Die approbierte Originalversion dieser Dissertation ist in der Hauptbibliothek der Technischen Universität Wien aufgestellt und zugänglich.
http://www.ub.tuwien.ac.at

The approved original version of this thesis is available at the main library of the Vienna University of Technology.
http://www.ub.tuwien.ac.at/eng

DISSERTATION

Local Projections for High-dimensional Data Analysis

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors
der technischen Wissenschaften

unter der Leitung von
Univ.-Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser,
Institut für Stochastik und Wirtschaftsmathematik (E105)

eingereicht an der Technischen Universität Wien an der Fakultät fr Mathematik and
Geoinformation von
Dipl.-Ing. Thomas Ortner

Matrikelnummer 0326473

Diese Dissertation haben begutachtet:

Peter Filzmoser TU Wien	Matthias Templ ZHAW School of Engineering	Marco Riani Universita' degli Studi di Parma

Wien, 10. September 2017

Thomas Ortner

Erklärung zur Verfassung der Arbeit

Dipl.-Ing. Thomas Ortner
Kleine Pfarrgasse 31/5
A-1020 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 10. September 2017

Thomas Ortner

Acknowledgements

I would like to thank my supervisor Prof. Peter Filzmoser for his support and guidance during my research. I further want to express my gratitude for the supportive, encouraging discussions and insights provided from my colleagues and friends at the computational statistics research group and the *famous* research project.

I am especially grateful for the continuous ever-lasting support of my parents and my family.

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.

Kurzfassung

Die Entwicklung im Bereich der Datenaquisemethoden der vergangenen Jahre führt zu immer größer werdenden Datensätzen. Dabei explodiert sowohl die Anzahl an Beobachtungen, als auch die Anzahl an Variablen. Klassische statistische Ansätze sind nicht dafür ausgelegt, um mit dieser neuen Situation adäquat umgehen zu können. Insbesondere flache Datenstrukturen mit mehr Variablen als Beobachtungen stellen ein erhebliches Problem dar, da diese Situation zu Singularitäten im Rahmen der Berechnung von statistischen Schätzern führt. Große Anzahlen an Variablen führen aber nicht nur zu Problemen sondern eröffnen auch neue Möglichkeiten in der Datenanalyse.

Der am häufigsten angewandte Ansatz im Zusammenhang mit hochdimensionalen Daten ist die Reduktion der Dimension durch Variablenselektion oder Methoden wie z.B. Hauptkomponentenanalyse. Die Mehrheit dieser Ansätze berücksichtigt dabei die Information des Komplementes der Projektion nicht, obwohl dort im Allgemeinen ein Teil oder auch die Mehrheit der nützlichen Daten zu finden ist. Nur wenige Ansätze (z.B. Hubert et al., 2005; Kriegel et al., 2012) erkennen diesen Aspekt an.

Wir entwickeln einen alternativen Projektionsansatz, der beide Informationen, die Distanzen zwischen Beobachtungen, sowie die Distanz zum Projektionsraum berücksichtigt. Zusätzlich vermeiden wir ein allgemeines Modell, das alle Daten gleichzeitig beschreibt, sondern entwickeln eine Serie von Projektionen, die die lokale Datenstruktur beschreibt. Diese Serie wird daher *lokale Projektionen* genannt. Wir stellen eine Reihe von Anwendungsmöglichkeiten dieser lokalen Projektionen aus den Bereichen Datentransformationen, Darstellungsmethoden zur Erkennung von Datenstrukturen, Ausreißererkennung und Klassifikationsanalyse vor. Jeder Ansatz verwendet die Möglichkeiten, die sich aus lokalen Projektionen und den Distanzen innerhalb der Projektion und zur Projektion ergeben, auf seine eigene Art und Weise.

Abstract

The development of data collection methods over the last decades led to increasingly larger numbers of observations and variables. Classical statistical methods have not been designed to deal with this new situation. Especially flat data structures, where more variables than observations are present, pose the problem of singularities during the computation of statistical estimators required for data analysis approaches. Large numbers of variables do not just pose a problem in data analysis but also open up new opportunities.

The most common practice in the context of high-dimensionality is the reduction of dimension by variable selection or other projection approaches like principal components analysis. The majority of those approaches does not take the information of the complement of the projection into account which typically still yields some if not the majority of the useful information. Few approaches (e.g. Hubert et al., 2005; Kriegel et al., 2012) acknowledge this aspect of high-dimensional data analysis.

We propose an alternative to projection methods taking both information, the distance between observations within the projection space as well as the distance to the projection space into account. In addition, instead of using one overall model, we use a series of projections, locally describing the data structure. Therefore, our projection approach is named *local projections*. Several possibilities including data transformations, diagnostics for groups in the data structure, outlier detection and supervised classification methods based on local projections are presented. Each approach uses the opportunities of learning from the within-projection and to-projection distance in a unique way.

Contents

Kurzfassung **vii**

Abstract **ix**

Contents **xi**

1 Introduction **1**

1.1 High-dimensional and flat data analysis 2
1.2 The concept of local projections 7
1.3 Outline of the thesis 12

2 Guided projections for analyzing the structure of high-dimensional data **15**

2.1 Introduction 16
2.2 Guided projections 18
2.3 Simulations 29
2.4 Real-world data sets 38
2.5 Conclusions and outlook 41

3 Local projections for high-dimensional outlier detection **45**

3.1 Introduction 46
3.2 Local projections 48
3.3 Interpretation and utilization of local projections 54
3.4 Evaluation setup 56
3.5 Simulation results 58
3.6 Application on real-world datasets 61
3.7 Discussion of runtime 64

3.8 Conclusions 66

4 Multigroup discrimination based on weighted local projections 69

4.1 Introduction 69
4.2 Methodology 71
4.3 Visualization of the discrimination 76
4.4 Evaluation 80
4.5 Conclusions and outlook 89

List of Figures 91

List of Tables 97

Bibliography 99

Index 107

CHAPTER 1

Introduction

Statistics and data analysis methods for high-dimensional data have frequently been in the focus of scientific endeavours over the last decades. The term high-dimensional refers to large numbers of variables or features, describing the objects of interest. The interest in high-dimensional data analysis can mainly be explained due to the vast amount of data of all kinds which is collected nowadays, reaching from DNA Microarrays, over spectra analysis data to social networks.

In traditional statistics many observations with few, well chosen variables are assumed to be available. The trend shows that this assumption does not hold any longer. During the data collection process it is not clear, what variables can be useful for applications which are not yet developed. Even more often, the questions are not defined before the data collection process starts. This leads to the problems of high dimensional data which are nowadays rule rather than exception (Bühlmann and Van De Geer, 2011). We experience examples where the objects are described by thousands or even millions of properties while only few objects are available for study. In extreme cases where fewer variables than observations are available, we speak of flat data. Classical statistics is not designed to deal with such datasets as pointed out by Donoho et al. (2000).

In this work we are concerned with possibilities to overcome the effects of high-dimensional data on analysis approaches. Typically, statistical and machine learning methods are concerned with the questions of supervised and unsupervised classification, outlier detection etc. For each of those questions a variety of machine learning methods are available where the vast majority uses distance measures, mostly Euclidean distances, to separate classes, clusters and outliers.

In this chapter we first discuss the challenges of high-dimensional data-analysis in Section 1.1 addressing dimensionality-based problems for statistical estimators as well as possibilities emerging from the dimensionality. Section1.2 describes the concept of local projections which provide the basis for this thesis as well as the connections to the related work. We conclude this chapter with Section 1.3, outlining the rest of the thesis containing information on submitted publications.

1.1 High-dimensional and flat data analysis

When discussing the challenges for classical statistics in the context of high-dimensional and possibly flat data we often encounter two concepts: The curse and the blessing of high-dimensionalty. One is perceived as a challenge to the concepts of data analysis by masking the differences between observations, the other as a chance based on additional observations. Before addressing the curse and blessing in detail, we introduce more generalized problems based on the example of regression analysis.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ denote n observations of a univariate numeric outcome measurement and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ the n observations of p variables, with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$, for $i = 1, \ldots, n$. Assuming we want to use a linear model

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \qquad i = 1, \ldots, n \tag{1.1}$$

to describe the relations between $\boldsymbol{y}$ and $\boldsymbol{X}$, and we want to predict the outcome y_0 for a future observation $\boldsymbol{x}_0$. Then we will experience several shortcomings for high-dimensional data in general and specifically in the presence of flat data structures.

The first problem is related to the estimation $\hat{\boldsymbol{\beta}}$ of the parameters $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$. In a case where $p > n$ holds, the estimators are unambiguously defined as we need to estimate p parameters based on n observations. The same problem occurs during the computation of covariance matrices which is regularly required by statistical approaches. Therefore a dimension reduction is commonly applied. Various methods, varying from feature selection (e.g. Guyon and Elisseeff, 2003) to principal component analysis (e.g. Abdi and Williams, 2010) are available for this task. By compressing the information to at most $n - 1$ linearly independent variables, the coefficients $\hat{\boldsymbol{\beta}}$ as well as the inverse covariance matrix are unambiguously defined. Nevertheless such a compression always comes at the price of removing information.

This directly leads to the second shortcoming which is the interpretability of the coefficients $\hat{\boldsymbol{\beta}}$. Large numbers of variables often occur jointly with random correlations between variables or groups of variables. Therefore, due to multicollinearity occurring by chance, the importance of variables cannot directly be determined from the coefficients any more. Again, the same issue exists in classification and outlier detection problems. If one is interested in identifying the reasons for separation, using analysis on the full dimensional space is often not useful.

The third issue might be of the highest importance. The predictive ability of high-dimensional models is limited due to problems of overfitting based on random effects added with each additional variable. With an increasing number of variables, methods can not distinguish between random and non-random properties of observations. Again, the problem is multicollinearity by chance leading to models describing a specific dataset rather than a class of datasets.

It is important to note that several methods have been developed in order to deal with those three problems, especially in the context of linear regression, including Ridge regression (Hoerl and Kennard, 1970) and Lasso based regression (Tibshirani, 1996). We want to offer an alternative approach to the problems of high-dimensionality. Before doing so, we provide a description of the concepts of the curse and the blessing of high-dimensionality leading to insights on critical aspects of high-dimensional data analysis.

The curse of high-dimensionality

The term *curse of dimensionality* which was first introduced by Bellman (1961) is often used in a vague form, indicating that the differences in distances between observations in high-dimensional spaces become insignificant. While a vast number of publications has addressed the curse of high-dimensionality and ways to overcome it (e.g. Keogh and Mueen, 2011; Indyk and Motwani, 1998; Aggarwal, 2005), it is more difficult to identify a clear definition. One of the most precise definitions covering the core of the problem is provided in Beyer et al. (1999).

Let $\boldsymbol{X}$ denote a data matrix containing of n observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, where all observations are drawn from the same p-dimensional random vector X_p. d_{max} further denotes the largest distance between two observations of $\boldsymbol{X}$ and d_{min}, the minimal distance between two observations from $\boldsymbol{X}$.

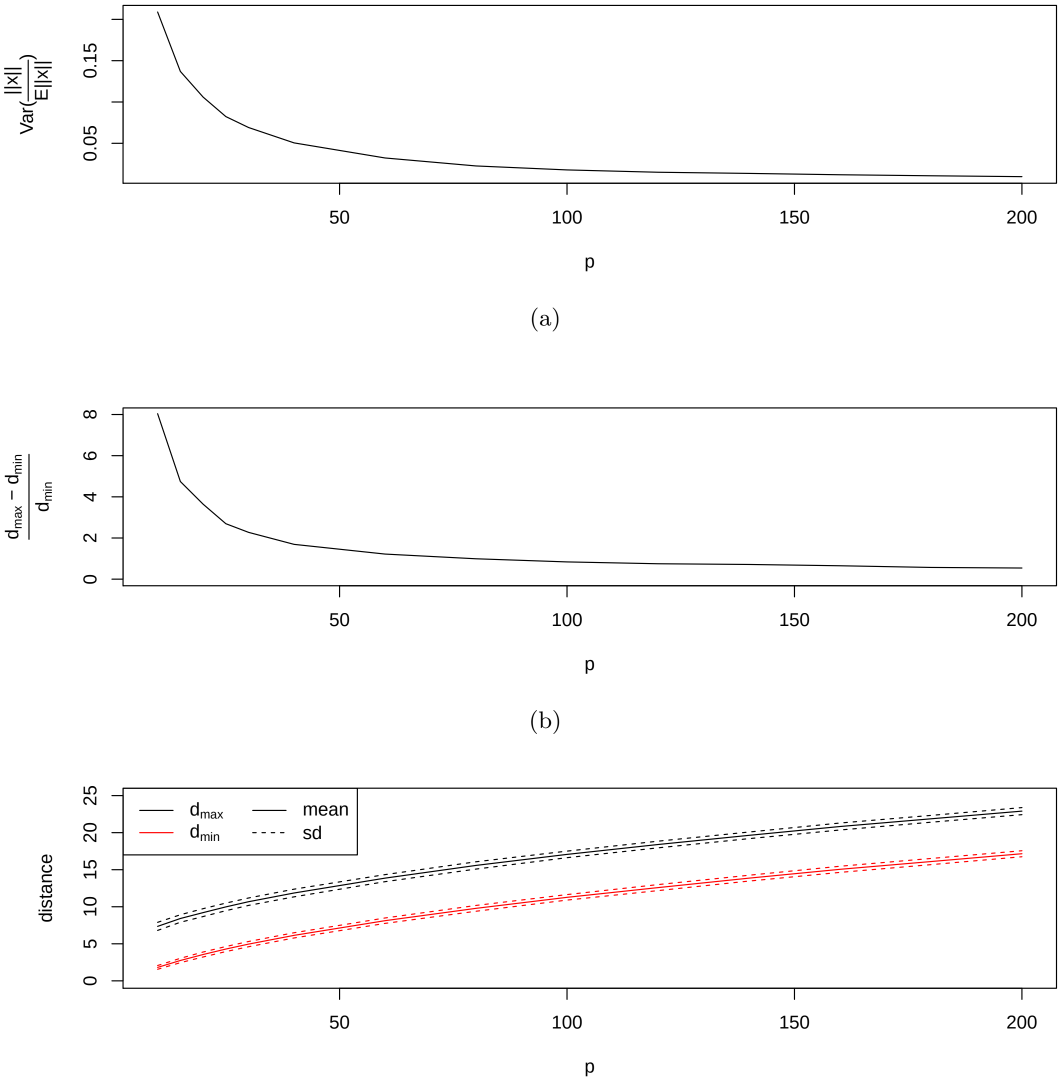
$$\begin{aligned} \text{If} \quad & \lim_{p\to\infty} Var\left(\frac{||X_p||}{E||X_p||}\right) = 0 \\ \Rightarrow \quad & \frac{d_{max}-d_{min}}{d_{min}} \underset{p\to\infty}{\rightarrow} 0 \end{aligned} \tag{1.2}$$

We use Figure 1.1 to visualize the implication in Figure 1.1a and the convergence in Figure 1.1b. For both figures we simulate p independent standard normally distributed variables. We see that only few variables are sufficient to reduce the variance close to 0. The aspect we would like to emphasize is that it is not the high dimensionality itself which causes problems. As long as each additional variable contains the same information, there is not reduction in the overall ability for separation. The term $d_{max} - d_{min}$ itself is not just not converging to 0. In Figure 1.1c we note that the difference between the minimal and the maximal distance remains the same. Also the variance is not increasing.

