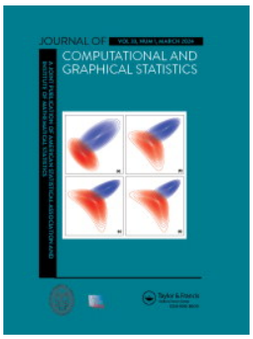

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ucgs20

Spatially Smoothed Robust Covariance Estimation for Local Outlier Detection

Patricia Puchhammer & Peter Filzmoser

To cite this article: Patricia Puchhammer & Peter Filzmoser (21 Dec 2023): Spatially Smoothed Robust Covariance Estimation for Local Outlier Detection, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2023.2277875

To link to this article: https://doi.org/10.1080/10618600.2023.2277875

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2023, VOL. 00, NO. 0, 1–13
https://doi.org/10.1080/10618600.2023.2277875

Spatially Smoothed Robust Covariance Estimation for Local Outlier Detection

Patricia Puchhammer and Peter Filzmoser[a]

Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria

ABSTRACT

Most multivariate outlier detection procedures ignore the spatial dependency of observations, which is present in many real datasets from various application areas. This article introduces a new outlier detection method that accounts for a (continuously) varying covariance structure, depending on the spatial neighborhood of the observations. The underlying estimator thus constitutes a compromise between a unified global covariance estimation, and local covariances estimated for individual neighborhoods. Theoretical properties of the estimator are presented, in particular related to robustness properties, and an efficient algorithm for its computation is introduced. The performance of the method is evaluated and compared based on simulated data and for a dataset recorded from Austrian weather stations. Supplemental materials to the article are available online.

ARTICLE HISTORY
Received June 2023
Accepted October 2023

KEYWORDS
Anomaly detection; Breakdown point; Minimum regularized covariance determinant estimator; Pairwise Mahalanobis distance; Spatial data

1. Introduction

The identification of multivariate outliers is probably one of the most important tasks in multivariate data analysis. A need to find outliers in order to make further analyses more reliable, or the direct interest in the outliers themselves motivate the numerous approaches available for multivariate outlier detection. The identified outliers are supposed to deviate to a certain extent from the main trend or structure of the data majority, and thus they are also called "global outliers" (Filzmoser, Ruiz-Gazen, and Thomas-Agnan 2013). In contrast, the term "local outliers" refers to a setting where additional information regarding some kind of neighborhood is available, for example provided by spatial coordinates of the observations. Then, local outliers are observations which clearly differ from the multivariate measurements of their spatial neighbors indicating local anomalies that spark interest and make further analysis essential. Nevertheless, the values themselves might still be in an ordinary range of the dataset, and thus, the observation would not be outlying in a global sense.

Existing statistical approaches for multivariate local outlier detection are often based on a distance measure and neighborhood structure. A neighborhood a is defined as a subset of the set of observation indexes, say $\{1, \ldots, n\}$. A p-variate observation $\boldsymbol{x}_i$, for $i \in \{1, \ldots, n\}$, is defined to be in neighborhood a if and only if $i \in a$. The decision if some observation $\boldsymbol{x}$ is in a neighborhood is typically based on its spatial coordinates $s(\boldsymbol{x})$. One way to construct the spatial neighborhood is to take a spatial k-nearest-neighborhood of each point $\boldsymbol{x}$, where $k \in \mathbb{N}$. For a fixed $\boldsymbol{x}$, the spatial distance to another point $\boldsymbol{y}$ is defined as the Euclidean distance

$$d_{\boldsymbol{x}}(\boldsymbol{y}) = \left|\left|s(\boldsymbol{x}) - s(\boldsymbol{y})\right|\right| = \left[\left(s(\boldsymbol{x}) - s(\boldsymbol{y})\right)' \left(s(\boldsymbol{x}) - s(\boldsymbol{y})\right)\right]^{1/2}.$$

A spatial neighborhood $a_k(\boldsymbol{x})$ can then be defined as the set of the k many spatial nearest observations,

$$a_k(\boldsymbol{x}) = \{\boldsymbol{x}_j : \ d_{\boldsymbol{x}}(\boldsymbol{x}_j) \leq d_{\boldsymbol{x}(k)}\}, \tag{1}$$

where $d_{\boldsymbol{x}(1)} \leq d_{\boldsymbol{x}(2)} \leq \cdots \leq d_{\boldsymbol{x}(k)} \leq d_{\boldsymbol{x}(n)}$ are the sorted distances to all observations $\boldsymbol{x}_i, i = 1, \ldots, n$. Regarding local outlier detection, a distance measure often used to evaluate outlyingness is the pairwise Mahalanobis distance (MD). For a neighborhood a with covariance Σ_a and an observation $\boldsymbol{x}_i$ in neighborhood a, the pairwise MD is defined as

$$\mathrm{MD}_{\Sigma_a}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left[(\boldsymbol{x}_i - \boldsymbol{x}_j)' \Sigma_a^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)\right]^{1/2} \quad \text{for all } j \in a.$$

In general, the MD describes the distance between two observations, where the Euclidean distance in the feature space is adapted according to local distribution properties.

The estimation method used to determine Σ_a for all neighborhoods is key to good and reliable results. It is essential that outlying observations themselves are not affecting the estimation, since this could possibly mask outliers, leaving them undetected. Thus, a robust covariance estimation on the neighborhood level is necessary. One of the most widely used robust estimators for the covariance is the Minimum Covariance Determinant (MCD) estimator (Rousseeuw 1984, 1985), where one has to identify the h sub-sample of observations (where h is fixed e.g., to half of the observations) that minimize the determinant of its sample covariance. The MCD covariance estimator is then given by the sample covariance of the h subset, multiplied by a consistency factor (Croux and Haesbroeck 1999). For its computation, the fastMCD algorithm developed by Rousseeuw and Driessen (1999) introduces an iterative concentration step, so-called C-step, that guarantees a decrease of the objective function until convergence to a (local) minimum, making the MCD

CONTACT Patricia Puchhammer patricia.puchhammer@tuwien.ac.at Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

estimator faster and even more popular. The global minimum is then approximated by iterating for a number of random starting values and choosing the smallest local minimum. By selecting a small number of good deterministic starting values for the fastMCD, the detMCD algorithm from Hubert, Rousseeuw, and Verdonck (2012) improves run-time even more. In spite of these excellent features of the MCD estimator, as well as affine equivariance and high robustness, one drawback is that the concept is not applicable in case of singularity of the sample covariance matrix of the h subset, which can easily occur. As for many methods, regularity of the estimated covariance is also needed to compute Mahalanobis distances. Especially in a setting where we are restricted to local neighborhoods consisting of a possibly small subset of observations, we might have a situation where regularity cannot be achieved and thus an inversion of the local covariance matrix is not possible. One solution is to base the local estimation on regularized robust covariance estimators, such as the recently developed Minimum Regularized Covariance Determinant (MRCD) estimator from Boudt et al. (2020) (or also on the Fritsch estimator, Fritsch et al. 2012). One of the many attractive properties of the MRCD is that a slightly adapted fastMCD algorithm based on C-steps is also applicable.

Existing methods for local outlier detection have different ways to define the covariance matrix in order to ensure regularity. The method of Filzmoser, Ruiz-Gazen, and Thomas-Agnan (2013) is dealing with regularity issues by using the MCD estimator calculated on the whole dataset, that is, $\Sigma_a = \Sigma$, thus, imposing a global covariance structure. They have shown that for iid Gaussian random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ with mean $\boldsymbol{\mu}$ and covariance matrix Σ, the conditional distribution of the pairwise squared $\mathrm{MD}_\Sigma(\boldsymbol{x}_i, \boldsymbol{x}_j)$, for $j = 1, \ldots, n$, given $\boldsymbol{x}_i$, is a noncentral Chi-square distribution with p degrees of freedom and noncentrality parameter $\mathrm{MD}^2_\Sigma(\boldsymbol{x}_i, \boldsymbol{\mu})$. Instead of a fixed cut-off value for the pairwise MD, different sophisticated visual approaches are used. By plotting a degree of isolation based on the pairwise MD and quantiles of the noncentral Chi-square distribution, suspicious and highly isolated observations can be discovered and analyzed in more detail. Quite contrary to Filzmoser, Ruiz-Gazen, and Thomas-Agnan (2013), the method of Ernst and Haesbroeck (2016) uses a very local covariance estimation by taking individual k-nearest-neighbor (kNN) neighborhoods for each point separately into account. To tackle the regularity issues, they use a regularized covariance estimation (originally the estimator from Fritsch et al. (2012), for the MRCD see also the adaptation made in Bellino et al. 2019) for each individual kNN neighborhood. Additionally, they introduce the concept of the "next distance", which is also MD based, and use the upper fence of the adjusted boxplot of Hubert and Vandervieren (2008) of all next distances as a nonparametric cut-off value for detecting outliers.

Since it is not necessary to use a MD concept to find local outliers, a short detour to machine learning techniques might be interesting. One of the most prominent approaches for detecting multivariate local outliers in machine learning is the local outlier factor (LOF) developed by Breunig et al. (2000). Initially, locality refers to multivariate values and Euclidean distances in the feature space but this method can also be canonically adapted to spatial local outlier detection. In Schubert, Zimek, and Kriegel (2012) this adaptation and further LOF-based approaches are discussed. Interestingly, also LOF and its variants are based on a concept of distance and neighborhoods.

The existing methods have shortcomings in various ways that have not yet been properly addressed. The rather global nature of the method of Filzmoser, Ruiz-Gazen, and Thomas-Agnan (2013) leads to a reliable and robust estimation of the covariance. Nevertheless, it is somewhat questionable if the estimated covariance is applicable and representative for the covariance structure on the local level. Trying to solve this issue of missing locality in the estimation, Ernst and Haesbroeck (2016) resort to a very local approach by recalculating the covariance matrix for each observation separately. Although more locality is achieved, the method is not taking into account that covariance matrices are not likely to change abruptly from one neighborhood to a next one. Also, the number of estimated parameters is extremely high and based on rather few observations, even if the local covariance structure is abruptly changing. A more global estimation of the local covariance matrices might be more stable and reliable and might also avoid repetitive calculations. It seems that until now there are only two extremes regarding locality of the covariance estimation available.

