Econometrics and Statistics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

Multivariate outlier explanations using Shapley values and Mahalanobis distances

Marcus Mayrhofer*, Peter Filzmoser

Institute of Statistics and Mathematical Methods in Economics, TU Wien, Wiedner Hauptstraße 8-10, Vienna 1040, Austria

ARTICLE INFO

Article history:
Received 20 October 2022
Revised 14 April 2023
Accepted 14 April 2023
Available online xxx

Keywords:
Shapley value
Anomaly detection
Cellwise outliers
Mahalanobis distance

ABSTRACT

For the purpose of explaining multivariate outlyingness, it is shown that the squared Mahalanobis distance of an observation can be decomposed into outlyingness contributions originating from single variables. The decomposition is obtained using the Shapley value, a well-known concept from game theory that became popular in the context of Explainable AI. In addition to outlier explanation, this concept also relates to the recent formulation of cellwise outlyingness, where Shapley values can be employed to obtain variable contributions for outlying observations with respect to their "expected" position given the multivariate data structure. In combination with squared Mahalanobis distances, Shapley values can be calculated at a low numerical cost, making them an even more attractive tool for outlier interpretation. Simulations and real-world data examples demonstrate the usefulness of these concepts.

1. Introduction

Multivariate outlier detection is a topic of unabated popularity in statistics and computer science. Not only does there exist a wide variety of approaches but also the terminology varies; anomaly detection, novelty detection, or fraud detection all refer to the problem of identifying unusual behavior (Zimek and Filzmoser, 2018). In a dataset with n observations measured at p variables, one is interested in identifying observations that do not conform to their expected behavior according to the remaining (neighboring) observations (Chandola et al., 2009; Grubbs, 1969).

A widespread tool for the detection of multivariate outliers in statistics is based on the Mahalanobis distance (Mahalanobis, 1936). Generally, for an observation vector $\boldsymbol{x} = (x_1, \ldots, x_p)'$ from a population with expectation vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ and covariance matrix $\boldsymbol{\Sigma}$, the squared Mahalanobis distance of $\boldsymbol{x}$ to $\boldsymbol{\mu}$ with respect to $\boldsymbol{\Sigma}$ is given as

$$\mathrm{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \tag{1}$$

and will be denoted as $\mathrm{MD}^2(\boldsymbol{x})$. To specify the outlyingness of an observation from a given sample, the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ need to be estimated, with their estimators being denoted as $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$. If the underlying distribution is a multivariate normal distribution, it is common to use the 0.975 quantile of a chi-square distribution with p degrees of freedom $\chi^2_{p;0.975}$

* Corresponding author.
E-mail addresses: marcus.mayrhofer@tuwien.ac.at (M. Mayrhofer), peter.filzmoser@tuwien.ac.at (P. Filzmoser).

https://doi.org/10.1016/j.ecosta.2023.04.003

as a cutoff value (Rousseeuw and Zomeren, 1990). Observations with a squared Mahalanobis distance exceeding this cutoff are identified as multivariate outliers. It is evident that for outlier detection, the estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ themselves must be robust against such outliers. Many different proposals for robust estimation of multivariate location and covariance can be found throughout the literature, with one of the most popular being the minimum covariance determinant (MCD) estimator (Rousseeuw, 1985).

The squared Mahalanobis distance from Eq. (1) can also be written as

$$\mathrm{MD}^2(\boldsymbol{x}) = \sum_{j=1}^{p}\sum_{k=1}^{p}(x_j - \mu_j)(x_k - \mu_k)\omega_{jk}, \tag{2}$$

where ω_{jk} denotes the element (j, k) of the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. This outlyingness measure collects distance contributions of all pairwise variable combinations, weighted by ω_{jk}, resulting in a single number. However, this value cannot be interpreted in the sense of contributions from individual variables, which would be vital for determining the effect of the single variables on the overall outlyingness.

Analyzing the contributions of individual variables is also of major interest in Explainable Artificial Intelligence, which is often referred to as Interpretable Machine Learning. For example, suppose a "black-box" classifier has been trained on a dataset; it is often essential to know how and why the individual variables of an observation contribute to the model's decision to assign an observation to a particular class (Ribeiro et al., 2016). Various tools have been established for this purpose, and Shapley values are among the most popular ones. Although the Shapley value (Shapley, 1953) was initially proposed in the context of game theory in 1953, it was applied much later in the context of machine learning by Štrumbelj and Kononenko (2010, 2014) and its popularity increased greatly after the publications of Lundberg and Lee (2017); Lundberg et al. (2018, 2020). We refer to Molnar (2019) and Biecek and Burzykowski (2021) for a more exhaustive discussion of these methods.

In this paper, we propose using the Shapley value for multivariate outlier explanation, which will be directly based on the squared Mahalanobis distance. Our method allows us to determine the individual variable contributions to the outlyingness and to answer the question *why* an observation is flagged as a multivariate outlier. The arguably most critical disadvantage of the Shapley values in a general setting is their high computational complexity, which exponentially increases with the number of variables. However, we will show that the Shapley values resulting from our approach can be expressed as a simplified problem, substantially facilitating their computation, even in a higher dimension. In addition, we present an extension of this concept that enables the assignment of outlyingness scores to pairs of variables, allowing the evaluation of interaction effects.

It should be mentioned that an alternative approach to answer which variables contribute the most to the multivariate outlyingness of an observation has been presented by Debruyne et al. (2019), who estimate the univariate direction of maximum outlyingness using sparse regression. Nevertheless, this method does not result in an additive decomposition of the squared Mahalanobis distance.

Another approach closely related to outlier explanation is called cellwise outlier detection. For an overview of this relatively recent research field, we refer to Raymaekers and Rousseeuw (2021). Its main idea is to investigate the outlyingness of each cell of a data matrix instead of focusing on entire observations. In general terms, cellwise outlyingness is based on the difference of the actual value of a cell compared to the value we would have expected.

Computing the amount by which a cell is anomalous is also related to multivariate outlier explanation. However, the approach is somehow reversed: To obtain explanations for the outlyingness of a single row, we decompose its squared Mahalanobis distance into outlyingness contributions for every cell using the Shapley value. In comparison, cellwise outlier detection methods must first identify a subset of clean cells for every row, which is used to derive the value a cell should have had. This is commonly done using the conditional expectation, and the resulting outlyingness scores are then based on the difference between a cell's actual value and its conditional expectation.

The remainder of this paper is structured as follows: In Section 2 we introduce Shapley values before we derive in detail how to apply them for multivariate outlier explanation using squared Mahalanobis distances. Moreover, we outline how to combine those results with the concept of cellwise outlier detection, leading to the cellwise robust outlier explanation algorithms described in Section 3. The performance of those outlier explanation tools for cellwise outlier detection is demonstrated via the numerical experiments presented in Section 4. In Section 5 we analyze the performance of our method on real-world examples. The final Section 6 summarizes the key points of our findings.

2. Shapley values for outlier explanation

In the following, we propose a method for the interpretation of multivariate outliers that combines squared Mahalanobis distances with Shapley values (Shapley, 1953). The concept of Shapley values is briefly introduced based on its nascent field of research, namely cooperative game theory (Peters, 2008).

2.1. Shapley values and cooperative game theory

In cooperative game theory, players can form coalitions that produce a payoff and decide how their coalitions' proceeds are distributed among them.

Definition 2.1.1. A coalitional (cooperative) game with transferable utility (TU-game) (T, v) is given by a set of players $T = \{1, 2, \ldots, t\}$ and the characteristic function v, which assigns the worth $v(S) \in \mathbb{R}$ to each coalition $S \subseteq T$, such that $v(\emptyset) = 0$.

In other words, the function v tells us how much collective payoff a coalition S of players can gain by cooperating. A payoff distribution for the grand coalition T is given by $\varphi(v) = (\varphi_1(v), \ldots, \varphi_t(v))'$, where $\varphi_j(v) \in \mathbb{R}$ is the payoff to player j. There are several proposals on how the payoff should be assigned to the players $j \in T$ to obtain a *fair* distribution. While there are different concepts and notions of fairness in the literature, we will focus on the one introduced by Shapley (1953). The *Shapley value* $\phi(v)$, with coordinates

$$\phi_j(v) = \sum_{S \subseteq T \setminus \{j\}} \frac{|S|!\,(t - |S| - 1)!}{t!} (v(S \cup \{j\}) - v(S)), \tag{3}$$

is the *unique* payoff distribution that fulfills the following conditions (Young, 1985):

- *Efficiency:* The payoff to individual players $\varphi_j(v)$ must add up to the worth of the grand coalition $v(T)$, hence $\sum_{j=1}^{p} \varphi_j(v) = v(T)$.
- *Symmetry:* If $v(S \cup \{j\}) = v(S \cup \{k\})$ holds for all $S \subseteq T \setminus \{j, k\}$ for two players j and k, then $\varphi_j(v) = \varphi_k(v)$.
- *Monotonicity:* If for any two games (T, v_1) and (T, v_2) and all $S \subseteq T$ the condition

  $$v_1(S \cup \{j\}) - v_1(S) \geq v_2(S \cup \{j\}) - v_2(S)$$

  is satisfied, then $\phi_j(v_1) \geq \phi_j(v_2)$.

Therefore, the Shapley value permits the definition of a fair payoff distribution for the grand coalition T. The term $v(S \cup \{j\}) - v(S)$ describes the marginal contribution of player j to a coalition S. The corresponding Shapley value $\phi_j(v)$ is then given as the weighted mean of the marginal contributions formed over all possible coalitions.

2.2. Linking Shapley value and Mahalanobis distance

Let us consider an observation vector $\boldsymbol{x} = (x_1, \ldots, x_p)'$ from a population with expectation vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ and covariance matrix $\boldsymbol{\Sigma}$. We would like to investigate the contribution of the j-th coordinate x_j to the outlyingness of $\boldsymbol{x}$. The set of players is denoted as $P = \{1, \ldots, p\}$, and it contains the indices of all variables. A coalition S is formed by a subset of P. We define the characteristic function v mentioned above as the squared Mahalanobis distance

$$\mathrm{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\hat{\boldsymbol{x}}^S) = \mathrm{MD}^2(\hat{\boldsymbol{x}}^S) \tag{4}$$

with $\hat{\boldsymbol{x}}^S = (\hat{x}_1^S, \ldots, \hat{x}_p^S)'$ and

$$\hat{x}_j^S := \begin{cases} x_j & \text{if } j \in S \\ \mu_j & \text{if } j \notin S \end{cases}, \tag{5}$$

which fulfills $\mathrm{MD}^2(\hat{\boldsymbol{x}}^S) = 0$, if $S = \emptyset$ is the empty set.

In this setting, the k-th coordinate of the Shapley value from Eq. (3) is given as the weighted average of the marginal contributions

$$\Delta_k \mathrm{MD}^2(\hat{\boldsymbol{x}}^S) := \mathrm{MD}^2(\hat{\boldsymbol{x}}^{S \cup \{k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^S)$$

over all 2^{p-1} subsets $S \subseteq P \setminus \{k\}$. This suggests an exponential computational complexity, which becomes costly, especially if p is large. However, the following theorem shows that this highly demanding problem can be reduced to linear complexity.

Theorem 2.2.1. *Given two vectors $\boldsymbol{x}, \boldsymbol{\mu} \in \mathbb{R}^p$ and a non-singular matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$, the contribution of the k-th variable to the squared Mahalanobis distance $\mathrm{MD}^2(\boldsymbol{x})$ based on the Shapley value is given by*

$$\phi_k(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) := \sum_{S \subseteq P \setminus \{k\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \Delta_k \mathrm{MD}^2(\hat{\boldsymbol{x}}^S) \tag{6}$$

$$= (x_k - \mu_k) \sum_{j=1}^{p} (x_j - \mu_j) \omega_{jk}, \tag{7}$$

with $\boldsymbol{\Sigma}^{-1} =: \boldsymbol{\Omega} = (\omega_{jk})_{j,k=1,\ldots,p}$ and $\hat{\boldsymbol{x}}^S$ as in Eq. (5).

Proof. The proof of this theorem is given in Appendix A. □

Indeed, we can compute the expression of Eq. (7) as an intermediate result when we compute the squared Mahalanobis distance, see Eq. (2).