Based on the conclusions of Equation (1.2) and the observations from Figure 1.1 we would like to introduce the concept of informative and non-informative or noise variables in the context of data analysis. Let $X = (X_1, \ldots, X_p)$ and $Y = (Y_1, \ldots, Y_p)$ denote two p-dimensional random variables where the distributions of X_i and Y_i are given by F_{X_i} and F_{Y_i} respectively.

Any variable, where $F_{X_i} = F_{Y_i}$ holds is classified as **non-informative** or as **noise** variable. If $F_{X_i} \neq F_{Y_i}$ holds, we call the variable **informative** as the variable can contribute to the distinction between X and Y.

Using this definition we can now clearly see the curse of dimensionality. Let all X_i and all Y_i be independently distributed. Let further $X_1 \sim N(0,1)$ and $Y_1 \sim N(3,1)$ and all X_i and all $Y_i, i > 1$ be standard normally distributed. The expected distance of Y to the center of X in the first variable is large enough to separate samples of the two random variables with a high probability. By adding additional noise variables, this difference becomes less relevant though as the variance from the noise variables begins to dominate and therefore masks the difference from the first variable. Figure 1.2 visualizes this effect. The dashed black lines represent the distribution of $||X||$, which changes with increasing p. The red solid line shows the expected value of $||X - Y||$. We see that only few noise variables are required to completely mask the differences between $||X||$ and $||X - Y||$. Starting from approximately 15 noise variables approximately 10% of the observations from X will be further away from the expected center of X than observations drawn from Y. Thus an observation from Y is likely not to be detectable as an object, not drawn from X, even though the first variable is highly different.



(a)

(b)

(c)

Figure 1.1: Figure (a) shows the convergence of the variance of $\frac{||X_p||}{E||X_p||}$ of a multivariate p-dimensional independent identically normally distributed variable. Figure (b) shows the relative difference between the maximum distance and the minimum distance between observations. Figure (c) shows the individual behaviour of d_{min} and d_{max}. Note that the behaviour in (b) and (c) is strongly influenced by the sample size n as well.

The blessing of high-dimensionality

The context of informative and noise variables further enables the possibility to analyze the concept of the blessing of high-dimensionality, which is considered a lot less frequently

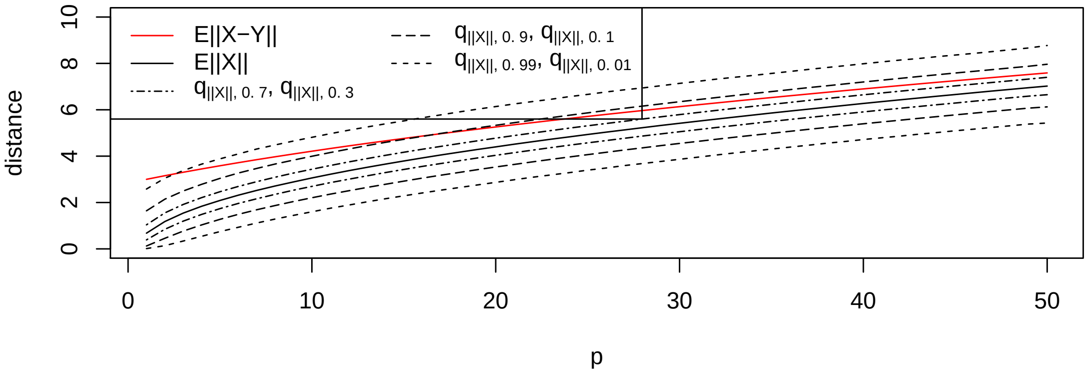


Figure 1.2: The expected distance of Y to X as well as the distribution of $||X||$ using various quantiles is visualized for varying p.

than the curse, e.g. by Chen et al. (2013a); Korn et al. (2001). The main idea of the blessing of high-dimensionality is that by repeatedly adding information from the same covariance structure the distance of all observations of a sample to the center of the random variable will converge against a constant due to the law of large numbers. Let $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote a multivariate normally distributed random variable. The Mahalanobis distance (e.g. De Maesschalck et al., 2000) of an observation $\boldsymbol{x}$ drawn from X to the location $\boldsymbol{\mu}$ of X is defined as follows:

$$MD_X(\boldsymbol{x}) = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})} \tag{1.3}$$

For high-dimensional data, normalized Mahalanobis distances converge against a constant as shown in Hall et al. (2005).

$$\frac{MD_X(\boldsymbol{x})}{\sqrt{p}} \underset{p \to \infty}{\to} 1 \tag{1.4}$$

If observations from a second random variable with a slightly different covariance structure and the same location parameter are present, the Mahalanobis distances for these observations will converge against a different constant. This behaviour is visualized in Figure 1.3. Two distributions are visualized for varying p, the distribution of $||X||$, with $X \sim N(\mathbf{0}, \mathbf{1})$, where $\mathbf{1}$ represents a p-dimensional diagonal matrix with 1-entries and $||Y||$, with $Y \sim N(\mathbf{0}, 1.25 \cdot \mathbf{1})$. As both random variables are located in the origin, a separation of the two groups is not possible with a low number of variables, but for each additional variable the mean increase in Mahalanobis distances is larger for observations drawn from Y than for observations drawn from X. This effect is visualized in Figure

1.3: p is set to values between 1 and 200, 15 samples of X and one sample of Y are drawn. Along the x-axis we start by considering the first variable only and increase the number of considered variables up to 200. All black observations slowly converge against 1, while the red observation, which initially can not be destinguished from the other observations, converges against a higher value. Note that even though the expected value of X and Y are equal, the blessing of high-dimensionality still allows us to perfectly separate the two groups.

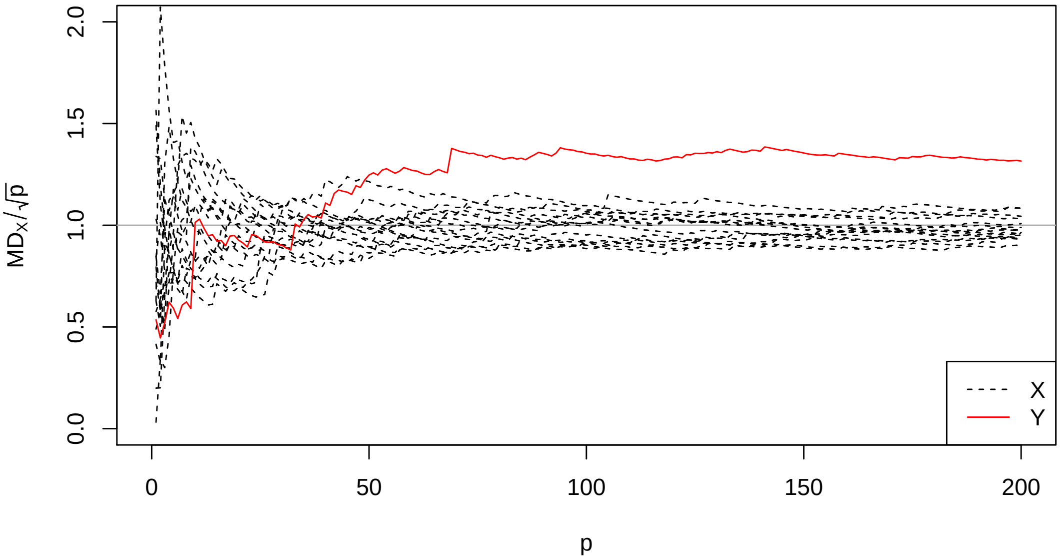


Figure 1.3: The development of Mahalanobis distances for 15 observations from $X \sim N(\mathbf{0}, \mathbf{1})$ (black) and one observation from $Y \sim N(\mathbf{0}, 1.25 \cdot \mathbf{1})$ (red) with increasing p are visualized. All Mahalanobis distances are computed based on the covariance structure of X. Therefore the observations of X converge against 1, while the observation of Y converges against a different constant.

1.2 The concept of local projections

This thesis is built around a method for dealing with problems of high-dimensionality in data analysis for data structures drawn from complex distributions such as multi-group and non-normal distributions. Here, especially flat data spaces are of interest. We do so by combining methods and concepts from different statistical and machine learning approaches.

As pointed out in Section 1.1, in order to be able to apply statistical methods, the ability to inverse covariance matrices is of critical importance. Therefore, we always want to project the data onto a lower dimensional subspace. This concept is very common and used in a variety of different approaches like principal component analysis (e.g. Abdi and Williams, 2010), projection pursuit (Friedman and Tukey, 1974), random projections (e.g. Achlioptas, 2003) and many more. Most methods project the data onto a subspace and discard the information lost due to the projection. One of few exceptions to this approach is RobPCA by Hubert et al. (2005), which is the primary influence on this work. In Hubert et al. (2005) a robust covariance estimation is performed in order to model the majority of observations. 50% of all observations are used for this robust estimation. Based on this covariance, the Mahalanobis distances to the robust estimation of location and the Euclidean distance of observations to the projection space are considered as measures for outlyingness. This approach is highly useful but lacks the ability to deal with more complex structures like subgroups in the data.

In order to be able to analyze subgroups, we need to take the local distribution into account. Especially knn-based methods (e.g. Zhang et al., 2006, 2009) perform well in terms of local density estimation. A combination of RobPCA and k-nearest neighbours leads to the fundamental idea of local projections:

Let $\boldsymbol{X}$ denote a matrix of n rows of observations of p variables where the observations have been drawn from one or multiple random variables. We explicitly consider the possibility of $p \geq n$. Therefore, we cannot properly describe the local density based on covariance estimations. Bringing together the ideas of knn-based estimations and RobPCA we use a set of k instead of $n/2$ observations, locally representing the data structure, and project all information onto the affine subspace spanned by these k observations. We call such a projection, locally describing the data structure, a *local projection* and any set of k observations used to define a local projection a *core* of a projection. Consequently we call the space spanned by the core observations a **core space**. A projection onto the core space is well defined by the right singular vectors of a singular value decomposition after centering and scaling the core observations. Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ denote a core:

$$\hat{\boldsymbol{\mu}} = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{x}_i \tag{1.5}$$

$$\hat{\boldsymbol{\sigma}} = \left(\sqrt{\hat{Var}(x_{11},\ldots,x_{k1})},\ldots,\sqrt{\hat{Var}(x_{1p},\ldots,x_{kp})}\right)' \tag{1.6}$$

$$\tilde{\boldsymbol{core}} = (\tilde{\boldsymbol{x}}_1,\ldots,\tilde{\boldsymbol{x}}_k)' \tag{1.7}$$