We bridge the gap between the fully global and the fully local approach by providing a covariance estimator based on the MRCD that addresses the missing locality on the one hand and the missing spatial smoothness on the other. By providing the possibility to set the amount of spatial smoothing and the size of the neighborhoods we get a generalization of the two detection methods, with the goal that good outlier detection properties based on the new local covariance estimations are achieved. Moreover, the covariance estimate can be seen as a generalization of the MRCD when the dataset has additional sub-structures.

The article is organized as follows. In Section 2 we introduce the new covariance estimator, derive its properties as well as properties of the original MRCD and establish the methodology to detect local outliers. An algorithm and the derivation of a generalized C-step are discussed in Section 3. Section 4 provides simulation results regarding run time, convergence and outlier detection, while in Section 5, a real dataset is analyzed using the newly developed local outlier detection method. Finally, the main results are presented and summarized in the conclusions.

2. Methodology

2.1. Spatially Smoothed MRCD Estimators

Assume that the p-dimensional observations $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$, for $i = 1, \ldots, n$, are arranged as rows in the $n \times p$ matrix $\boldsymbol{X}$ with $n > 2p$. Furthermore, each observation has spatial information available, for example spatial coordinates, and is assigned to one of N many neighborhoods, defined by the index sets $a_1, \ldots, a_N$ of size $n_1, \ldots, n_N$.

The goal of the proposed method is to obtain local covariance estimates for each neighborhood that are suitable for calculating the pairwise MDs and to some extent smooth among nearby neighborhoods. Since the MCD estimator requires at least $n > 2p$ to provide a regular covariance matrix, which is a severe limitation especially for small neighborhoods, we focus on its regularized extension, the MRCD estimator. Instead of mini-

mizing the determinant of the sample covariance matrix as in the MCD case, the minimization objective is the determinant of a convex combination of the sample covariance and a symmetric and positive definite matrix, the so-called target matrix $\boldsymbol{T}$. Boudt et al. (2020) suggest a data-driven approach based on the condition number of the covariance matrix to set the degree of regularization ρ which is used in the convex combination. Regarding the target matrix $\boldsymbol{T}$, it is sensible to choose a robust and regular covariance, for example, a diagonal matrix based on univariate robust scale estimates.

In the following, we adapt the idea of the MRCD estimator to our setting of local and smooth covariance estimation. Let $\mathcal{H} = (H_1, \ldots, H_N)$ be subsets of the index sets $a_1, \ldots, a_N$ defining the neighborhoods. The size of each subset H_i is $h_i = |H_i| = \lceil \alpha n_i \rceil$, for $i = 1, \ldots, N$, where α is selected in the interval $[0.5, 1]$, and $\lceil \cdot \rceil$ is the ceiling function, rounding up to the next integer. A smaller value of α will result in more robustness against outliers, and it would also be possible to adjust this value to each neighborhood individually. The observations of subset H_i are written as matrix $\boldsymbol{X}_{H_i}$, with dimensionality $h_i \times p$. Let the neighborhood specific MRCD-based covariance matrix $\boldsymbol{K}_i(\mathcal{H})$, for $i = 1, \ldots, N$, be defined as

$$\boldsymbol{K}_i(\mathcal{H}) = \rho_i \boldsymbol{T} + (1 - \rho_i) c_\alpha \text{cov}(\boldsymbol{X}_{H_i}), \tag{2}$$

with $\text{cov}(\boldsymbol{Y})$ being the sample covariance matrix of $\boldsymbol{Y}$, and c_α a consistency factor for the proportion α (see Croux and Haesbroeck 1999). The regularization parameter ρ_i is set individually for each neighborhood, and it could also be chosen as zero if the estimated covariance matrix is already invertible. Finally, since we want to smooth the covariance matrices, it seems counter intuitive to choose neighborhood specific target matrices, which would also require more parameter estimations. Therefore, we assume a global target matrix $\boldsymbol{T}$, taken as a robust and regular covariance matrix estimated based on the full dataset $\boldsymbol{X}$. Since we assume $n > 2p$ we propose to use the MCD estimator for $\boldsymbol{X}$ as target matrix.

We want to find the combination of subsets in $\mathcal{H}$ that minimizes the objective function

$$f(\mathcal{H}) = \sum_{i=1}^{N} \det \left((1 - \lambda) \boldsymbol{K}_i(\mathcal{H}) + \lambda \sum_{j=1, j \neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}) \right). \tag{3}$$

The tuning parameter $\lambda \in [0, 1]$ is used to balance the influence of an individual local neighborhood and the remaining neighborhoods in the covariance estimations. In case of $\lambda = 0$, there is no spatial influence at all which is equivalent to the estimation of the MRCD for each neighborhood separately while using a global target matrix. For the other extreme $\lambda = 1$, the covariance matrix in a specific neighborhood is an average over the surrounding covariance estimates without adding local information from the neighborhood itself. Moreover, it is possible to interpret the second part in the determinant as a penalization term. Due to the minimization of the determinant, observations from a_i that match well with the main trend of observations in neighborhoods with positive weights ω_{ij} are more likely to be in the optimal H-set if λ is increased. The weights ω_{ij} are supposed to be nonnegative, and we set $\omega_{ii} = 0$. The elements of the weight vector $\boldsymbol{\omega}_i = (\omega_{i1}, \ldots, \omega_{iN})'$ indicate the relative influence that the estimated covariances of other neighborhoods have on the covariance estimation of the ith neighborhood. Also, each weight vector has to sum up to one, $\sum_{j=1}^{N} \omega_{ij} = 1$ for all $i = 1, \ldots, N$. All these weights need to be pre-specified, for example based on inverse geographical distances, and are collected as rows in the weighting matrix $\boldsymbol{W} \in \mathbb{R}^{N \times N}$.

Note that for the objective function (3) a global minimum $\mathcal{H}^* = (H_i^*)_{i=1,\ldots,N}$ exists since its domain consists of a finite number of subset combinations. For this global minimum, the estimated covariance matrix for each neighborhood a_i is

$$\hat{\boldsymbol{\Sigma}}_{SSM,i} = (1 - \lambda) \boldsymbol{K}_i(\mathcal{H}^*) + \lambda \sum_{j=1, j \neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^*), \tag{4}$$

and the location estimate $\hat{\boldsymbol{\mu}}_{SSM,i}$ is the sample mean of the selected observations $\boldsymbol{X}_{H_i^*}$. We call these estimators the spatially smoothed MRCD (ssMRCD) location and covariance estimators.

Although the neighborhood structure and the value of λ will often depend on the data at hand, there are some sensible and natural choices for $\boldsymbol{W}$. If we have a neighborhood structure that has no further meaning and might just be used to divide the spatial space into subsets, an inverse-distance based weight matrix, also used in Sections 4 and 5, might be a good choice. Other possibilities include binary matrices with ones if neighborhoods share a border and zero otherwise, with rows scaled appropriately. Moreover, the regularity parameters can be set by default. For a neighborhood a_i we suggest to set the regularization parameter ρ_i as the data driven value that is proposed by the MRCD algorithm in Boudt et al. (2020), when interpreting the neighborhood as its own dataset.

2.2. Theoretical Properties

In the following we will show that the spatially smoothed MRCD estimators proposed here are—in contrast to the original MRCD estimator—affine equivariant, and we derive its breakdown point. For a dataset $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, a location and covariance estimator are called affine equivariant for all neighborhoods if for any nonsingular matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$, any vector $\boldsymbol{b} \in \mathbb{R}^p$, and all $i = 1, 2, \ldots, N$, it holds that

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{SSM,i}(\boldsymbol{X}\boldsymbol{A}' + \boldsymbol{1}_n \boldsymbol{b}') &= \hat{\boldsymbol{\mu}}_{SSM,i}(\boldsymbol{X})\boldsymbol{A}' + \boldsymbol{b}', \\ \hat{\boldsymbol{\Sigma}}_{SSM,i}(\boldsymbol{X}\boldsymbol{A}' + \boldsymbol{1}_n \boldsymbol{b}') &= \boldsymbol{A} \hat{\boldsymbol{\Sigma}}_{SSM,i}(\boldsymbol{X}) \boldsymbol{A}'. \end{aligned} \tag{5}$$

Theorem 1 (Affine equivariance). Let $\boldsymbol{T}$ be any robust, regular and affine equivariant estimate of the covariance for the dataset $\boldsymbol{X}$, here denoted as $\boldsymbol{T}(\boldsymbol{X})$. Then, the spatially smoothed MRCD estimators of location and covariance with target matrix $\boldsymbol{T}(\boldsymbol{X})$ are affine equivariant.

Proof. The proof is given in the supplement. □

The assumptions of Theorem 1 for $\boldsymbol{T}(\boldsymbol{X})$ can be fulfilled by the MCD applied to the full dataset $\boldsymbol{X}$ for $n > 2p$. Nevertheless, any robust estimator satisfying the assumptions can be used, for example, S-estimators (Rousseeuw and Leroy 1987). Note

that for local outlier detection tasks we typically have enough observations globally to get regularity with standard robust estimators. As a remedy if regularity is not achievable, the MRCD (or e.g., the OGK estimator of Maronna and Zamar 2002) can be used. Assuming that the target matrix can be estimated in a robust, regular and affine equivariant way representing a covariance, the MRCD would also be affine equivariant. However, this might question the usefulness of the MRCD since we already have a robust, regular and affine equivariant covariance estimator available, namely $\boldsymbol{T}(\boldsymbol{X})$. Therefore, in typical application scenarios, the ssMRCD is affine equivariant whereas the original MRCD is not. Anyhow, in the case of global high-dimensionality, that is $n \leq p+1$, affine equivariance can only be achieved by the non-robust sample mean and covariance (Tyler 2010). Neither the MRCD nor the ssMRCD can then be affine equivariant.

