arXiv:1708.01550v1 [stat.ME] 4 Aug 2017

Local projections for high-dimensional outlier detection

Thomas Ortner[1]*, Peter Filzmoser[1], Maia Zaharieva[1], Sarka Brodinova[2] and
Christian Breiteneder[2]

TU Wien
[1]Institute of Statistics and Mathematical Methods in Economics
[2]Institute of Software Technology and Interactive Systems

August 7, 2017

Abstract

In this paper, we propose a novel approach for outlier detection, called local projections, which is based on concepts of Local Outlier Factor (LOF) (Breunig et al., 2000) and RobPCA (Hubert et al., 2005). By using aspects of both methods, our algorithm is robust towards noise variables and is capable of performing outlier detection in multi-group situations. We are further not reliant on a specific underlying data distribution.

稠密

For each observation of a dataset, we identify a local group of dense nearby observations, which we call a core, based on a modification of the k-nearest neighbours algorithm. By projecting the dataset onto the space spanned by those observations, two aspects are revealed. First, we can analyze the distance from an observation to the center of the core within the projection space in order to provide a measure of quality of description of the observation by the projection. Second, we consider the distance of the observation to the projection space in order to assess the suitability of the core for describing the outlyingness of the 观察的边远性 observation. These novel interpretations lead to a univariate measure of outlyingness based 单因素 on aggregations over all local projections, which outperforms LOF and RobPCA as well as other popular methods like PCOut (Filzmoser et al., 2008) and subspace-based outlier detection (Kriegel et al., 2009) in our simulation setups. Experiments in the context of real-word applications employing datasets of various dimensionality demonstrate the advantages of local projections.

1 Introduction

Classical outlier detection approaches in the field of statistics are experiencing multiple problems in the course of the latest developments in data analysis. The increasing number of variables,

[*] thomas.ortner@tuwien.ac.at

无信息的噪声变量

奇异协方差矩阵

especially non-informative noise variables, combined with complex multivariate variable distributions makes it difficult to compute classical critical values for flagging outliers. This is mainly due to singular covariance matrices, distorted distribution functions and therefore skewed critical values (e.g. Aggarwal and Yu, 2001). At the same time, outlier detection methods from the field of computer science, which do not necessarily rely on classical assumptions such as normal distribution, enjoy an increase in popularity even though their application is commonly limited due to large numbers of variables or flat data structures (more variables than observations). These observations motivated the proposed approach for outlier detection incorporating aspects from two popular methods: the Local Outlier Factor (LOF) (Breunig et al., 2000), originating in the computer science, and RobPCA, a robust principal component analysis-based (PCA) approach for outlier detection coming from the field of robust statistics (Hubert et al., 2005). The core aim of the proposed approach is to measure the outlyingness of observations avoiding any assumptions on the underlying data distribution and being able to cope with high-dimensional datasets with fewer observations than variables (flat data structures).

结合

偏差

影响很小

LOF avoids any assumptions on the data distribution by incorporating a k-nearest neighbour algorithm. Within groups of neighbours, it evaluates whether or not an observation is located in a similar density as its neighbours. Therefore, multi-group structures, skewed distributions, and other obstacles have minor impact on the method as long as there are enough observations for modelling the local behaviour. On the contrary, RobPCA uses a robust approach for modelling the majority of observations, which are assumed to be normally distributed. It uses a projection on a subspace based on this majority. In contrast to most other approaches, RobPCA does not only investigate this subspace but also the orthogonal complement, which reduces the risk of missing outliers due to the projection procedure.

重新启动

The proposed approach aims at combining these two aspects by defining projections based on the local neighbourhood of an observation where no reliable assumption about the data structure can be made and by considering the concept of the orthogonal complement similar to RobPCA. The approach of local projections is an extension of *Guided projections for analyzing the structure of high-dimensional data* (Ortner et al., 2017). We identify a subset of observations locally, describing the structure of a dataset in order to evaluate the outlyingness of other nearby observations. While guided projections create a sequence of projections by exchanging one observation by another and re-project the data onto the new selection of observations, in this work, we re-initiate the subset selection in order to cover the full data structure as good as possible with n local descriptions, where n represents the total number of observations. We discuss how outlyingness can be interpreted with regard to local projections, why the local projections are suitable for describing the outlyingness of an observation, and how to combine those projections in order to receive an overall outlyingness estimation for each observation of a dataset.

The procedure of utilizing projections linked to specific locations in the data space has the crucial advantage of avoiding any assumptions about the distribution of the analyzed data as utilized by other knn-based outlier detection methods as well (e.g. Kriegel et al., 2009). Furthermore, multi-group structures do not pose a problem due to the local investigation.

We compare our approach to related and well-established methods for measuring outlyingness. Besides RobPCA and LOF, we consider PCOut (Filzmoser et al., 2008), an outlier detection method focusing on high-dimensional data from the statistics, KNN (Campos et al., 2016), since our algorithm incorporates knn-selection similar to LOF, subspace-based outlier detection

(SOD) (Kriegel et al., 2009), a popular subspace selection method from the computer science and Outlier Detection in Arbitrary Subspaces (COP) (Kriegel et al., 2012), which follows a similar approach but has difficulties when dealing with flat data structures. Our main focus in this comparison is exploring the robustness towards an increasing number of noise variables.

The paper is structured as follows: Section 2 provides the background for a single local projection including a demonstration example. We then provide an interpretation of outlyingness with respect to a single local projection and a solution for aggregating the information based on series of local projections in Section 3. Section 4 describes all methods used in the comparison, which are then applied in two simulated settings in Section 5. Finally, in Section 6, we show the impact on three real-world data problems of varying dimensionality and group structure before we provide a brief discussion on the computation time in Section 7. We conclude with a discussion in Section 8.

2 Local projections

Let $\boldsymbol{X}$ denote a data matrix with n rows (observations) and p columns (variables) drawn from a p-dimensional random variable X, following a non-specified distribution function F_X. We explicitly consider the possibility of $p > n$ to emphasize the situation of high-dimensional low sample size data referred to as flat data, which commonly emerges in modern data analysis problems. We assume that F_X represents a mixture of multiple distributions $F_{X_1}, \dots, F_{X_q}$, where the number of sub-distributions q is unknown. The distributions are unspecified and can differ from each other. However, we assume that the distributions are continuous. Therefore, no ties are present in the data, which is a reasonable assumption especially for a high number of variables. In the case of ties, a preprocessing step, excluding ties can be applied in order to meet this assumption. An outlier in this context is any observation, which deviates from each of the groups of observations associated with the q sub-distributions.

明确的

多种分布的混合

Our approach for evaluating the outlyingness of observations is based on the concept of using robust approximations of F_X, which do not necessarily need to provide a good overall estimation of F_X on the whole support but only of the local neighborhood of each observation. Therefore, we aim at estimating the local distribution around each observation $\boldsymbol{x}_i$, for $i = 1, \dots, n$, not by all available observations but by a small subset, which is located close to $\boldsymbol{x}_i$.

近似值

We limit the number of observations included in the local description in order to avoid the influence of inhomogenity in the distribution (e.g. multimodal distributions or outliers being present in the local neighbourhood) of the underlying random variable.

不均匀性

For complex problems, especially high-dimensional problems, such approximations are difficult to find. We use projections onto groups of observations locally describing the distribution. Therefore, we start by introducing the concept of a local projection, which will then be used as one such approximation before describing a possibility of combining those local approximations. In order to provide a more practical incentive, we demonstrate the technical idea in a simulated example throughout the section.

激励

2.1 Definition of local projections

Let $\boldsymbol{y}$ denote one particular observation of the data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$, where $\boldsymbol{x}_i = (x_{i1} \ldots x_{ip})'$. For any such $\boldsymbol{y}$, we can identify its k nearest neighbours using the Euclidean distance欧式距离 between $\boldsymbol{y}$ and $\boldsymbol{x}_i$, denoted by $d(\boldsymbol{y}, \boldsymbol{x}_i)$ for all $i = 1, \ldots, n$:

$$knn(\boldsymbol{y}) = \{\boldsymbol{x}_i : d(\boldsymbol{y}, \boldsymbol{x}_i) \leq d_k\}, \tag{1}$$

where d_k is the k-smallest distance from $\boldsymbol{y}$ to any other observation in the dataset.