$\hat{Var}$ denotes the sample variance and $\tilde{\boldsymbol{x}}_i$ the scaled and centered core observations using the estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$. Based on this notation, the projection $\boldsymbol{V}'$ is given by

$$\tilde{\boldsymbol{core}} = \boldsymbol{UDV}'. \tag{1.8}$$

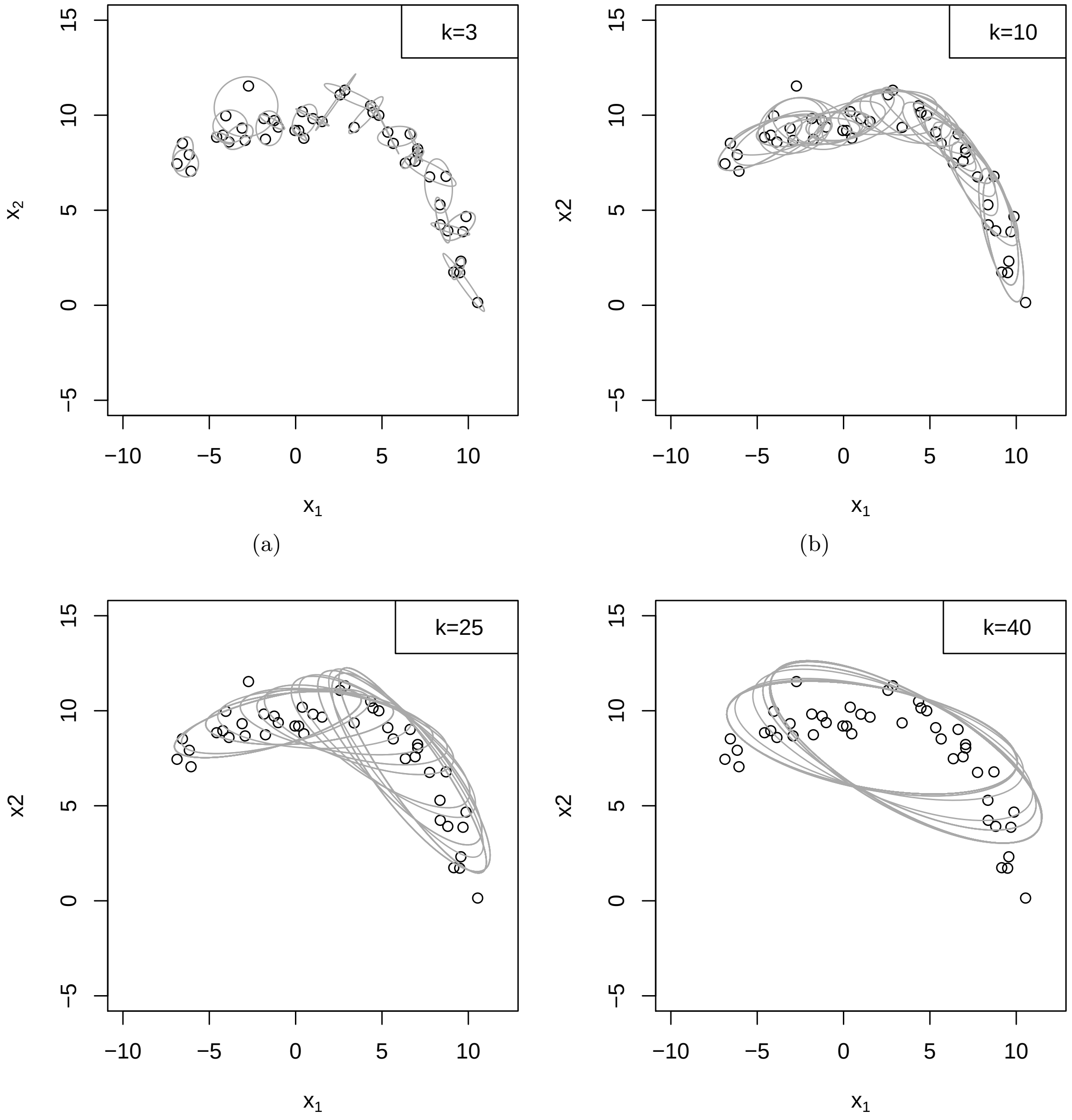
Using $\boldsymbol{V}'$, new observations can be projected onto the core space, after centering and scaling them with respect to the core-based location and scatter estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$. Besides providing a local description, a core space allows for inverting the covariance matrix estimated by the core observations and therefore for computing Mahalanobis distances. Thus we can utilize the two measures of outlyingness used by Hubert et al. (2005), the orthogonal distance (OD) and the score distance which in accordance with the core we call core distance (CD):

$$OD_{\tilde{\boldsymbol{core}}}(\boldsymbol{x}) = ||\tilde{\boldsymbol{x}} - \boldsymbol{VV}'\tilde{\boldsymbol{x}}||, \tag{1.9}$$

$$CD_{\tilde{\boldsymbol{core}}}(\boldsymbol{x}) = \sqrt{\frac{(\boldsymbol{VV}'\tilde{\boldsymbol{x}})'\boldsymbol{D}(\boldsymbol{VV}'\tilde{\boldsymbol{x}})}{k-1}} \tag{1.10}$$

where $\tilde{\boldsymbol{x}}$ denotes a scaled and centered observation. Both aspects, the core distance within the core space and the orthogonal distance to the core space yield important information which should not be ignored (e.g. Kriegel et al., 2012). Depending on the data structure and the research question, local projections provide a framework for combining core and orthogonal distances.

As any set of k observations can be interpreted as a core we do not deal with one local projection but with a series of projections. Different possibilities for finding cores and combining cores are described throughout this thesis. One thing they all have in common in the concept, that a core should always describe the local density. The degree of locality can be adjusted by varying k. An example for the effect of different k (3, 10, 25 and 40) is visualized in Figure 1.4. We use the $k-$nearest neighbours of each observation to estimate the local covariance structure and location. The local description is then provided by the respective confidence ellipse for a 0.75 confidence level.



(a)

(b)

(c)

(d)

Figure 1.4: The degree of locality by using different k in the nearest neighbor based estimation of location and covariance is visualized. For $k = 3$ the data is highly overfitted, while it is underfitted for $k = 40$. The optimal choice of k depends on the data structure as well as on the research question.

The optimal choice of k depends on the data structure. In Figure 1.4 we place 45 observations along an ark combined with a normally distributed error. In Figure 1.4a the 3 nearest observations are used to estimate the local data structure. We see that the description is not homogeneous. There are several breaks between the ellipses and it is not possible to identify any *extreme* observations. Overall, the structure is clearly overfitted. For $k = 10$ in Figure 1.4b the arc becomes apparent and visually extreme observations are not located within the tolerance ellipses any more. Thus, $k = 10$ is a good choice to describe this specific data structure. With increasing k in Figure 1.4c and 1.4d the data is increasingly underfitted by the overall model. In general, k will be optimized based on cross-validation approaches.

Comparison of local projections and RobPCA

The orthogonal and score distances in RobPCA are based on a robust covariance estimation using the Minimum Covariance Determinant (MCD) (Rousseeuw, 1985; Rousseeuw and Driessen, 1999) approach. The covariance matrix with the smallest determinant, based on $(n-p+1)\cdot\alpha$ observations are used. α is a robustness parameter set between 0.5 and 1. In the case of RobPCA it is proposed to set α to 0.5. Based on this covariance matrix, the first h principal components are used to define the projection space, where h is set in a way that a certain percentage, usually 80% of the variance is contained in the projection space.

From the perspective of local projections, the $(n-p+1)\cdot\alpha$ observations form a core, defining one local projection. While RobPCA uses a smaller number of principal components we use the full affine subspace. Therefore, the generalized concept of cores of local projections can be interpreted as a generalization of RobPCA while the flexible number of principal components in RobPCA can be interpreted as a generalization of local projections. RobPCA and local projections are clearly highly related to each other. Nevertheless, these differences are based on severe differences in the assumptions:

Firstly the idea of selecting the observations obviously differs. The concept of modeling the majority of observations is based on the assumption, that only one main group of observations containing some outliers is present. We rather reduce the number of observations below the minimal expected group size in order to locally model the data structure. This approach is based on the assumption that it is easier to identify a small group of observations from the same subgroup than modelling the full data space at once.

Secondly, the reduction of the projection space to the first h principal components is performed on the assumption that the later principal components contain no useful information. In order to perform such a reduction, a good covariance estimation is required. In a flat data space, any covariance estimation will be singular and deviating from the true structure with an increasing number of variables. Thus, for the setup of high-dimensional and flat data we rather use the full affine subspace. By doing so the direction of each core observation from the center of the core obtains the same degree of importance. This aspect is fully intended as it represents the idea that differences between distances of observations drawn from the same distribution are insignificant in high-dimensional spaces.

1.3 Outline of the thesis

The remainder of the thesis consists of three applications of local projections on different data analysis problems. In Chapter 2 we create a series of projections by sequentially exchanging one core observation after another. That way a functional data structure is created which can be interpreted as a data transformation. Based on the transformed data structure we analyse the degree of separation using various validation measures as well as the performance of hierarchical clustering approaches. In addition the functional approach leads to the advantage of having diagnostic plots for the transformed data available. Chapter 2 has been submitted to the Journal of Computational and Graphical Statistics:

Ortner, T., Filzmoser, P., Zaharieva, M., Breiteneder, C., and Brodinova, S., Guided projections for analysing the structure of high-dimensional data. Submitted to *Journal of Computational and Graphical Statistics.*

Chapter 3 is dealing with the problems of outlier detection in high-dimensional and flat data. A local estimation of the density close to each observation is performed using a robustified core estimation. We then use the core distances to measure the relevance of the projection space for each potential outlier and then weight the orthogonal distance, used as measure of outlyingness with the measured relevance. Chapter 3 has been submitted to the Journal of Statistical Analysis and Data Mining:

Ortner, T., Filzmoser, P., Zaharieva, M., Breiteneder, C., and Brodinova,

S., Guided projections for analysing the structure of high-dimensional data. Submitted to *Statistical Analysis and Data Mining.*

Chapter 4 uses local projections in the context of supervised classification. For each projection, an LDA classification model is computed. Based on the quality of the projection-based-model we aggregate all models to receive an overall classification. As the concept of visualization of LDA models can not be aggregated we propose an alternative way of visualizing the classification results which can be applied to most other classification results. Chapter 4 has been submitted to the Journal of Computational and Graphical Statistics:

Ortner, T., Hoffmann, I., Filzmoser, P., Zaharieva, M., Breiteneder, C., and Brodinova, S., "Multigroup discrimination based on weighted local projections. Submitted to *Journal of Computational and Graphical Statistics.*

CHAPTER 2

Guided projections for analyzing the structure of high-dimensional data

Abstract

A powerful data transformation method named guided projections is proposed creating new possibilities to reveal the group structure of high-dimensional data in the presence of noise variables. Utilizing projections onto a space spanned by a selection of a small number of observations allows measuring the similarity of other observations to the selection based on orthogonal and score distances. Observations are iteratively exchanged from the selection creating a non-random sequence of projections which we call guided projections. In contrast to conventional projection pursuit methods, which typically identify a low-dimensional projection revealing some interesting features contained in the data, guided projections generate a series of projections that serve as a basis not just for diagnostic plots but to directly investigate the group structure in data. Based on simulated data we identify the strengths and limitations of guided projections in comparison to commonly employed data transformation methods. We further show the relevance of the transformation by applying it to real-world data sets.

Keywords: dimension reduction, data transformation, diagnostic plots, informative variables

2.1 Introduction

One of the most frequent problems in classical data analysis is the high dimensionality of data sets. In this paper we propose a novel method for data transformations, called *guided projections*, in order to reveal structure in high-dimensional, potentially flat data (more variables than observations). The presented approach uses subsets of observations to locally describe the data structure close to the subsets and measures similarity of all observations to these subsets utilizing the projection onto such subsets. Exchanging observations one by one, we continuously change the location of the local description. By *guiding* the way these subsets are selected, we receive a sequence of projections which can be directly used as a data transformation, as well as a method for visualizing group structure in high-dimensional data. In this paper we present some theoretical background and properties of the proposed guided projections and focus on the general separation between groups in data and how this separation, measured by various validation indices, is affected by the transformation. Furthermore, we compare with existing methods and discuss the strengths and the limitations of guided projections in experiments on both synthetic and real-world data. An implementation of the proposed methodology is publicly available in form of the R-package *lop* at https://github.com/tortnertuwien/lop

Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ denote a data matrix, with p variables and n observations. We further assume that some unknown group structure is present in the observations. In particular we want to consider the possibility that p is larger than n. A large number of variables leads to two main problems we would like to address: First, the cost of computational effort for computing all pairwise distances is $O(n^2 p)$. While we cannot directly influence n, a reduction in p will directly affect computation time. Second, in general, not all p variables hold relevant information about the underlying group structure (Hung and Tseng, 2013). Assume that the data contain some inherent group structure. In accordance to Hung and Tseng (2013) we call variables contributing to a group separation *informative* and variables not contributing to a group separation *non-informative* variables. Accordingly, let us assume $p = p_1 + p_2$, where p_1 denotes the number of informative variables, and p_2 denotes the number of non-informative variables. If p_1 increases, a dimension reduction can considerably reduce the computational burden. If, however, p_2 increases, the variance from non-informative variables will mask the separation provided from informative variables. One possible solution to deal with this masking effect is the application of a data transformation to reveal the group structure in a lower dimensional space. The analysis of effects of such data transformations is the focus of this paper.

A variety of data transformations has been proposed in the past. We present a small selection of commonly employed methods before proposing a novel approach for data transformation.

Classical *variable selection* methods rely on selecting a subset of features which are useful for identifying group structures in data (Guyon and Elisseeff, 2003). A dimension reduction to a small subset of variables, based on some statistic on the distribution of the variables usually provides a suboptimal framework for the analysis of present group structures. One example is the commonly applied method of selecting the 5% of variables with the largest variance for gene expression data. From the variance itself, in general, it can not be concluded whether or not variables are informative.

With the focus on computation time, *Random Projections* (RP) (Achlioptas, 2003) randomly project $\boldsymbol{X}$ onto $\mathbb{R}^{n \times k}$, $k < p$, preserving the expected pairwise distances. There are different ways to identify the required projection matrices. In this paper we use iid normally distributed coefficients as proposed in Li et al. (2006). Such random projections always contain contributions in the same proportion from non-informative variables as from informative variables though.