The Shapley value of an observation $\boldsymbol{x}$ resulting from Theorem 2.2.1 is given by the vector

$$\phi(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\phi_1(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}), \ldots, \phi_p(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}))' \tag{8}$$

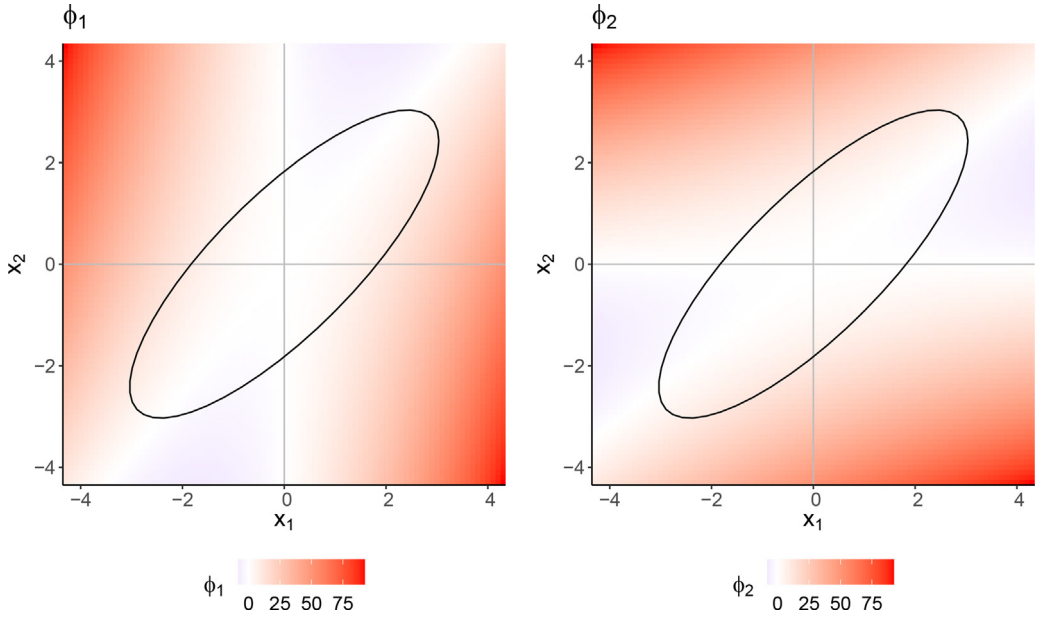

Fig. 1. Plots illustrating a two-dimensional visualization of the Shapley values $\phi(\boldsymbol{x})$ for $\boldsymbol{x} \in [-4, 4] \times [-4, 4]$ with mean $\boldsymbol{\mu} = (0, 0)'$ and covariance matrix $\boldsymbol{\Sigma}$, with elements $\sigma_{12} = \sigma_{21} = 0.8$, and $\sigma_{11} = \sigma_{22} = 1$. The graphs are colored according to the components $\phi_1(\boldsymbol{x})$ and $\phi_2(\boldsymbol{x})$ of the Shapley value, respectively, and both panels show the 99-percentile confidence ellipse.

and we will simply denote it as $\phi(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), ..., \phi_p(\boldsymbol{x}))'$, whenever the (robustly estimated) mean and covariance matrix are employed for its computation. Considering Theorem 2.2.1, it is straightforward to see that

$$\phi(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}) \circ \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), \tag{9}$$

where $\circ$ denotes the element-wise product.

Since $\phi(\boldsymbol{x})$ is based on the Shapley value, it is the only decomposition of the squared Mahalanobis distance with the characteristic function defined in Eq. (4) that fulfills the following properties:

- *Efficiency:* The contributions $\phi_j(\boldsymbol{x})$, for $j = 1, \ldots, p$, sum up to the squared Mahalanobis distance of $\boldsymbol{x}$, hence

$$\sum_{j=1}^{p} \phi_j(\boldsymbol{x}) = \mathrm{MD}^2(\boldsymbol{x}). \tag{10}$$

- *Symmetry:* If $\mathrm{MD}^2(\hat{\boldsymbol{x}}^{S\cup\{j\}}) = \mathrm{MD}^2(\hat{\boldsymbol{x}}^{S\cup\{k\}})$ holds for all subsets $S \subseteq P \setminus \{j, k\}$ for two coordinates j and k, then $\phi_j(\boldsymbol{x}) = \phi_k(\boldsymbol{x})$.
- *Monotonicity:* Let $\boldsymbol{\mu}, \tilde{\boldsymbol{\mu}} \in \mathbb{R}^p$ be two vectors and $\boldsymbol{\Sigma}, \tilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{p\times p}$ be two non-singular matrices. If

$$\mathrm{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\hat{\boldsymbol{x}}^{S\cup\{j\}}) - \mathrm{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\hat{\boldsymbol{x}}^{S}) \geq \mathrm{MD}^2_{\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}}(\hat{\boldsymbol{x}}^{S\cup\{j\}}) - \mathrm{MD}^2_{\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}}(\hat{\boldsymbol{x}}^{S})$$

holds for all subsets $S \subseteq P$, then $\phi_j(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \geq \phi_j(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$.

A single coordinate $\phi_j(\boldsymbol{x})$ of the Shapley value defined in Theorem 2.2.1 can be interpreted as the average marginal contribution of the j-th variable to the squared Mahalanobis distance of an individual observation $\boldsymbol{x}$. While the squared Mahalanobis distance aggregates all distance contributions and results in an outlyingness measure for an entire observation, Eq. (7) reveals that a coordinate $\phi_j(\boldsymbol{x})$ of the Shapley value only accounts for those distance contributions that are related to the j-th variable. This also implies that the outlyingness contribution of the j-th variable is connected to all other variables since a large distance of another variable to its mean influences the contribution of the j-th variable. However, the weighting based on the precision matrix alleviates this issue, since only variables that are not conditionally independent influence the score, at least in the case of elliptically distributed data (Baba et al., 2004). The efficiency property stated in Eq. (10) indicates that we obtain an additive decomposition of the squared Mahalanobis distance into variable contributions, where a large value of $\phi_j(\boldsymbol{x})$ indicates a large contribution of the j-th coordinate to $\mathrm{MD}^2(\boldsymbol{x})$. It should be noted that the contributions can also be negative, as illustrated in Fig. 1.

Remark: The definition given in Eq. (5), where $\hat{x}_j^S = \mu_j$ if $j \notin S$, could also be modified. In the literature, it is often suggested to use the conditional expectation of x_j, given all other variables with index contained in S, instead of the expected value (Lundberg and Lee, 2017). However, evaluating the conditional expectation explicitly requires us to impose distributional assumptions or to apply approximation techniques, and this may also lead to high computational complexity, as for every coordinate $j \in P$ there are 2^{p-1} possible subsets S. Our definition of $\hat{\boldsymbol{x}}^S$ results in two major advantages:

1. The computational complexity of computing the Shapley value reduces from an exponential to a linear one; see Theorem 2.2.1.
2. For any S and the resulting $\hat{\boldsymbol{x}}^S$, the definition of Eq. (5) results in the fact that $\mathrm{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\hat{\boldsymbol{x}}^S)$ is identical to $(\boldsymbol{x}_S - \boldsymbol{\mu}_S)'\boldsymbol{\Omega}_S(\boldsymbol{x}_S - \boldsymbol{\mu}_S)$, the squared Mahalanobis distance of $\boldsymbol{x}_S = (x_j)_{j\in S}$, where $\boldsymbol{x}_S$ and $\boldsymbol{\mu}_S$ only consist of the coordinates of $\boldsymbol{x}$ and $\boldsymbol{\mu}$ contained in the set S, respectively, and $\boldsymbol{\Omega}_S$ is the submatrix of the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ with rows and columns included in S. Therefore, analyzing the outlyingness of $\hat{\boldsymbol{x}}^S$ using the squared Mahalanobis distance is equivalent to an analysis of the outlyingness of the lower dimensional version $\boldsymbol{x}_S$, see also Eq. (2).

2.3. Shapley interaction index

The analysis of interactions between players is often of interest in cooperative game theory, and one of the first proposals for such an analysis is due to Owen (1972). In more recent developments Grabisch and Roubens (1999); Fujimoto et al. (2006) introduced a framework that allows for an axiomatic generalization of the Shapley value to the so-called Shapley interaction index, which is also used in the field of Explainable AI (Lundberg et al., 2018). Applying this method within our framework, we can investigate pairwise outlyingness contributions of variables.

Using the notation of cooperative game theory, as in Section 2.1, the Shapley interaction index for S, with fixed $|S| = s$, is given by

$$I_{Sh}(v, S) = \sum_{T\subseteq P\setminus S} \frac{t!(p-t-s)!}{(p-s+1)!}\Delta_S v(T), \tag{11}$$

with $t = |T|$, and $\Delta_S v(T) = \sum_{L\subseteq S}(-1)^{s-l}v(T\cup L), l = |L|$, also known as the *discrete* or *set function derivative* (Grabisch, 2016). We refer to the previously mentioned articles of Grabisch and Roubens (1999); Fujimoto et al. (2006) for more details regarding the theory and properties connected to this concept.

As before, we decompose the squared Mahalanobis distance using the characteristic function defined in Eq. (4). Moreover, we only focus on the *pairwise* Shapley interaction index ($|S| = 2$) because higher order Shapley interaction indices ($|S| \geq 3$) turn out to be zero in this setting (see Appendix B1 for a proof).

Theorem 2.3.1. *Given two vectors $\boldsymbol{x}, \boldsymbol{\mu} \in \mathbb{R}^p$ and a non-singular matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, the pairwise contributions of the variable pair (j, k) of an observation $\boldsymbol{x}$ to the squared Mahalanobis distance $\mathrm{MD}^2(\boldsymbol{x})$, based on the Shapley interaction index as defined in Eq. (11), are collected in the matrix $\boldsymbol{\Phi}(\boldsymbol{x}) = \boldsymbol{\Phi}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the off-diagonal elements are given by*

$$\Phi_{jk}(\boldsymbol{x}) := \sum_{T\subseteq P\setminus\{j,k\}} \frac{t!(p-t-2)!}{(p-1)!}\Delta_{\{j,k\}}\mathrm{MD}^2(\hat{\boldsymbol{x}}^T) \tag{12}$$

$$= 2(x_j - \mu_j)(x_k - \mu_k)\omega_{jk}, \tag{13}$$

with

$$\Delta_{\{j,k\}}\mathrm{MD}^2(\hat{\boldsymbol{x}}^T) = \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j,k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{k\}}) + \mathrm{MD}^2(\hat{\boldsymbol{x}}^T). \tag{14}$$

The diagonal elements are defined as

$$\Phi_{jj}(\boldsymbol{x}) := \phi_j(\boldsymbol{x}) - \sum_{k\neq j}\Phi_{jk}(\boldsymbol{x}) \tag{15}$$

$$= (x_j - \mu_j)^2\omega_{jj} - (x_j - \mu_j)\sum_{k\neq j}(x_k - \mu_k)\omega_{jk}, \tag{16}$$

where $\phi_j(\boldsymbol{x})$ is the j-th coordinate of the Shapley value as in Theorem 2.2.1.

Proof. The proof of this theorem is given in Appendix B.1. □

To gain a better understanding of what the Shapley interaction index measures, we start by rewriting Eq. (14) as

$$\begin{aligned}\Delta_{\{j,k\}}\mathrm{MD}^2(\hat{\boldsymbol{x}}^T) = &(\mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j,k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^T))\\ &- (\mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^T))\\ &- (\mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^T)).\end{aligned}$$

This reveals that $\Delta_{\{j,k\}}\mathrm{MD}^2(\hat{\boldsymbol{x}}^T)$ measures the difference in squared Mahalanobis distance between simultaneous, pairwise and individual, marginal replacement of the variables x_j and x_k with their means μ_j and μ_k, respectively. The Shapley interaction index $\Phi_{jk}(\boldsymbol{x})$ then aggregates the pairwise differences $\Delta_{\{j,k\}}\mathrm{MD}^2(\hat{\boldsymbol{x}}^T)$ across all 2^{p-2} subsets $T \subseteq P\setminus\{j,k\}$ and measures the average effect of a pairwise versus a marginal replacement. Theorem 2.3.1 shows that the Shapley interaction index can be simplified, such that $\Phi_{jk}(\boldsymbol{x})$ only depends on the deviation of the j-th and the k-th coordinate from their

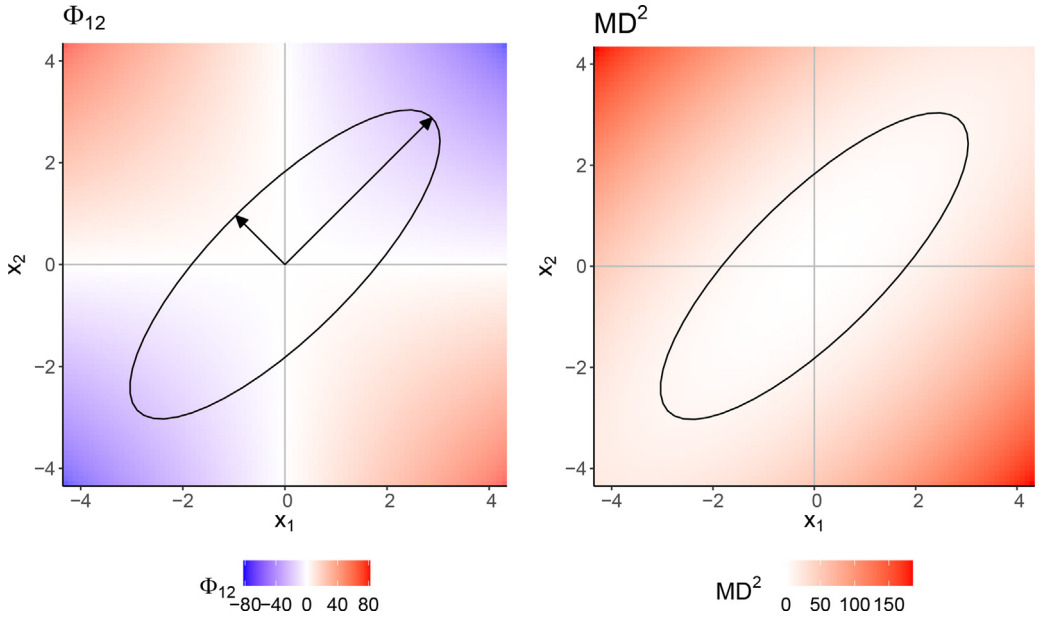

Fig. 2. Using the same setup as for the example described in Fig. 1, the pairwise contributions Φ_{12} of x_1 and x_2 to the squared Mahalanobis distance are visualized in the left panel. In this simple two-dimensional example, we can see that the pairwise contributions are highest in the direction of the eigenvector of $\boldsymbol{\Sigma}$ with the smallest eigenvalue. Hence, observations with a large multivariate outlyingness and a small univariate outlyingness are assigned high pairwise outlyingness scores Φ_{12}. In the right graph, we display the squared Mahalanobis distance, and both panels include the 99-percentile confidence ellipse.

mean, weighted by the corresponding entry of the precision matrix. This discloses that the Shapley interaction index $\Phi_{jk}(\boldsymbol{x})$ isolates the outlyingness contribution of the variable pair (j, k), while the Shapley value $\phi_j(\boldsymbol{x})$ accounts for all marginal contributions in which the j-th variable is involved. Fig. 2 provides an illustration of the Shapley interaction index between the first and the second variable in a simple two-dimensional example.