Using the strategy of robust estimation, we consider a subset of $\lceil \alpha \cdot k \rceil$ observations from $knn(\boldsymbol{y})$ for the description of the local distribution, where α represents a trimming 修剪 parameter describing the 比例 proportion of observations, which are assumed to be non-outlying in any knn. Here, $\lceil c \rceil$ denotes the smallest integer $\geq c$. The parameter α is usually set to 0.5 in order to avoid neighbors that are 异质 heterogeneous (e.g. due to outliers) but it can be adjusted if additional information about the specific dataset is available. By doing so, we reduce the influence of outlying observations, which would distort our estimation. The idea is to get the most dense group of $\lceil \alpha \cdot k \rceil$ observations, which we call the *core* of the projection, initiated by $\boldsymbol{y}$, not including $\boldsymbol{y}$ itself. The center of this core is defined by

$$\boldsymbol{x}_0 = arg \min_{\boldsymbol{x}_i \in knn(\boldsymbol{y})} \{d_{(\lceil \alpha \cdot k \rceil)}(\boldsymbol{x}_i)\}, \tag{2}$$

where $d_{(\lceil \alpha \cdot k \rceil)}(\boldsymbol{x}_i)$ represents the $\lceil \alpha \cdot k \rceil$-largest distance between $\boldsymbol{x}_i$ and any other observation from $knn(\boldsymbol{y})$. The observation $\boldsymbol{x}_0$ can be used to define the *core* of a local projection initiated by $\boldsymbol{y}$:

$$\begin{aligned} core(\boldsymbol{y}) = \{\boldsymbol{x}_i :& d(\boldsymbol{x}_0, \boldsymbol{x}_i) < d_{(\lceil \alpha \cdot k \rceil)}(\boldsymbol{x}_0) \wedge \\ & \boldsymbol{x}_i \in knn(\boldsymbol{y}) \wedge \boldsymbol{x}_i \neq \boldsymbol{y}\} \end{aligned} \tag{3}$$

In order to provide an 直观的 intuitive access to the proposed approach, we explain the concept of local projections for a set of simulated observations. In this example, we use 200 observations drawn from a two-dimensional normal distribution. The original observations and the procedure of selecting the $core(\boldsymbol{y})$ are visualized in Figure 1: The red observation was manually selected to initiate our local projection process and refers to $\boldsymbol{y}$. It can be exchanged by any other observation. However, in order to emphasize the necessity of the second step of our procedure, we selected an observation off the center. The blue observations are the $k = 20$ nearest neighbours of $\boldsymbol{y}$ and the filled blue circles represent the core of $\boldsymbol{y}$ using $\alpha = 0.5$. We note that the observations of $core(\boldsymbol{y})$ tend to be closer to the center of the distribution than $\boldsymbol{y}$ itself, since we can expect an increasing density towards the center of the distribution, which likely leads to more dense groups of observations.

A projection onto the space, spanned by the observations contained in $core(\boldsymbol{y})$, provides a description of the similarity between any observation and the core, which is especially of interest for $\boldsymbol{y}$ itself. Such a projection can efficiently be computed using the singular value decomposition (SVD) of the matrix of observations in $core(\boldsymbol{y})$, centered and scaled with respect to the core itself. In order to estimate the location and scale parameters 比例参数 for scaling the data 缩放数据, we can apply classical estimators on the core preserving robustness properties, since the observations have been included

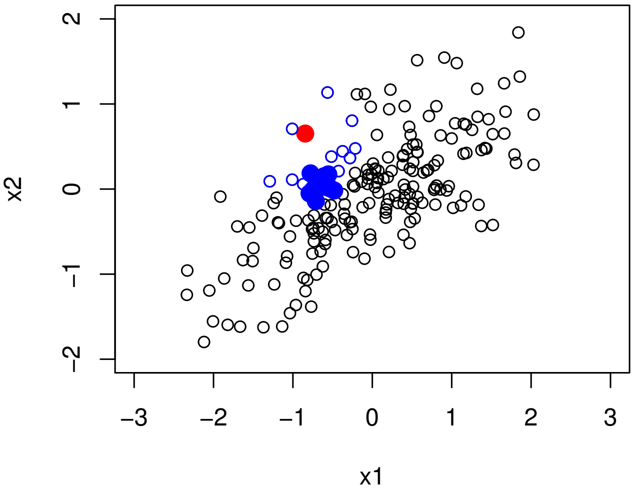


Figure 1: Visualization of the *core*-selection process. The red observation represents the initiating observation $\boldsymbol{y}$. The blue observations represent $knn(\boldsymbol{y})$ and the filled blue observations represent $core(\boldsymbol{y})$. x_0 itself is not visualized but it is known to be an element of $core(\boldsymbol{y})$.

into the core in a robust way.

$$
\begin{aligned}
\boldsymbol{X}_{core(\boldsymbol{y})} =& (\boldsymbol{x}_{\boldsymbol{y},1}, \dots, \boldsymbol{x}_{\boldsymbol{y},\lceil\alpha\cdot k\rceil})' \qquad (4)\\
& \boldsymbol{x}_{\boldsymbol{y},j} \in core(\boldsymbol{y}) \qquad \forall j \in \{1, \dots, \lceil\alpha\cdot k\rceil\}\\
\hat{\boldsymbol{\mu}}_{\boldsymbol{y}} =& \frac{1}{\lceil\alpha\cdot k\rceil} \sum_{\boldsymbol{x}_i \in core(\boldsymbol{y})} \boldsymbol{x}_i \qquad (5)\\
\hat{\boldsymbol{\sigma}}_{\boldsymbol{y}} =& \Bigg(\sqrt{Var(x_{\boldsymbol{y},11}, \dots, x_{\boldsymbol{y},\lceil\alpha\cdot k\rceil 1})}, \dots,\\
& \sqrt{Var(x_{\boldsymbol{y},1p}, \dots, x_{\boldsymbol{y},\lceil\alpha\cdot k\rceil p})}\Bigg)' \qquad (6)\\
=& (\hat{\sigma}_{\boldsymbol{y},1}, \dots, \hat{\sigma}_{\boldsymbol{y},p})',
\end{aligned}
$$

样本方差

where Var denotes the sample variance. Using $\hat{\boldsymbol{\mu}}_{\boldsymbol{y}}$, the centered observations are given by

$$
\boldsymbol{x}^c_{\boldsymbol{y}} = (x^c_{\boldsymbol{y},1}, \dots x^c_{\boldsymbol{y},p})' = \boldsymbol{x}_{\boldsymbol{y}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{y}}, \qquad (7)
$$

逐列缩放数据矩阵

which can be used to provide the centered and column-wise scaled data matrix with respect to the core of $\boldsymbol{y}$:

$$
\begin{aligned}
\tilde{\boldsymbol{X}}_{\boldsymbol{y}} = &\Bigg(\left(\frac{x^c_{\boldsymbol{y},11}}{\hat{\sigma}_{\boldsymbol{y},1}}, \dots, \frac{x^c_{\boldsymbol{y},1p}}{\hat{\sigma}_{\boldsymbol{y},p}}\right)', \dots,\\
&\left(\frac{x^c_{\boldsymbol{y}\lceil\alpha\cdot k\rceil 1}}{\hat{\sigma}_{\boldsymbol{y},1}}, \dots, \frac{x^c_{\boldsymbol{y},\lceil\alpha\cdot k\rceil p}}{\hat{\sigma}_{\boldsymbol{y},p}}\right)'\Bigg)' \qquad (8)
\end{aligned}
$$

Based on $\tilde{\boldsymbol{X}}_{\boldsymbol{y}}$, we provide a projection onto the space spanned by the observations of $core(\boldsymbol{y})$ by $\boldsymbol{V}_{\boldsymbol{y}}$ from the SVD of $\tilde{\boldsymbol{X}}_{\boldsymbol{y}}$,

$$
\tilde{\boldsymbol{X}}_{\boldsymbol{y}} = \boldsymbol{U}_{\boldsymbol{y}} \boldsymbol{D}_{\boldsymbol{y}} \boldsymbol{V}'_{\boldsymbol{y}}. \qquad (9)
$$