An approach different from random projections and variable selection is *Principal Component Analysis* (PCA) (e.g. Abdi and Williams, 2010) which is likely the most studied data transformation method. PCA identifies $k < p$ linear combinations of variables, maximizing the variances of each resulting component under the restriction of orthogonality. Such components are called principal components. Classical PCA is subject to restrictions like identifying linear subspaces only. Furthermore, the differences in distances remain masked, since the principal components contain an increasing portion of the non-informative variables with an increasing number of such variables. The problem of linearity has been addressed in several publications (Gorban et al., 2008; De Leeuw, 2011). We will consider *Diffusion maps* (DIFF) (Coifman and Lafon, 2006) as one possible modification, where PCA is performed on the transformed data, based on distances measured by random walk processes. We will further consider *Sparse Principal Component Analysis* (SPC) (Zou and Hastie, 2005; Zou et al., 2006; Witten et al., 2009), since the goal of sparse PCA is to avoid the second problem we addressed, namely the presence of non-informative variables, by downweighting the non-informative variables.

A more general projection approach is *Projection Pursuit* (Friedman and Tukey, 1974) where a projection onto a low-dimensional subspace is identified, maximizing a measure of interest like non-normality. This approach can further be generalized to similarities between estimated and general density functions (Cook et al., 1993) and visualized using so called guided tours (Cook et al., 1995). There are also proposals for modifications of the projection pursuit index in order to cope with high-dimensional data (Lee and Cook, 2010). With the main intention of visualization and visual analysis of projections, the dimension of the projection pursuit is mostly limited between one and three.

After performing such a data transformation, one hopes to yield more information about the underlying group structure of the data. Such information can be measured in terms of performance with respect to a subsequent application of outlier detection methods, discriminant analysis, clustering methods, and other related methods.

The paper is structured as follows. The methodology and properties of our approach is presented in Section 2.2 providing insight on the effects of the transformation as well as a possibility for diagnostic plots. We define synthetic setups for the comparison of the newly introduced method with existing data transformation methods in Section 2.3 and report the results of the performed comparison. In Section 2.4 we apply the methods to two real-world data sets to illustrate the relevance of guided projections. Finally, we provide conclusions and an outlook on possible extensions and applications of the proposed method in Section 2.5.

2.2 Guided projections

Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ denote the data matrix to be analyzed. We further assume, that the observations $\boldsymbol{x}_i$, $i \in \{1 \dots n\}$, are randomly drawn from one of the distributions $F_1, \dots, F_m$, $m < n$. Therefore, up to m groups are present in our data structure.

The basic concept of guided projections is to find a non-random series of projections providing directions where differences between occurring groups are present. Each projection will be described by a selection of observations spanning the projection space. Any such selection describes the data structure close to the selected observations. The sequence of projections starts in a dense region of the data distribution, describing a certain cluster, and alters the observations used for the projections such that another dense region can be reached. By using a small number of observations for the projections, we avoid the masking effects of outlying observations on the description. In this context an outlying observation is an observation which is likely to be from a different group. This concept is visualized in Figure 2.1 for a two dimensional space, using Mahalanobis distances as a representative for the similarity between observations. Since we assume a high-dimensional flat data space, we describe the properties of observations with respect to each specific projection. Therefore we use two distance measures described in Hubert et al. (2005), the *orthogonal distance* and the *score distance*. Using these distances, we iteratively identify a series of observations leading to the series of projections (guided projections).

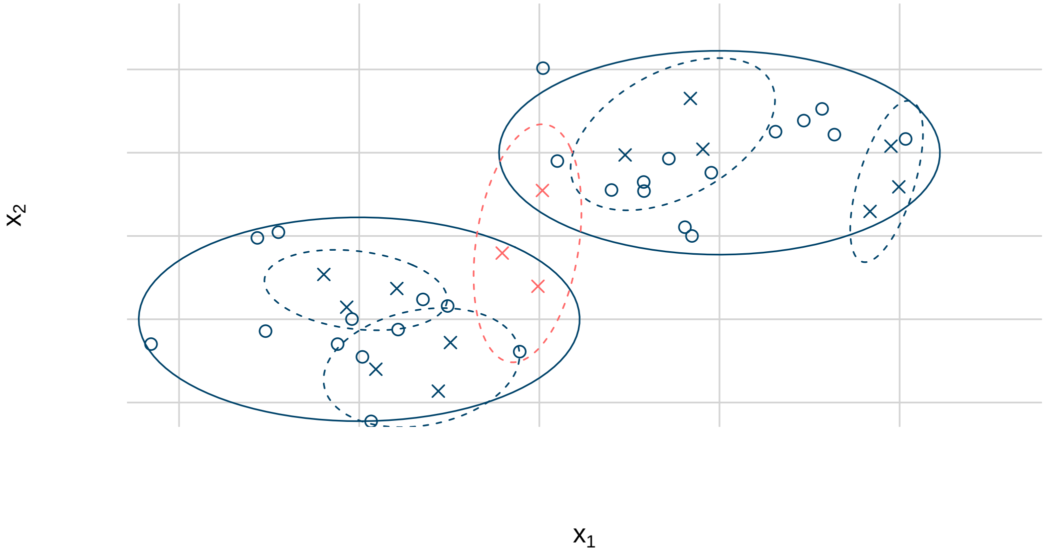


Figure 2.1: This plot demonstrates the concept of guided projections. The figure shows two group structures and the corresponding true covariance structures described by solid ellipses. Each small subset of three observations, represented by x-symbols, will provide a local approximation of this group structure as visualized by the dashed ellipses. The aim of the proposed guided projections approach is to provide a series of such selections, offering a good overall description of the present group structures. Each blue/dark set of x-symbols represents selections from the same group, providing useful information about the group separation, the red/light group represents a mixed selection, where the group structure is masked, i.e. observations from both groups are present inside the ellipse.

Our assumption that a selection of observations from a specific group can describe the remaining observations of this group better than the observations from other groups becomes especially interesting in the high dimensional case. While the overall distances of all observations become more and more similar with an increasing number of noise variables, we can utilize the projections onto the selected observations. By analyzing the proportion of distances in the projection space and in the orthogonal complement we can reduce the impact of the curse of dimensionality.

Therefore, after introducing the slightly adapted version of orthogonal and score distances from Hubert et al. (2005) in Section 2.2, in Section 2.2 we present a method for identifying a first starting projection. We then introduce a decision process, exchanging observations one by one depending on the similarity based on the introduced distances before discussing the properties of the projection sequence in Section 2.2.

Orthogonal and score distances

Let $\mathbb{P}$ denote the set of all orthogonal projections P from $\mathbb{R}^p$ onto $\mathbb{R}^{q-1}$, where p is the number of variables in the original space and $q-1$ the fixed dimension of the projected space. Each projection P can be represented by its projection matrix $\boldsymbol{V}'_P$, where $\boldsymbol{V}_P \in \mathbb{R}^{p \times q-1}, P \in \mathbb{P}$:

$$\forall P \in \mathbb{P} : \exists \boldsymbol{V}_P \in \mathbb{R}^{p \times q-1} : P(\boldsymbol{x}) = \boldsymbol{V}'_P \boldsymbol{x} \quad \forall \boldsymbol{x} \in \mathbb{R}^p \tag{2.1}$$

Given a projection $P \in \mathbb{P}$, we define the *orthogonal distance* (OD_P) of an observation $\boldsymbol{x} \in \mathbb{R}^p$ to a projection space defined by P, given a location $\boldsymbol{\mu}$ as

$$OD_P(\boldsymbol{x}) = ||\boldsymbol{x} - \boldsymbol{\mu} - V_P V'_P(\boldsymbol{x} - \boldsymbol{\mu})||, \tag{2.2}$$

and the *score distance* (SD_P) of $\boldsymbol{x}$, given the location $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}_P$ of the distribution in the projection space as

$$SD_P(\boldsymbol{x}) = \sqrt{(\boldsymbol{V}'_P(\boldsymbol{x} - \boldsymbol{\mu}))' \boldsymbol{\Sigma}_P^{-1} (\boldsymbol{V}'_P(\boldsymbol{x} - \boldsymbol{\mu}))}, \tag{2.3}$$

where $||.||$ stands for the Euclidean norm.

This definition slightly differs from the original concept presented in Hubert et al. (2005). Originally, the orthogonal and score distances intend to identify outliers from one main group of observations. Therefore, robust estimators of location and scatter are used to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_P$. Thus, the orthogonal and score distances are always interpreted with respect to the center and covariance structure of the majority of observations. The larger those distances get, the less likely the evaluated observation belongs to the same group. While the original work is based on the assumption of one main group of observations and a small subset of outliers, we assume the presence of multiple groups. In the latter situation, robust estimators calculated from less than 50% of the observations is not appropriate because in robust statistics a majority of observations has to be considered. Therefore, we alter the location and scatter estimates and estimate them from a small subset of observations where we try to select the observations from the same group.

Since $SD_P(\boldsymbol{x})$ and $OD_P(\boldsymbol{x})$ are both measures for similarity with respect to a location and covariance matrix, we define

$$\begin{aligned} OSD_P(\boldsymbol{x}) &= f(SD_P(\boldsymbol{x}), OD_P(\boldsymbol{x})) \qquad\qquad \boldsymbol{x} \in \mathbb{R}^p, \\ & f : \mathbb{R}^2 \rightarrow \mathbb{R} \\ & f \text{ monotonically increasing in } OD_P \text{ and } SD_P \end{aligned} \tag{2.4}$$

as a new univariate measure for similarity, always to be interpreted in reference to a location, a covariance matrix, and the dimensionality q of the projection space, which in case of Hubert et al. (2005) is given by the number of components used for the robust principal component analysis. Examples for such functions f are provided in Pomerantsev (2008).

We utilize a subclass of the presented projections defined by $\mathbb{P}$. Let $\mathcal{I}$ denote a set of q indices $\mathcal{I}_1, \dots, \mathcal{I}_q$ of $\boldsymbol{X}$, $\mathcal{I} \in \mathcal{P}(1, \dots, n) : |\mathcal{I}| = q$, where $\mathcal{P}$ is the power set. $\boldsymbol{X}_{\mathcal{I}}$ defines the matrix of scaled and centred selected observations. To scale and centre the observations, we use a location estimator

$$\hat{\boldsymbol{\mu}}_{\mathcal{I}} = \bar{\boldsymbol{x}}_{\mathcal{I}} = \frac{1}{q} \sum_{i \in \mathcal{I}} \boldsymbol{x}_i \tag{2.5}$$

and a scale estimator

$$\begin{aligned} \hat{\boldsymbol{\sigma}}_{\mathcal{I}} &= \left(\sqrt{Var(x_{\mathcal{I}_1 1}, \dots, x_{\mathcal{I}_q 1})}, \dots, \sqrt{Var(x_{\mathcal{I}_1 p}, \dots, x_{\mathcal{I}_q p})}\right)' \\ &= (\hat{\sigma}_{\mathcal{I}1}, \dots, \hat{\sigma}_{\mathcal{I}p})', \end{aligned} \tag{2.6}$$

where $\boldsymbol{x}_{\mathcal{I}_k} = (x_{\mathcal{I}_k 1}, \dots, x_{\mathcal{I}_k p})'$ denotes the k-th selected observation and Var is the empirical variance. $\boldsymbol{x}^c_{\mathcal{I}_k}$ denotes the centred observation $\boldsymbol{x}_{\mathcal{I}_k}$:

$$\boldsymbol{x}^c_{\mathcal{I}_k} = \boldsymbol{x}_{\mathcal{I}_k} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mathcal{I}}} = (x^c_{\mathcal{I}_k 1}, \dots x^c_{\mathcal{I}_k p})' \tag{2.7}$$

$$\boldsymbol{X}_{\mathcal{I}} = \left(\left(\frac{x^c_{\mathcal{I}_1 1}}{\hat{\sigma}_{\mathcal{I}1}}, \dots, \frac{x^c_{\mathcal{I}_1 p}}{\hat{\sigma}_{\mathcal{I}p}} \right)', \dots, \left(\frac{x^c_{\mathcal{I}_q 1}}{\hat{\sigma}_{\mathcal{I}1}}, \dots, \frac{x^c_{\mathcal{I}_q p}}{\hat{\sigma}_{\mathcal{I}p}} \right)' \right)' \tag{2.8}$$

The matrix $\boldsymbol{X}_{\boldsymbol{\mathcal{I}}}$ can be represented via a singular value decomposition:

$$\boldsymbol{X}_{\mathcal{I}} = \boldsymbol{U}_{\mathcal{I}} \boldsymbol{D}_{\mathcal{I}} \boldsymbol{V}'_{\mathcal{I}} \tag{2.9}$$

Note that the centring of the observations reduces the rank of the data matrix by one. Therefore, under the assumption of $q < p$, the rank of $\boldsymbol{V}'_{\mathcal{I}}$, which is equal to the rank of $\boldsymbol{X}_{\mathcal{I}}$, is $q-1$. This assumption is reasonable due to the focus on high-dimensional data. If $q < p$ does not hold, the dimension of the space is small enough such that a data transformation is not required. $\boldsymbol{V}'_{\mathcal{I}}$ from the decomposition in Equation (2.9) provides a projection matrix onto the space spanned by the q observations selected in $\mathcal{I}$. $\boldsymbol{V}'_{\mathcal{I}}$ represents an element of $\mathbb{P}$ since the dimension of the projection space is equal to the rank of $\boldsymbol{V}'_{\mathcal{I}}$ which is $q-1$. For such a projection, we can measure the similarity of any observation from $\mathbb{R}^p$ to the selected observations using the location estimation from Equation (2.5) and covariance matrix describing the covariance structure in the projection space, provided by the selection itself as follows:

$$\hat{\boldsymbol{\Sigma}}_{\mathcal{I}} = \frac{1}{q-1}(\boldsymbol{V}_{\mathcal{I}}\boldsymbol{X}'_{\mathcal{I}})(\boldsymbol{V}_{\mathcal{I}}\boldsymbol{X}'_{\mathcal{I}})' \tag{2.10}$$