The definition of the diagonal elements $\Phi_{jj}(\boldsymbol{x})$ is chosen such that a generalization of the *Efficiency* property given in Eq. (10) is possible:

$$\phi_j(\boldsymbol{x}) = \sum_{k=1}^{p} \Phi_{jk}(\boldsymbol{x}) \quad \text{and} \quad \mathrm{MD}^2(\boldsymbol{x}) = \sum_{j=1}^{p}\sum_{k=1}^{p} \Phi_{jk}(\boldsymbol{x}).$$

Thus, the Shapley values for every variable can be decomposed into pairwise interactions with the remaining variables. Since the covariance matrix only contains information about the pairwise, linear relationship between variables, it is quite intuitive that no further decomposition is possible.

It is worth mentioning that there are other suggestions on how to generalize the Shapley value such that an explicit definition of $\Phi_{jj}(\boldsymbol{x}), j = 1, \ldots, p$, is not necessary (e.g. Sundararajan et al., 2020, *Shapley-Taylor interaction index*).

3. Cellwise robust outlier explanation

Cellwise outlier detection focuses on identifying unusual *cells* rather than rows in a data matrix. Such a procedure is particularly justified when dealing with datasets containing many variables: If only individual cells of an observation are contaminated, then the majority of non-contaminated cells still contains valuable information that should not be discarded. Moreover, already a small proportion of outlying cells spread out over the whole data matrix could, in a rowwise treatment, soon lead to a setting where the majority of observations would have to be considered as traditional rowwise outliers. However, rowwise robust methods can only deal with settings where at least half of the observations are not corrupted. To deal with such settings, Alqallaf et al. (2009) formalized the cellwise contamination model. Several papers that build on this concept are referred to in Raymaekers and Rousseeuw (2021), and they also introduce a novel procedure for cellwise outlier identification.

As already outlined in Section 1, the key objective of this work concerns the explanation of multivariate outliers based on the Shapley value for given or appropriately estimated parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. To obtain cellwise robust covariance estimates, the 2SGS approach of Agostinelli et al. (2015), the DDC method of Rousseeuw and Bossche (2018), or the cellMCD estimator of Raymaekers and Rousseeuw (2022) can be used. Since the Shapley value enables an additive decomposition of the squared Mahalanobis distance, it can be used to identify outlying cells. However, these contributions do not inform

about the supposed cell values under the assumption that they were not contaminated. Apart from detecting outlying cells, estimating the values the cells were supposed to have is of major importance when handling cellwise outliers. In this section, we outline how to combine the ideas of cellwise outlier detection and multivariate outlier explanation to obtain *cellwise robust outlier explanations.*

3.1. SCD (Shapley Cell Detector) algorithm

As a starting point, we take another look at the decomposition derived in Theorem 2.2.1, where we obtain the average marginal contributions of each component to the squared Mahalanobis distance. Eqs. (5) and (6) allow us to interpret said contributions in more detail: The value of $\phi_j(\boldsymbol{x})$ represents the average change in $\mathrm{MD}^2(\boldsymbol{x})$ across all 2^{p-1} possible variations of other variables, when the j-th component of $\boldsymbol{x}$ is replaced by its mean. Hence, positive values of $\phi_j(\boldsymbol{x})$ indicate that replacing x_j with μ_j would lead to an average reduction in $\mathrm{MD}^2(\boldsymbol{x})$, whereas negative values indicate that such a replacement would have the opposite effect.

The information provided by the Shapley value can now be used to design an algorithm for identifying outlying cells and replacing their values. We propose a stepwise procedure, which is described in detail in Algorithm 1. We call this method *Shapley Cell Detector*, abbreviated as SCD. The set R is updated in each step and will finally contain the indices of all cells of an observation $\boldsymbol{x}$ which are marked as outlying. The set of outlying coordinates R can be related to Eq. (5), where $S = \bar{R} = P \setminus R$ denotes the cells which are not replaced. In the course of each iteration, we replace the coordinates of $\boldsymbol{x}$ that have the highest scores according to the Shapley value $\boldsymbol{\phi}(\boldsymbol{x})$ until the modified observation $\tilde{\boldsymbol{x}}$ is no longer a multivariate outlier. As is common in multivariate outlier detection, a cutoff based on the chi-square distribution is used, which entails the same distributional assumptions discussed in the introduction. The replaced value does not directly correspond to the mean but rather to a value towards the direction of the mean whereby the magnitude of the correction is controlled by a step size parameter $\delta \in (0, 1]$. This is done for each set R until a score resulting from the complement $\bar{R} := P \setminus R$ of R is larger than one obtained from the set R. Here, the r-dimensional subvector $\tilde{\boldsymbol{x}}_R = (\tilde{x}_j)_{j \in R}$ of the modified observation $\tilde{\boldsymbol{x}}$ consists of the replaced values, which are dependent on $\boldsymbol{\mu}_R = (\mu_j)_{j \in R}$. Note that the maximum in line 6 of Algorithm 1 is usually unique, implying that $k = 1$ and only one index is added to R per iteration.

Example 3.1.1. We illustrate the working principle of Algorithm 1 by considering a 5-dimensional observation $\boldsymbol{x} = (0, 1, 2, 2.2, 2.5)'$ from a population with mean $\boldsymbol{\mu} = (0, 0, 0, 0, 0)'$ and covariance matrix $\boldsymbol{\Sigma}$, with elements $\sigma_{jk} = 0.9$, $j \neq k$, and $\sigma_{jj} = 1$. Here, $\boldsymbol{x}$ would be marked as a multivariate outlier since $\mathrm{MD}^2(\boldsymbol{x}) = 44.90 > 15.09 = \chi^2_{5,0.99}$ and we can employ the Shapely value of Theorem 2.2.1 to explain this multivariate outlier, resulting in $\boldsymbol{\phi}(\boldsymbol{x}) = (0, -5.07, 9.87, 15.26, 24.84)'$. Those outlyingness scores are then used in Algorithm 1 to flag outlying cells, and, for simplicity, we analyze the case where $\delta = 1$. In this scenario, the coordinate x_5 is identified first, followed by x_4, and then x_3. Each variable in turn is replaced by μ_5, μ_4, and μ_3, respectively. This results in an altered version $\tilde{\boldsymbol{x}}$ of the original observation $\boldsymbol{x}$, which is no longer outlying, and therefore the algorithm stops.

It should be noted that in this example, we have no information about which cells are truly outlying or have been manipulated. However, in general, it seems desirable to keep the number of modified coordinates as small as possible.

Algorithm 1 is easy to implement and fast to compute. The discrepancy between the original and replaced cells indicates the outlyingness in the particular variables. However, this simplicity results from our definition of the Shapley value in Theorem 2.2.1, which leads to a replacement by a value towards the mean in Algorithm 1.

Algorithm 1 Shapley Cell Detector (SCD)

1: **procedure** SCD($\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta$)
2: $\tilde{\boldsymbol{x}} \leftarrow \boldsymbol{x}$
3: $R \leftarrow \emptyset$
4: $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)' \leftarrow \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\phi_1(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}), \ldots, \phi_p(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}))'$
5: **while** $\mathrm{MD}^2(\tilde{\boldsymbol{x}}) > \chi^2_{p,0.99}$ **do**
6: $R \leftarrow R \cup \{j_1, \ldots, j_k\}$, where $(\phi_{j_l})_{l=1,\ldots,k} = \max_{i=1,\ldots,p} \phi_i$
7: **while** $\max_{j \in R} \phi_j > \max_{j \in \bar{R}} \phi_j$ **do**
8: $\tilde{\boldsymbol{x}}_R \leftarrow \tilde{\boldsymbol{x}}_R - (\tilde{\boldsymbol{x}}_R - \boldsymbol{\mu}_R)\delta$
9: $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi}(\tilde{\boldsymbol{x}}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$
10: **end while**
11: **end while**
12: **return** $\tilde{\boldsymbol{x}}$
13: **end procedure**

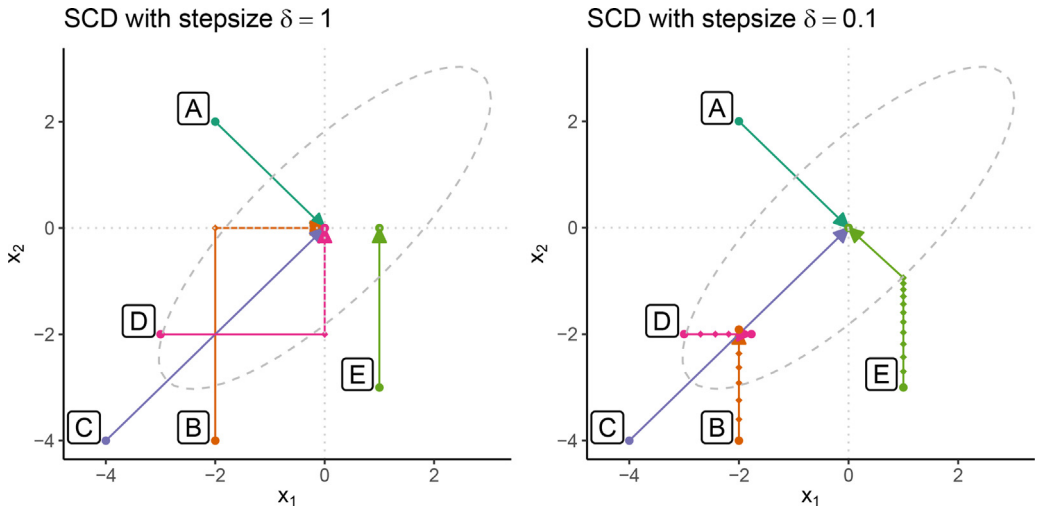

Fig. 3. In this figure, two graphs are displayed to illustrate the operating principle of Algorithm 1 in a two-dimensional setting. Both plots show the position of the five outlying points A to E and their replacements. The plot on the left side shows the results when cells are directly replaced by their corresponding mean, while the right side illustrates the stepwise approach.

Fig. 3 provides a further illustration of the SCD procedure for a two-dimensional example. It schematically displays five specific observations, denoted by A to E, to which Algorithm 1 is applied. The left plot shows the result when setting $\delta = 1$ in the algorithm, while the right plot corresponds to $\delta = 0.1$. The points in the plots highlight the individual computation steps of the algorithm, and the ellipse indicates the stopping criterion $\chi^2_{2,0.99}$. While for $\delta = 1$ the algorithm uses at most two steps, this behavior changes for the case of $\delta = 0.1$. Using a smaller step size leads to different replacement values for the points B, D, and E. Comparing the computation steps for points B and D, the results in the right plot seem more meaningful since they avoid increasing the Mahalanobis distance during the computation, and the final replacement is more similar to the original points.

Until now, we only considered a replacement of outlying cells by the mean or by a value towards the direction of the mean. However, the algorithm is only stopped by a sufficient reduction of the squared Mahalanobis distance. Therefore, the task at hand can thus be redefined further: Find the optimal replacements for outlying cells to achieve the highest possible reduction in squared Mahalanobis distance. As before, the Shapley value should determine the outlyingness of the cells.

3.2. MOE (Multivariate Outlier Explainer) algorithm

Based on the definition of the Shapley value in Eq. (7), a coordinate has a low outlyingness contribution if it is close to its mean. Consequently, it is unlikely that this cell is flagged as outlying. This center-outward ordering is induced by the squared Mahalanobis distance computed with respect to the mean, and thus it explains the *global outlyingness* of an observation. However, the described procedure might not be optimal for detecting cellwise outliers, where *local outlyingness* is emphasized, because the information contained in the regular cells of an observation could be incorporated to define an optimal replacement. For this purpose, an alternative approach to using the mean as the center for computing Mahalanobis distances and Shapley values is outlined in the following paragraphs. We call the newly proposed center parameter *reference point*. This new Shapley value will also be used later for an outlier replacement strategy.