Any observation $\boldsymbol{x}$ can be projected onto the projection space by centering with $\hat{\boldsymbol{\mu}}_{\boldsymbol{y}}$, scaling with $\hat{\boldsymbol{\sigma}}_{\boldsymbol{y}}$, and applying the linear transformation $\boldsymbol{V}'_{\boldsymbol{y}}$. The projection of the whole dataset is given by $\tilde{\boldsymbol{X}}_{\boldsymbol{y}}\boldsymbol{V}_{\boldsymbol{y}}$. We refer to the projected observations as the representation of observations in the *core space* of $\boldsymbol{y}$. Since the dimension of the core space is limited by $\lceil \alpha \cdot k \rceil$, in any case where $p > \lceil \alpha \cdot k \rceil$ holds and $\boldsymbol{X}_{core(\boldsymbol{y})}$ is of full rank, a non-empty orthogonal complement of this core space exists. Therefore, any observation $\boldsymbol{x}$ consists of two representations, the core representation $\boldsymbol{x}_{core(\boldsymbol{y})}$ given the core space,

$$\boldsymbol{x}_{core(\boldsymbol{y})} = \boldsymbol{V}'_{\boldsymbol{y}} \left(\frac{x_1^c}{\hat{\sigma}_{\boldsymbol{y},1}}, \ldots, \frac{x_p^c}{\hat{\sigma}_{\boldsymbol{y},p}} \right)', \tag{10}$$

where $\boldsymbol{x}^c = (x_1^c, \ldots, x_p^c)'$ is computed as defined in Equation (7) and the orthogonal representation $\boldsymbol{x}_{orth(\boldsymbol{y})}$ given the orthogonal complement of the core space,

$$\boldsymbol{x}_{orth(\boldsymbol{y})} = \boldsymbol{x}^c - \boldsymbol{V}_{\boldsymbol{y}} \boldsymbol{x}_{core(\boldsymbol{y})}. \tag{11}$$

Figure 2a shows the representation of our 200 simulated observations in the core space. Note that in this special case, the orthogonal representation is constantly $\boldsymbol{0}$ due to the non-flat data structure of the core observations ($p < k$). We further see that the center of the core is now located in the center of the coordinate system. 坐标系

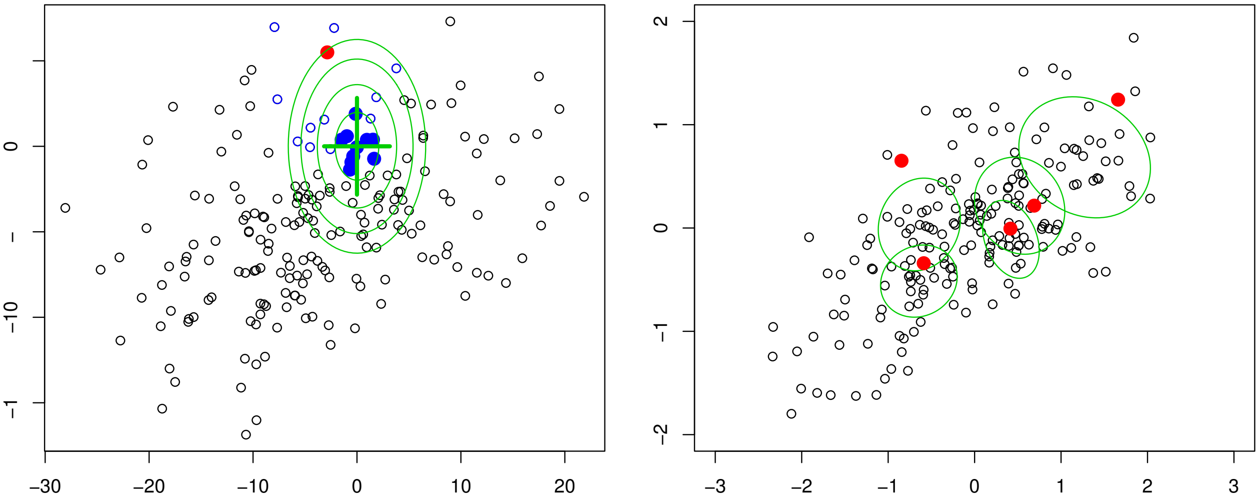
Given a large enough number of observations and a small enough dimension of the sample space, we can approximate F_X with arbitrary accuracy given any desired neighborhood. However, in practice, the quality of this approximation is limited by a finite number of observations. Therefore, it depends on various aspects like the size of d_k and $d_{\lceil \alpha \cdot k \rceil}$ and, thus, the approximation is always limited by the restrictions imposed by the properties of the dataset. Especially the behavior of the core observations will, in practice, significantly deviate from the expected distribution with increasing $d_{\lceil \alpha \cdot k \rceil}$.

In order to take this local distribution into account, it is useful to include the properties of the core observations in the core space into the distance definition within the core space. A more advantageous way to measure the deviation of core distances from the center of the core 偏差 than using Euclidean distances is the usage of Mahalanobis distances (e.g. De Maesschalck et al., 2000). For the projection space, an orthogonal basis is defined by the left eigenvectors of the SVD from Equation (9), while the singular values given by the diagonal of the matrix $\boldsymbol{D}_{\boldsymbol{y}}$ provide the standard deviation for each direction of the projection basis. Therefore, weighting the directions of the Euclidean distances with the inverse singular values directly leads to Mahalanobis distances in the core space, which take the variation of each direction into account:

$$CD_{\boldsymbol{y}}(\boldsymbol{x}) = \sqrt{\frac{\boldsymbol{x}'_{core(\boldsymbol{y})} \boldsymbol{D}_{\boldsymbol{y}}^{-1} \boldsymbol{x}_{core(\boldsymbol{y})}}{min(\lceil \alpha \cdot k \rceil - 1, p)}} \tag{12}$$

.

The computation of core distances can be derived from Figure 2a. The green cross in the center of the coordinate system refers to the (projected) left singular vectors of the SVD. We note that the scale of the two axes in Figure 2 differ appreciably. The green ellipses represent Mahalanobis distances based on the variation of the two orthogonal axes, which provide a more suitable measure for describing the distribution locally.



(a)

(b)

Figure 2: Plot (a) provides a visualization of the transformed observations from Figure 1. The red observation represents the initiating observation $\boldsymbol{y}$. The blue observations represent $knn(\boldsymbol{y})$ and the filled blue observations represent $core(\boldsymbol{y})$. The green ellipses represent the covariance structure estimated by the core observations representing the local distribution. Plot (b) uses the same representation as Figure 1 but shows the concept of multiple local projections initiated by different observations marked as red dots. Each of the core distances represented by green ellipses refers to the same constant value taking the different covariance structures of the different cores into account.

The distances of the representation in the orthogonal complement of the core cannot be rescaled as in the core space. All observations from the core, which are used to estimate the local structure, i.e. to span the core space, are fully located in the core space. Therefore, their orthogonal complement is equal to $\mathbf{0}$:

$$\boldsymbol{x}_{orth(\boldsymbol{y})} = \mathbf{0} \quad \forall \boldsymbol{x} \in core(\boldsymbol{y}) \tag{13}$$

Since no variation in the orthogonal complement is available, we cannot estimate the rescaling parameters for the orthogonal components. Therefore, we directly use the Euclidean distances in order to describe the distance from any observation $\boldsymbol{x}$ to the core space of $\boldsymbol{y}$. We will refer to this distance as orthogonal distance (OD).

$$OD_{\boldsymbol{y}}(\boldsymbol{x}) = ||\boldsymbol{x}_{orth(\boldsymbol{y})}|| \tag{14}$$

The two measures for similarity, CD and OD, are inspired by the score and the orthogonal distance of Hubert et al. (2005). In contrast to Hubert et al. (2005), we do not try to elaborate critical values for CD and OD to directly decide if an observation is an outlier. Such critical values always depend on an underlying normal distribution and on the variation of the core and the orthogonal distances of the core observations. Instead, we aim at providing multiple local projections in order to be able to estimate the degree of outlyingness for observations in any location of a data set. A core and its core distances can be defined for every observation. Therefore, a total of n projections with core and orthogonal distances are available for analyzing the data structure. Figure 2b visualizes a small number (5) of such projections in order to demonstrate how the concept works in practice. The red observations are used as the initiating observations, the green ellipses represent core distances based on each of the 5 cores. Each core distance refers to the same constant value considering the different covariance estimations of each core. We see that observations closer to the boundary of the data are described less adequately by their respective core, while other observations, close to the center of the distribution, are well described by multiple cores.