Using the provided definitions and notation, we can define a univariate measure $OSD_{\mathcal{I}}(\boldsymbol{x})$ for similarity between an observation $\boldsymbol{x} \in \mathbb{R}^p$ and a set of observations, defined by $\mathcal{I}$:

$$OSD_{\mathcal{I}}(\boldsymbol{x}) = f(SD_{\mathcal{I}}(\boldsymbol{x}), OD_{\mathcal{I}}(\boldsymbol{x})), \qquad \boldsymbol{x} \in \mathbb{R}^p \tag{2.11}$$

$$SD_{\mathcal{I}}(\boldsymbol{x}) = \sqrt{(\boldsymbol{V}'_{\mathcal{I}}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{\mathcal{I}}))'\hat{\boldsymbol{\Sigma}}_{\mathcal{I}}^{-1}(\boldsymbol{V}'_{\mathcal{I}}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{\mathcal{I}}))}. \tag{2.12}$$

$$OD_{\mathcal{I}}(\boldsymbol{x}) = ||\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{\mathcal{I}} - \boldsymbol{V}_{\mathcal{I}}\boldsymbol{V}'_{\mathcal{I}}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{\mathcal{I}})|| \tag{2.13}$$

Guided projections algorithm

To create a sequence of non-random projections, we aim to identify a set of q observations, project all observations onto the space spanned by those q observations, and use $OSD_{\mathcal{I}}$ to measure the similarity between an observation $\boldsymbol{x} \in \mathbb{R}^p$ and the selected group of observations. In general, q is a configuration parameter which needs to be adjusted based on the data set to be analysed. Depending on both the expected number of observations in groups in the data structure and on the sparsity of the data set, we typically select q between 10 and 25. Out of the selected group of observations, we replace one observation after another by a new observation and therefore get a new projection space leading to new measures for similarity.

To identify a set q of starting observations, we exploit the Euclidean distances between all observations. Let d_{ij} denote the Euclidean distance $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = ||\boldsymbol{x}_i - \boldsymbol{x}_j||$

between observation $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. $d_{i(k)}$ denotes the k^{th} smallest distance from $\boldsymbol{x}_i$:

$$\min_{j\in\{1,\dots,n\}} d_{ij} = d_{i(1)} \leq \cdots \leq d_{i(n)} = \max_{j\in\{1,\dots,n\}} d_{ij} \tag{2.14}$$

Similar to the k-nearest-neighbor approach (e.g. Altman, 1992), we identify a dense group of q observations given by their indices $\mathcal{I}_1^0, \dots, \mathcal{I}_q^0$. Let $i_0 = \arg \min_{i\in\{1,\dots,n\}} d_{i(q)}$ denote the index of the observation with the smallest distance to the q^{th}-closest observation and $\boldsymbol{X}_{\mathcal{I}^0}$ the centered and scaled matrix of observations as defined in Equation (2.8):

$$\mathcal{I}^0 = \{\mathcal{I}_1^0, \dots, \mathcal{I}_q^0\} = \{j : d_{i_0 j} \leq d_{i_0(q)}\} \tag{2.15}$$

Note that in Equation (2.15) we assume that the number of observations in $\mathcal{I}^0$ is equal to q even though the second equality does not hold in general. In the case of ties, more than q observations may fulfill the criterion $d_{i_0 j} \leq d_{i_0(q)}$ of Equation (2.15). In such a case, we randomly select from the tied observations to be added to $\mathcal{I}^o$, such that q observations are selected.

During the determination of the sequence of projections, we always add the observation with the smallest OSD to the set of selected observations. To keep the dimensionality of the projected space constant, which ensures comparability of OSDs, we remove one observation each time we add an observation. Assuming the observations are ordered in a certain sense, each observation remains in the group of selected observations for q projections before it is removed again.

To identify the observation $\boldsymbol{x}_{i_1}$ to be added in the first step, we solely need to consider $OSD_{\mathcal{I}^0}$ defined in Equation (2.11). The set of observations available to be selected is defined by A^0:

$$A^0 = \{1, \dots, n\} \backslash \mathcal{I}^0 \tag{2.16}$$

$$i_1 = \arg \min_{i\in A^0} OSD_{\mathcal{I}^0}(\boldsymbol{x}_i) \tag{2.17}$$

To identify the observation to be removed, we need to provide an order of $\mathcal{I}^0$ first, which is determined by using leave-one-out distances (LOD). Sorting all elements from $\mathcal{I}^0$ decreasingly according to LOD provides the sorted starting observations and the first selected observation i_1 defined by I^1:

$$LOD_{\mathcal{I}^0}(j, i_1) = OSD_{\{\mathcal{I}^0\backslash\{j\}\}\cup\{i_1\}}(\boldsymbol{x}_j) \qquad \forall j \in \mathcal{I}^0 \tag{2.18}$$

$$I^1 = (j_1, \ldots, j_q, i_1) = (\iota_1^1, \ldots, \iota_{q+1}^1) \qquad j_k \in \mathcal{I}^0, k = 1, \ldots, q \tag{2.19}$$

$$LOD_{\mathcal{I}^0}(j_1, i_1) \geq \cdots \geq LOD_{\mathcal{I}^0}(j_q, i_1)$$

$$A^1 = A^0\backslash i_1 = \{1, \ldots, n\}\backslash I^1 \tag{2.20}$$

$$\mathcal{I}^1 = \{\mathcal{I}^0\backslash j_1\} \cup \{i_1\} \tag{2.21}$$

$\mathcal{I}^1$ and A^1 again denote the index sets of observations selected in the first step and the remaining observations available for selection after the first step, respectively. After this first step, for any following step, in general for the s^{th} step, two projections, represented by $\mathcal{I}_L$ and $\mathcal{I}_R$ are relevant for selecting a new observation:

$$\mathcal{I}_L = \{\iota_1^1, \ldots \iota_{q-1}^1\} \tag{2.22}$$

$$\mathcal{I}_R = \{\iota_2^1, \ldots \iota_q^1\} \tag{2.23}$$

The notation L and R comes from the left and right end of the series of indexes in I^1 representing the first and the last q observations.

The reason to consider multiple projections is based on the assumption that we start from a dense region of the data distribution. By adding one observation we move away from this dense region in one direction. Once the observations at the border of this direction have been reached, the remaining observations are far away from the selection, yet close to the initially selected observations in the center. Figure 2.2 visualizes this issue.

Since we aim at a series of projections as consistent as possible, we always select the projection with the smallest distance. In the showcase in Figure 2.2 we show the selection of $\mathcal{I}^0$ and the first observation i_1 in plot (a). Plot (b) to (f) represent the steps 1 to 5 of our procedure. The two ellipses represent the OSD, based on $\mathcal{I}_L$ and $\mathcal{I}_R$ respectively. The choice of observation to be added is marked as a red/light dot. Starting from plot (d) we notice that the selection $\mathcal{I}_R$, represented by the observations marked with an R, requires a large $OSD_{\mathcal{I}_R}$ to add an additional observation. Therefore, starting from (d) we add observations to the left end of the series I^s. In general it makes sense to consider all previous projections. However, to create a series of projections where we can look for structural changes and visualize a development, we limit ourselves to $\mathcal{I}_L$ and $\mathcal{I}_R$.

Depending on the smallest OSD to either $\mathcal{I}_L$ or $\mathcal{I}_R$, the newly added observation, the new set of sorted observations I^s, and the new set of available observations for future

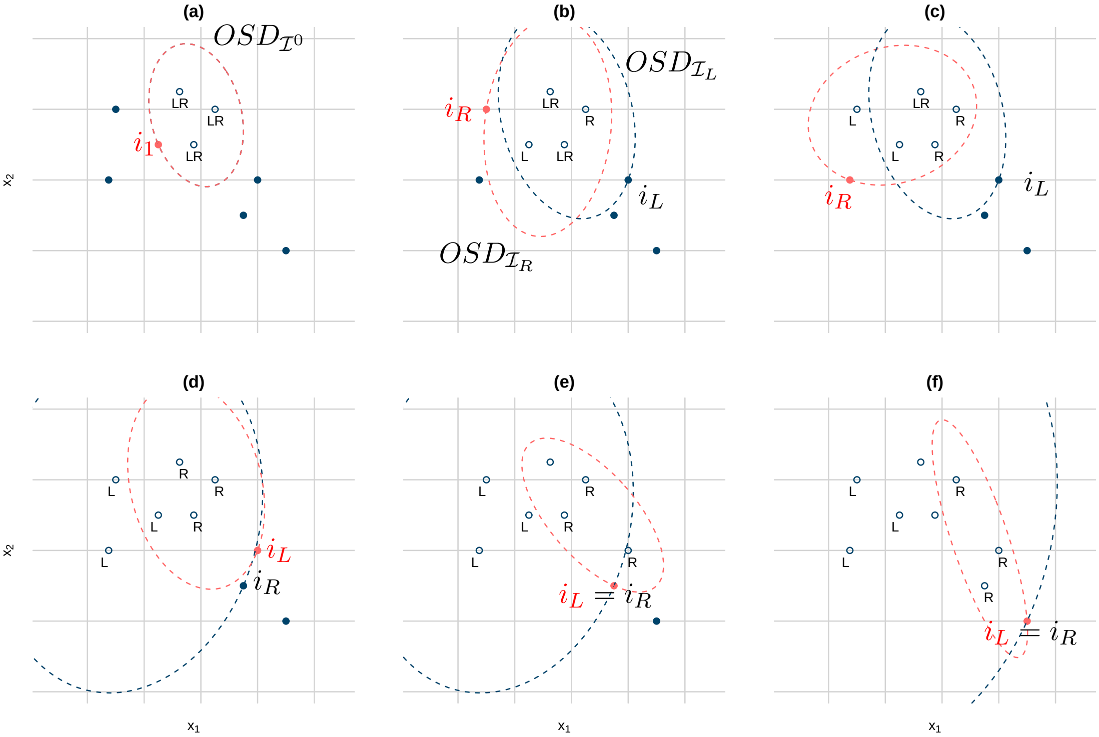


Figure 2.2: Visualization of the selection procedure. To keep the observations in a constant location for each plot we use a two-dimensional space. The distances $OSD_{\mathcal{I}}$ to a selection of observations $\mathcal{I}$ are represented by dashed ellipses. The red/light ellipse represents the smaller distance and therefore the choice for the next observation to be selected. If an observation is part of $\mathcal{I}_L$ or $\mathcal{I}_R$, it is marked with an L or R respectively. Solid points represent observations which have not been selected so far, empty circles have been selected before or are part of a current selection. The next observation to be added to the sequence is marked by a red/light dot.

projections A^s are determined for the s^{th} step, provided $s \geq 2$ holds:

$$i_L = \arg \min_{i \in A^{s-1}} OSD_{\mathcal{I}_L}(\boldsymbol{x}_i) \tag{2.24}$$

$$i_R = \arg \min_{i \in A^{s-1}} OSD_{\mathcal{I}_R}(\boldsymbol{x}_i) \tag{2.25}$$

$$I^s = \begin{cases} (i_L, \iota_1^{s-1}, \ldots, \iota_{s-1+q}^{s-1}), & OSD_{\mathcal{I}_L}(\boldsymbol{x}_{i_L}) \leq OSD_{\mathcal{I}_R}(\boldsymbol{x}_{i_R}) \\ (\iota_1^{s-1}, \ldots, \iota_{s-1+q}^{s-1}, i_R), & \text{else} \end{cases} \tag{2.26}$$

$$= (\iota_1^s, \ldots, \iota_{s+q}^s)$$

$$A^s = \{1, \ldots, n\} \backslash I^s \tag{2.27}$$

I^s is a superset of I^{s-1} for all $s \geq 1$ and provides all information about the sequence of previous projections. In total, there are $n-q+1$ projections available which are determined after $n-q$ steps. Therefore, we can define the guided projections GP based on I^{n-q} alone.

$$GP(\boldsymbol{x}) = (GP_1(\boldsymbol{x}), \dots, GP_{n-q+1}(\boldsymbol{x})) \tag{2.28}$$

$$GP_j(\boldsymbol{x}) = OSD_{\{\iota_j^{n-q}, \dots, \iota_{j+q-1}^{n-q}\}}(\boldsymbol{x}) \qquad j \in 1, \dots, n-q+1 \tag{2.29}$$

As a result, we receive one series of measures for each observation. Whenever the measure is small, the observation is likely from the same group as the respective selected observations. Thus, structures in data can be identified by looking for similar behaviour in $GP(\boldsymbol{x})$.