The question of how to best replace cells of an observation to minimize the squared Mahalanobis distance has been addressed in Raymaekers and Rousseeuw (2021). Here we assume that the set R of outlying cells is fixed (and $R \neq \emptyset$) for an observation $\boldsymbol{x} = (x_1, \ldots, x_p)'$, and the cells x_j should be shifted to the values $\tilde{x}_j$, for $j \in R$. Explicitly we can write this as $\boldsymbol{x} - \boldsymbol{E}_R\boldsymbol{\beta}$ where $\boldsymbol{E}_R$ denotes the $p \times r$ matrix with the standard basis vectors $\boldsymbol{e}_j$, $j \in R$ as columns. The squared Mahalanobis distance of this expression can now be rewritten as follows,

$$\begin{aligned}
\mathrm{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x} - \boldsymbol{E}_R\boldsymbol{\beta}) &= (\boldsymbol{x} - \boldsymbol{\mu} - \boldsymbol{E}_R\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu} - \boldsymbol{E}_R\boldsymbol{\beta}) \\
&= \left\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu} - \boldsymbol{E}_R\boldsymbol{\beta})\right\|_2^2 \\
&= \left\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}) - \boldsymbol{\Sigma}^{-1/2}\boldsymbol{E}_R\boldsymbol{\beta}\right\|_2^2.
\end{aligned}$$

Minimizing this expression corresponds to a least-squares problem, which leads to the least-squares estimator

$$\hat{\boldsymbol{\beta}}(R) = \underset{\boldsymbol{\beta}\in\mathbb{R}^s}{\operatorname{argmin}}\, \mathrm{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x} - \boldsymbol{E}_R\boldsymbol{\beta}) = (\boldsymbol{E}_R'\boldsymbol{\Sigma}^{-1}\boldsymbol{E}_R)^{-1}\boldsymbol{E}_R'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), \tag{17}$$

and the replaced values are given by $\tilde{\boldsymbol{x}}_R = \boldsymbol{x}_R - \hat{\boldsymbol{\beta}}(R)$, which are equal to the conditional means under multivariate normality, *i.e.* $\tilde{\boldsymbol{x}}_R = \mathbb{E}[\boldsymbol{x}_R|\boldsymbol{x}_{\bar{R}}]$ (Raymaekers and Rousseeuw, 2021).

If R consists of only one element, say $R = \{j\}$, for $j \in P$, then the solution of Eq. (17) simplifies to

$$\hat{\boldsymbol{\beta}}(j) = \frac{1}{\omega_{jj}}(\omega_{j1}, \ldots, \omega_{jp})(\boldsymbol{x} - \boldsymbol{\mu}), \tag{18}$$

where ω_{ij} denotes the element (i, j) of $\boldsymbol{\Sigma}^{-1}$, and the modification for observation $\boldsymbol{x}$ is given by $\tilde{x}_j = x_j - \hat{\boldsymbol{\beta}}(j)$.

Building on those findings, we can now define the new reference point $\tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R)$ for a fixed set of outlying cells R, by setting each coordinate to

$$\tilde{\mu}_j(\boldsymbol{x}, R) = x_j - \hat{\boldsymbol{\beta}}_{(j)}(R \cup \{j\}), \tag{19}$$

where $\hat{\boldsymbol{\beta}}_{(j)}(R \cup \{j\})$ is the component of $\hat{\boldsymbol{\beta}}(R \cup \{j\})$ corresponding to the index j. To determine the set R, we adapt the SCD procedure by incorporating $\tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R)$ as a reference point for the Mahalanobis distance and updating it in each iteration. We refer to this procedure as Multivariate Outlier Explainer (MOE) and outline its general workflow in Algorithm 2.

The MOE procedure is initialized by computing the reference point $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R)$, with $R = \emptyset$. For the initial computation of $\hat{\boldsymbol{\beta}}$ we can simply apply Eq. (18) to each coordinate of $\boldsymbol{x}$, which can be done in one step by matrix multiplication. Using this initial reference point, we obtain the squared Mahalanobis distance $\mathrm{MD}^2_{\tilde{\boldsymbol{\mu}},\boldsymbol{\Sigma}}(\tilde{\boldsymbol{x}})$, which is in turn used to define the corresponding Shapley value $\boldsymbol{\phi}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ according to Eq. (9). We want to emphasize that the properties of the Shapley value listed in Section 2 remain unchanged, particularly the *Efficiency* property: The sum of the coordinates of the Shapley value $\boldsymbol{\phi}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ equals the squared Mahalanobis distance with respect to the new reference point $\tilde{\boldsymbol{\mu}}$. Outlying cells are then identified based on the Shapley value and corrected in the direction of their corresponding entries of $\tilde{\boldsymbol{\mu}}$, resulting in the modified observation $\tilde{\boldsymbol{x}}$. The process of updating the reference point $\tilde{\boldsymbol{\mu}}$, identifying outlying cells based on their Shapley values, and correcting them in the direction of $\tilde{\boldsymbol{\mu}}$, is then repeated until the vector $\tilde{\boldsymbol{x}}$ is no longer marked as outlying. Aside from using the reference point $\tilde{\boldsymbol{\mu}}$ in the MOE procedure instead of $\boldsymbol{\mu}$, the concept of the algorithm is similar to the SCD procedure, but there are two other important distinctions:

- The outlier cutoff value used in line is adapted to the new reference point. Filzmoser et al. (2014) have shown that for a sample $\boldsymbol{x}$ drawn from a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the conditional distribution of the squared Mahalanobis distance $\mathrm{MD}^2_{\tilde{\boldsymbol{\mu}},\boldsymbol{\Sigma}}(\boldsymbol{x})$ given $\tilde{\boldsymbol{\mu}}$ is a non-central chi-square distribution with p degrees of freedom and non-centrality parameter $\lambda = \mathrm{MD}^2(\tilde{\boldsymbol{\mu}})$, denoted as $\chi^2_p(\lambda)$. Therefore, the 0.99 quantile of this distribution is taken as the cutoff value to exit the loop.
- Since the goal of this procedure is cellwise outlier detection, we want to avoid flagging coordinates that were only shifted by a negligible amount. Therefore, we monitor the distance $\boldsymbol{d}$ by which each cell of $\boldsymbol{x}$ is shifted in the direction of $\tilde{\boldsymbol{\mu}}$. Initially, this distance is set to $d_j = 0$, $j = 1, \ldots, p$, followed by an iterative update of the distance variable in line . Moreover, we adjust $\boldsymbol{d}$ such that the distances are independent of the scale of the single coordinates. We then update the set of outlying coordinates R by only choosing coordinates for which $d_j > \eta \max_{l=1,\ldots,p} d_l$, with $\eta \in [0, 1]$. In simulations not included in this work, $\eta = 0.2$ resulted in a good trade-off between the recall, meaning the fraction of correctly identified cells among all contaminated cells, and the precision, meaning the fraction of correctly identified cells among all detected cells, of the procedure and is therefore chosen as a default value. Finally, we amend $\tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R)$, $\boldsymbol{\phi}(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$, and $\tilde{\boldsymbol{x}}$ according to the updated set R.

Algorithm 2 allows us to detect and impute cellwise outliers, and it also yields a local explanation of the outlyingness. Furthermore, the Shapley values computed with respect to the reference point $\tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R)$ can be used to explain the results of other cellwise outlier detection procedures. To this end, we merely need to compute $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R)$ for a given set of outlying cells R of an observation $\boldsymbol{x}$. By subsequently determining the Shapley value $\boldsymbol{\phi}(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$, we can therefore explain *why* the observation is outlying.

Example 3.2.1. We reiterate Example 3.1.1 with the MOE procedure, using a step size of δ of 0.1. The first two coordinates x_1 and x_2 are marked as outlying, resulting in $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}(\boldsymbol{x}, \{1, 2\}) = (2.19, 2.19, 2.27, 2.13, 2.04)$ and $\boldsymbol{\phi}(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = (34.89, 7.07, -0.86, 1.28, 4.88)$.

Comparing the results of Algorithms 1 and 2, it can be seen that the sets of outlying cells for the two algorithms are disjoint. Therefore, the interpretations of the results are different. The reason for this discrepancy is mainly that we no longer decompose $\mathrm{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x})$, but instead the squared Mahalanobis distance of the amended reference point $\tilde{\boldsymbol{\mu}}$, $\mathrm{MD}^2_{\tilde{\boldsymbol{\mu}},\boldsymbol{\Sigma}}(\boldsymbol{x})$. While the Shapley value $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ used in Algorithm 1 explains the global outlyingness, the Shapley value $\boldsymbol{\phi}(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ used in Algorithm 2 provides us with a local understanding of the outlyingness, which is better suited to the setting of cellwise outlyingness.

In Fig. 4, we compare the final Shapley values yielded by the SCD and MOE algorithms. In Fig. 5, we show the Shapley values computed during each iteration for both algorithms, using a step size $\delta = 0.1$. Both figures indicate the squared Mahalanobis distance (black bar) and the corresponding (non-)central chi-square quantile (dotted line).

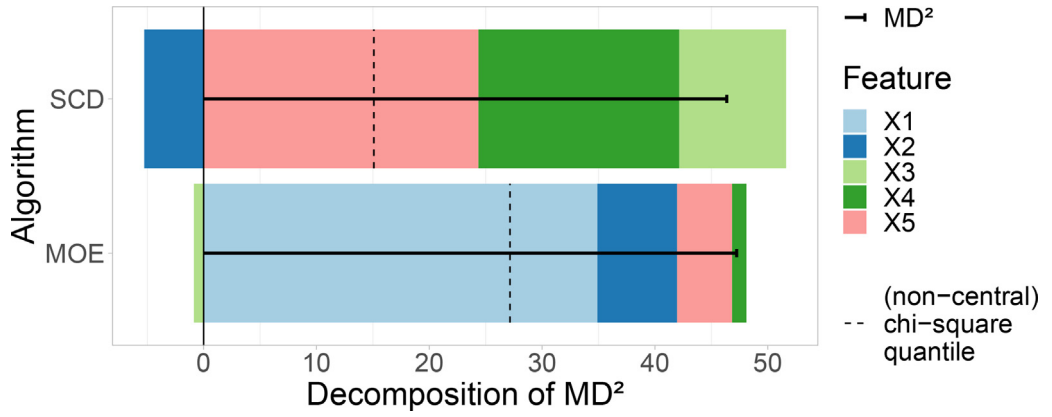

Fig. 4. Comparison of the Shapley values $\phi(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ used in Algorithm 1 to explain the *global* outlyingness, and $\phi(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$, used in Algorithm 2 to gain *local* insights on the outlyingness, with the input values defined in Example 3.1.1. The SCD procedure identifies the three coordinates x_3, x_4, and x_5, which are furthest from the mean $\boldsymbol{\mu}$. On the other hand, the MOE algorithm uses the alternative reference point $\tilde{\boldsymbol{\mu}}$ to identify variables x_1 and x_2.

Algorithm 2 Multivariate Outlier Explainer (MOE)

procedure MOE($\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta, \eta$)
  $\tilde{\boldsymbol{x}} \leftarrow \boldsymbol{x}$
  $R \leftarrow \emptyset$
  $\boldsymbol{d} = (d_1, \ldots, d_p)' \leftarrow (0, \ldots, 0)'$
  $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_p)' \leftarrow \tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R) = (x_1 - \hat{\beta}(1), \ldots, x_p - \hat{\beta}(p))'$
  $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)' \leftarrow \boldsymbol{\phi}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = (\phi_1(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}), \ldots, \phi_p(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}))'$
  while $\mathrm{MD}^2_{\tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}}(\tilde{\boldsymbol{x}}) > \chi^2_{p, 0.99}(\mathrm{MD}^2(\tilde{\boldsymbol{\mu}}))$ **do**
    $R \leftarrow R \cup \{j_1, \ldots, j_k\}$, where $(\phi_{j_l})_{l=1,\ldots,k} = \max_{i=1,\ldots,p} \phi_i$
    while $\max_{j \in R} \phi_j > \max_{j \in \bar{R}} \phi_j$ **do**
      $\boldsymbol{c} \leftarrow (\tilde{\boldsymbol{x}}_R - \tilde{\boldsymbol{\mu}}_R)\delta$
      $\boldsymbol{d}_R \leftarrow \boldsymbol{d}_R + \boldsymbol{c}$
      $\tilde{\boldsymbol{x}}_R \leftarrow \tilde{\boldsymbol{x}}_R - \boldsymbol{c}$
      $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$
    end while
    $\tilde{\boldsymbol{\mu}} \leftarrow \tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R)$
  end while
  $\boldsymbol{d} = (d_1, \ldots, d_p)' \leftarrow (d_1/\sqrt{\sigma_{11}}, \ldots, d_p/\sqrt{\sigma_{pp}})'$
  $R \leftarrow \{j_1, \ldots, j_m\}$, for which $(d_{j_l})_{l=1,\ldots,m} > \eta \max_{i=1,\ldots,p} d_i$
  $\tilde{\boldsymbol{\mu}} \leftarrow \tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R)$
  $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi}(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$
  $\tilde{\boldsymbol{x}} \leftarrow \boldsymbol{x}$
  $\tilde{\boldsymbol{x}}_R \leftarrow \tilde{\boldsymbol{\mu}}_R$
  return $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\phi}$
end procedure

4. Simulations

The simple numerical example from the previous section has illustrated that the SCD and MOE algorithms can lead to quite different outcomes. However, it needs to be emphasized that their purposes also differ: While the SCD procedure aims at global outlier explanation, i.e. with respect to the distribution of the entire dataset, the MOE procedure is locally applicable and builds on the local information contained in the regular cells of an individual observation. Nevertheless, it can be interesting to compare both procedures in terms of their ability to identify cellwise outliers and, in particular, to examine their performance in comparison to a reference method, namely the cellHandler procedure introduced by Raymaekers and Rousseeuw (2021). We choose standard parameters for all three procedures, meaning that both the SCD and MOE algorithms are set up with a step size of $\delta = 0.1$, and the MOE procedure additionally uses a detection threshold $\eta = 0.2$.

In our analysis, we compare two different mechanisms for generating outliers and analyze the effects of various parameter configurations, which are summarized in Table 1 and described in more detail in the following paragraphs. For each specific parameter combination, we repeat the simulations 50 times and compute averages of the resulting measures Recall, Precision, and F-Score.