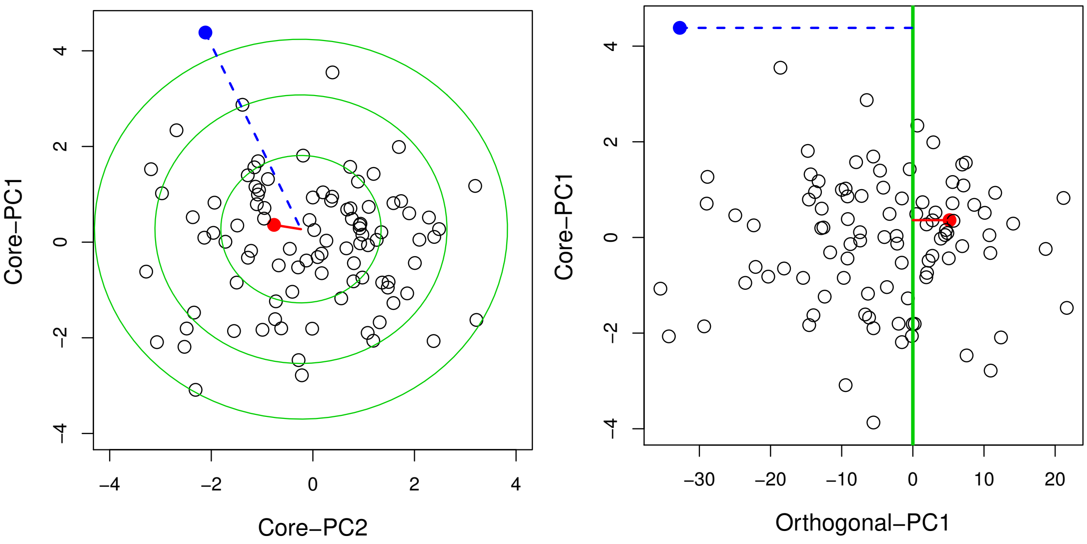
3 Interpretation and utilization of local projections

Most subspace-based outlier detection methods, including PCA-based methods such as *PCOut* (Filzmoser et al., 2008) and projection pursuit methods (e.g. Henrion et al., 2013), focus on the outlyingness of observations within a single subspace only. The risk of missing outliers due to the subspace selection by the applied method is evident as the critical information might remain outside the projection space. RobPCA (Hubert et al., 2005) is one of the few methods considering the distance to the projection space in order to monitor this risk as well.

We would like to use both aspects, distances within the projection space and to the projection space, to evaluate the outlyingness of observations as follows: The projection space itself is often used as a model, employed to measure the outlyingness of an observation. Since we are using a local knn-based description, we can not directly apply this concept as our projections are bound to a specific location defined by the cores. The core distance from the location of our projection rather describes whether an observation is close to the selected core. If this is the case, we can assume that the model of description (the projection represented by the projection space) fits the observation well. Therefore, if the observation is well-described, there should be little information remaining in the orthogonal complement leading to small orthogonal distances.

We visualize this approach in Figure 3 in two plots. Plot (a) shows the first two principal components of the core space and plot (b) the first principal component of the core and the orthogonal space respectively. In order to retrace our concept of interpreting core distances as the quality of the local description model and the core distances as a measure of outlyingness with respect to this description, we look at the two observations marked in red and blue. While the red observation is close to the center of our core as seen in plot (a), the blue one is located far off. Therefore, the blue observation is not as well described by the core as the red observation, which becomes evident when looking at the first principal component of the orthogonal complement in plot (b), where the blue observation is located far off the green line representing the projection space.

Note that this interpretation does not hold for core observations. This is due to the fact that the full information of core distances is located in the core space. With increasing p, the distance of all observations from the same group converges to a constant as shown in e.g. Filzmoser et al. (2008) for multivariate normal distributions. While this distance is completely represented in the core space for core observations, a proportion of distances from non-core observations will be represented in the orthogonal complement of the core space. Therefore, the probability of the core distance of a core observation being larger than the core distance of any other observation from



(a) (b)

Figure 3: Visualization of orthogonal and core distances for a local projection of a multivariate 100-dimensional normal distribution. Plot (a) describes the core space by its first two principal components. The measurement of the core distances is represented by the green ellipses. Plot (b) includes the orthogonal distance. The vertical green line represents the projection space.

the same group converges to 1 with increasing p:

$$\lim_{p\to\infty} P(CD_{\boldsymbol{y}}(\boldsymbol{x}_i) > CD_{\boldsymbol{y}}(\boldsymbol{z})) = 1, \quad \begin{array}{l} \boldsymbol{x}_i \in core(\boldsymbol{y}), \\ \boldsymbol{z} \notin core(\boldsymbol{y}) \end{array} \tag{15}$$

So far we considered a single projection, where we deal with a total of n projections. Let $\mathcal{X}$ denote a set of n observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. Therefore, for each initializing observation $\boldsymbol{x} \in \mathcal{X}$, the core distance $CD_{\boldsymbol{x}}$ and the orthogonal distance $OD_{\boldsymbol{x}}$ are well-defined for all observations from $\mathcal{X}$. As motivated above, we want to measure the quality of local description using the core distances and the local outlyingness using the orthogonal distances. The smaller the core distance of an observation for a specific projection is, the more relevant this projection is for the overall evaluation of the outlyingness of this observation. Therefore, we downweight the orthogonal distance based on the inverse core distances. In order to make the final outlyingness score comparable, we scale these weights by setting the sum of weights to 1 for each local projection initiating observation $\boldsymbol{y}$:

$$w_{\boldsymbol{y}}(\boldsymbol{x}) = \begin{cases} 0, & \boldsymbol{x} \in core(\boldsymbol{y}) \\ \dfrac{\frac{1}{CD_{\boldsymbol{y}}(\boldsymbol{x})} - \min\limits_{\tilde{\boldsymbol{z}} \in \mathcal{X}} \left(\frac{1}{CD_{\tilde{\boldsymbol{z}}}(\boldsymbol{x})} \right)}{\sum\limits_{\boldsymbol{z} \in \mathcal{X}} \left(\frac{1}{CD_{\boldsymbol{z}}(\boldsymbol{x})} - \min\limits_{\tilde{\boldsymbol{z}} \in \mathcal{X}} \left(\frac{1}{CD_{\tilde{\boldsymbol{z}}}(\boldsymbol{x})} \right) \right)}, & else \end{cases} \tag{16}$$

The scaled weights $w_{\boldsymbol{y}}$ make sure, that the sum of contributions by all available projections remains the same. Therefore, the sum of weighted orthogonal distances, corresponding to local outlyingness through local projections (LocOut),

$$LocOut(\boldsymbol{x}) = \sum_{\boldsymbol{y} \in \mathcal{X}} \left(w_{\boldsymbol{y}}(\boldsymbol{x}) \cdot OD_{\boldsymbol{y}}(\boldsymbol{x}) \right), \tag{17}$$

provides a useful, comparable measure of outlyingness for each observation.

Note that this concept of outlyingness is limited to high-dimensional spaces. Whenever we analyze a space where $p \leq \lceil \alpha \cdot k \rceil$ holds, the full information of all observations will be located in the core space of each local projection. Therefore, for varying core distances, the orthogonal distance will always remain zero. Thus, the weighted sum of orthogonal distances can not provide any information on outlyingness unless there is information available in the orthogonal representation of observations.

4 Evaluation setup

In order to evaluate the performance of our proposed methodology, we compare it with related algorithms, namely LOF (Breunig et al., 2000), RobPCA (Hubert et al., 2005), PCOut (Filzmoser et al., 2008), COP (Kriegel et al., 2012), KNN (Ramaswamy et al., 2000), and SOD (Kriegel et al., 2009). Each of those algorithms tries to identify outliers in the presence of noise variables. Some methods use a configuration parameter describing the dimensionality of the resulting subspace or the number of neighbours in a knn-based algorithm. In our algorithm, we use $\lceil \alpha \cdot k \rceil$ observations to create a subspace, which we employ for assessing the outlyingness of observations. In order to provide a fair comparison, the configuration parameters of each method are adjusted individually for each dataset: We systematically test different configuration values and report the best achieved performance for each method. Instead of outlier classification, we rather use each of the computed measures of outlyingness since not all methods provide cutoff values. The performance itself is reported in terms of the commonly used area under the ROC Curve (AUC) (Fawcett, 2006).

4.1 Compared methods

Local Outlier Factor (LOF) (Breunig et al., 2000) is one of the main inspirations for our approach. The similarity of observations is described using ratios of Euclidean distances to k-nearest observations. Whenever this ratio is close to 1, there is a consistent group of observations and, therefore, no outliers. As for most outlier detection methods, no explicit critical value can be provided for LOF (e.g. Campos et al., 2016; Zimek et al., 2012). In order to optimize the performance of LOF, we estimate the number of k-nearest neighbours for each evaluation. We used the R-package *Rlof* (Hu et al., 2015) for the computation of LOF.

The second main inspiration for our approach is the **RobPCA** algorithm by Hubert et al. (2005). The approach employs distances (similar to the proposed core and orthogonal distances) for describing the outlyingness of observations with respect to the majority of the underlying data. This method should work fine with one consistent majority of observations. In the presence of a multigroup structure, we would expect it to fail since the majority of data cannot be properly described with a model of a single normal distribution. RobPCA calculates two outlyingness scores, namely orthogonal and score distances[1]. RobPCA usually flags observations as outliers if

[1] For a multivariate interpretation of outlyingness based on those two scores, we refer to Pomerantsev (2008).

either the score distance or the orthogonal distance exceed a certain threshold. This threshold is based on transformations of quantiles of normal and χ^2 distributions. We use the maximum quantile of each observation for the distributions of orthogonal and score distances as a measure for outlyingness in order to stay consistent with the original outlier detection concept of RobPCA. The dimension of the subspace used for dimension reduction is dynamically adjusted. We used the R-package *rrcov* (Todorov and Filzmoser, 2009) for the computation of RobPCA.