Additional insight on guided projections

Choice for OSD: A variety of useful OSDs can be defined for guided projections. Some possibilities to combine orthogonal and score distances to a univariate measure are presented in Pomerantsev (2008). The best choice for OSD depends on the distribution of the data structure. When dealing with high-dimensional data, especially sparse data where groups are best described by different variables, the orthogonal distance contributes more to the group separation than the score distance. When dealing with low-dimensional data, the opposite is true. Therefore, the decision on the most appropriate OSD needs to be met for each analysis individually depending on the underlying data characteristics. Given the fact, that we deal with high-dimensional data and for reasons of simplicity we restrict the choice of OSD for this work to the orthogonal distance, utilizing the properties of the complement of the projection space which is often ignored (e.g. Gattone and Rocci, 2012; Ilies and Wilhelm, 2010).

$$OSD_{\mathcal{I}}(\boldsymbol{x}) = OD_{\mathcal{I}}(\boldsymbol{x}) \tag{2.30}$$

Two-dimensional visualization of guided projections: Each projection results in a representation of all observations by orthogonal and score distances which can be visualized in a two-dimensional plane. The series of projections $GP(\boldsymbol{x}) = (GP_1(\boldsymbol{x}), \dots, GP_{n-q+1}(\boldsymbol{x}))$ typically starts with observations from one group. Therefore, the observations to be selected in the following steps are observations which are similar to the selected observations and thus likely from the same group. By replacing only one observation per projection, we achieve a high correlation between OSDs created by consecutive projections. Each step represents a slight rotation of the two-dimensional

OD-SD-plane, the observations are projected onto. This behaviour is represented in Figure 2.3 where the projection space is always spanned by 10 observations.

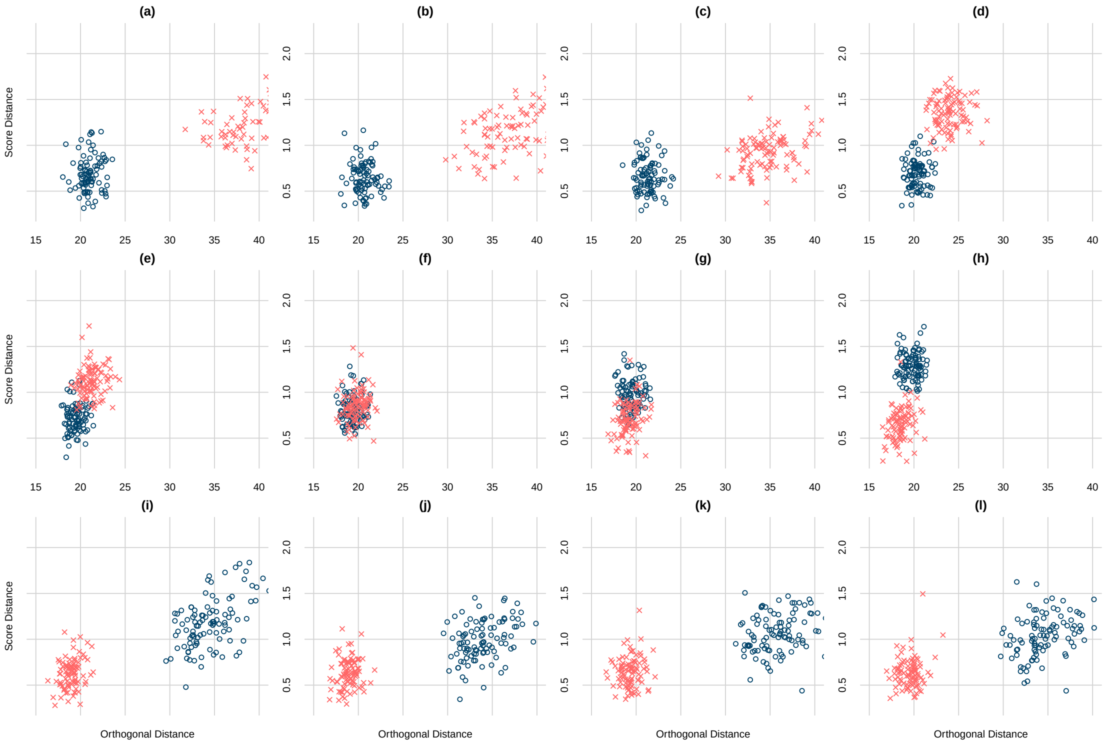


Figure 2.3: Subset of the series of projections for simulated data, consisting of two groups with 100 observations each, generated from two different fifty-dimensional normal distributions. The groups are visualized with red circles and blue plus symbols. Each plot represents one step of guided projections, where all observations are projected onto the space spanned by 10 selected observations.

In Figure 2.3, the plots (a) to (d) show projections where all selected observations are taken from the blue (circles) group. Figure (e) shows the first time where an observation from the red (x-symbols) group is selected. Therefore, the distances for the red group start decreasing. In plot (g) the majority of selected observations is taken from the red group. In plot (h) only one blue observation remains in the selection. Starting from plot (i) in the third row, the groups are separated again since all observations for the projection are selected from the red group.

Specific behaviour of OD and SD for guided projections: Assume one of the projection matrices $\boldsymbol{V}_{\mathcal{I}^s}$, where $\mathcal{I}^s$ represents the selected observations in the s^{th} step. Let us consider plot (a) of Figure 2.3 as an example. One could argue that critical values can be directly provided separating the red from the blue group for this projection, making the rest of the sequence obsolete. Details for the determination of those critical values for orthogonal distances and score distances are provided in Olkin (1992) and Pomerantsev (2008). The problem with this argument can be described as follows.

The possibility of separating two or more groups is based on the assumption that all selected observations are taken from the same group and an estimation of location and the covariance matrix based on this group only can be provided. Therefore, such a decision needs to be made after the initial selection. Thus, only q observations are available for the required estimation of location and covariance in the $q-1$ dimensional space. This estimation cannot be provided due to the following properties for all $s \in \{0,\dots,n-q+1\}$:

$$OD_{\mathcal{I}^s}(\boldsymbol{x}) = 0, \quad \Longleftrightarrow \quad \boldsymbol{x} \in span(\{\boldsymbol{x}_i : i \in \mathcal{I}^s\}) \tag{2.31}$$

$$SD_{\mathcal{I}^s}(\boldsymbol{x}_i) = \frac{q-1}{\sqrt{q}}, \quad \forall i \in \mathcal{I}^s \text{ and } q = |\mathcal{I}^s|, s \in \{0,\dots,n-q+1\} \tag{2.32}$$

The proof of these statements can be found in the Appendix. Since there is no variation in the orthogonal and score distance for the selected observations for $\mathcal{I}^s$, the parameters for the critical values, which are based on the variation, cannot be derived. The orthogonal and score distances for observations of $\mathcal{I}^s$ are extremely distorted and do not follow the expected theoretical distribution of $OD_{\mathcal{I}^s}$ and $SD_{\mathcal{I}^s}$.

Visualization of guided projections

Guided projections can be visualized in a diagnostic plot. In such a plot, the series of $OSDs$ is shown for each observation. As an example, consider the data set used in Figure 2.3. Due to Equation (2.31), any selected observation will have an orthogonal distance of zero for certain projections, and therefore in our application an OSD of zero, as defined in Equation (2.30).

Figure 2.4 shows the change in OSD by modifying the projection direction, which is achieved by substituting one observation in the selection spanning the projection space. Each observation is selected once. Therefore, for each projection, one observation drops to zero from a non-zero level and one observation goes up to a non-zero level.

Note that this diagnostic plot contains a visualization of the transformed data. As the dimensionality of the transformed space might remain high – depending on the number of observations – a visualization of individual coordinates is often not helpful. In our case we additionally have a very high dependence between variables. Therefore, we use line plots to track each observation through all variables/projections, assuming that similar observations from the same group will follow a similar structure in a functional context.

Given the 200 observations, selecting 10 observations for each projection results in a total number of 191 projections. For the first 85 projections, all observations are selected from group one (blue dashed lines). During this procedure, no significant changes occur. Starting with the 86^{th} projection though, which is the same projection as plot (e) of Figure 2.3, we see some mixed projections and a structural change in OSD for both groups. The $OSDs$ of the observations from one group drop to a lower level while the $OSDs$ of the observations from the other group increase.

Such a structural change in guided projections clearly indicates the presence of a second group in the analyzed data structure. In general, observations whose OSD stays close to each other during the whole sequence of projections are expected to belong to the same group.

2.3 Simulations

The aim of this section is to measure the effect of data transformations on the separation of present groups in simulated data. We consider the data transformation approaches introduced in Section 2.1: Classical PCA [PCA], Sparse PCA [SPC], Diffusion Maps [DIFF], and Random Projections [RP]. We use two simulated multivariate normally distributed data setups to measure the impact of noise variables as well as the impact of differences in covariance structures. The effects themselves are measured by a selection of common cluster validity measures.

Evaluation Measures

An overview of internal evaluation indices is presented in Desgraupes (2013). All measures can be directly accessed through the R-package *clusterCrit* (Desgraupes, 2016). The provided indices depend on various measures like total dispersion, within-group scatter and between-group scatter. Some of those measures heavily depend on the dimensionality of the transformation space. Thus, depending on the design of the validity measures, a lower dimensional space is often preferred over a high-dimensional space even though the quality of separation decreases with decreasing dimensionality. We use two simulations visualized in Figure 2.5 to demonstrate this aspect. In the first setup we generate k simulated independent normally distributed variables. Group one uses a

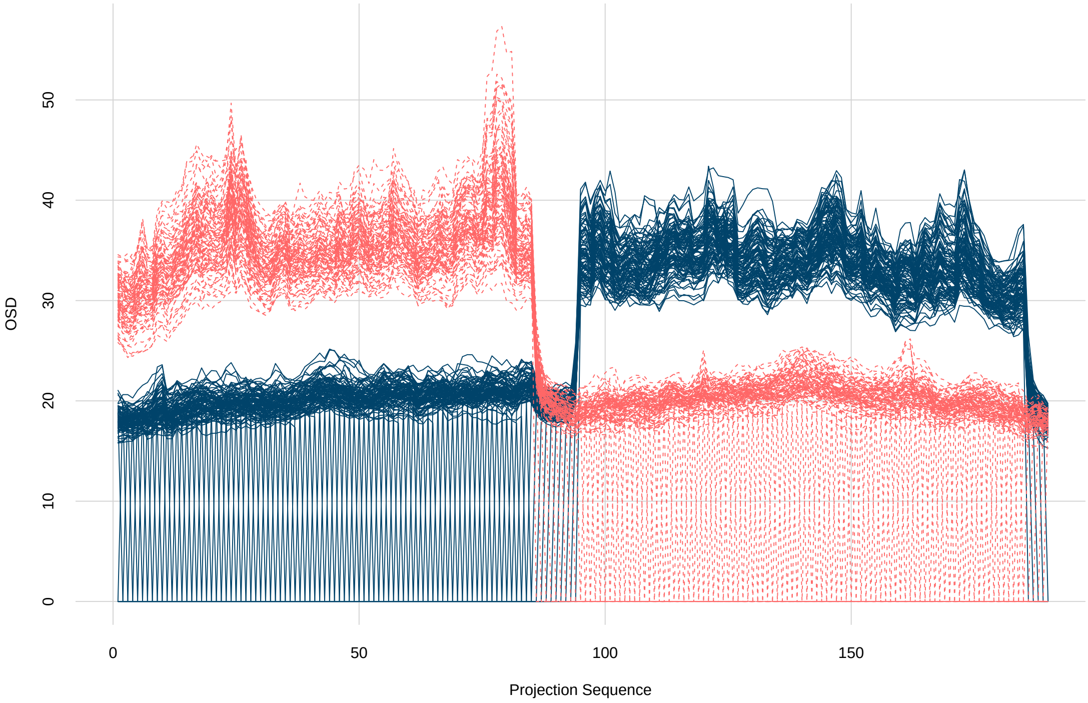


Figure 2.4: Diagnostic plot utilizing guided projections for the simulated data from Figure 2.3. The colors represent the two clusters, originally located in a fifty-dimensional space. The projection index on the x-axis stands for the index j of $GP_j(\boldsymbol{x})$ of Equation (2.29). For each observation we can follow the change in OSD while slightly changing the projection direction. Similar observations are represented in parallel lines, close to each other.

mean value of 1, while group two uses mean values of -1. The more variables are used, the better the expected separation should be. The second simulation setup always uses 50 of those variables and in addition adds k normally iid variables with mean value of zero for both groups. Those non-informative variables theoretically reduce the quality of the group separation. For a selection of popular validation measures we simulate those two setups, varying k between 1 and 350. Note that not all original measures should be maximized. Therefore we transformed all measures which should be minimized, like the Banfeld Raftery index, in such a way that they are to be maximized to simplify Figure 2.5.