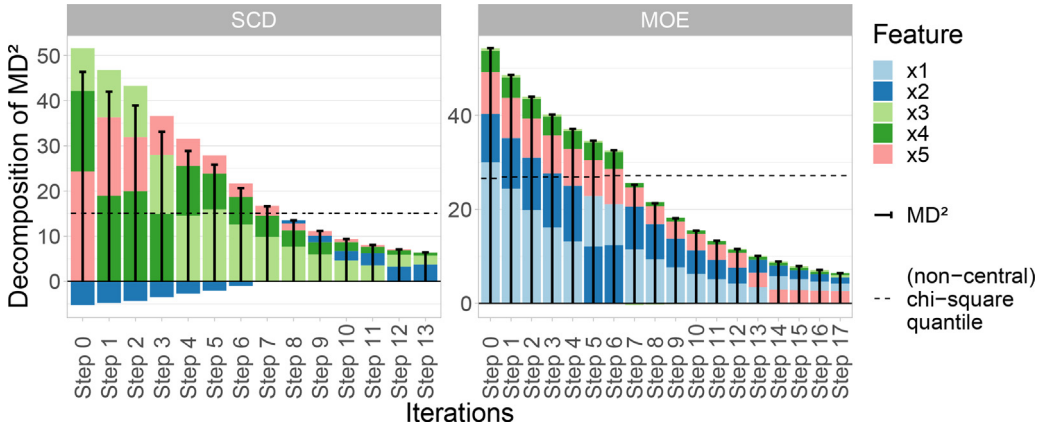

Fig. 5. Comparison between the Shapley values calculated for each iteration of Algorithm 1 (left) and Algorithm 2 (right), respectively, for Example 3.1.1. While the outlyingness is monotonically decreasing in both cases, the sets of identified variables are disjoint. Both the SCD and MOE procedures reduce the outlyingness by iteratively shifting the identified variables toward the corresponding coordinates of $\boldsymbol{\mu}$ or $\tilde{\boldsymbol{\mu}}$, respectively.

Table 1
Summary of the parameters used for the two simulation scenarios on cellwise outlier detection discussed in Section 4.

Parameters	Shift outliers	Structured outliers
Dimension, p	5, 10, 20, 30, 40	5, 10, 20, 30, 40
Covariance, $\boldsymbol{\Sigma}$	$\boldsymbol{C}_{\text{mix}}$, $\boldsymbol{C}_{\text{low}}$, $\boldsymbol{C}_{\text{mod}}$	$\boldsymbol{C}_{\text{mix}}$, $\boldsymbol{C}_{\text{low}}$, $\boldsymbol{C}_{\text{mod}}$
Fraction of outlying columns, ϵ_1	0.1, 0.2, 0.3, 0.4	-
Fraction of outlying rows, ϵ_2	0.1, 0.2, 0.3, 0.4	-
Fraction of outlying cells, ϵ_3	-	0.1, 0.2, 0.3, 0.4
Magnitude of outlyingness, γ	1, 2, 3	2, 3, 4, 5, 6
Total combinations	720	300

For both outlier generation procedures, we generate data matrices with p columns and $n = 20p$ rows from multivariate normal distributions with mean $\boldsymbol{\mu} = \mathbf{0}$ and three different types of covariance matrices $\boldsymbol{\Sigma}$, namely $\boldsymbol{C}_{\text{mod}}$, $\boldsymbol{C}_{\text{mix}}$, and $\boldsymbol{C}_{\text{low}}$. In all three cases, the diagonal elements are set to 1. For $\boldsymbol{C}_{\text{mod}}$, the off-diagonal elements are chosen as 0.5, resulting in moderate correlations. The off-diagonal elements of $\boldsymbol{C}_{\text{mix}}$ correspond to $(-0.9)^{|j-k|}$, $j \neq k$, yielding both high and low correlations. For $\boldsymbol{C}_{\text{low}}$, the off-diagonal elements are randomly generated as described in Agostinelli et al. (2015), generally resulting in low correlations.

To analyze the effect of highly correlated shift-outliers, we randomly select $\lceil n\epsilon_2 \rceil$ rows, and for each of those rows, we replace $r = \lceil p\epsilon_1 \rceil$ randomly selected cells by r-variate outliers. Those follow a Gaussian distribution with mean $\boldsymbol{\mu} = (\gamma, \ldots, \gamma)'$ and covariance matrix $\tilde{\boldsymbol{\Sigma}}$, with elements $\tilde{\sigma}_{jk} = 0.7$, $j \neq k$, and $\tilde{\sigma}_{jj} = 1$. The magnitude of the outliers is determined by the value γ, which is selected according to Table 1. Following this approach, the fraction of outlying cells ranges between 0.01 and 0.16.

For the second scenario, outliers are generated such that they are structurally outlying but have low univariate outlyingness, as proposed by Raymaekers and Rousseeuw (2021). For this purpose, $n\epsilon_3$ cells are selected randomly in each column. Like this, each row contains a subset $K \subseteq P$ of cells $\boldsymbol{x}_K$ which are subsequently replaced by the vector $\gamma\sqrt{k}\boldsymbol{u}' / \text{MD}_{\boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K}(\boldsymbol{u})$, where $k = |K|$, and $\boldsymbol{u}$ is the eigenvector of $\boldsymbol{\Sigma}_K$ that corresponds to the smallest eigenvalue.

We summarize the overall results in Table 2, comparing Precision, Recall, and F-Score. The performance metrics are averaged over all parameters not listed in the table (p, ϵ_1, ϵ_2, ϵ_3, and γ) and all replications. Regarding Precision, the MOE algorithm exhibits the best results in 4 out of 6 settings. Concerning Recall, the SCD procedure performs best when the correlations are low to moderate, while the cellHandler procedure performs best when the correlations are moderate or mixed. Finally, when comparing the F-Score, we see that each algorithm outperforms the remaining two at least once. However, the results listed in the table are averaged over a wide range of parameter settings. Therefore, we study the individual effects of the different parameters in more detail in the following.

In Fig. 6, we analyze the effect of the dimension p on the cellwise outlier detection performance. We focus on the case of highly correlated shift outliers, with fixed $\epsilon_1 = \epsilon_2 = 0.4$ and $\gamma = 3$. This results in a situation with many moderately contaminated cells. We observe an increase in Precision for all three algorithms and covariance structures as p increases. The SCD procedure shows the most substantial increase and the highest overall Precision in case of low and moderate correlations. For the mixed correlations, the MOE procedure exhibits the highest Precision. Moving on to Recall, we see an initial increase followed by a very slight decline for all three methods for mixed and moderate correlations. For low correlations, the MOE procedure shows a severe drop in Recall, while the other two procedures only show a slight decline

Table 2
Summary of the results of the simulations described in Section 4. The performance metrics Precision, Recall, and F-Score listed in this table are averaged over all replications and parameter combinations.

		Shift outliers			Structured outliers		
$\boldsymbol{\Sigma}$	Algorithm	Precision	Recall	F-Score	Precision	Recall	F-Score
$\boldsymbol{C}_{mix}$	SCD	0.690	0.737	0.708	0.546	0.551	0.540
$\boldsymbol{C}_{mix}$	MOE	0.894	0.707	**0.782**	0.916	0.545	**0.668**
$\boldsymbol{C}_{mix}$	cellHandler	0.760	0.743	0.741	0.854	0.564	0.667
$\boldsymbol{C}_{low}$	SCD	0.713	0.510	**0.574**	0.767	0.715	**0.729**
$\boldsymbol{C}_{low}$	MOE	0.678	0.396	0.478	0.880	0.597	0.695
$\boldsymbol{C}_{low}$	cellHandler	0.599	0.473	0.508	0.900	0.630	0.722
$\boldsymbol{C}_{mod}$	SCD	0.767	0.405	0.507	0.859	0.530	0.627
$\boldsymbol{C}_{mod}$	MOE	0.808	0.421	**0.528**	0.954	0.476	0.599
$\boldsymbol{C}_{mod}$	cellHandler	0.649	0.471	0.522	0.917	0.513	**0.634**

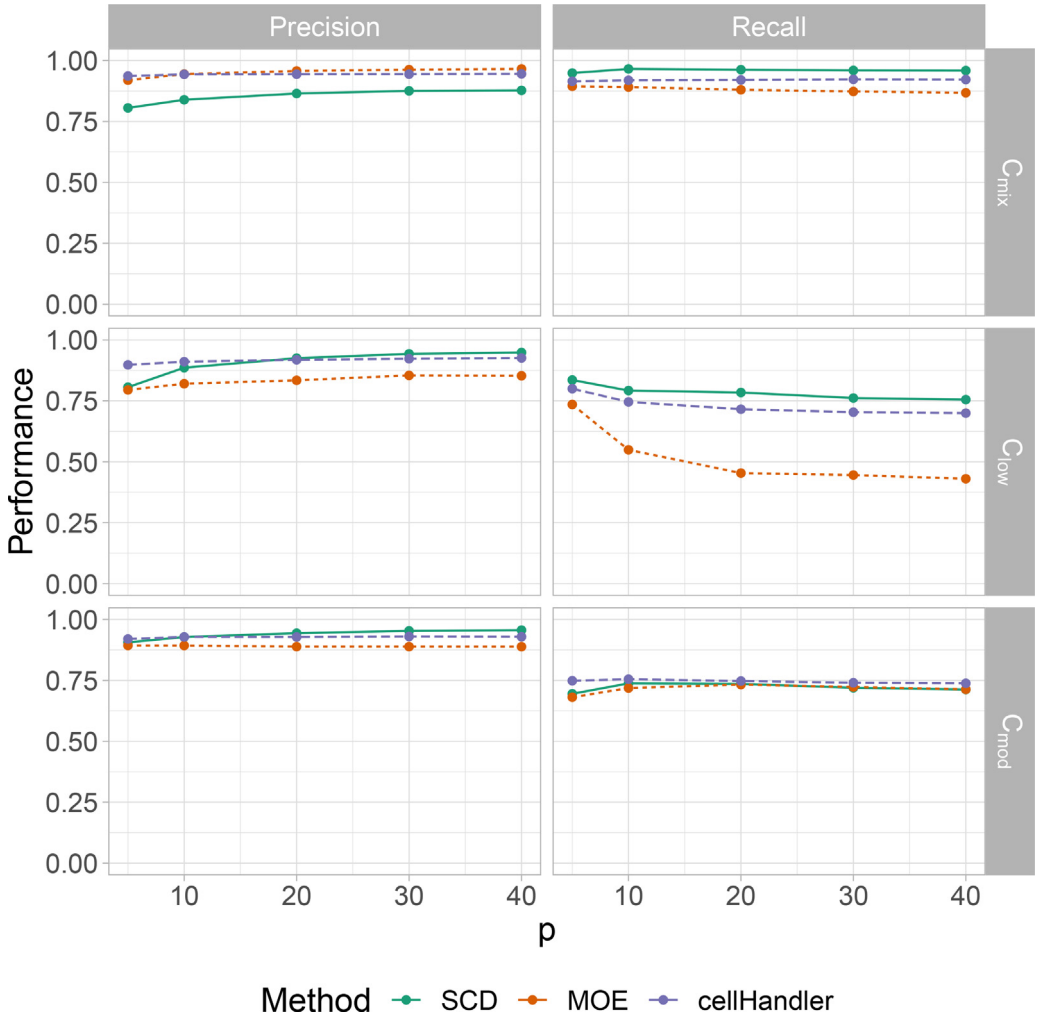

Fig. 6. Comparison between the SCD, MOE, cellHandler procedures in the simulation setting of cellwise shift outliers outlined in Section 4, with simulation parameters $\epsilon_1 = \epsilon_2 = 0.4$ and $\gamma = 3$. The performance scores Precision (left) and Recall (right) of the individual algorithms are listed separately for each type of covariance structure.

in performance. This is related to the default choice of the tuning parameter $\eta = 0.2$ for the MOE procedure, and the Recall could be improved by choosing a smaller value of η. While the Recall is similar for all methods in case of moderate and mixed correlations, we observe that the SCD procedure has the highest Recall in case of low correlations, followed by the cellHandler algorithm.