In addition to LOF and RobPCA, we compare the proposed local projections with **PCOut** by Filzmoser et al. (2008). PCOut is an outlier detection algorithm where location and scatter outliers are identified based on robust kurtosis and biweight measures of robustly estimated principal components. The dimension of the projection space is automatically selected based on the proportion of the explained overall variance. A combined outlyingness weight is calculated during the process, which we use as an outlyingness score. The method is implemented using the R-package *mvoutlier* (Filzmoser and Gschwandtner, 2015).

Another method included in our comparison is the **subspace-based outlier detection (SOD)** by Kriegel et al. (2009). The method is looking for relevant dimensions parallel to the axis in which outliers can be identified. The identification of those subspaces is based on knn, where k is optimized in a way similar to LOF and local projections. We used the implementation of SOD in the ELKI framework (Achtert et al., 2008) for performance reasons.

All three methods, LocOut, LOF and SOD, implement knn-estimations in their respective procedures. Therefore, it is reasonable to monitor the performance of **k-nearest neighbors (KNN)**, which can be directly used for outlier detection as suggested in Ramaswamy et al. (2000). The performance is optimized over all reasonable k between 1 and the minimal number of non-outlying observations of a group which we take from the ground truth in our evaluation. We used the R-package *dbscan* for the computation of KNN.

Similar to the proposed local projections, **Outlier Detection in Arbitrary Subspaces (COP)** by Kriegel et al. (2012) locally evaluates the outlyingness of observations. The k-nearest neighbours of each observation are used to estimate the local covariance structure and robustly compute the representation of the evaluated observation in the principal component space. The last principal components are then used to measure the outlyingness, while the number of principal components to be cut off is dynamically computed. Although the initial concept looks similar to our proposed algorithm, it contains some disadvantages. The number of observations used for the knn estimation needs to be a lot larger than the number of variables. A proportion of observations to variables of three to one is suggested. Therefore, the method can not be employed for flat data structures, which represent the focus of the proposed approach for outlier analysis. While COP performed competitive for simulations with no or a very small number of noise variables, the computation of COP is not possible in flat data settings. As the non-flat settings only represent a minor fraction of the overall simulations, we did not include COP in the simulated evaluation but only in the low-dimensional real data evaluation of Section 6.

5 Simulation results

We used two simulation setups to evaluate the performance of the methods for increasing number of noise variables in order to determine their usability for high-dimensional data. We do that by starting with 50 informative variables and 0 noise variables, increasing the number of noise variables up to 5000. We use three groups of observations with 150, 150, and 100 observations. Starting from 350 noise variables, the data structure becomes flat, which we expect to lead to performance drops as the estimation of the underlying density becomes more and more problematic. Each of the three groups of observations is simulated based on a randomly rotated covariance matrix $\boldsymbol{\Sigma}^i$ as performed in Campello et al. (2015),

$$\boldsymbol{\Sigma}^i = \begin{pmatrix} \boldsymbol{\Sigma}^i_{inf} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{noise} \end{pmatrix} \qquad \boldsymbol{\Sigma}^i_{inf} = \boldsymbol{\Omega}_i \begin{pmatrix} 1 & \rho^i & \dots & \rho^i \\ \rho^i & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho^i \\ \rho^i & \dots & \rho^i & 1 \end{pmatrix} \boldsymbol{\Omega}_i', \tag{18}$$

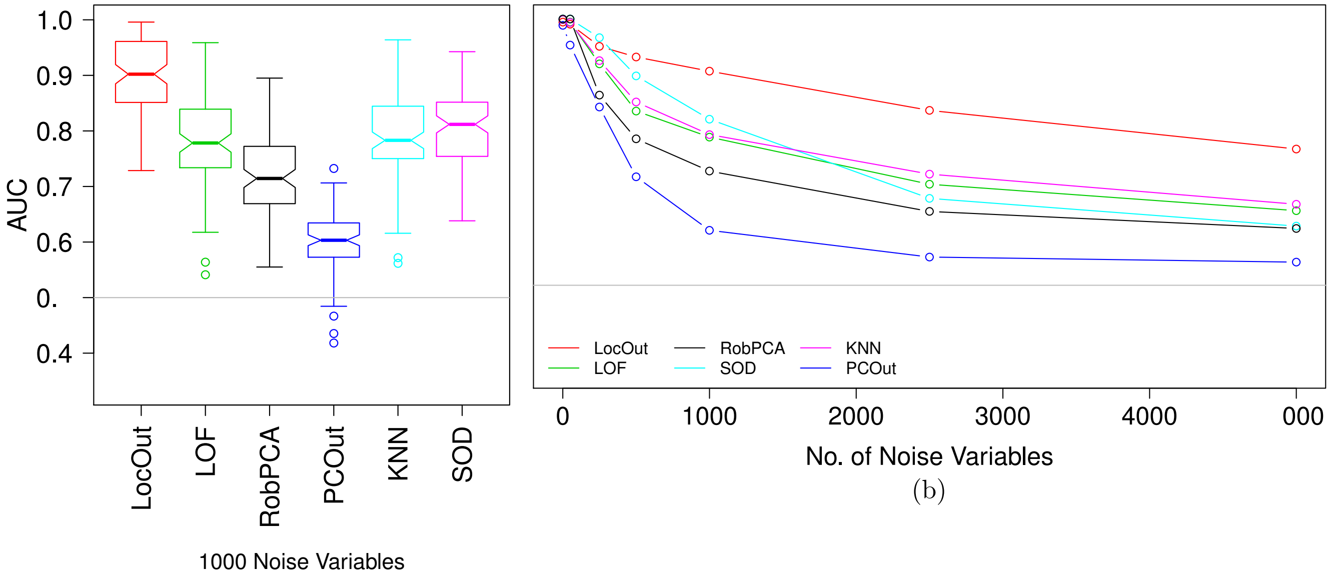
for $i = 1, 2, 3$, where $\boldsymbol{I}_{noise}$ is an identity matrix describing the covariance of uncorrelated noise variables and $\boldsymbol{\Sigma}^i_{inf}$ the covariance matrix of informative variables, which are variables containing information about the separation of present groups where ρ^i is randomly selected between 0.1 and 0.9, $\rho^i \sim U[0.1, 0.9]$. $\boldsymbol{\Omega}^i$ represents the randomly generated orthonormal rotation matrix. For our simulation setups we always consider the dimensionality of $\boldsymbol{\Sigma}^i_{inf}$ to be 50. During the simulation, we evaluate the impact of such noise variables and therefore perform the simulation for a varying number of noise variables. While the mean values of the noise variables are fixed to zero for all groups, the mean values of the informative variables are set as follows:

$$(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3) = \begin{pmatrix} \mu & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & \mu \\ \mu & 0 & 0 \\ 0 & \mu & 0 \\ \vdots & \vdots & \vdots \end{pmatrix}. \tag{19}$$

Therefore, for each informative variable, one group can be distinguished from the two other groups. The degree of separation, given by μ, is randomly selected from a uniform distribution $U_{[-6,-3]\cup[3,6]}$. The first simulation setup uses multivariate normally distributed groups of observations using the parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}^i_{inf}$, for $i \in \{1, 2, 3\}$, and the second setup uses multivariate log-normally distributed groups of observations with the same parameters. Note that noise variables can be problematic for several of the outlier detection methods, and skewed distributions can create difficulties for methods relying on elliptical distributions.

After simulating the groups of observations, scatter outliers are generated by replacing 5% of the observations of each group with outlying observations. Therefore, we use the same location parameter $\boldsymbol{\mu}_i$, but their covariance matrix is a diagonal matrix with constant diagonal elements σ which are randomly generated between 3 and 9, $\sigma \sim U[3, 9]$, for informative variables. The reason for using scatter outliers instead of location outliers (changed $\boldsymbol{\mu}_i$) is the advantage, that outliers will not form a separate group but will stick out of their respective group in random directions.

The outcome of the first simulation setup based on multivariate normal distribution is visualized in Figure 4. Figure 4a shows the performance for 100 repetitions with 1000 noise variables as boxplots measured by the AUC value. We note that local projections (LocOut) outperform all other methods, while LOF, SOD, and KNN perform approximately at the same level. For smaller numbers of noise variables, especially SOD performs better than local projections. This becomes clear in Figure 4b, showing the median performance of all methods with a varying number of noise variables. We see that the performance of SOD drops quicker than other methods, while local projections are effected the least by an increasing number of noise variables. The horizontal grey line corresponds to a performance of 0.5 which refers to random outlier scores.