Another important property of robust estimators is the finite sample breakdown point, which is defined as the minimal fraction of points that need to be exchanged in order to make the estimators useless. Before considering the spatially smoothed MRCD we have to derive the breakdown point of the original MRCD without prior scaling of the observations, from now on called *raw MRCD*. The breakdown point of a location estimator $\hat{\boldsymbol{\mu}}_n$ is formally defined as

$$\epsilon_n^*(\hat{\boldsymbol{\mu}}_n; \boldsymbol{X}_n) = \frac{1}{n} \min\{m : \sup ||\hat{\boldsymbol{\mu}}_n(\boldsymbol{X}_{n,m}) - \hat{\boldsymbol{\mu}}_n(\boldsymbol{X}_n)|| = +\infty\},$$

where $\boldsymbol{X}_{n,m}$ is the data matrix $\boldsymbol{X}_n$ with m-many observations exchanged with arbitrary values (Maronna, Martin, and Yohai 2006).

For the covariance estimate $\hat{\boldsymbol{\Sigma}}_n$ the finite sample breakdown point is defined as

$$\epsilon_n^*(\hat{\boldsymbol{\Sigma}}_n; \boldsymbol{X}_n) = \frac{1}{n} \min\{m : \sup_j \max | \ln(\lambda_j(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_{n,m}))) - \ln(\lambda_j(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_n)))| = +\infty\},$$

with $\lambda_1(\boldsymbol{\Sigma}), \ldots, \lambda_p(\boldsymbol{\Sigma})$ denoting the eigenvalues of a matrix $\boldsymbol{\Sigma}$ in decreasing order. Since the eigenvalues are sorted, we only have to consider the biggest eigenvalue $\lambda_1(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_{n,m}))$ which might explode when exchanging observations with arbitrary values (*explosion breakdown point*) and the smallest eigenvalue $\lambda_p(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_{n,m}))$ which might become zero (*implosion breakdown point*) and thus implies singularity (Maronna, Martin, and Yohai 2006).

Theorem 2. Consider the raw MRCD estimator with fixed $\rho > 0$, regular and fixed $\boldsymbol{T} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$ and the data matrix $\boldsymbol{X}_n$. Then, the following statements hold:

a. The MRCD location estimator $\hat{\boldsymbol{\mu}}_n$ has the finite sample breakdown point $\min(h, n-h+1)/n$.
b. The MRCD covariance estimator $\hat{\boldsymbol{\Sigma}}_n$ has the finite sample explosion breakdown point $(n-h+1)/n$.
c. The MRCD covariance estimator $\hat{\boldsymbol{\Sigma}}_n$ has the finite sample implosion breakdown point 1.

Proof. The proof is given in the supplement. □

Regarding the finite sample breakdown point of location and covariance estimators with multiple estimators let us define the finite sample breakdown point ϵ_n^* as the minimal fraction of points of the same arbitrary neighborhood that need to be exchanged in order to make at least one of the estimators useless. For the location estimators $\hat{\boldsymbol{\mu}}_{SSM,n,i}$, $i = 1, \ldots, N$, the formal definition is

$$\epsilon_n^*((\hat{\boldsymbol{\mu}}_{SSM,n,i})_{i=1}^N; \boldsymbol{X}_n) = \min_{i=1,\ldots,N} \frac{1}{n_i} \min\{m : \sup ||\hat{\boldsymbol{\mu}}_{SSM,n,i}(\boldsymbol{X}^i_{n,m}) - \hat{\boldsymbol{\mu}}_{SSM,n,i}(\boldsymbol{X}_n)|| = +\infty\},$$

where $\boldsymbol{X}^i_{n,m}$ is the matrix $\boldsymbol{X}_n$ with m-many observations of neighborhood a_i exchanged with arbitrary values.

For the breakdown points of the ssMRCD estimator, we restrict the values of the parameters $\boldsymbol{W}$ and λ to exclude the following special case. In general it is possible that all entries of one column of the weight matrix are zero for (at least) one neighborhood a_i ($\omega_{ij} = 0$ for all $j = 1, \ldots, N$) meaning that neighborhood a_i does not contribute to spatial smoothing for any other neighborhood. If additionally $\lambda = 1$, the observations and the MRCD-based covariance matrix of neighborhood a_i (see (2)) do also not contribute to the estimation of the ssMRCD-covariance estimate of neighborhood a_i itself. Thus, the values of the observations in neighborhood a_i do not affect the estimation in any way implying that it is not sensible to include this neighborhood in the calculation of the breakdown point. Therefore, in the following theorems of this section we assume that there is at least one nonzero entry per column of $\boldsymbol{W}$ if $\lambda = 1$.

Theorem 3. The location estimators $(\hat{\boldsymbol{\mu}}_{SSM,n,i})_{i=1}^N$ of the spatially smoothed MRCD have the finite sample breakdown point

$$\epsilon_n^*((\hat{\boldsymbol{\mu}}_{SSM,n,i})_{i=1}^N; \boldsymbol{X}_n) = \min_{i=1,\ldots,N} \min(n_i - h_i + 1, h_i)/n_i.$$

Proof. The proof is given in the supplement. □

For the covariance estimates $\hat{\boldsymbol{\Sigma}}_{SSM,n,i}$, $i = 1, \ldots, N$, the finite sample breakdown point is defined accordingly,

$$\epsilon_n^*\left(\left(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}\right)_{i=1}^N; \boldsymbol{X}_n\right) = \min_{i=1,\ldots,N} \frac{1}{n_i} \min\{m : \sup_j \max | \ln(\lambda_j(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}(\boldsymbol{X}^i_{n,m}))) - \ln(\lambda_j(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}(\boldsymbol{X}_n)))| = +\infty\}.$$

Again, we can differentiate between explosion and implosion breakdown point.

Theorem 4. Given a fixed and regular target matrix $\boldsymbol{T}$, the finite sample implosion breakdown point of the spatially smoothed MRCD covariance estimators $\left(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}\right)_{i=1}^N$ is equal to

$$\epsilon_n^*\left(\left(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}\right)_{i=1}^N; \boldsymbol{X}_n\right) = 1.$$

The finite sample explosion breakdown point is

$$\epsilon_n^*\left(\left(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}\right)_{i=1}^N; \boldsymbol{X}_n\right) = \min_{i=1,\ldots,N} (n_i - h_i + 1)/n_i.$$

Proof. The proof is given in the supplement. □

Note that for all the breakdown properties of the original and the spatially smoothed MRCD, the target matrix $\boldsymbol{T}$ is assumed to be regular and fixed. In applications the target matrix would be some covariance estimator $\boldsymbol{T}(\boldsymbol{X})$ with its own breakdown point. Then, the explosion breakdown point of the ssMRCD with estimated target matrix is the minimum of the two breakdown points, and the implosion breakdown point is 1 (see online appendix, after proof of Theorem 4).

2.3. Local Outlier Detection

The final step for detecting outliers is to decide how the spatially smoothed covariances available for neighborhoods a_i, $i = 1, \ldots, N$, will be linked to the pairwise MD.

The method is based on Ernst and Haesbroeck (2016). In order to compare each observation $\boldsymbol{x}$ with its local neighbors we need a second neighborhood structure that provides spatially close neighbors in contrast to the structural neighborhoods a_i that are used for the smoothed covariance estimation. Thus, we select some $k \in \mathbb{N}$ and calculate the spatial k nearest neighbors, $b_k(\boldsymbol{x})$, see also Definition (1), where a typical value might be $k = 10$. However, when applying the method, the spatial structure of the dataset should also be considered.

For each observation $\boldsymbol{x} \in a_i$, the *next distance* is defined as

$$d(\boldsymbol{x}) = min_{\boldsymbol{y} \in b_k(\boldsymbol{x})} \left[(\boldsymbol{x} - \boldsymbol{y})' \hat{\boldsymbol{\Sigma}}_{SSM,i}^{-1} (\boldsymbol{x} - \boldsymbol{y}) \right]^{1/2},$$

which is the minimum of all pairwise MDs based on the covariance matrix $\hat{\boldsymbol{\Sigma}}_{SSM,i}$ with all observations in the spatial k nearest neighborhood $b_k(\boldsymbol{x})$. The neighborhood $b_k(\boldsymbol{x})$ is not necessarily a subset of a_i. However, due to the spatial smoothing of the covariance matrices and the spatial correlation of regular observations, an observation $\boldsymbol{y}$ spatially close to $\boldsymbol{x}$ should be similar enough to not be classified as outlier even if the covariance matrix of another but close neighborhood a_i is used. In the case of a strong difference between $\boldsymbol{x}$ and $\boldsymbol{y}$, the observation $\boldsymbol{y}$ would still be classified as outlying.

The next distance $d(\boldsymbol{x})$ is used as a measure of outlyingness. If the next distance is high, none of the observations in the spatial k nearest neighborhood is similar to the observation $\boldsymbol{x}$, which means that $\boldsymbol{x}$ would be flagged as a local outlier. As a notion what values for the next distance are considered as high, a nonparametric cut-off value can be used based on the upper fence of the adjusted boxplot (Hubert and Vandervieren 2008) using all next distances from all neighborhoods a_i, $i = 1, \ldots, N$. Observations above the cut-off value are considered to be local outliers.

Possible further extensions like the restriction to homogeneous neighborhoods as suggested in Ernst and Haesbroeck (2016) are not included but are interest of future research.

3. Algorithm and C-step

For computing the spatially smoothed MRCD location and covariance estimators we need to minimize the objective function (3). However, since the number of possible combinations of subsets is comparable with the MCD or the MRCD, it is in general not feasible to just calculate the value of the objective function for all these combinations and select the best one. Instead, the strategy of C-steps (concentration steps) introduced for the MCD estimator by Rousseeuw and Driessen (1999) will be adapted to this problem setting. Given an index set H_1 corresponding to h observations of the data matrix $\boldsymbol{X}$, the C-step chooses the subsequent subset H_2 where the h observations with the smallest Mahalanobis distances, based on the arithmetic mean and sample covariance matrix of the observations from H_1, are taken. Rousseeuw and Driessen (1999) have shown that this procedure converges to a local minimum. The C-step idea has also been extended for the MRCD (Boudt et al. 2020), and we will now adapt the generalized C-step to our setting.