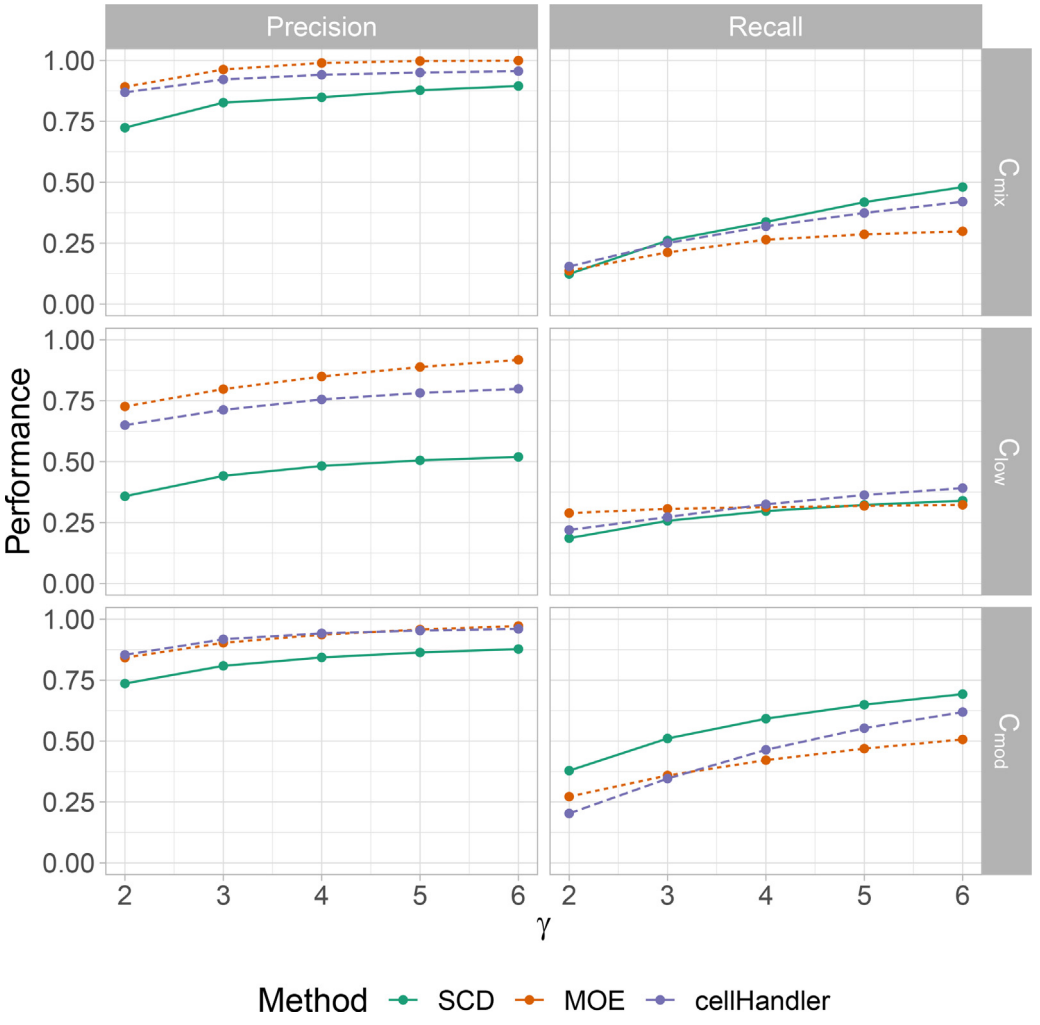

Fig. 7. Comparison between the SCD, MOE, cellHandler procedures in the simulation setting of structured cellwise outliers outlined in Section 4, with simulation parameters $\epsilon_3 = 0.4$ and $p = 30$. The performance scores Precision (left) and Recall (right) of the individual algorithms are listed separately for each type of covariance structure.

For the structured outliers, we illustrate the influence of γ for fixed $\epsilon_3 = 0.4$ and $p = 30$ in Fig. 7. As expected, Precision and Recall are increasing as the magnitude of outlyingness, controlled by γ, increases. The MOE procedure shows the highest overall Precision. However, regarding Recall, the SCD procedure performs better for mixed and high correlations. For low correlations, the cellHandler procedure exhibits the steepest increase in Recall as γ increases.

In conclusion, these simulations show that our approaches based on the Shapley value, particularly the MOE procedure, yield comparable results to one of the current state-of-the-art methods, namely the cellHandler procedure. While cellwise outlier detection presents the focus of the latter method, our approach is instead based on utilizing cellwise outlier detection specifically to enhance and robustify the outlyingness scores based on Theorem 2.2.1, with respect to an observation's "expected" position, as outlined in Eqs. (17) and (19).

5. Applications

While the simulations shown in the previous section have demonstrated the performance of the methods and algorithms introduced in Sections 2 and 3 on simulated datasets, we now apply them to two real-world data. To this end, we analyze the *Top Gear* dataset from Alfons (2021) and the *Weather in Vienna* dataset from Stadt Wien (2022).

5.1. Top Gear

The *Top Gear* dataset comprises measurements of 11 numerical attributes (see Fig. 8) of 245 complete data instances of cars featured on the website of the BBC television series. We apply a logarithmic transformation to five variables for data

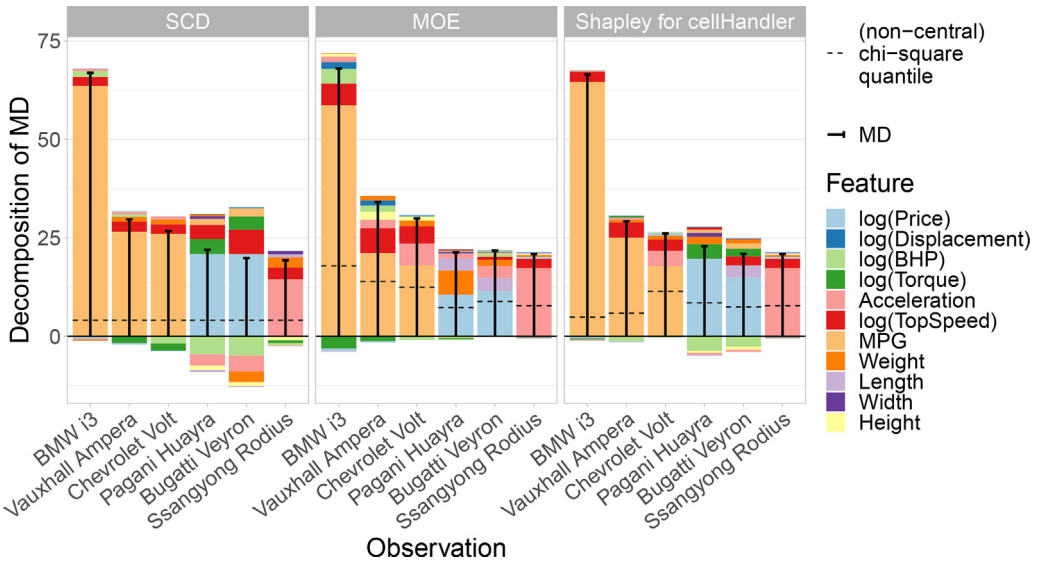

Fig. 8. Comparison of the outlyingness scores resulting from the SCD (left), MOE (center), and cellHandler (right) procedures. Each graph shows a visualization of the Shapley values for the six most outlying observations.

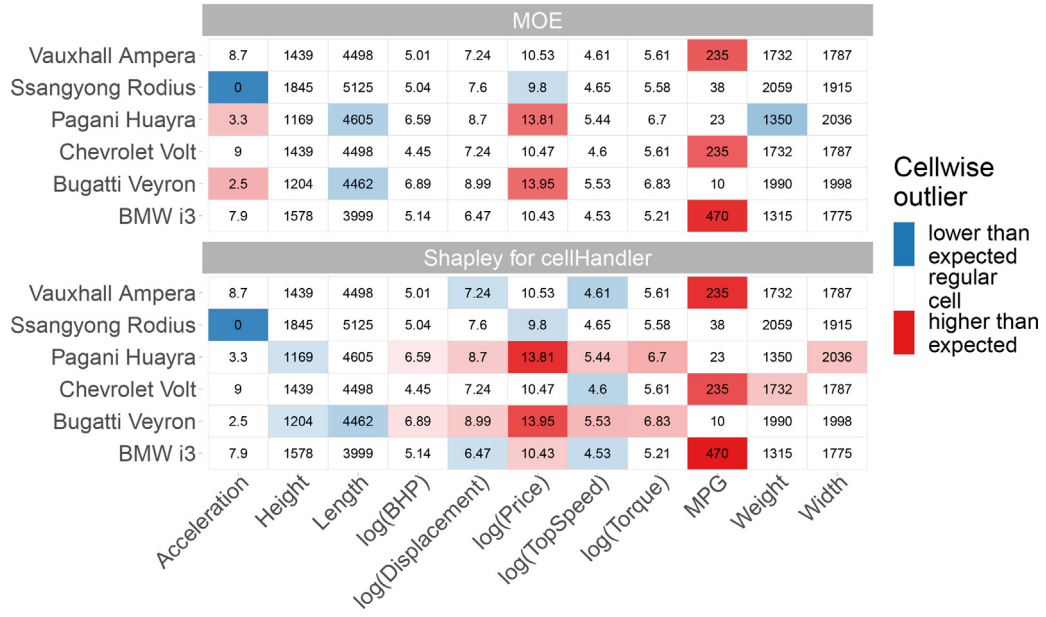

Fig. 9. Outlying cells according to Algorithm 2 (top) and the cellHandler procedure (bottom). Each cell shows the original value from the dataset, color coding indicates whether those values were higher (red) or lower (blue) than the imputed values, and the color intensity is based on the magnitude of the Shapley value.

preprocessing to obtain more symmetrical marginal distributions. Additionally, each column is robustly centered and scaled based on the median and the MAD. Furthermore, we estimate the covariance using the MCD estimator before applying the SCD, MOE, and cellHandler procedures.

In the following, we use three different types of plots to analyze the results of all three tested algorithms on this dataset: Fig. 8 summarizes the Shapley values, Fig. 9 shows the outlying cells, and Fig. 10 displays the Shapley interaction indices, respectively.

In detail, Fig. 8 consists of three graphs, each displaying the outlyingness decompositions according to the applied algorithm of the six cars with the highest Mahalanobis distance. In the left panel, we see the results generated using the SCD procedure, where we use the center of the data as a reference point. In the center panel, we show the results of using the MOE algorithm with the non-central chi-square cutoff. For both procedures, we use a step size $\delta = 0.1$, and the MOE

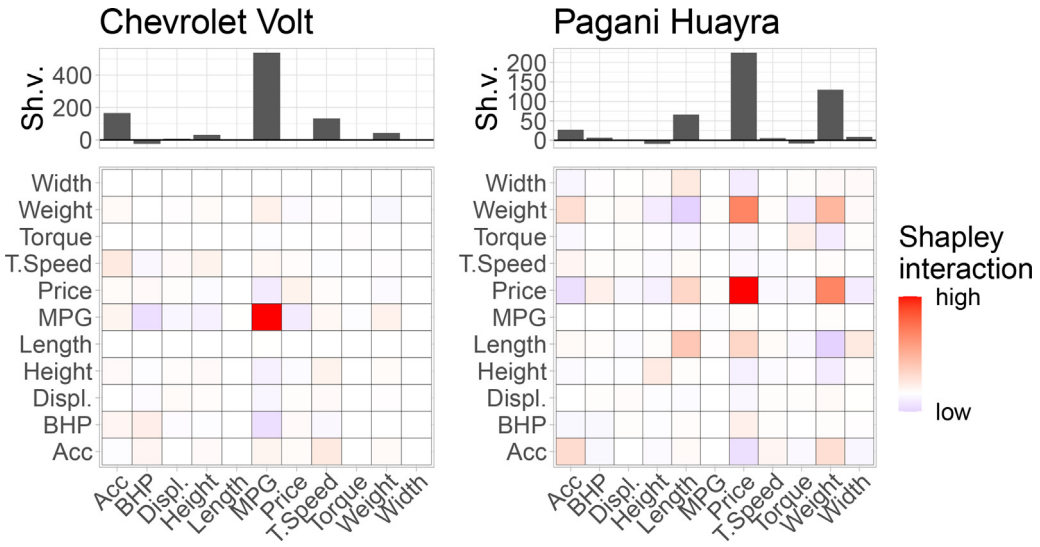

Fig. 10. The two graphs in the lower portion of this figure show the Shapley interaction indices $\Phi(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}(\boldsymbol{x}, S), \Sigma)$ for the `Chevrolet Volt` and `Pagani Huayra`, which are computed with respect to the reference point provided by Algorithm 2. The corresponding Shapley values are displayed above the heatmaps.

algorithm's detection threshold is $\eta = 0.2$. In the right panel, we show the results of using the cellHandler procedure to flag outlying cells, and then employing the Shapley value, with reference point $\tilde{\boldsymbol{\mu}}(\boldsymbol{x}, S)$ according to Eq. (17), to enhance the interpretability of the results, as outlined in Section 3. Since we are analyzing multiple observations with large differences in squared Mahalanobis distance, plotting the squared distance is ineligible, and we display the square root instead. However, since we are decomposing the squared distance in Theorem 2.2.1, we must scale the outlyingness scores. For this reason, we derive each variable's proportional contribution to the squared distance and multiply it by the (not-squared) Mahalanobis distance. While this results in a somewhat distorted graph, this workflow enables us to analyze and compare multiple observations using a stacked bar chart.

Analyzing Fig. 8, we first want to focus on the three cars with the highest outlyingness. For those cars, the main contribution to the squared Mahalanobis distance in the three graphs is caused by the variable `MPG`. Considering that these specific models are hybrid vehicles, it seems reasonable that their fuel consumption differs strongly from that of gasoline and diesel cars. All three methods lead to similar results in this case. For the two sports cars `Bugatti Veyron` and `Pagani Huayra`, we see that the `Price` variable is contributing the most to the outlyingness, which is again visible in the results of all three methods. For these two cars, most characteristics are similar to a certain extent, except for their weight: The Bugatti weighs 1990 kg while the Pagani has only a weight of 1350 kg. This fact becomes clearly visible when applying the MOE algorithm, where the `Weight` variable has a high contribution to the squared Mahalanobis distance of the Pagani but not for the Bugatti. Again, the three procedures agree for the `Ssangyong Rodius`, where `Acceleration` contributes the most. In fact, the listed value for `Acceleration` is 0, which is clearly an error in the published dataset itself.

In Fig. 9, we show the results of applying the MOE procedure (top) to the TopGear dataset, as well as the Shapley values based on the cellHandler procedure (bottom). In these plots, the original values of the variables are displayed in each cell. White rectangles represent regular cells, while outlying cells are colored red or blue, depending on whether the cell's original value is higher (red) or lower (blue) than the replacement. The color intensity is given according to the Shapley values of the cells. The biggest differences between the MOE and the cellHandler algorithm can be seen between the two sports cars `Bugatti Veyron` and `Pagani Huayra`, where the cellHandler procedure results in many more outlying cells. However, it is surprising that the `Acceleration` parameter is not flagged since both cars have an exceptionally fast acceleration.