(a)

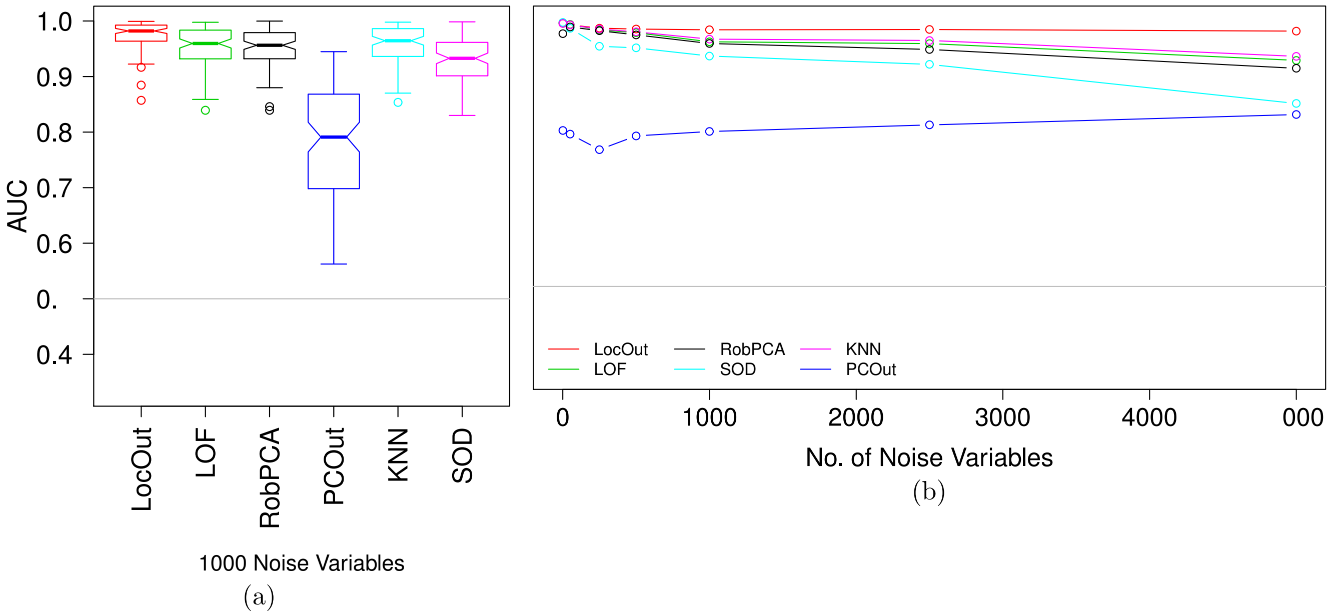
(b)

Figure 4: Evaluation of outliers in three multivariate normally distributed groups with a varying number of noise variables. 5% of the observations were replaced by outliers. Plot (a) shows boxplots for the setup with 1000 noise variables. Each setup was repeatedly simulated 100 times. Plot (b) shows the median performance of each method for various numbers of noise variables.

Setup 2, visualized in Figure 5, shows the effect of non-normal distributions on the outlier detection methods. The same parameters used for log-normal distributions as in the normally distributed setup, make it easier for all methods to identify outliers. Nevertheless, the order of performance changes since the methods are affected differently. SOD is stronger affected than LOF, since it is easier for SOD to identify useful spaces for symmetric distributions while LOF does not benefit from such properties. LocOut still shows the best performance, at least for an increasing number of noise variables. The most notable difference is the effect on RobPCA, which heavily depends on the violated assumption of normal distribution.

6 Application on real-world datasets

In order to demonstrate the effectiveness of local projections in real-world applications, we analyze three different datasets, varying in the number of groups, the dimension of the data space, and the separability of the groups. We always use observations from multiple groups as non outlying observations and a small number of one additional group to simulate outliers in the dataset.



(a)

(b)

Figure 5: Evaluation of outliers in three multivariate log-normally distributed groups with a varying number of noise variables. 5% of the observations were replaced by outliers. Plot (a) shows boxplots for the setup with 1000 noise variables. Each setup was repeatedly simulated 100 times. Plot (b) shows the median performance of each method for various numbers of noise variables.

6.1 Olive Oil

The first real-world dataset consists of 120 samples of measurements of 25 chemical compositions (fatty acids, sterols, triterpenic alcohols) of olive oils from Tuscany, Italy (Armanino et al., 1989). The dataset is used as a reference dataset in the R-package *rrcovHD* (Todorov, 2016) for robust multivariate methods for high-dimensional data and consists of four groups of 50, 25, 34, and 11 observations, respectively. We use observations from the smallest group with 11 observations to sample 5 outliers 50 times.

In our context, this dataset represents a situation where the distribution can be well-estimated due to its non-flat data structure. Therefore, it is possible to include COP in the evaluation. It is important to note that at least 26 observations must be used by COP in order to be able to locally estimate the covariance structure, while there will always be a smallest group of 25 observations at most present for each setup. Thus, we would assume, that COP has problems distinguishing between outliers and observations from this smallest group which does not yield enough observations for the covariance estimation.

We show the performance of the compared outlier detection methods based on the AUC values in Figure 6. We note that all methods but PCOut and COP perform at a very high level. For KNN, SOD and LocOut, there is only a non-significant difference in the median performance.

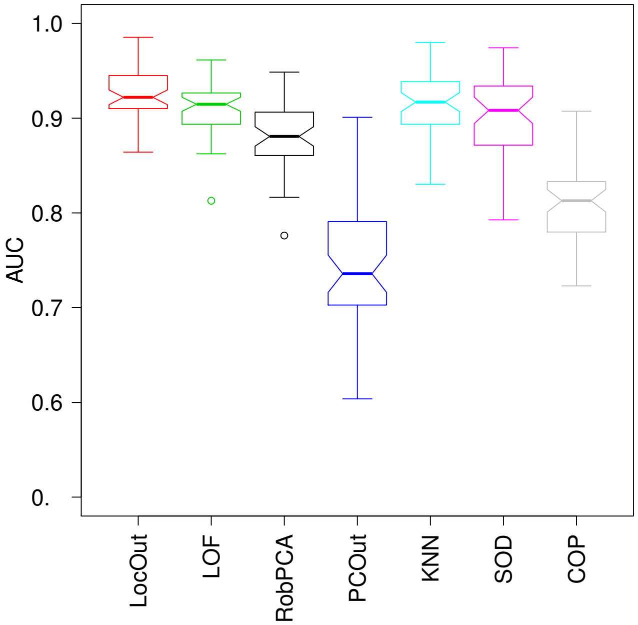


Figure 6: Performance of different outlier detection methods for the 25 dimensional olive oil dataset measured by the area under the ROC curve (AUC). For each method the configuration parameters are optimized based on the ground truth.

6.2 Melon

The second real-world dataset used for the evaluation is a fruit data set, which consists of 1095 observations in a 256 dimensional space corresponding to spectra of the different melon cultivars. The observations are documented as members of three groups of sizes 490, 106 and 499, but in addition, during the cultivation processes different illumination systems have been used leading to subgroups. The dataset is often used to evaluate robust statistical methods (e.g. Hubert and Van Driessen, 2004).

We sample 100 observations from two randomly selected main groups to simulate a highly inconsistent structure of main observations and add 7 outliers, randomly selected from the third remaining group. We repeatedly simulate such a setup 150 times in order to make up for the high degree of inconsistency. As Figure 7 shows, the identification of outliers is extremely difficult for this dataset. A combination of properly reducing the dimensionality and modelling the existing sub-groups is required. LocOut outperforms the compared methods, followed by LOF and PCOut.

