}; \boldsymbol{X})^2.$$

□

A.2 Proof of Theorem 1

Proof We know that, by the theory of ordinary least squares regression,

$$\boldsymbol{\theta}_\varepsilon = (\boldsymbol{X}_{w,\varepsilon}^T\boldsymbol{X}_{w,\varepsilon})^{-1}\boldsymbol{X}_{w,\varepsilon}^T\boldsymbol{y}_{w,\varepsilon}^{n+1}$$

and by the definition of our weighted covariance matrix, $\hat{\boldsymbol{\Sigma}}_{w,\varepsilon} = \frac{1}{n_{w,\varepsilon}-1}\boldsymbol{X}_{w,\varepsilon}^T\boldsymbol{X}_{w,\varepsilon}$, we can write

$$\boldsymbol{\theta}_\varepsilon = ((n_{w,\varepsilon} - 1)\hat{\boldsymbol{\Sigma}}_{w,\varepsilon})^{-1}\boldsymbol{X}_{w,\varepsilon}^T\boldsymbol{y}_{w,\varepsilon}^{n+1}.$$

We know that $((n_{w,\varepsilon} - 1)\hat{\boldsymbol{\Sigma}}_{w,\varepsilon})^{-1} = \frac{1}{n_{w,\varepsilon}-1}\hat{\boldsymbol{\Sigma}}_{w,\varepsilon}^{-1}$ and it is easy to see that $\boldsymbol{X}_{w,\varepsilon}^T\boldsymbol{y}_{w,\varepsilon}^{n+1} = \sqrt{\varepsilon}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{w,\varepsilon})$, if we look at the definitions of $\boldsymbol{X}_{w,\varepsilon}$ and $\boldsymbol{y}_{w,\varepsilon}^{n+1}$. Thus we have that

$$\boldsymbol{\theta}_\varepsilon = \frac{\sqrt{\varepsilon}}{n_{w,\varepsilon} - 1}\hat{\boldsymbol{\Sigma}}_{w,\varepsilon}^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{w,\varepsilon}).$$

Since ε is strictly larger than zero, we have that

$$\frac{\boldsymbol{\theta}_\varepsilon}{\|\boldsymbol{\theta}_\varepsilon\|} = \frac{\hat{\boldsymbol{\Sigma}}_{w,\varepsilon}^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{w,\varepsilon})}{\|\hat{\boldsymbol{\Sigma}}_{w,\varepsilon}^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{w,\varepsilon})\|}.$$

Then we get that

$$\lim_{\varepsilon\to 0}\frac{\boldsymbol{\theta}_\varepsilon}{\|\boldsymbol{\theta}_\varepsilon\|}=\frac{\hat{\boldsymbol{\Sigma}}_w^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_w)}{\|\hat{\boldsymbol{\Sigma}}_w^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_w)\|}=\boldsymbol{a}(\boldsymbol{x})$$

since $\lim_{\varepsilon\to 0} n_{w,\varepsilon}=n_w$, $\lim_{\varepsilon\to 0}\hat{\boldsymbol{\mu}}_{w,\varepsilon}=\hat{\boldsymbol{\mu}}_w$ and $\lim_{\varepsilon\to 0}\hat{\boldsymbol{\Sigma}}_{w,\varepsilon}^{-1}=\hat{\boldsymbol{\Sigma}}_w^{-1}$. □

References

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

Alfons, A.: robusthd: Robust methods for high-dimensional data. R package version **01** (2012)

Bibby, J., Kent, J., Mardia, K.: Multivariate Analysis. Academic Press, London (1979)

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

Candès, E., Tao, T.: The dantzig selector: statistical estimation when p is much larger than n. Ann. Stat. **35**, 2313–2351 (2007)

Cerioli, A.: Multivariate outlier detection with high-breakdown estimators. J. Am. Stat. Assoc. **105**(489), 147–156 (2010)

Chun, H., Keleş, S.: Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **72**(1), 3–25 (2010)