Theorem 5. For each $j = 1, \ldots, N$, let $\rho_j \in (0, 1)$ and H_j^0 be any starting subset of a_j of respective size $h_j \in (n_j/2, n_j)$. Let α_j be the corresponding fraction of the observations used, $\alpha_j = h_j/n_j$. Let $\mathcal{H}^0 = (H_1^0, \ldots, H_N^0)$ be the combination of the subsets and $\lambda \in [0, 1)$ fixed. The target matrix $\boldsymbol{T}(\boldsymbol{X})$ is assumed to be positive definite, symmetric and fixed, and $\boldsymbol{K}_j(\mathcal{H}^0)$ is defined as in (2) with $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$. Calculate the sample mean for each neighborhood $a_j, j = 1, \ldots, N$, based on the respective subset, $\boldsymbol{m}_j(\mathcal{H}^0) = \frac{1}{h_j} \sum_{k \in H_j^0} \boldsymbol{x}_k$.

Fix neighborhood a_i. For $\boldsymbol{x} \in a_i$, let the MD-based measure with the subset given by $\mathcal{H}^0$ be defined as

$$\begin{aligned} d(\boldsymbol{x}; \mathcal{H}^0) = {} & (\boldsymbol{x} - \boldsymbol{m}_i(\mathcal{H}^0))' \\ & \times \left[(1 - \lambda) \boldsymbol{K}_i(\mathcal{H}^0) + \lambda \sum_{j=1, j \neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^0) \right]^{-1} \\ & \times (\boldsymbol{x} - \boldsymbol{m}_i(\mathcal{H}^0)). \end{aligned}$$

For a new subset H_i^1 of a_i of size h_i with

$$\sum_{k \in H_i^1} d(\boldsymbol{x}_k; \mathcal{H}^0) \leq \sum_{k \in H_i^0} d(\boldsymbol{x}_k; \mathcal{H}^0),$$

denote $\tilde{\mathcal{H}} = (H_1^0, \ldots, H_i^1, \ldots, H_N^0)$ (note that $\boldsymbol{K}_j(\tilde{\mathcal{H}}) = \boldsymbol{K}_j(\mathcal{H}^0)$ for $j \neq i$). Then,

$$\begin{aligned} & \det \left((1 - \lambda) \boldsymbol{K}_i(\tilde{\mathcal{H}}) + \lambda \sum_{j=1, j \neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\tilde{\mathcal{H}}) \right) \\ & \leq \det \left((1 - \lambda) \boldsymbol{K}_i(\mathcal{H}^0) + \lambda \sum_{j=1, j \neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^0) \right) \end{aligned}$$

with equality if and only if $\boldsymbol{K}_i(\tilde{\mathcal{H}}) = \boldsymbol{K}_i(\mathcal{H}^0)$ and $\boldsymbol{m}_i(\tilde{\mathcal{H}}) = \boldsymbol{m}_i(\mathcal{H}^0)$.

Proof. The proof is given in the supplement. □

The C-step theorem states that the objective function will decrease with every step as long as the other covariance estimators stay fixed. In the implemented algorithm described below, this will in general not be the case. However, the theorem and its proof should be sufficient to motivate and provide a reason for the algorithm proposed in the following.

The algorithm makes use of the C-step to solve the minimization problem based on Boudt et al. (2020). Suppose that we can

estimate the target matrix $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ by the affine equivariant MCD estimator, then we also obtain affine equivariance for the spatially smoothed MRCD. Thus, we can ignore the scaling step in Boudt et al. (2020) and reduce the number of parameter estimations. Using an eigen-decomposition of $\boldsymbol{T} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$, with $\boldsymbol{Q}$ containing the eigenvectors as columns, and $\boldsymbol{\Lambda}$ being the diagonal matrix of positive eigenvalues, we transform the observations,

$$\boldsymbol{z}_i = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}'\boldsymbol{x}_i, \tag{6}$$

for $i = 1, \ldots, n$. Thus, in the next steps we can use $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)'$ as data matrix and $\boldsymbol{T} = \boldsymbol{I}_p$ as target matrix, which numerically simplifies the data-driven selection procedure for ρ_j.

In order to start the iteration process by making use of the C-steps, we need starting values, which should be robust and regular covariance estimates for each neighborhood. As proposed for the original MRCD estimator, we will also use the deterministic MCD algorithm of Hubert, Rousseeuw, and Verdonck (2012) for each neighborhood separately. This approach results in six estimates for location and scatter for each neighborhood, which show especially good convergence properties. Furthermore, for each neighborhood a_j, the value of ρ_j is calculated using the data-driven selection procedure based on a maximal condition number according to steps 3.2 to 3.4 from Boudt et al. (2020). The set of six deterministic starting values for each neighborhood is restricted to those with a sufficiently small condition number (for more details see step 3.5. in Boudt et al. 2020).

One new issue that arises is the number of possible combinations of initial subsets: for N neighborhoods we end up with up to 6^N subset combinations as possible starting values. Since a high number of starting values might not be essential for a good approximation, we will consider an upper limit of $6N$ starting values in the following, and they will be selected at random out of the possible combinations that are left after the ρ-selection step. While accuracy can be increased with additional starting values, the restriction to $6N$ leads to computational feasibility for the algorithm and still provides reliable estimates as can be seen in the performance of local outlier detection in Section 4.

Suppose now that we start the procedure with an initial subset $\mathcal{H}^0 = (H_1^0, \ldots, H_N^0)$, then we apply the C-step for each neighborhood a_i and obtain a new subset H_1^i. The combination of these subsets $\mathcal{H}^1 = (H_1^1, \ldots, H_N^1)$ is used as starting point for the next iteration step, etc. After there is no change in the subsets, the iteration process stops (see also Figure 8 in the online supplements). After applying the C-step iterations for all starting values, we choose the subset combination with the smallest objective function value as the final subset combination $\mathcal{H}^* = (H_1^*, \ldots, H_N^*)$.

Although Theorem 5 is not proving that the objective function decreases with every step due to the additional covariance matrices being adapted separately for each neighborhood after each iteration step, simulation results show that the algorithm provides stable results and good monotonic behavior in most cases (see Section 4).

After receiving the final combination of subsets $\mathcal{H}^*$ for each neighborhood, the matrices $\boldsymbol{K}_i^Z$ are back-transformed to

$$\boldsymbol{K}_i^* = \boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}\left[\boldsymbol{K}_i^Z(\mathcal{H}^*)\right]\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}'. \tag{7}$$

Algorithm 1: Algorithm for the spatially smoothed MRCD estimator.

1 **Step 1.1:** Calculation of target matrix $\boldsymbol{T}$ using the MCD estimator on $\boldsymbol{X}$;
2 **Step 1.2:** Eigen-decomposition of $\boldsymbol{T} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$ and transform observations according to (6);
3 **Step 2:** Initialization step according to steps 3.1 to 3.5 (without C-step) from Boudt et al. (2020) for each neighborhood;
4 **for** $i = 1, \ldots, N$ **do**
5 Fix neighborhood a_i;
6 Get 6 initial deterministic sets of h_i observations from a_i according to Hubert, Rousseeuw, and Verdonck (2012);
7 Calculate 6 initial covariance matrices and mean estimates;
8 Select neighborhood specific ρ_i via data driven approach;
9 Filter subset of initial starting estimates according to condition number (step 3.5 from Boudt et al. 2020);
10 **end**
11 Select set of initial h-set combinations as starting values at random;
12 **Step 3.1:** C-step: For each initial combination of subsets iterate until convergence;
13 **Step 3.2:** Select best combination of subset based on objective function value;
14 **Step 4:** Calculate $\boldsymbol{K}_i(\mathcal{H}^*)$ $\forall i$ and use (7) and (8) to get final estimates;

The covariance estimate for neighborhood a_i is then

$$\hat{\boldsymbol{\Sigma}}_{SSM,i} = (1-\lambda)\boldsymbol{K}_i^* + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij}\boldsymbol{K}_j^*, \tag{8}$$

and the location estimate is the arithmetic mean of the optimal subset $\boldsymbol{X}_{\mathcal{H}_i^*}$. The detailed numerical procedure is summarized as pseudo-code in Algorithm 1.

Regarding the tuning parameter λ there is no standard procedure to get a good value that is similarly automated like the calculation procedure for the regularization parameters ρ_i. However, there are multiple possibilities demonstrated in Section 5 that can be used to choose λ in applications.

4. Numerical Simulations

In order to test the new method, two simulation setups are constructed which also incorporate the neighborhood structures necessary for the covariance estimation and outlier detection. For both setups we simulate covariance matrices which depend on a parameter δ, denoted by $\boldsymbol{\Sigma}(\delta)$, where the entries are defined as $(\boldsymbol{\Sigma}(\delta))_{jk} = \delta^{(j-k)}$, for $j, k \in \{1, \ldots, p\}$, leading to positive definiteness and symmetry.