6.3 Archaeological glass vessels

The observations of the glass vessels dataset, described e.g. in Janssens et al. (1998) refer to archaeological glass vessels, which have been excavated in Antwerp (Belgium). In order to distinguish between different types of glass, which have either been imported or locally produced, 180 glass samples were chemically analyzed using electron probe X-ray microanalysis (EPXMA). By measuring a spectrum at 1920 different energy levels corresponding to different chemical elements for each of those vessels, a high-dimensional data set for classifying the glass vessels is created. A

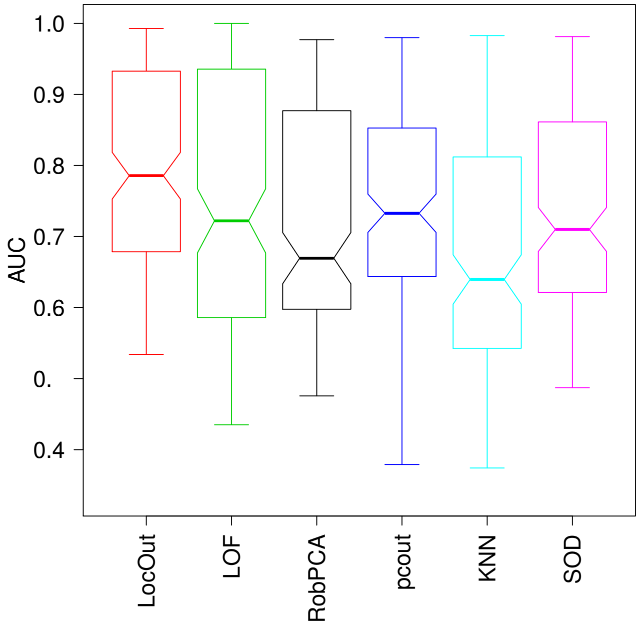


Figure 7: Evaluation of the performance of the outlier detection algorithms on the fruit data set, showing boxplots of the performance of 150 repetitions of outlier detection measured by the AUC.

few (11) of those variables/energy levels contain no variation and are therefore removed from our experiments in order to avoid problems during the computation of outlyingness.

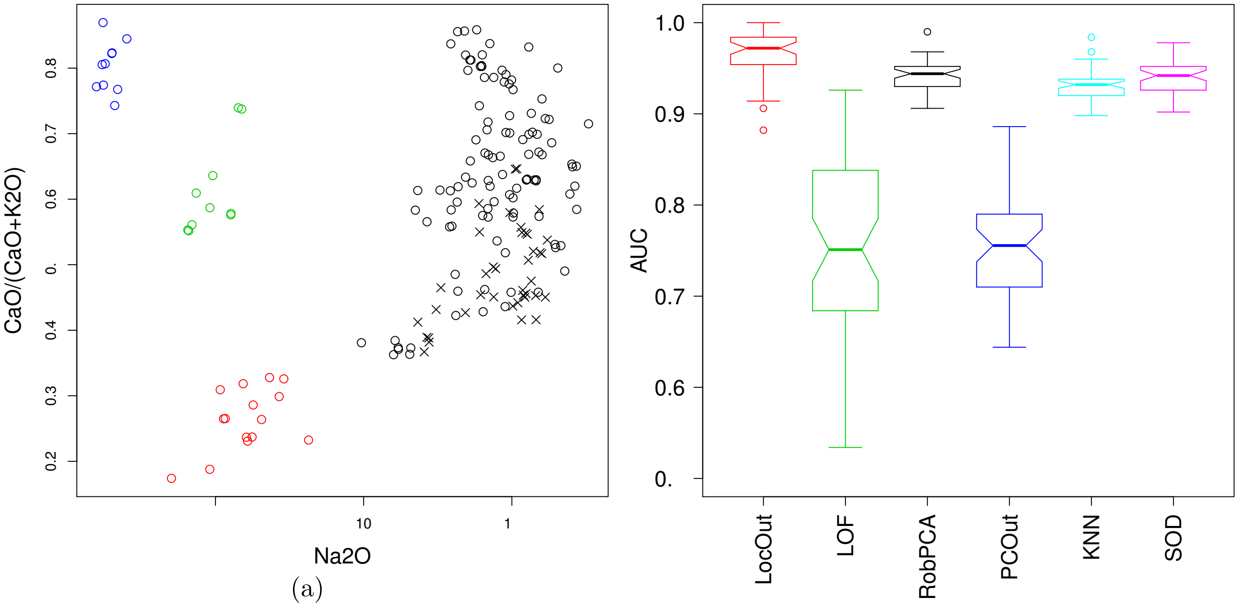
While performing PLS regression analysis, Lemberge et al. (2000) realized that some vessels had been measured at a different detector efficiency and, therefore, removed those spectra from the dataset. We do not remove those observations, since from an outlier detection perspective they represent bad leverage points as indicated by Serneels et al. (2005), which we want to be able to identify. These leverage points are visualized in Figure 8a with x-symbols. By including these observations as part of the main groups, it becomes especially difficult to identify outliers sampled from the green group (potasso-calic). We sample 100 observations from the non-potasso-calic group 50 times and add 5 randomly selected potasso-calic observations as outliers. The performance is visualized in Figure 8b. Again, LocOut outperforms all compared methods, while LOF and PCOut have problems to deal with this data setup.

7 Discussion of runtime

The algorithm for local projections was implemented in an R-package which is publicly available[2]. The package further includes the glass vessels data set used in Section 6. Based on this R-package, we performed simulations to test the required computational effort for the proposed algorithm and the impact of changes in the number of observations and the number of variables.

For each projection, the first step of our proposed algorithm is based on the k-nearest neighbours concept. Therefore, we need to compute the distance matrix for all n available p-dimensional observations leading to an effort of $O(n(n-1)p/2)$, where $n(n-1)/2$ represents the combinations

[2] http://www.applied-statistics.at/locout_1.0.tar.gz



(a)

(b)

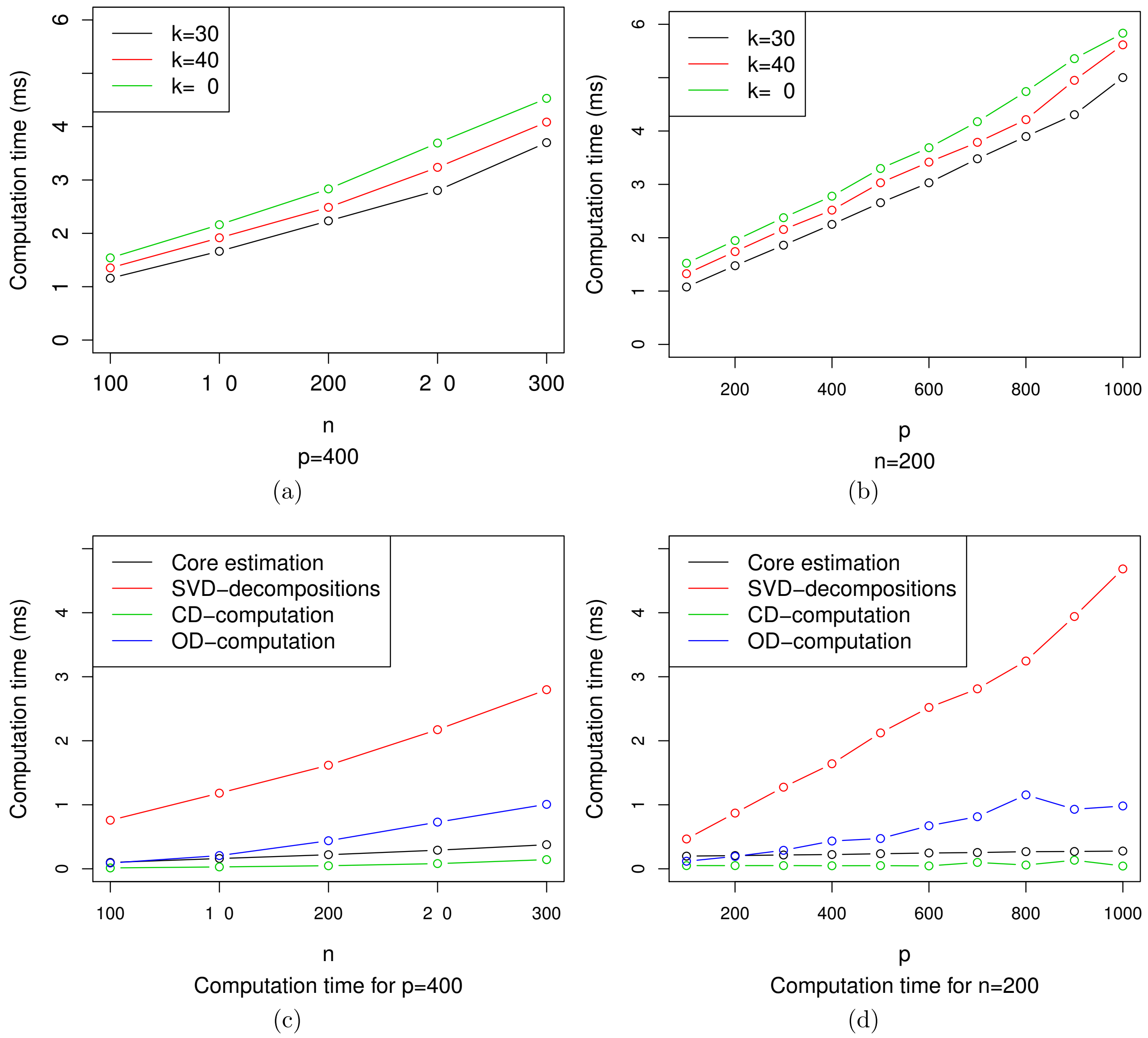
Figure 8: Evaluation of the performance of the outlier detection algorithms on the glass vessels data set. Plot (a) shows the classification of the group structure based on the chemical composition. Plot (b) shows boxplots of the performance of 50 repetitions of outlier detection measured by the AUC.

of observations and p, being the dimension of the data space, reflects the effort for the computation of each Euclidean distance.