Finally, Fig. 10 consists of heatmaps displaying the Shapley interaction indices and barplots showing the corresponding Shapley values for the `Chevrolet Volt` (left) and `Pagani Huayra` (right). The Shapley values and interaction indices are based on the reference point obtained from Algorithm 2. For the Chevrolet, we see a single outstanding index for `MPG`. On the other hand, the Pagani not only shows a high index for `Price` but also for the pairwise outlyingness score between `Weight` and `Price`, which indicates that for an expensive sports car, it is unexpectedly lightweight.

5.2. Weather in Vienna

As a second real-world example, we analyze monthly weather data from the weather station "Hohe Warte" in Vienna (Stadt Wien, 2022). Therefore we consider 16 numerical attributes, which are described in Table C.3 in Appendix C, over a time period spanning from 1955 to 2022. Furthermore, we restrict our investigation to the three summer months, June, July,

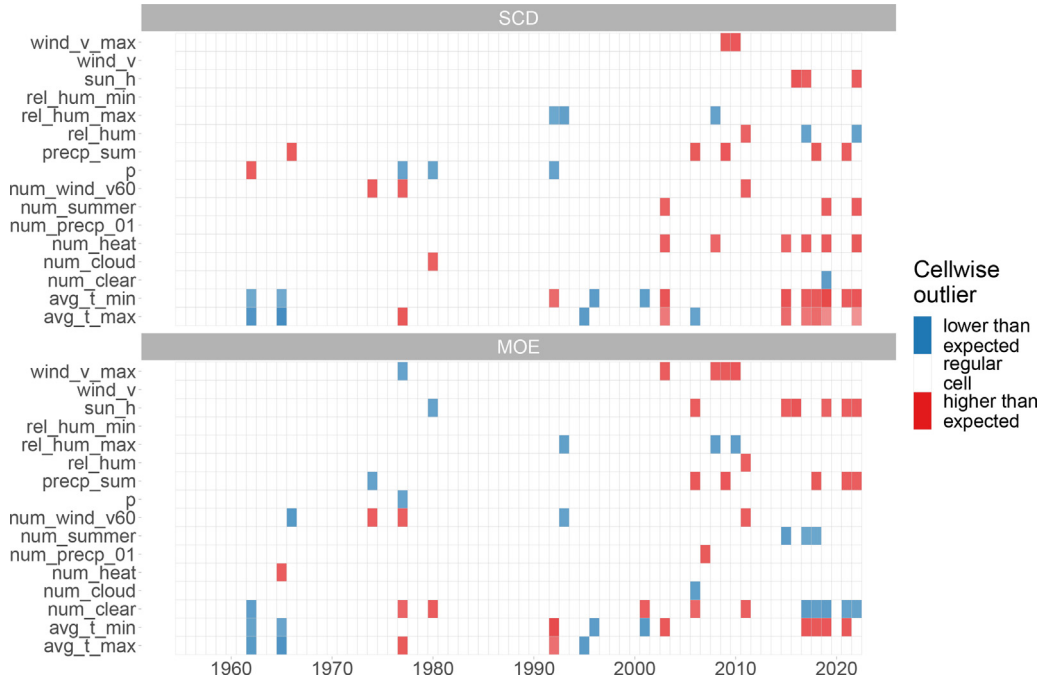

Fig. 11. Comparison of outlying cells according to Algorithm 1 (top) and Algorithm 2 (bottom) for the weather data of Vienna. It is visible in the results of both procedures that the number of anomalies is increasing over the years.

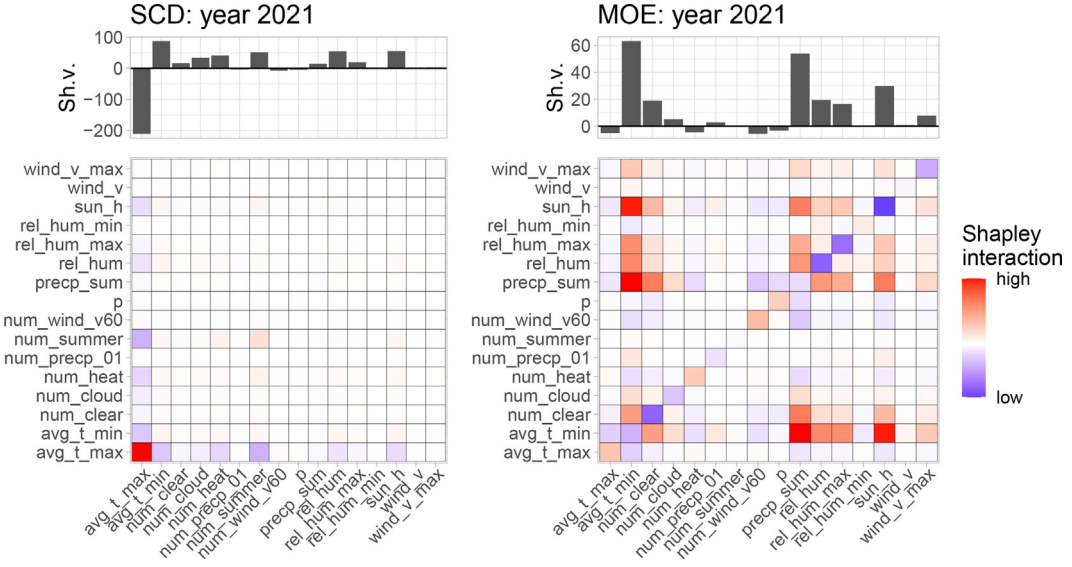

Fig. 12. The two graphs in the lower panel show the Shapley interaction indices of the year 2021 for the SCD procedure (left) and the MOE procedure (right). The corresponding Shapley values are displayed above the heatmaps.

and August, and compute average values for the considered variables, which yields 68 annual observations for each variable. As for the previous example, we center and scale the data using median and MAD and estimate the covariance using the MCD estimator before applying the SCD and MOE algorithms using the same setup as before.

Fig. 11 displays the outlying cells of the entire 68 years of measurements: The top panel shows the results from the SCD algorithm, and the bottom panel displays those from the MOE algorithm. Both panels reveal that the number of detected anomalies has increased over the years. The SCD procedure further yields results that we would expect to find given that we are currently experiencing an anthropogenic climate change, such as an increasing number of hot days over the years or an increased minimum and maximum mean daily temperature. We emphasize that the SCD procedure results in a global

outlyingness measure with respect to the overall mean. On the other hand, the MOE algorithm acts as a local measure: With given values of the regular cells in a particular year, the outlyingness in the remaining variables is determined.

A more detailed analysis of the results can be made by comparing the Shapley values and pairwise outlyingness scores obtained from each procedure. Such an analysis is representatively carried out for the year 2021. The results are displayed in Fig. 12, where we can observe a clear distinction between the results of the SCD and MOE procedures, respectively. Both algorithms detect anomalies in the average temperature minimum (`avg_t_min`) and total precipitation (`precp_sum`). However, using the local reference point enables the MOE procedure to detect outliers in the number of sun hours (`sun_h`) and the number of clear days (`num_clear`). According to the results of the local MOE procedure given in Figs. 11 and 12, the weather of Vienna in 2021 was unusually hot, with more rain than we would expect. When considering the trend of increasing temperature over the years at this specific weather station, we would generally expect fewer sun hours and more clear days than observed in 2021.

6. Discussion and conclusions

This paper introduced Shapley values in connection with Mahalanobis distances for multivariate outlier explanation. The Mahalanobis distance is commonly employed for multivariate outlier detection in statistics. Then again, the Shapley value is a concept that originated in cooperative game theory and recently gained popularity in the field of Explainable AI. There it is used to explain the predictions of complex machine learning models by providing information about the contributions of the individual features to a model's prediction. Combining the Shapley value with the squared Mahalanobis distance enables us to derive outlyingness scores for each coordinate of an observation. Those scores consider all 2^p possible combinations of p variables of a single instance and allow us to additively decompose the squared Mahalanobis distance into contributions originating from the individual variables. Without further simplification, the computation would entail evaluating the squared Mahalanobis distance for those 2^p combinations, which would pose a substantial computational challenge. However, we showed that our approach leads to a much simpler and computationally efficient form of the Shapley value. Moreover, the Shapley interaction indices generalize Shapley values and can be used to derive outlyingness scores for pairs of variables.

Outlier explanation, and thus identifying the contributions of a variable to the outlyingness of a particular observation, is closely related to cellwise outlyingness, where one aims to identify unusual cells instead of entire observations. We have adopted cellwise outlyingness into the framework of Shapley values and have proposed two procedures for simultaneous outlier detection and explanation. First, we introduced the SCD procedure as a straightforward implementation of Shapley values for cellwise outlier detection. This algorithm is iteratively replacing anomalous cells with a value towards their mean until the observation is no longer outlying. The more sophisticated MOE procedure takes the information of the non-outlying cells into account and determines a local reference point based on this added input. As a result, one again obtains an additive decomposition of the squared Mahalanobis distance, but with contributions that explain the *local* outlyingness of an observation.

The performance of the two cellwise outlier detection and explanation procedures has been evaluated in simulations and on real-world datasets. It has further been compared to the recently published cellHandler procedure. However, we want to emphasize that the goal of our work is clearly defined as outlier explanation rather than cellwise outlier detection. In particular, Mahalanobis distances rely on a robustly estimated covariance matrix, which has not been in focus in this paper.

We believe that Shapley values are a powerful tool for providing humanly interpretable explanations that allow us to gain further insights into the results of models and methods used in statistics and computer science. They show great potential for further use in this area, especially when a simplification of the computation is possible, as is the case when combining them with Mahalanobis distances. Possible extensions of Shapley values for outlier detection in functional data analysis will be the subject of our future research.

Software and data availability: The methods introduced in this work are available in the R package `ShapleyOutlier` on CRAN, including the weather dataset and a vignette to reproduce the examples presented in Section 5.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors want to thank the editor and associate editor for providing guidance and support throughout the publication process and the reviewers for their constructive feedback that helped improve the quality of the paper.

This work is part of the AI4CSM project and has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 101007326. The JU receives support from the European Union's Horizon 2020 research and innovation programme and national authorities. This work is also funded by the Austrian IKT der Zukunft programme via the Austrian Research Promotion Agency (FFG) and the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK) under project No 884070. The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Programme.

Appendix A. Proof of Theorem 2.2.1

Lemma A.0.1. *The contributions* $\Delta_k \mathrm{MD}^2(\hat{\boldsymbol{x}}^S) = \mathrm{MD}^2(\hat{\boldsymbol{x}}^{S\cup\{k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^S)$ *can be expressed as*

$$\Delta_k \mathrm{MD}^2(\hat{\boldsymbol{x}}^S) = 2(x_k - \mu_k)\left(\sum_{j\in S\cup\{k\}} (x_j - \mu_j)\omega_{jk}\right) - (x_k - \mu_k)^2\omega_{kk}, \tag{A.1}$$

for any subset $S \subseteq P \setminus \{k\}$.

Proof.

$$\begin{aligned}
\Delta_k \mathrm{MD}^2(\hat{\boldsymbol{x}}^S) &= \mathrm{MD}^2(\hat{\boldsymbol{x}}^{S\cup\{k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^S) \\
&= (\hat{\boldsymbol{x}}^{S\cup\{k\}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{x}}^{S\cup\{k\}} - \boldsymbol{\mu}) - (\hat{\boldsymbol{x}}^S - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{x}}^S - \boldsymbol{\mu}) \\
&= \sum_{j=1}^{p}\sum_{l=1}^{p}(\hat{x}_j^{S\cup\{k\}} - \mu_j)(\hat{x}_l^{S\cup\{k\}} - \mu_l)\omega_{jl} - \sum_{j=1}^{p}\sum_{l=1}^{p}(\hat{x}_j^S - \mu_j)(\hat{x}_l^S - \mu_l)\omega_{jl} \\
&= \sum_{j\in S\cup\{k\}}\sum_{l\in S\cup\{k\}}(x_j - \mu_j)(x_l - \mu_l)\omega_{jl} - \sum_{j\in S}\sum_{l\in S}(x_j - \mu_j)(x_l - \mu_l)\omega_{jl} \\
&= \sum_{j\in S\cup\{k\}}(x_k - \mu_k)(x_j - \mu_j)\underbrace{\omega_{kj}}_{=\omega_{jk}} + \sum_{j\in S}(x_k - \mu_k)(x_j - \mu_j)\omega_{jk} \\
&= (x_k - \mu_k)^2\omega_{kk} + 2(x_k - \mu_k)\sum_{j\in S}(x_j - \mu_j)\omega_{jk} = \text{(A.1)}
\end{aligned}$$

□

Now that we have derived a simpler form for the contributions $\Delta_k \mathrm{MD}^2(\hat{\boldsymbol{x}}^S)$, we can use this result to rewrite Eq. (6) for the k-th component of the Shapley value $\phi_k(\boldsymbol{x})$. We apply Lemma A.0.1 in the first step of the proof below, and for a simpler notation, we write

$$w(|S|) := \frac{|S|!(p - |S| - 1)!}{p!},$$

for which $\sum_{S\subseteq P\setminus\{k\}} w(|S|) = 1$ holds.