Setup 1: The first setup is inspired by the original idea of covariance matrices smoothly transforming over space. We start by setting up the (two-dimensional) coordinates (s_i^1, s_i^2) of the observations $\boldsymbol{x}_i$ with $n_{sim} = 41$ observations per coordinate axis,

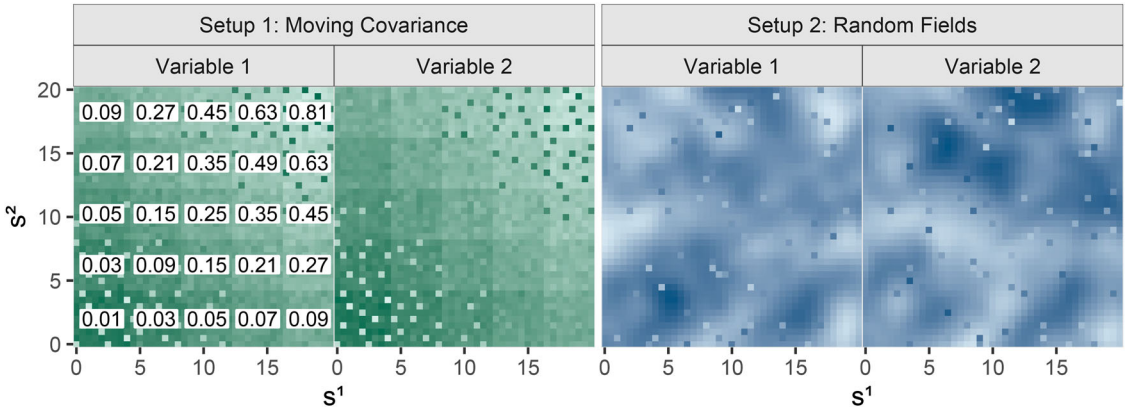

Figure 1. Simulation scenarios with $p = 2$ and a 5% contamination rate. On the left hand side the simulation setup 1 is presented with contamination achieved through the swapping process described in Ernst and Haesbroeck (2016), $N_{sim} = 25$ and $n_{sim} = 41$. The values printed on the left-most panel are corresponding to the parameter δ_{lm}. On the right hand side, setup 2 with $\nu = 3$ is shown with completely random swapping.

evenly spread between 0 and 20, resulting in $n = n_{sim}^2 = 1681$ data points in total.

For N_{sim} many areas a_{lm}, $l, m \in \{1, \ldots, \sqrt{N_{sim}}\}$, of similar observations we construct a second spatial grid with each cell consisting of n_{sim}/N_{sim} observations on average. The borders of the areas for the first coordinate are defined as $b_l^1 = l\frac{20}{\sqrt{N_{sim}}}$ for $l = 0, \ldots, \sqrt{N_{sim}}$. Analogously, the borders for the second coordinate are defined as $b_m^2 = m\frac{20}{\sqrt{N_{sim}}}$ for $m = 0, \ldots, \sqrt{N_{sim}}$, leading the an evenly spaced grid. Thus, the area a_{lm} consists of observations $\{\boldsymbol{x}_i : s_i^1 \in (b_{l-1}^1, b_l^1] \cap s_i^2 \in (b_{m-1}^2, b_m^2]\}$ for $l, m \neq 0$. For either l or m equal to 1, the left edge of the interval (which would be zero) is included. The coordinate centers of the area a_{lm} are defined as $c_{lm}^1 = \frac{b_{l-1}^1 + b_l^1}{2}$ and $c_{lm}^2 = \frac{b_{m-1}^2 + b_m^2}{2}$.

The observed values of observations in area a_{lm} are then randomly drawn from a p-dimensional normal distribution $\mathcal{N}\left(\boldsymbol{\mu}_{lm}, \boldsymbol{\Sigma}\left(\delta_{lm}\right)\right)$, where $\boldsymbol{\mu}_{lm} := ((c_{lm}^1 + c_{lm}^2)/2, \ldots, (c_{lm}^1 + c_{lm}^2)/2) \in \mathbb{R}^{p\times 1}$, thus, having entry values between 0 and 20. For the covariance matrix $\boldsymbol{\Sigma}\left(\delta_{lm}\right)$ for areas a_{lm} we use the structure described above with parameter δ_{lm} defined as

$$\delta_{lm} = \left(0.1 + l\frac{0.9 - 0.1}{\sqrt{N_{sim}}}\right)\left(0.1 + m\frac{0.9 - 0.1}{\sqrt{N_{sim}}}\right) \in [0.01, 0.81],$$

increasing smoothly from the left bottom to the right upper corner. A simulated dataset with $p = 2$ is presented in Figure 1 (left) where the grid structure and the change of the mean are clearly visible. The resulting δ_{lm} for each area is shown as well.

Setup 2: To get a more flexible simulation setup, a random field, specifically the parsimonious multivariate Matérn model (see Gneiting, Kleiber, and Schlather 2010; Ernst and Haesbroeck 2016), is used as suggested in Ernst and Haesbroeck (2016) and Harris et al. (2013). Instead of the constructed matrices used in Ernst and Haesbroeck (2016) and Harris et al. (2013) we again choose a matrix structure $\boldsymbol{\Sigma}\left(\delta\right)$ as described above with $\delta = 0.7$, since it can be extended for higher dimensions. For δ being set to 0.7 we get a range of high to low correlations reflecting different relationships between variables. The spatial smoothness is assumed to be the same for all variables, and it is regulated by one smoothness parameter ν, which is taking values in $\{0.5, 1.5, 3\}$. A higher ν leads to more spatial smoothness and in general more distinct outliers after contamination. The spatial scale parameter a of the Matérn model is set to one. A grid structure for the coordinates of the observations with values ranging from 0 to 20 with grid size 0.5 is imposed, leading to $41^2 = 1681$ observations overall, similar to the standard setting in setup 1.

Lastly, contamination with outliers is achieved by swapping coordinates of observations with each other. Filzmoser, Ruiz-Gazen, and Thomas-Agnan (2013) exchange observations that are completely randomly chosen, whereas Ernst and Haesbroeck (2016) propose to swap the most extreme observations regarding the first score of the global robust principal components. In order to avoid the problem of exchanging whole areas of observations with each other due to high spatial correlation, once observations are swapped, their 15 closest neighbors are removed from the swapping process. This leads to a clear distinction of outlying observations without the possibility of other outliers being close (see also Figure 1). Thus, swapping according to Ernst and Haesbroeck (2016) should in general result in a better performance for all considered methods. Both swapping approaches in both setups will be analyzed with a varying contamination level β between 1% and 15%.

For the ssMRCD covariance estimation we impose a grid based neighborhood structure. Similar to the description of setup 1 we use evenly spaced borders and assign the observations to neighborhoods n_i, for $i = 1, \ldots, N$. Thus, the case $N = N_{sim}$ in setup 1 depicts a perfect match of the neighborhoods selected for the ssMRCD and the real underlying covariance structure. For $N_{sim} > N$ the ssMRCD uses less neighborhoods leading to more smoothness of the covariance estimation. The weighting matrix $\boldsymbol{W}$ is based on the inverse (Euclidean) distance of the centers, that are defined equivalently to setup 1.

We will focus on the suitability of the ssMRCD covariances for local outlier detection and refer to the online supplement for additional analysis regarding computational efficiency and convergence behavior of the algorithm described in Section 3. We compare its performance to the local outlier detection methods of Filzmoser, Ruiz-Gazen, and Thomas-Agnan (2013) (F), Ernst and Haesbroeck (2016)(EH) and the local outlier factor methodology of Breunig et al. (2000), canonically adapted to spatial neighborhoods as described in Schubert, Zimek, and

Kriegel (2012) (LOF). Both simulation setups vary in the parameters p, ν and N_{sim}, respectively, and both swapping processes are used, each combination is simulated 100 times. All methods considered compare each observation to k many of its neighboring observations which we will assume to be equal to 10 for all methods. For the ssMRCD we will assume the default values ($\lambda = 0.5$, $N = 25$, and $\boldsymbol{W}$ based on inverse-distances) arrived from the parameter sensitivity analysis included in the online supplement.

The outlier classification method of Filzmoser, Ruiz-Gazen, and Thomas-Agnan (2013) has a parameter β_F which is the percentage of neighboring observations that a local outlier is allowed to be similar to. Here we use the value $\beta_F = 0.1$, as proposed in Ernst and Haesbroeck (2016), meaning that $0.1k = 1$ observation is allowed to be similar within the 10 nearest neighbors. For inliers the expected value of the isolation degree is β_F. If the actual degree of isolation is higher than the expected value, this signals local outlyingness. As cutoff for classifying an observation as local outlier we use twice the value of β_F, so 20%. This cutoff is less strict than in the simulation setup from Ernst and Haesbroeck (2016) who take a cutoff of three times the expected value, that is, 30%. Note that the methodology of Filzmoser, Ruiz-Gazen, and Thomas-Agnan (2013) mostly focuses on visual outlier detection tools, so the cutoff value chosen here might not be optimal.

For the regularized spatial detection technique by Ernst and Haesbroeck (2016), the parameter β_{EH}, which gives the fraction of the most homogeneous neighborhoods included in the outlier detection procedure, is set to one. In the simulations we are interested in all outliers, and the heterogeneity in the simulated data should be comparable for all of the observations. Thus, only considering a fraction leads to non comparable results. Moreover, the simulation results of Ernst and Haesbroeck (2016) show that over all considered setups, $\beta_{EH} = 1$ is also optimal. As regular covariance matrix estimator we use the MRCD with the default target matrix (equi-correlated target matrix) and $\alpha = 75\%$.

Last but not least, the nonparametric LOF, which calculates a local outlier factor for each observation based on a comparison of the so-called "local reachability density" with its k-nearest neighbors, needs a cutoff value. Since there is no fixed rule on how to choose a cutoff value, a local outlier factor above 1.5 determines an outlier. This value is also used in the original paper of Breunig et al. (2000).

The results for both simulation setups and the completely random switching are shown in Figures 2 and 3. For the results regarding the swapping method of Ernst and Haesbroeck (2016) shown in Figures 13 and 14 as well as a comparison of the four methods in terms of computational efficiency we refer to the online supplement.

Starting with the false positive rate (FPR), we see that the method of Ernst and Haesbroeck (2016) has some issues with classifying too many normal observations as outliers in nearly all settings. This is likely due to the very local covariance estimation which might be too strict in general leading to a strong swamping effect, especially in settings where there is no strong spatial correlation of the observed values. Interestingly, the behavior of the FPR for the method of Filzmoser, Ruiz-Gazen, and Thomas-Agnan (2013) depends on the data simulation setup. For the moving matrix scenarios, the FPR is rather high, for random fields it is very low. The LOF and the ssMRCD-based outlier detection method have reliable low FPR for all scenarios.