The decision on which indices to consider for the evaluation is based on the simulation results. Validity measures with a non-monotonous development for the second setup (Xie Beni, Dunn Index and GDI) are excluded. Also measures with a decreasing

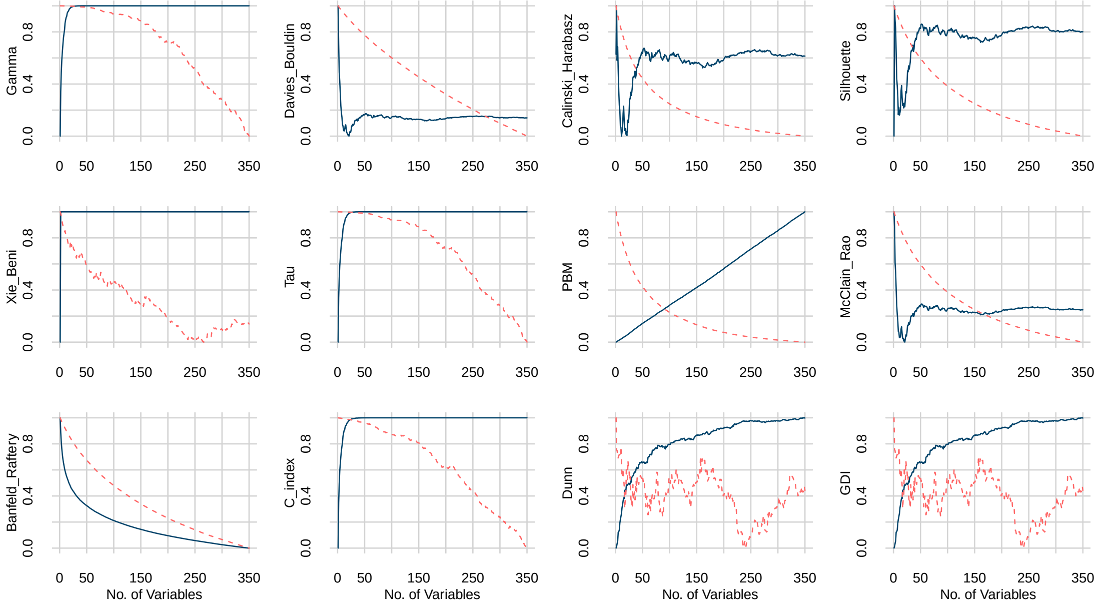


Figure 2.5: The solid (black) line refers to the previously described setup one (informative variables only), the dashed (red) line to setup two (including non-informative variables). The transformed validity measures for both setups have been independently scaled to the interval $[0, 1]$ for a better visualization. Both lines are depending on the number of variables related to the respective setup. In total, 1000 observations are simulated for each simulation setup and group to evaluate the considered measures.

development in the first setup (Davies Bouldin and Banfield Raftery) or a large fluctuation range in setup 1 (Calinski Harabasz and McClain Rao) have been excluded. Among the remaining validation measures, based on their popularity we decided to include the *Gamma* index (Baker and Hubert, 1975), the *Silhouette* index (Rousseeuw, 1987), and the *C index* (Hubert and Schultz, 1976) for the evaluation of the group structure of data transformations.

In addition to the selected validity measures, we are interested in the effect of data transformations before applying clustering procedures. Therefore, we perform hierarchical Ward clustering (Ward Jr, 1963) after applying the data transformations and evaluate the clustering result using the *F-measure* (Larsen and Aone, 1999).

Parameter optimization

A number of data transformations has been presented in Section 2.1. Each of them is depending on one or more configuration parameters, leading to different quality of the projections and thus directly affecting the validation measures.

All methods are optimized for each data set individually. For each parameter we set upper and lower boundaries in which we optimize the parameters for each specific data transformation method and validation measure. This way we make the methods comparable since a specific parameter set might work better for one transformation than for another providing an unfair advantage for one method. The same is true for specific validation measures. The optimization itself is performed by allowing a discrete number of parameters within their boundaries and performing and evaluating each combination of parameters. Hereinafter we present parameters to be optimized for the compared data transformations.

PCA: For principal component analysis the only parameter that needs to be adjusted is the proportion of variance of $\boldsymbol{X}$ which should be represented in the projection space. This can be translated to the number of components considered to span the projection space. This dimension is optimized for any number between 1 and the rank of $\boldsymbol{X}$, which is the maximum number of possible components.

SPC: The considered sparse principal component analysis by Witten et al. (2009) uses two optimization parameters. The first parameter is the number of sparse components, the second parameter the degree of sparsity defined by the sum of absolute values of elements of the first right singular vector of the data matrix. The number of components is optimized equivalently to PCA. The sparsity parameter is optimized between 1 and the square root of the number of columns of the data as recommended in Witten et al. (2009).

DIFF: Diffusion maps utilize an ϵ-parameter to describe the degree of localness in the diffusion weight matrix. A recommended starting point is $2med_{knn}^2$, where med_{knn}^2 represents the squared median of the k^{th} nearest neighbour. By varying k between 0.5% and 3.5% of the number of observations, which extends the recommended 1% to 2%, we adjust the ϵ-parameter. The number of components to describe the transformation space is adjusted in the same way as for PCA.

RP: For random projections we repeatedly project the observations on a k dimensional projection space 500 times. We optimise k between 1 and k_{max}. The upper limit k_{max} is the maximum number of components available in PCA for real data and the number of informative variables for simulated data.

GP: For guided projections, only one parameter needs to be adjusted, namely the number of observations in each projection. We propose to optimize this number between 5 and 30.

While performing hierarchical clustering, the number of clusters emerges as an additional configuration parameter. To provide a fair comparison, we allow any possible number of clusters between 1 and the number of observations, and report the best possible result. Figure 2.6 visualizes the optimization for the Gamma index and the F-measure for an exemplary data set for SPC.

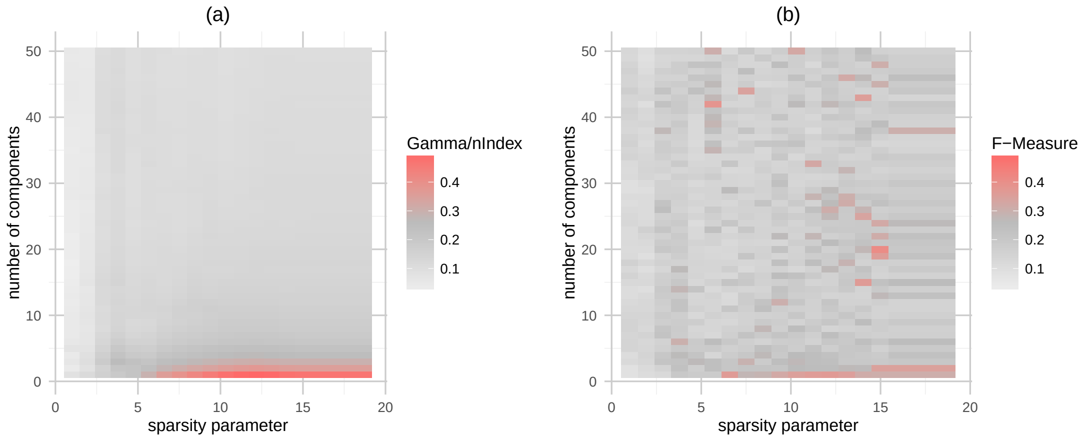


Figure 2.6: The optimization procedure for SPC is visualized. On the x-axis the sparsity parameter is presented, on the y-axis the number of sparse components. The quality of each parameter combination is presented by the color of the respective combination. Red/dark corresponds to a high value of the considered validity measure, grey/light to a low value. Figure (a) shows the optimization for the Gamma index, Figure (b) for the F-measure. For each index, the individual optimum is selected. The sparsity parameter for the F-measure is selected slightly larger than for the Gamma index. The optimal F-measure requires 20 sparse principal components while the Gamma index uses one.

Note that we do not compare with projection pursuit, since the aim of this approach is to identify a low-dimensional projection (one to three dimensions) revealing the group structure of the data. We evaluated the final projection of a guided tour from Wickham et al. (2011) and found no significant difference to the performance of random projections. Such an evaluation is unfair though since two-dimensional projections are being compared with methods that incorporate multiple or higher dimensional projections. Therefore, projection pursuits are not considered for the full evaluation.

First simulation setting

The first simulated data setup consists of two groups of observations, where the observations are drawn from different multivariate normal distributions $X_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and

$X_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. The parameters are as follows:

$$\boldsymbol{\mu}_1 = (\mathbf{0}_{50}, \mathbf{0.5}_{50}, \mathbf{0}_{250})' \tag{2.33}$$

$$\boldsymbol{\mu}_2 = (\mathbf{0}_{r}, -\mathbf{0.5}_{50}, \mathbf{0}_{300-r})' \tag{2.34}$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} \boldsymbol{I}_{50} & 0 & 0 \\ 0 & \boldsymbol{\Sigma}_{50}^{rand_2} & 0 \\ 0 & 0 & \boldsymbol{I}_{250} \end{pmatrix} \tag{2.35}$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} \boldsymbol{I}_{r} & 0 & 0 \\ 0 & \boldsymbol{\Sigma}_{50}^{rand_2} & 0 \\ 0 & 0 & \boldsymbol{I}_{300-r} \end{pmatrix} \tag{2.36}$$

In (2.33) to (2.36), $\mathbf{0}_r$ and $\mathbf{0.5}_r$ denote a row-vector of length r with 0 or 0.5 entries, respectively. $\boldsymbol{I}_r$ denotes an r-dimensional unit matrix and $\boldsymbol{\Sigma}_{50}^{rand_1}$ and $\boldsymbol{\Sigma}_{50}^{rand_2}$ represent randomly generated, fifty-dimensional covariance matrices. In order to simulate these matrices we use the R package *clusterGeneration* by Qiu and Joe (2015).

By varying r we modify the subspace where the informative variables are located. For $r = 51$, a 50 dimensional informative subspace is present but this subspace is informative for both present groups. For other values of r, the informative variables of X_2 are getting shifted away from the informative variables from X_1. An interesting aspect of this setup is the fact that the expected difference between the two groups changes with r. The expected distance between X_1 and X_2 is based on the number of informative variables as well as on the expected distance for each informative variables. In fact, the expected distances turn out to be

$$E(||X_1 - X_2||) = \sqrt{50 - \frac{1}{2} min(50, |51 - r|)}. \tag{2.37}$$

This distance is maximized for $r = 51$ and is decreasing with any changes in r leading to the expectation of a maximized separation for $r = 51$. For each r between 1 and 100, we repeatedly simulate the setup 25 times. For each simulated data set we report optimized validation measures.

Each plot in Figure 2.7 shows a similar individual behaviour for each method. The performance of principal component based methods (columns one, two and five) increases with increasing expected distance between X_1 and X_2, which is described in Equation (2.37), while the quality of guided projections increases with additional informative variables and especially with an increase in the shift of informative variables. This behaviour

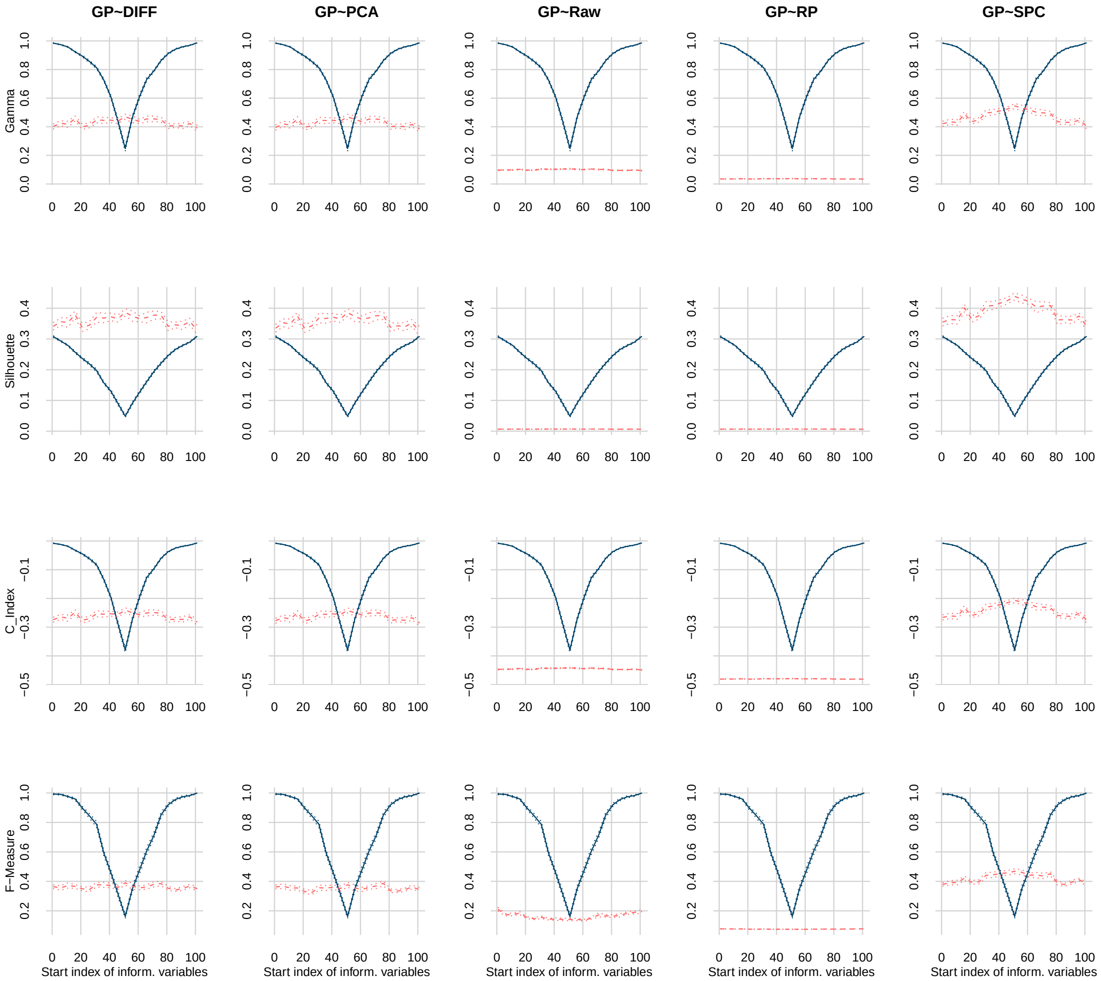


Figure 2.7: For each selected validation measure, we show the mean performance of guided projections (solid blue/dark line) and one other transformation (dashed red/light line) as well as their respective standard errors (dotted lines). The performance of no transformation is shown by the *Raw* category. The start index of the informative variable on the x-axis refers to the parameter r of Equation (2.34) and (2.36).

by guided projections occurs due to the following properties: When observations from the same group are selected, the subspace spanned by those observations describes the informative variables of those observations. Therefore, if the second group consists of different informative variables, the difference in orthogonal distances increases, which are used here for OSD. If the informative variables are the same though, the differences in the orthogonal space are expected to be the small. Since we completely ignore the score distances, guided projections are outperformed by principal component based methods in this case. This feature is visible for all considered validation measures. Most validation measures indicate that guided projections clearly outperform the other projection methods if the number of informative (shifted) variables increases. An exception is the Silhouette index, which declares guided projections as the worst method. However, this might be quite specific in a two-group setting.