Proof of Theorem 2.2.1.

$$\begin{aligned}
\phi_k(\boldsymbol{x}) &= \sum_{S\subseteq P\setminus\{k\}} w(|S|)\Delta_k \mathrm{MD}^2(\hat{\boldsymbol{x}}^S) \\
&= \sum_{S\subseteq P\setminus\{k\}} w(|S|)\left((x_k - \mu_k)^2\omega_{kk} + 2(x_k - \mu_k)\sum_{j\in S}(x_j - \mu_j)\omega_{jk}\right) \\
&= (x_k - \mu_k)^2\omega_{kk}\left(\underbrace{\sum_{S\subseteq P\setminus\{k\}} w(|S|)}_{=1}\right) + 2(x_k - \mu_k)\sum_{S\subseteq P\setminus\{k\}}\left(w(|S|)\sum_{j\in S}(x_j - \mu_j)\omega_{jk}\right) \\
&= (x_k - \mu_k)^2\omega_{kk} + 2(x_k - \mu_k)\sum_{s=1}^{p-1}\left(w(s)\sum_{\substack{S\subseteq P\setminus\{k\}\\|S|=s}}\sum_{j\in S}(x_j - \mu_j)\omega_{jk}\right) \\
&= (x_k - \mu_k)^2\omega_{kk} + 2(x_k - \mu_k)\sum_{s=1}^{p-1}\left(w(s)\binom{p-2}{s-1}\sum_{j\in P\setminus\{k\}}(x_j - \mu_j)\omega_{jk}\right) \\
&= (x_k - \mu_k)^2\omega_{kk} + 2(x_k - \mu_k)\sum_{s=1}^{p-1}\left(\frac{s}{p(p-1)}\sum_{j\in P\setminus\{k\}}(x_j - \mu_j)\omega_{jk}\right) \\
&= (x_k - \mu_k)^2\omega_{kk} + (x_k - \mu_k)\sum_{j\in P\setminus\{k\}}(x_j - \mu_j)\omega_{jk} \\
&= (x_k - \mu_k)\sum_{j\in P}(x_j - \mu_j)\omega_{jk} = (x_k - \mu_k)\left(\sum_{j=1}^{p}(x_j - \mu_j)\omega_{jk}\right)
\end{aligned}$$

□

Appendix B. Proof of Theorem 2.3.1

Proof of Theorem 2.3.1. To derive the off-diagonal elements defined in Eq. (12), we start with rewriting $\Delta_{\{j,k\}}\,\mathrm{MD}^2(\hat{\boldsymbol{x}}^T)$, $T \subseteq P \setminus \{j,k\}$, by applying Lemma A.0.1:

$$\begin{aligned}
\Delta_{\{j,k\}}\,\mathrm{MD}^2(\hat{\boldsymbol{x}}^T) =& \left[\mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j,k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j\}})\right] - \left[\mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T})\right] \\
=& 2(x_k-\mu_k)\left(\sum_{l\in T\cup\{j\}}(x_l-\mu_l)\omega_{jk} - \sum_{l\in T\cup\{k\}}(x_l-\mu_l)\omega_{jk}\right) + 2(x_k-\mu_k)^2\omega_{kk} \\
=& 2(x_k-\mu_k)\big((x_j-\mu_j)\omega_{jk} - (x_k-\mu_k)\omega_{kk}\big) + 2(x_k-\mu_k)^2\omega_{kk} \\
=& 2(x_k-\mu_k)(x_j-\mu_j)\omega_{jk}
\end{aligned}$$

Moving on, we plug the result into the formula for Φ_{jk} for $j \neq k$, given in Eq. (12), and we obtain

$$\begin{aligned}
\Phi_{jk} &= \sum_{T\subseteq P\setminus\{j,k\}} \frac{t!(p-t-2)!}{(p-1)!}\Delta_{\{j,k\}}\,\mathrm{MD}^2(\hat{\boldsymbol{x}}^T) \\
&= \sum_{T\subseteq P\setminus\{j,k\}} \frac{t!(p-t-2)!}{(p-1)!} 2(x_k-\mu_k)(x_j-\mu_j)\omega_{jk} \\
&= 2(x_k-\mu_k)(x_j-\mu_j)\omega_{jk},
\end{aligned}$$

where the last equality is obtained by following the same structure as in the proof of Theorem 2.2.1. Finally, we have to derive the diagonal elements Φ_{jj} given by

$$\begin{aligned}
\Phi_{jj} &= \phi_j - \sum_{k\neq j}\Phi_{jk} \\
&= (x_j-\mu_j)\sum_{k=1}^{p}(x_k-\mu_k)\omega_{jk} - 2(x_j-\mu_j)\sum_{k\neq j}(x_k-\mu_k)\omega_{jk} \\
&= (x_j-\mu_j)^2\omega_{jj} - (x_j-\mu_j)\sum_{k\neq j}(x_k-\mu_k)\omega_{jk}. \qquad \square
\end{aligned}$$

Appendix B1. Higher order interactions

Proof. To show that all interactions of order three or higher are zero, it is sufficient to show that for the three-way interactions the *set function derivative* $\Delta_{\{j,k,l\}}\,\mathrm{MD}^2(\hat{\boldsymbol{x}}^T)$ is zero for all $T \subseteq P \setminus \{j,k,l\}$. This follows from the iterative definition of the set function derivative for $S \cap \{j\} = \emptyset$ (Grabisch, 2016), which is given by

$$\Delta_{S\cup\{j\}}\,\mathrm{MD}^2(\hat{\boldsymbol{x}}^T) = \Delta_S(\Delta_j\,\mathrm{MD}^2(\hat{\boldsymbol{x}}^T)).$$

Hence, to show that all Shapley interaction indices

$$I_{Sh}(v,S) = \sum_{T\subseteq P\setminus S}\frac{t!(p-t-s)!}{(p-s+1)!}\Delta_S v(T),$$

with $|S| \geq 3$ are zero, we only have to prove that $\Delta_{\{j,k,l\}}\,\mathrm{MD}^2(\hat{\boldsymbol{x}}^T) = 0, \forall T \subseteq P \setminus \{j,k,l\}$. For this purpose, we first rewrite the above expression and then apply Lemma A.0.1:

$$\begin{aligned}
\Delta_{\{j,k,l\}}\,\mathrm{MD}^2(\hat{\boldsymbol{x}}^T) =& -\mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j,k,l\}}) + \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j,k\}}) + \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j,l\}}) + \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{k,l\}}) \\
&- \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{l\}}) + \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T}) \\
=& -\left[\mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j,k,l\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{k,l\}})\right] + \left[\mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j,l\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{l\}})\right] \\
&+ \left[\mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j,k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{k\}}) + \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T})\right] \\
=& -\left[(x_j-\mu_j)^2\omega_{jj} + 2(x_j-\mu_j)\sum_{m\in T\cup\{k,l\}}(x_m-\mu_m)\omega_{jm}\right] \\
&+ \left[(x_j-\mu_j)^2\omega_{jj} + 2(x_j-\mu_j)\sum_{m\in T\cup\{l\}}(x_j-\mu_j)\omega_{jk}\right] + \left[2(x_k-\mu_k)(x_j-\mu_j)\right] \\
=& -2(x_k-\mu_k)(x_j-\mu_j) + 2(x_k-\mu_k)(x_j-\mu_j) = 0
\end{aligned}$$

$\square$

Table C.3
Description of the parameters of the *Weather in Vienna* dataset.

Parameter	Desciption
`avg_t_max`	Mean daily maximum air temperature in °C
`avg_t_min`	Mean daily minimum air temperature in °C
`num_summer`	Number of summer days (days with a temperature maximum $t_{max} \geq 25.0$ °C)
`num_heat`	Number of hot days (days with a temperature maximum $t_{max} \geq 30.0$ °C)
`p`	Daily mean air pressure in hPa (mean of all measurements at 7 a.m., 2 p.m., 7 p.m. CET; before 1971 9 p.m. instead of 7 p.m.)
`sun_h`	Monthly total sunshine duration in hours
`num_clear`	Number of clear days (daily mean cloudiness $< 20/100$)
`num_cloud`	Number of cloudy days (daily mean cloudiness $> 80/100$)
`rel_hum`	Daily mean relative humidity in percent (2 x RH7 mean + RH14 mean + RH19 mean)/4; before 1971 9 p.m. instead of 7 p.m.)
`rel_hum_max`	Relative humidity maximum in percent
`rel_hum_min`	Relative humidity minimum in percent
`wind_v`	Monthly average wind speed in km/h
`num_wind_v60`	Number of days with wind peaks ≥ 60 km/h
`wind_v_max`	Maximum wind speed in km/h
`precp_sum`	Monthly total precipitation in mm
`num_precp_01`	Number of days with precipitation ≥ 0.1 mm

Appendix C. Weather in Vienna - Parameters

Table C.3 shows the parameter descriptions for the *Weather in Vienna* dataset, which have been adapted and translated from Stadt Wien (2022).

References

Agostinelli, C., Leung, A., Yohai, V., Zamar, R., 2015. Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. Test 24, 441–461. doi:10.1007/s11749-015-0450-6.
Alfons, A., 2021. robustHD: An R package for robust regression with high-dimensional data. Journal of Open Source Software 6 (67), 3786. doi:10.21105/joss.03786.
Alqallaf, F., Van Aelst, S., Yohai, V.J., Zamar, R.H., 2009. Propagation of outliers in multivariate data. The Annals of Statistics 311–331.
Baba, K., Shibata, R., Sibuya, M., 2004. Partial correlation and conditional correlation as measures of conditional independence. Australian & New Zealand Journal of Statistics 46 (4), 657–664.
Biecek, P., Burzykowski, T., 2021. Explanatory Model Analysis. Chapman and Hall/CRC, New York. https://pbiecek.github.io/ema/
Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: A survey. ACM Comput. Surv. 41 (3). doi:10.1145/1541880.1541882.
Debruyne, M., Höppner, S., Serneels, S., Verdonck, T., 2019. Outlyingness: Which variables contribute most? Statistics and Computing 29, 707–723. doi:10.1007/s11222-018-9831-5.
Filzmoser, P., Ruiz-Gazen, A., Thomas-Agnan, C., 2014. Identification of local multivariate outliers. Statistical Papers 55. doi:10.1007/s00362-013-0524-z.
Fujimoto, K., Kojadinovic, I., Marichal, J.-L., 2006. Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. Games and Economic Behavior 55, 72–99. doi:10.1016/j.geb.2005.03.002.
Grabisch, M., 2016. Set Functions, Games and Capacities in Decision Making, 1st Springer Publishing Company, Incorporated.
Grabisch, M., Roubens, M., 1999. An axiomatic approach to the concept of interaction among players in cooperative games. International Journal of Game Theory 28, 547–565.
Grubbs, F.E., 1969. Procedures for detecting outlying observations in samples. Technometrics 11 (1), 1–21. http://www.jstor.org/stable/1266761
Lundberg, S.M., Erion, G., Chen, H., Degrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., et al, 2020. From local explanations to global understanding with explainable ai for trees. Nature Machine Intelligence 2 (1), 56–67. doi:10.1038/s42256-019-0138-9.
Lundberg, S.M., Erion, G.G., Lee, S.-I., 2018. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
Mahalanobis, P.C., 1936. On the generalized distance in statistics. Proceedings of the National Institute of Sciences (Calcutta) 2, 49–55.
Molnar, C., 2019. Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/
Owen, G., 1972. Multilinear extensions of games. Management Science 18 (5-part-2), 64–79.
Peters, H., 2008. Game Theory. Springer, Berlin Heidelberg.
Raymaekers, J., Rousseeuw, P., 2021. Handling cellwise outliers by sparse regression and robust covariance. Journal of Data Science, Statistics, and Visualisation 1. doi:10.52933/jdssv.v1i3.18.
Raymaekers, J., Rousseeuw, P.J., 2022. The cellwise minimum covariance determinant estimator. arXiv preprint arXiv:2207.13493 doi:10.48550/ARXIV.2207.13493.
Ribeiro, M., Singh, S., Guestrin, C., 2016. "why should i trust you?": Explaining the predictions of any classifier, pp. 97–101. doi:10.18653/v1/N16-3020.
Rousseeuw, P., 1985. Multivariate estimation with high breakdown point. Mathematical Statistics and Applications Vol. B 283–297. doi:10.1007/978-94-009-5438-0_20.
Rousseeuw, P., Zomeren, B., 1990. Unmasking multivariate outliers and leverage points. Journal of The American Statistical Association - J AMER STATIST ASSN 85, 633–639. doi:10.1080/01621459.1990.10474920.
Rousseeuw, P.J., Bossche, W.V.D., 2018. Detecting deviating data cells. Technometrics 60 (2), 135–145. doi:10.1080/00401706.2017.1340909.
Shapley, L.S., 1953. A value for n-person games. Contributions to the Theory of Games 2 (28), 307–317.
Stadt Wien, 2022. Monthly data from the weather station Hohe Warte since April 1872 - Vienna. https://www.data.gv.at/katalog/dataset/wetter-seit-1872-hohe-warte-wien.
Štrumbelj, E., Kononenko, I., 2010. An efficient explanation of individual classifications using game theory. Journal of Machine Learning Research 11, 1–18.
Štrumbelj, E., Kononenko, I., 2014. Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems 41, 647–665.

Sundararajan, M., Dhamdhere, K., Agarwal, A., 2020. The shapley taylor interaction index. In: International Conference on Machine Learning. PMLR, pp. 9259–9268.
Young, H., 1985. Monotonic solutions of cooperative games. International Journal of Game Theory 14, 65–72.
Zimek, A., Filzmoser, P., 2018. There and back again: Outlier detection between statistical reasoning and data mining algorithms. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8, e1280. doi:10.1002/widm.1280.