The outcome for the false negative rate (FNR) is quite different. While for Ernst and Haesbroeck (2016) the FNR is in many settings below all other methods, this might just be due to the high FPR. The method of Filzmoser, Ruiz-Gazen, and Thomas-Agnan (2013) has a very high FNR in most scenarios even in those with a high FPR. Regarding LOF, the simulation scenario has a strong effect on its performance. For the moving matrix setup we see a rather good performance compared to the other methods, while for random fields the FNR can hardly keep up with the other methods except the method of Filzmoser, Ruiz-Gazen, and Thomas-Agnan (2013). The ssMRCD method is somewhere in between. Although the corresponding FNRs for the moving matrix setup with completely random swapping are not overwhelming, they are still in a reasonable range for the switching method of Ernst and Haesbroeck (2016). Moreover, the FNRs for the ssMRCD outlier detection technique in the other scenarios are compellingly low.

Comparing the F1-scores of the different methods, the best method to use in general depends on the scenario. While the three selected methods seem to have pitfalls in at least one simulation scenario, the ssMRCD-based method is consistently showing reliable results and is mostly among the two best methods. Moreover, less extreme behavior occurs when it comes to the FNR and the FPR. Thus, the ssMRCD-based local outlier detection method could be the method of choice for standard outlier detection tasks. However, note that with increasing contamination it might become more beneficial to use a method that is generally more prone to flag observations as outliers, for example the technique of Ernst and Haesbroeck (2016), or to adapt the parameters of the ssMRCD to allow for more locality.

5. Example

In this section we consider a dataset provided by GeoSphere Austria (2022). It consists of monthly weather data for Austrian weather stations and is used to test and compare the different methods. The dataset contains measurements of air pressure [hPa] (p), relative air humidity [%] (rel), the monthly sum of sunshine duration [h] (s), wind velocity [m/s] (vv), air temperature in 2 meters above the ground [°C] (t), and the average daily sum of precipitation [mm] (rsum), averaged over all months in 2021, for $n = 183$ weather stations distributed all over Austria (see also Figure 4 or 7). The coordinates used for all methods are given in latitude and longitude.

We set $k = 10$ for all methods, that is, we want to compare one observation with its ten closest neighbors, independent of the methodology used. Although for the method of Ernst and Haesbroeck (2016) it is possible to remove observations with comparably high levels of heterogeneity among the neighbors, we want to include all observations, thus, setting $\beta = 1$. Even if there is increased heterogeneity among the neighbors, an observation might still be an interesting outlier clearly visible with the naked eye. Moreover, as mentioned in the prior section, the simulation results in Ernst and Haesbroeck (2016) show the

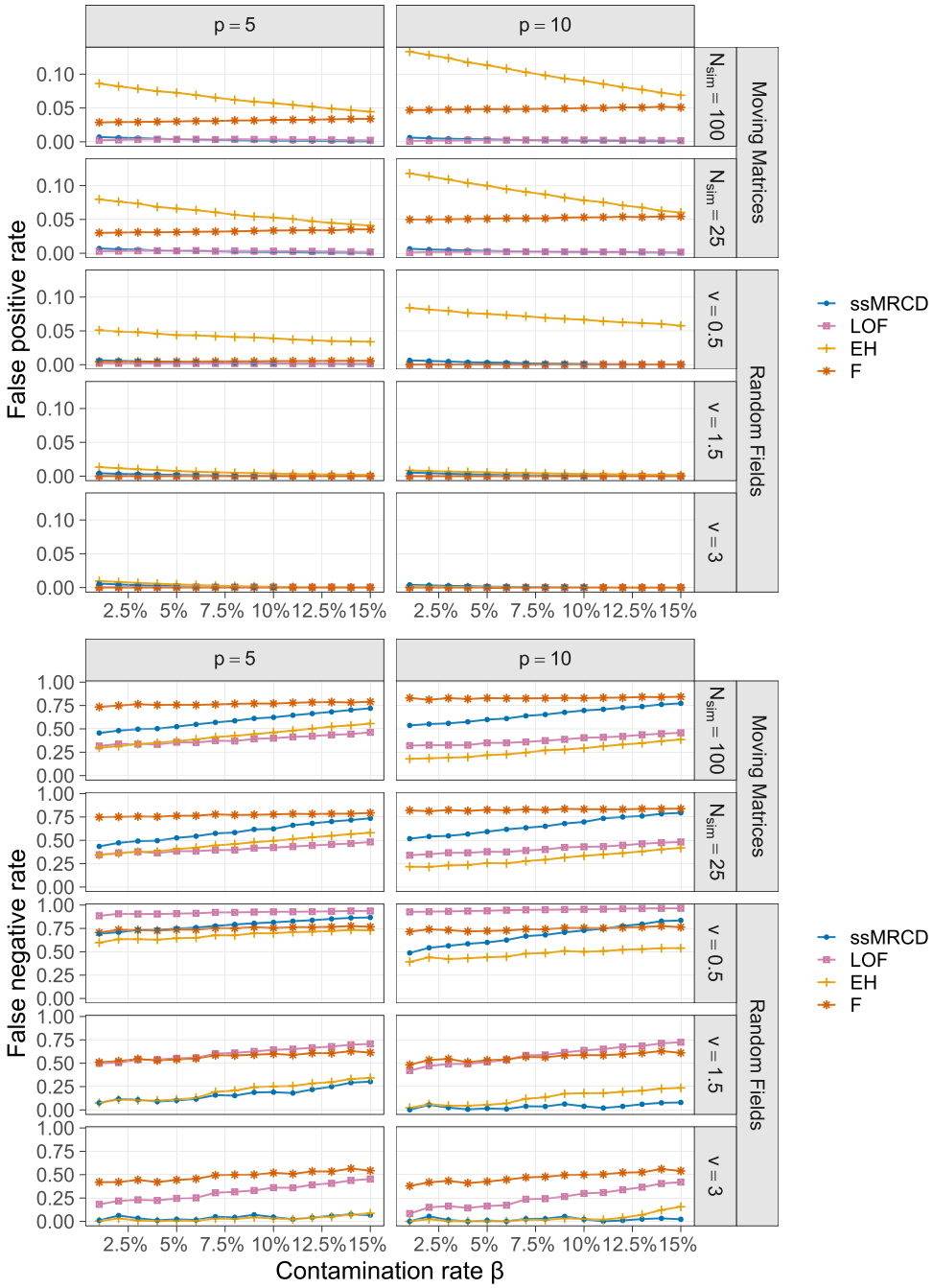

Figure 2. False positive and false negative rate for all four outlier detection methods with varying contamination levels achieved through completely random switching for different scenarios. Each point represents the mean of 100 repetitions.

best performance for high β. For the methodology of Filzmoser, Ruiz-Gazen, and Thomas-Agnan (2013) in accordance with the simulation setup we allow for one of the ten neighbors to be similar to the local outlier. The cutoff value is again set to 0.2. We use the same cutoff value of 1.5 for the LOF as in the simulations.

For the ssMRCD local outlier detection method we use a grid based neighborhood structure for the covariance estimation. Due to the Alpine landscape especially in the Western parts of Austria we aim at a rather local covariance structure, thus, choosing a rather fine grid with $N = 21$ neighborhoods and $n_i \approx 8.7$ observations per neighborhood on average. Other possible options could be based on underlying structures, for example, due to historical or political reasons, or on other classifying methods like clustering of the spatial coordinates. Furthermore, we use inverse-distance weights for the weighting matrix $\boldsymbol{W}$ between neighborhoods based on their center and select the default smoothing degree of $\lambda = 0.5$ to gain enough smoothing but still keep the locality of the fine grid structure.

As an alternative to using the default value of $\lambda = 0.5$ we can set up a simulation procedure. Assuming that the real data is uncontaminated, we can swap observations similar to the simulation studies in Section 4 and define them as local outliers. We can apply the outlier detection technique with the ssMRCD and different choices of λ (this can also be applied to other parameter settings of the ssMRCD), and then analyze the fraction of found outliers and the total number of outliers. Since

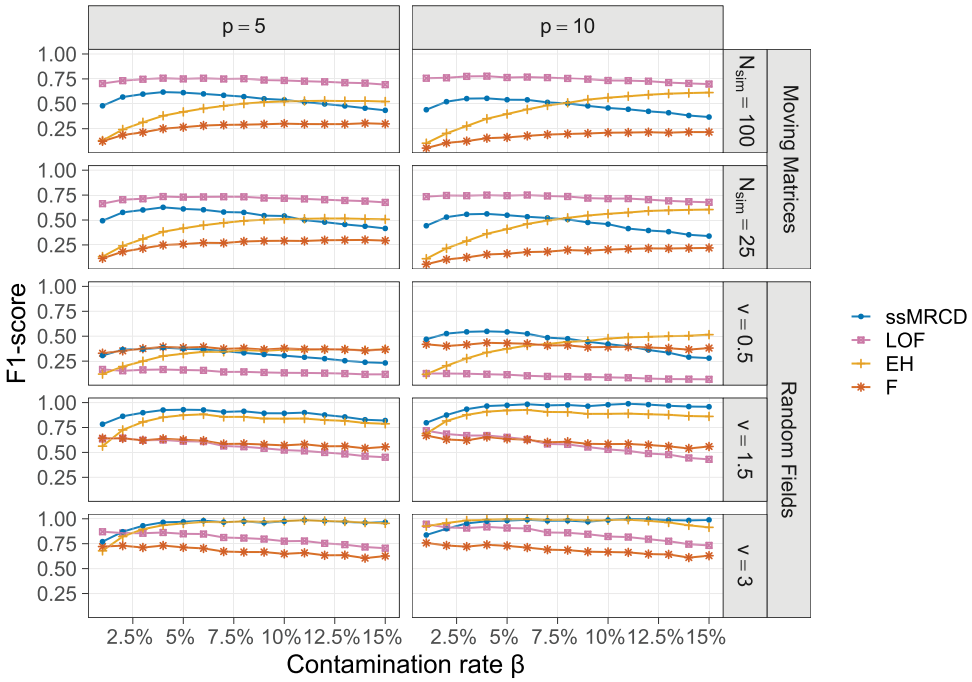

Figure 3. F1-Score for all four outlier detection methods with varying contamination levels achieved through completely random switching for different scenarios. Each point represents the mean of 100 repetitions.