After the basic distance computation, we need to compare those distances (which scales with n but only contributes negligibly to the overall effort) and scale the data based on the location and scale estimation for the selected core (which also does not significantly affect the computation time).

For all of the n cores, we perform an SVD decomposition leading to an effort of $O(p^2n^2 + n^4)$. Therefore, a total effort of $O(n^2p(1+p)+n^4)$ is expected for the computation of all local projections. In this calculation, reductions, such as the multiple computation of the projection onto the same core, are not taken into account. Such an effect is very common due to the presence of hubs in data sets (Zimek et al., 2012). Figure 9 provide an overview of the overall computation time decomposed into the different aspects of the computation algorithm.

We observe that the computation time increases approximately linearly with p, while it increases faster than a linear term with increasing n. There is an interaction effect between k and n visible in plot (a) of Figure 9 as well, due to the necessity of n knn computations. Plots (c) and (d) show that the key factors are the n SVDs. Especially the core estimation and the computation of the core distance are just marginally affected by increasing n and not affected at all by increasing p. The orthogonal distance computation is non-linearly affected by increasing n and p which however remains relatively small when being compared to the SVD estimations.



(a)

(b)

(c)

(d)

Figure 9: Visualisation of the computation time of the local projections. Plots (a) and (b) evaluate the development of the overall computation time for increasing n in plot (a) and increasing p in plot (b). Those evaluations are performed for varying k. Plot(c) and (d) focus on different components of the computation for a fixed $k = 40$ and increasing n and p.

8 Conclusions

We proposed a novel approach for evaluating the outlyingness of observations based on their local behaviour, named *local projections*. By combining techniques from the existing robust outlier detection RobPCA (Hubert et al., 2005) and from Local Outlier Factor (LOF) (Breunig et al., 2000), we created a method for outlier detection, which is highly robust towards large numbers of non-informative noise variables and which is able to deal with multiple groups of observations, not necessarily following any specific standard distribution.

These properties are gained by creating a local description of a data structure by robustly selecting a number of observations based on the k-nearest neighbours of an initiating observation and projecting all observations onto the space spanned by those observations. Doing so repeatedly, where each available observation initiates a local description, we describe the full space in which the data set is located. In contrast to existing subspace-based methods, we create a new concept for interpreting the outlyingness of observations with respect to such a projection by introducing the concept of quality of local description of a model for outlier detection. By aggregating the measured outlyingness of each projection and by downweighting the outlyingness with this quality-measure of local description, we define the univariate local outlyingness score, *LocOut*. *LocOut* 单变量 measures the outlyingness of each observation in comparison to other observations and results in a ranking of outlyingness for all observations. We do not provide cut off values for classifying observations as outliers and non-outliers. While at first consideration this poses a disadvantage, 漠视 it allows for disregarding any assumptions about the data distribution. Such assumptions would be required in order to compute theoretical critical values.

We showed that this approach is more robust towards the presence of non-informative noise variables in the data set than other well-established methods we compared to (LOF, SOD, PCOut, KNN, COP, and RobPCA). Additionally, skewed non-symmetric data structures have less impact than for the compared methods. These properties, in combination with the new interpretation of outlyingness allowed for a competitive analysis of high-dimensional data sets as demonstrated on three real-world application of varying dimensionality and group structure.

The overall concept of the proposed local projections utilized for outlier detection opens up possibilities for more general data analysis concepts. Any clustering method and discriminant analysis method is based on the idea of observations being an outlier for one group and therefore being part of another group. By combining the different local projections, a possibility for avoiding assumptions about the data distribution - which are in reality often violated - is provided. Thus, applying local projections on data analysis problems could not only provide a suitable method for analyzing high-dimensional problems but could also reveal additional information on method-influencing observations due to the quality of local description interpretation of local projections.

Acknowledgements

This work has been partly funded by the Vienna Science and Technology Fund (WWTF) through project ICT12-010 and by the K-project DEXHELPP through COMET - Competence Centers for Excellent Technologies, supported by BMVIT, BMWFW and the province Vienna. The COMET program is administrated by FFG.

References

Achtert, E., Kriegel, H.-P., and Zimek, A. (2008). Elki: a software system for evaluation of subspace clustering algorithms. In Scientific and statistical database management, pages 580–585. Springer.

Aggarwal, C. C. and Yu, P. S. (2001). Outlier detection for high dimensional data. In ACM Sigmod Record, volume 30, pages 37–46. ACM.

Armanino, C., Leardi, R., Lanteri, S., and Modi, G. (1989). Chemometric analysis of tuscan olive oils. Chemometrics and Intelligent Laboratory Systems, 5(4):343–354.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: identifying density-based local outliers. In ACM sigmod record, volume 29, pages 93–104. ACM.

Campello, R. J., Moulavi, D., Zimek, A., and Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Transactions on Knowledge Discovery from Data (TKDD), 10(1):5.

Campos, G. O., Zimek, A., Sander, J., Campello, R. J., Micenková, B., Schubert, E., Assent, I., and Houle, M. E. (2016). On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data Mining and Knowledge Discovery, 30(4):891–927.

De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. L. (2000). The mahalanobis distance. Chemometrics and intelligent laboratory systems, 50(1):1–18.

Fawcett, T. (2006). An introduction to roc analysis. Pattern recognition letters, 27(8):861–874.

Filzmoser, P. and Gschwandtner, M. (2015). mvoutlier: Multivariate outlier detection based on robust methods. R package version 2.0.6.

Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions. Computational Statistics & Data Analysis, 52(3):1694–1711.

Henrion, M., Hand, D. J., Gandy, A., and Mortlock, D. J. (2013). Casos: a subspace method for anomaly detection in high dimensional astronomical databases. Statistical Analysis and Data Mining: The ASA Data Science Journal, 6(1):53–72.

Hu, Y., Murray, W., Shan, Y., and Australia. (2015). Rlof: R Parallel Implementation of Local Outlier Factor(LOF). R package version 1.1.1.

Hubert, M., Rousseeuw, P. J., and Vanden Branden, K. (2005). Robpca: a new approach to robust principal component analysis. Technometrics, 47(1):64–79.

Hubert, M. and Van Driessen, K. (2004). Fast and robust discriminant analysis. Computational Statistics & Data Analysis, 45(2):301–320.

Janssens, K., Deraedt, I., Freddy, A., and Veekman, J. (1998). Composition of 15-17th century archeological glass vessels excavated in antwerp. belgium. Mikrochimica Acta. v15 iSuppl, pages 253–267.

Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. (2009). Outlier detection in axis-parallel subspaces of high dimensional data. Advances in knowledge discovery and data mining, pages 831–838.

Kriegel, H.-P., Kroger, P., Schubert, E., and Zimek, A. (2012). Outlier detection in arbitrarily oriented subspaces. In Data Mining (ICDM), 2012 IEEE 12th International Conference on, pages 379–388. IEEE.

Lemberge, P., De Raedt, I., Janssens, K. H., Wei, F., and Van Espen, P. J. (2000). Quantitative analysis of 16–17th century archaeological glass vessels using pls regression of epxma and μ-xrf data. Journal of Chemometrics, 14(5-6):751–763.

Ortner, T., Filzmoser, P., Zaharieva, M., Breiteneder, C., and Brodinova, S. (2017). Guided projections for analysing the structure of high-dimensional data. arXiv preprint arXiv:1702.06790.

Pomerantsev, A. L. (2008). Acceptance areas for multivariate classification derived by projection methods. Journal of Chemometrics, 22(11-12):601–609.

Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In ACM Sigmod Record, volume 29, pages 427–438. ACM.

Serneels, S., Croux, C., Filzmoser, P., and Van Espen, P. J. (2005). Partial robust m-regression. Chemometrics and Intelligent Laboratory Systems, 79(1):55–64.

Todorov, V. (2016). rrcovHD: Robust Multivariate Methods for High Dimensional Data. R package version 0.2-5.

Todorov, V. and Filzmoser, P. (2009). An object-oriented framework for robust multivariate analysis. Journal of Statistical Software, 32(3):1–47.

Zimek, A., Schubert, E., and Kriegel, H.-P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analysis and Data Mining: The ASA Data Science Journal, 5(5):363–387.