Croux, C., Ruiz-Gazen, A.: High breakdown estimators for principal components: the projection-pursuit approach revisited. J. Multivar. Anal. **95**, 206–226 (2005)

Davies, P., Gather, U.: The identification of multiple outliers. J. Am. Stat. Assoc. **88**, 782–792 (1993)

Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)

Farcomeni, A., Greco, L.: Robust Methods for Data Reduction. CRC Press, Boca Raton (2015)

Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)

Hoffmann, I., Serneels, S., Filzmoser, P., Croux, C.: Sparse partial robust m regression. Chemom. Intell. Lab. Syst. **149**, 50–59 (2015)

Hoffmann, I., Filzmoser, P., Serneels, S., Varmuza, K.: Sparse and robust PLS for binary classification. J. Chemom. **30**, 153–162 (2016)

Hubert, M., Rousseeuw, P.J., Vanden Branden, K.: ROBPCA: a new approach to robust principal components analysis. Technometrics **47**, 64–79 (2005)

Janssens, K.H., De Raedt, I., Schalm, O., Veeckman, J.: Composition of 15–17th century archæological glass vessels excavated in antwerp, belgium. Mikrochimica Acta **15**(Suppl.), 253–267 (1998)

Lemberge, P., De Raedt, I., Janssens, K.H., Wei, F., Van Espen, P.J.: Quantitative analysis of 16–17th century archæological glass vessels using pls regression of epxma and μ-xrf data. J. Chemom. **14**, 751–763 (2000)

Lopuhaä, H.: Multivariate τ-estimators for location and scatter. Can. J. Stat. **19**, 307–321 (1991)

Maronna, R., Zamar, R.: Robust estimates of location and dispersion for high-dimensional data sets. Technometrics **44**, 307–317 (2002)

Maronna, R., Martin, D., Yohai, V.: Robust statistics: theory and methods. Wiley, New York (2006)

Öllerer, V., Croux, C.: Robust high-dimensional precision matrix estimation. In: Modern nonparametric, robust and multivariate methods, pp. 325–350. Springer (2015)

Öllerer, V., Alfons, A., Croux, C.: The shooting s-estimator for robust regression. Comput. Stat. **31**, 829–844 (2016)

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

Rousseeuw, P.J.: Least median of squares regression. J. Am. Stat. Assoc. **79**, 871–880 (1984)

Rousseeuw, P.J., Van den Bossche, W.: Detecting deviating data cells. Technometrics (Accepted) (2017). https://doi.org/10.1080/00401706.2017.1340909

Rousseeuw, P.J., Croux, C.: Alternatives to the median absolute deviation. J. Am. Stat. Assoc. **88**(424), 1273–1283 (1993)

Rousseeuw, P.J., Leroy, A.: Robust regression and outlier detection. Wiley, New York (1987)

Rousseeuw, P.J., Van Driessen, K.: A fast algorithm for the minimum covariance determinant estimator. Technometrics **41**, 212–223 (1999)

Rousseeuw, P.J., Van Zomeren, B.: Unmasking multivariate outliers and leverage points. J. Am. Stat. Assoc. **85**, 633–651 (1990)

Serneels, S., Croux, C., Filzmoser, P., Van Espen, P.J.: Partial robust m-regression. Chemom. Intell. Lab. Syst. **79**, 55–64 (2005)

Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Methodol.) **58**(1), 267–288 (1996)

Willems, G., Joe, H., Zamar, R.: Diagnosing multivariate outliers detected by robust estimators. J. Comput. Gr. Stat. **18**(1), 73–91 (2009)

Wold, H.: Estimation of principal components and related models by iterative least squares. In: Krishnaiaah, P.R. (ed.) Multivariate Analysis, pp. 391–420. Academic Press, New York (1966)

Zhang, C.H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**(2), 894–942 (2010)

Zou, H.: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. **101**(476), 1418–1429 (2006)

Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **67**(2), 301–320 (2005)