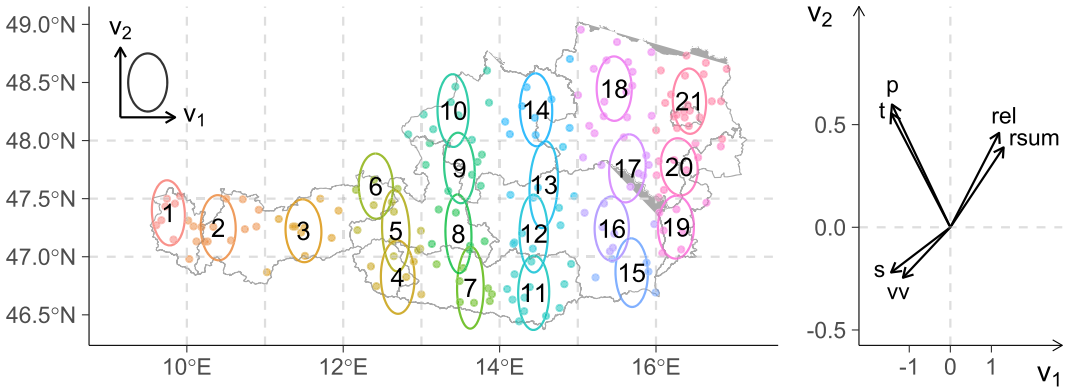

Figure 4. On the left hand side, the contours of Austria and its districts are shown together with the imposed grid structure for the ssMRCD neighborhoods. Here, singular points are assigned to a neighboring neighborhood. For each neighborhood a_i the index i is placed at the center, and the tolerance ellipses of corresponding correlation matrices based on the ssMRCD estimator are plotted along the first and second eigenvector coordinate of $\boldsymbol{T}$, which can be seen in the upper left corner as reference. The biplot of $\boldsymbol{T}$ is shown on the right hand side.

only focusing on the known FNR for the found outliers leads to an increased false positive rate, it is sensible to also take the total number of found outliers into account. A good value of λ is a tradeoff between a low FNR and a comparatively low number of found outliers overall. Interestingly, this procedure endorses the choice of $\lambda = 0.5$ in this dataset.

The resulting ssMRCD correlation matrices for each neighborhood can be seen in Figure 4. The observations and the tolerance ellipses of the ssMRCD correlation matrices are colorized according to their neighborhoods. Since we have dimension $p = 6$, we reduce the dimensionality to the first two eigenvectors v_1 and v_2 of the global MCD correlation matrix $\boldsymbol{T}$, which is displayed at the upper left corner, hoping to depict most of the relevant variance. Moreover, a biplot of $\boldsymbol{T}$ is added at the right hand side to link the correlation matrices to our weather data.

Applying all four outlier detection methods leads to 24 observations in total classified as outliers. The most outliers (21) are classified by the method of Ernst and Haesbroeck (2016), the least (3) by Filzmoser, Ruiz-Gazen, and Thomas-Agnan (2013), which is consistent with the simulation results regarding FPR and FNR, especially for the random fields setup. The distances which are used for outlier detection for each method and observation are shown in Figure 5. For further comparison of the results, the upper part of Figure 6 shows all 24 classified outliers with the corresponding ratio of distance value to cutoff value. Ratios above one are outliers. We can see that there are multiple weather stations that are classified as outliers only by the method of Ernst and Haesbroeck (2016) which lends itself to a notion consistent with the simulation results that there are some false positives among these weather stations. One example for a false positive could be panel b) in in the lower part of Figure 6. The station Feuerkogel (panel a)) was not detected by the method of Filzmoser, Ruiz-Gazen, and Thomas-Agnan (2013), also consistent to the simulation results for the

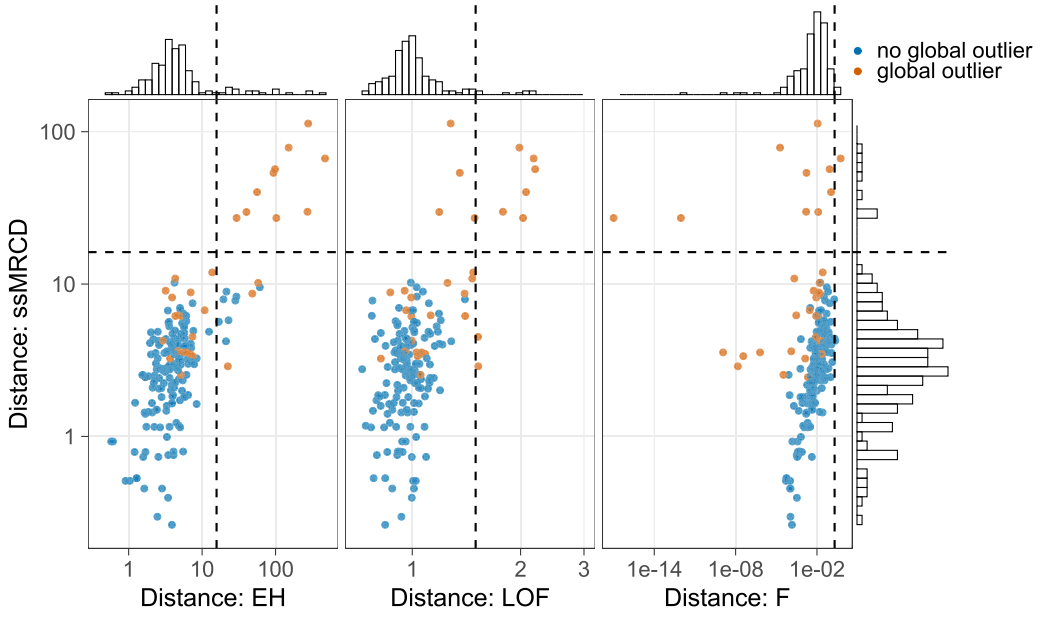

Figure 5. Distance-distance plots with the outlyingness scores of EH (next distance), of LOF (local outlier factor) and of F (isolation degree) against the next distance of the ssMRCD-based method. Observations are separated into global outliers based on the robust MD with the MCD as covariance estimator. At the margins the distribution of the different outlyingness scores are depicted by histograms.

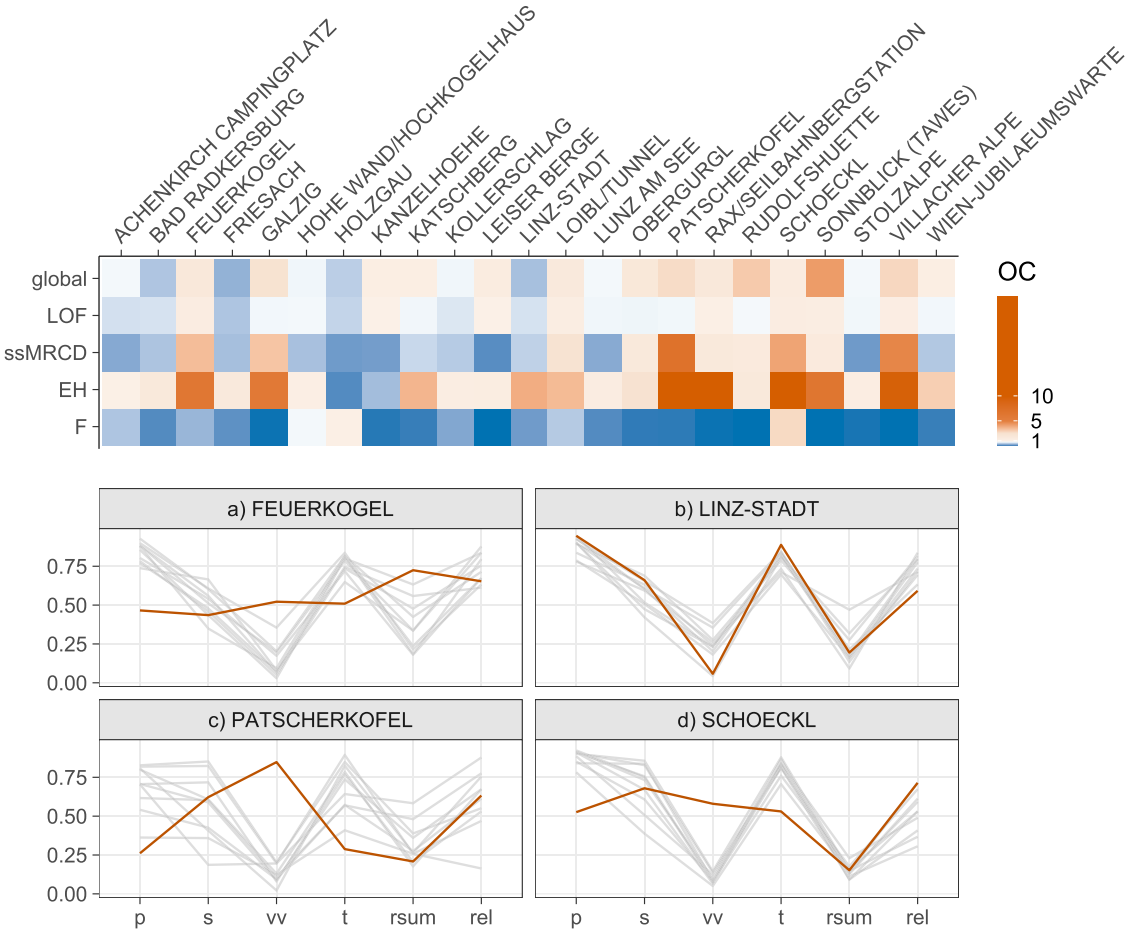

Figure 6. Upper part: Weather stations classified as outliers colorized according to their outlyingness score in relation to the cut-off value (OC) for all four methods and their global outlyingness. Lower part: in panels a)–d) four exemplary weather stations are selected to show differences on methodologies. Each variable is scaled to range [0, 1]. The values of the selected outliers are emphasized against the corresponding 10 nearest weather stations.

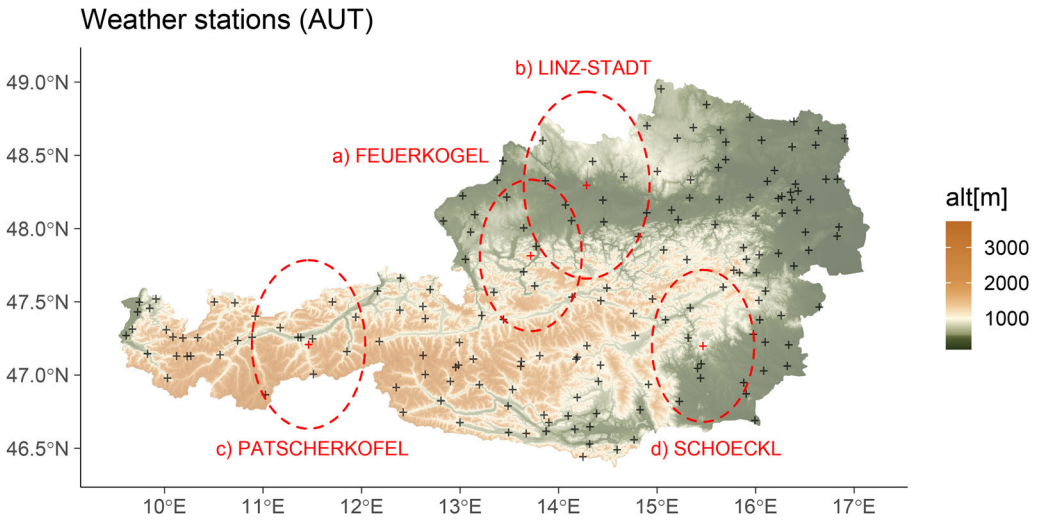

Figure 7. Altitude map of Austria with all weather stations and four selected outliers. Each dashed ellipse indicates the k nearest neighbors with whom the corresponding outlier is compared.

random fields setup and the generally high FNR. Interestingly, also LOF seems to have drawbacks and fails for example for the weather station Patscherkofel (panel c)), which was not detected as outlier. Nevertheless, the weather station Schoeckl (panel d)) was detected by all of them.

When looking at Figure 7 we can find some explanation for the local outlyingness for two of the three local outlier stations and why panel b) might not be outlying. While the stations Schoeckl and Feuerkogel are rather exposed on higher altitudes than most of the surrounding k-nearest stations which can easily lead to different patterns regarding weather, the station Linz-Stadt (panel b)) is in a rather flat area similar to its neighbors. The station Patscherkofel is already deep in the Alpine area and is surrounded by other stations in valleys but also on mountains. Although from panel c) in Figure 6 it is evident that Patscherkofel differs significantly in wind velocity, it is not clear why it differs so much also from stations with similar altitude and exposure.

6. Conclusions

In this article we enhance the limited toolbox for multivariate local outlier detection by extending the approaches of Filzmoser, Ruiz-Gazen, and Thomas-Agnan (2013), and Ernst and Haesbroeck (2016). The developed ssMRCD based on the MRCD (Boudt et al. 2020) bridges the gap between fully local and fully global covariance matrices used in the pairwise MD by exchanging the extremely local covariance matrices used in Ernst and Haesbroeck (2016) with spatially smooth estimates.

We define the ssMRCD by means of a minimization problem and prove theoretical properties of the estimator, such as equivariance and breakdown point. A heuristic is provided for the stable convergence property of the proposed algorithm under reasonable spatial changes in underlying covariance matrices. Moreover, the methods of Filzmoser, Ruiz-Gazen, and Thomas-Agnan (2013), Ernst and Haesbroeck (2016), and the ssMRCD outlier detection method are compared with the local outlier factor adapted for local outliers (Schubert, Zimek, and Kriegel 2012) regarding outlier detection performance and computational efficiency for simulated data and real world data from Austrian weather stations.

While we support the conclusion of Ernst and Haesbroeck (2016) that it is difficult to select the "best" method for outlier detection techniques, the ssMRCD-based outlier detection technique seems to be the only method providing reliable (but still improvable) results over all analyzed simulation scenarios. Note, that there might be non-analyzed scenarios where the ssMRCD-based outlier detection technique is not performing satisfactorily enough. Additionally, it is able to compete with the other methods regarding runtime even though the computation is quite complex. However, for a thorough real data analysis it is still preferable to use different outlier detection methods and compare the results in order to exploit all possible advantages of the available methods. Comparing results of multiple methodologies provides more insight in the data and significant local outliers can be classified with more reliability overall.

The ssMRCD covariance structure can be exploited also beyond local outlier detection. All covariance based methods that are sensible to adapt to spatial data can be extended by using the ssMRCD instead, for example spatial principal component analysis. A special case for the application of the ssMRCD might also be spatial data with structural breaks that need to be considered in the analysis. Finally, the presented ideas could also be transferred to a time series context, where the spatial dependency is replaced by the temporal dependency of multivariate time series, and the dependence structure could change over time. Such settings are usually quite challenging for outlier detection.

Supplementary Materials

R-Code: All code for reproducing results shown in the article. For further information we refer to the readme.txt file (supplements.zip).

Proof of theorems Proofs for Theorems 1–5.

C-step and convergence Convergence analysis of C-steps in the algorithm.

Runtime analysis Runtime of covariance estimation and local outlier detection methods.

Parameter sensitivity Sensitivity analysis for ssMRCD-based local outlier detection.

Outlier detection analysis Further analysis using the switching procedure by Ernst and Haesbroeck (2016).

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Data Availability Statement

An implementation of the methodology and the Austrian weather data is available in the R-package `ssMRCD` on CRAN.

Funding

Co-funded by the European Union (SEMACRET, Grant Agreement no. 101057741) and UKRI (UK Research and Innovation). The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Programme.

ORCID

Peter Filzmoser http://orcid.org/0000-0002-8014-4682

References

Bellino, A., Alfani, A., Riso, L. D., and Baldantoni, D. (2019), "Multivariate Spatial Analysis for the Identification of Criticalities and of the Subtended Causes in River Ecosystems," *Environmental Science and Pollution Research*, 27, 30969–30976. [2]

Boudt, K., Rousseeuw, P. J., Vanduffel, S., and Verdonck, T. (2020), "The Minimum Regularized Covariance Determinant Estimator," *Statistics and Computing*, 30, 113–128. [2,3,5,6,12]

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000), "LOF," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data - SIGMOD '00*. ACM Press. [2,7,8]

Croux, C., and Haesbroeck, G. (1999), "Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator," *Journal of Multivariate Analysis*, 71, 161–190. [1,3]

Ernst, M., and Haesbroeck, G. (2016), "Comparison of Local Outlier Detection Techniques in Spatial Multivariate Data," *Data Mining and Knowledge Discovery*, 31, 371–399. [2,5,7,8,10,12,13]

Filzmoser, P., Ruiz-Gazen, A., and Thomas-Agnan, C. (2013), "Identification of Local Multivariate Outliers," *Statistical Papers*, 55, 29–47. [1,2,7,8,9,10,12]

Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.-B., and Thirion, B. (2012), "Detecting Outliers in High-Dimensional Neuroimaging Datasets with Robust Covariance Estimators," *Medical Image Analysis*, 16, 1359–1370. [2]

GeoSphere Austria. (2022), *https://data.hub.zamg.ac.at*. [8]

Gneiting, T., Kleiber, W., and Schlather, M. (2010), "Matérn Cross-Covariance Functions for Multivariate Random Fields," *Journal of the American Statistical Association*, 105, 1167–1177. [7]

Harris, P., Brunsdon, C., Charlton, M., Juggins, S., and Clarke, A. (2013), "Multivariate Spatial Outlier Detection Using Robust Geographically Weighted Methods," *Mathematical Geosciences*, 46, 1–31. [7]

Hubert, M., Rousseeuw, P. J., and Verdonck, T. (2012), "A Deterministic Algorithm for Robust Location and Scatter," *Journal of Computational and Graphical Statistics*, 21, 618–637. [2,6]

Hubert, M., and Vandervieren, E. (2008), "An Adjusted Boxplot for Skewed Distributions," *Computational Statistics & Data Analysis*, 52, 5186–5201. [2,5]

Maronna, R., Martin, D., and Yohai, V. (2006), *Robust Statistics: Theory and Methods*, Chichester: Wiley. [4]

Maronna, R. A., and Zamar, R. H. (2002), "Robust Estimates of Location and Dispersion for High-Dimensional Datasets," *Technometrics*, 44, 307–317. [4]

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880. [1]

——— (1985), "Multivariate Estimation with High Breakdown Point," *Mathematical Statistics and Applications*, 8, 283–297. [1]

Rousseeuw, P. J., and Driessen, K. V. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, 41, 212–223. [1,5]

Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: Wiley. [3]

Schubert, E., Zimek, A., and Kriegel, H.-P. (2012), "Local Outlier Detection Reconsidered: A Generalized View on Locality with Applications to Spatial, Video, and Network Outlier Detection," *Data Mining and Knowledge Discovery*, 28, 190–237. [2,8,12]

Tyler, D. E. (2010), "A Note on Multivariate Location and Scatter Statistics for Sparse Data Sets," *Statistics & Probability Letters*, 80, 1409–1413. [4]