Second simulation setting

The second simulated data setup uses three groups drawn from multivariate normally iid stochastic variables $X_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $X_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ and $X_3 \sim N(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$ with the following parameters:

$$\boldsymbol{\mu}_1 = (\mathbf{1}_{25}, \mathbf{1}_{25}, \mathbf{0}_{25}, \mathbf{0}_r)' \tag{2.38}$$

$$\boldsymbol{\mu}_2 = (\mathbf{1}_{25}, \mathbf{0}_{25}, \mathbf{1}_{25}, \mathbf{0}_r)' \tag{2.39}$$

$$\boldsymbol{\mu}_3 = (\mathbf{0}_{25}, \mathbf{1}_{25}, \mathbf{1}_{25}, \mathbf{0}_r)' \tag{2.40}$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} \boldsymbol{\Sigma}_{25}^{rand_{1,1}} & \boldsymbol{\Sigma}_{25}^{rand_{1,2}} & 0 & 0 \\ \boldsymbol{\Sigma}_{25}^{rand_{1,3}} & \boldsymbol{\Sigma}_{25}^{rand_{1,4}} & 0 & 0 \\ 0 & 0 & \boldsymbol{I}_{25} & 0 \\ 0 & 0 & 0 & \boldsymbol{I}_r \end{pmatrix} \tag{2.41}$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} \boldsymbol{\Sigma}_{25}^{rand_{2,1}} & 0 & \boldsymbol{\Sigma}_{25}^{rand_{2,2}} & 0 \\ 0 & \boldsymbol{I}_{25} & 0 & 0 \\ \boldsymbol{\Sigma}_{25}^{rand_{2,3}} & 0 & \boldsymbol{\Sigma}_{25}^{rand_{2,4}} & 0 \\ 0 & 0 & 0 & \boldsymbol{I}_r \end{pmatrix} \tag{2.42}$$

$$\boldsymbol{\Sigma}_3 = \begin{pmatrix} \boldsymbol{I}_{25} & 0 & 0 & 0 \\ 0 & \boldsymbol{\Sigma}_{25}^{rand_{3,1}} & \boldsymbol{\Sigma}_{25}^{rand_{3,2}} & 0 \\ 0 & \boldsymbol{\Sigma}_{25}^{rand_{3,3}} & \boldsymbol{\Sigma}_{25}^{rand_{3,4}} & 0 \\ 0 & 0 & 0 & \boldsymbol{I}_r \end{pmatrix} \tag{2.43}$$

Similar as before, $\mathbf{0}_r$ and $\mathbf{1}_r$ represent vectors of length r with 0 and 1 entries, respectively. The matrices $\begin{pmatrix} \boldsymbol{\Sigma}_{25}^{rand_{i,1}} & \boldsymbol{\Sigma}_{25}^{rand_{i,2}} \\ \boldsymbol{\Sigma}_{25}^{rand_{i,3}} & \boldsymbol{\Sigma}_{25}^{rand_{i,4}} \end{pmatrix}$ from Equation (2.41) to (2.43) represent randomly created 50 dimensional covariance matrices. Therefore, $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$ represent covariance matrices too. The first 75 variables are informative variables, while the remaining r variables are non-informative. With increasing r, the separation between the present groups gets increasingly masked. The focus of this simulation setup is the robustness of data transformations towards non-informative variables.

The parameter r is varied between 0 and 1250 leading to a 75 to 1325 dimensional space. For each setup we compare three groups of 100 simulated observations per group. 25 repeated simulations are performed for each evaluated r by randomly creating different covariance matrices.

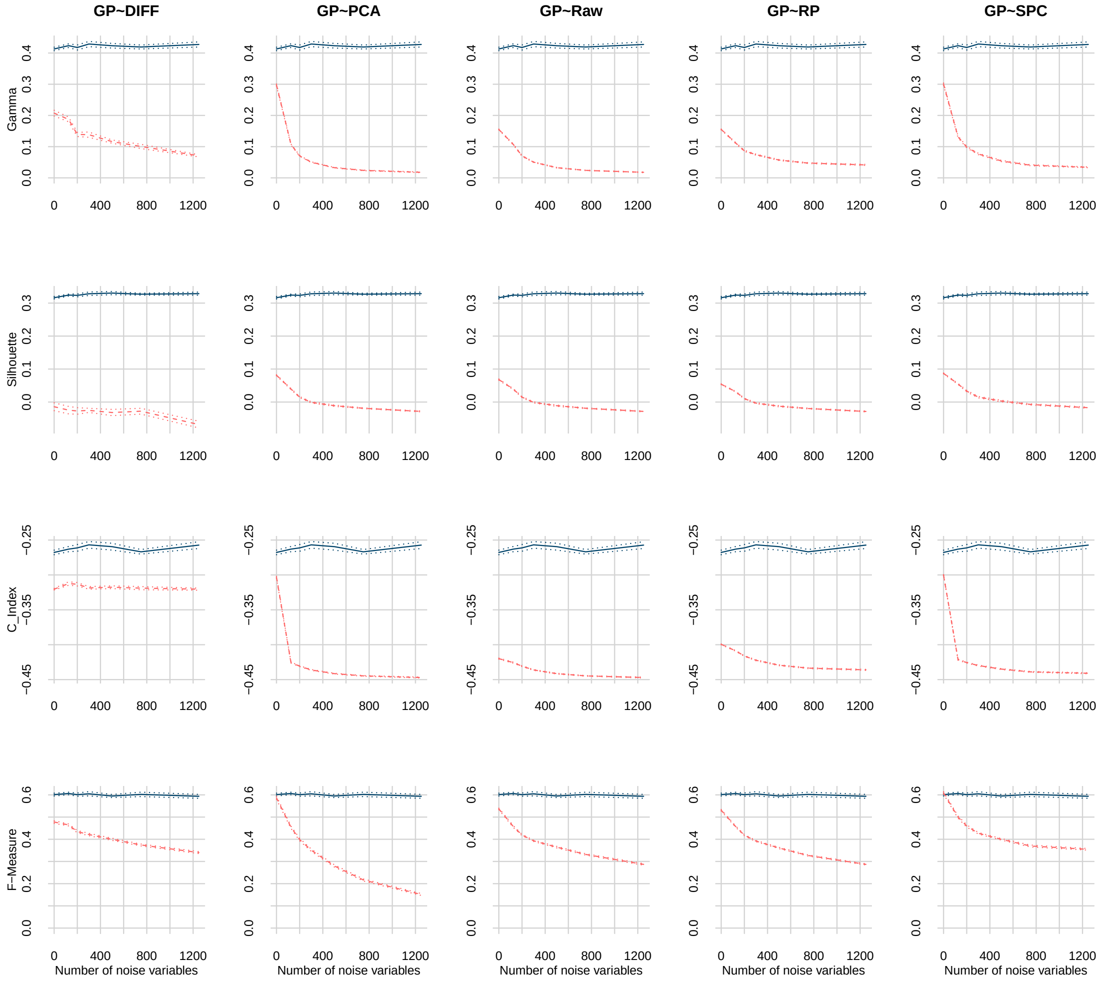


Figure 2.8: For the selected validation indices, we analyze the impact of additional noise variables. The mean optimal performance and the respective standard error is visualized for an increasing number of noise variables for guided projections (solid blue/dark line) and one compared method (dashed red/light line). In general we expect a decrease in quality with increasing noise variables.

Figure 2.8 shows the effect of increasing r non-informative variables on the quality of the considered data transformation, based on the different validation measures. The number of non-informative variables r refers to r in Equation (2.38) to (2.43). For each method and validation measure but guided projections for all measures and diffusion maps for C-index index we see the quality of transformations being affected in the same way as the level of separation is affected for the untransformed data. For guided projections though, there seems to be no impact from additional non-informative variables. Compared to setup 1 where only two groups were present, guided projection clearly outperform all other transformation regardless of the validation index.

Note that the data setup refers to a situation similar to the first simulated setup where a shift between the informative variables is present. Therefore, we deal with a *good* situation for guided projections due to the choice of $OSD_{\mathcal{I}}(\boldsymbol{x}) = OD_{\mathcal{I}}(\boldsymbol{x})$ from Equation (2.30). Assuming that the distances between the observations become more similar with an increasing number of variables, the proportion of the distance which is located in the projection space (described by OSD) becomes important. The different covariance structures combined with the shifted informative variables of the present groups lead to strong differences between subspaces associated with the different groups and therefore to larger differences in the proportion of distances represented in those subspaces. This effect is not addressed by the other transformations which mainly rely on differences in the mean values (see setup 1), and therefore are strongly affected by noise variables masking this effect.

2.4 Real-world data sets

The first real-world dataset we take into consideration is the fruit data set which is often used to demonstrate the stability of robust statistical methods (e.g. Hubert and Van Driessen, 2004). It consists of 1095 observations of spectra of three different types of melon labelled with D, M and HA, presented in a 256 dimensional space of wavelength. It is known that the groups consist of subgroups due to changed illumination systems and changed lamps while cultivating the plants. Since we do not have labels for the subgroups, we only consider the originally provided labels. For those labels we randomly select 100 observations per group repeatedly 50 times.

Figure 2.9 evaluates the separation of groups based on the Gamma index, the Silhouette measure, the C-index and the F-measure. Guided projections clearly outperforms all other transformations as well as the untransformed data situation. Only when measured with the C-index, diffusion maps perform better than guided projections. For all other validation measures though, diffusion maps perform below average.

In addition to showing that the presentation of the observations with guided projections leads to a better group separation, we can visualize the transformation using the diagnostic plot. Figure 2.10 visualizes the transformation for a selection of 100 randomly selected observation. We use this selection in order to reduce the number of overlapping lines to provide clearer insight into the group structure. First, we notice a strong fluctuation of OSD indicating the presence of additional group structure, which in the case of the red/solid grey group is obvious. Additionally we can identify outliers from the present groups, e.g. by blue/light dotted observation. The presence of outliers and additional group structure for this data set is well known (e.g. Hubert and Van Driessen, 2004). These subgroups, however, are not documented, and therefore an evaluation of the additionally observed group structure is not possible.

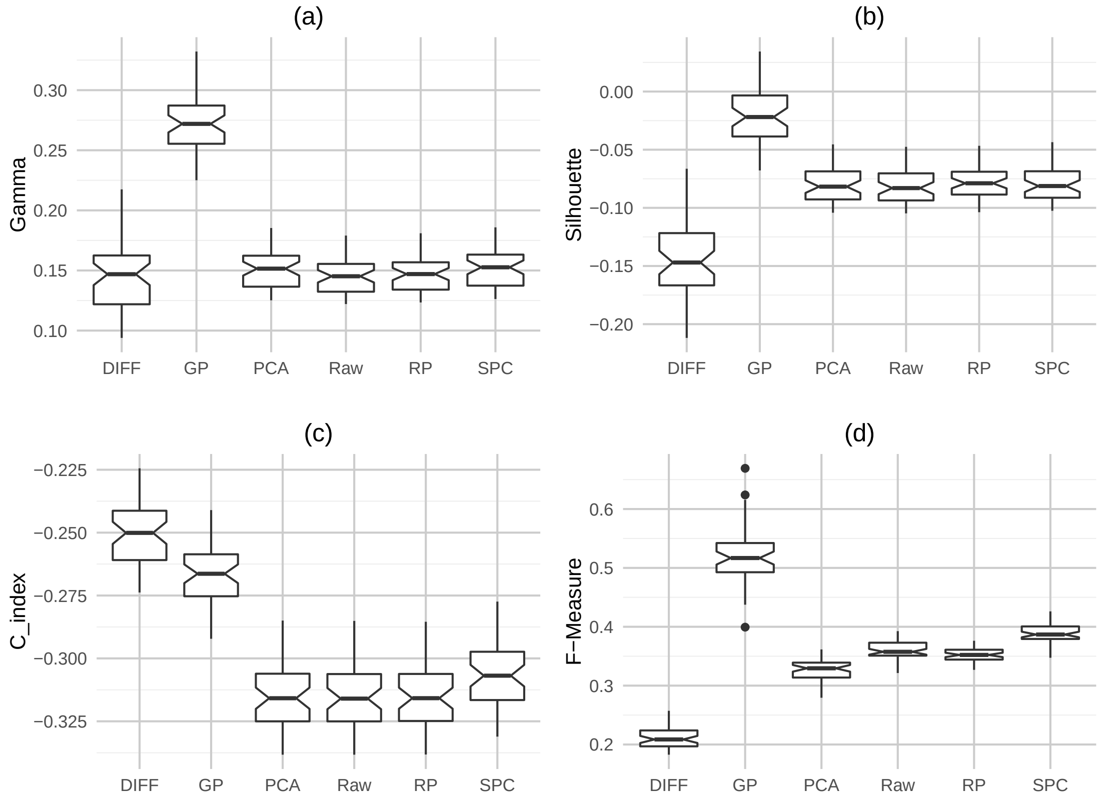


Figure 2.9: The performance of data transformations is measured by four different validation measures. 50 randomly selected subsets of the fruit data set are evaluated, based on the originally provided labels.

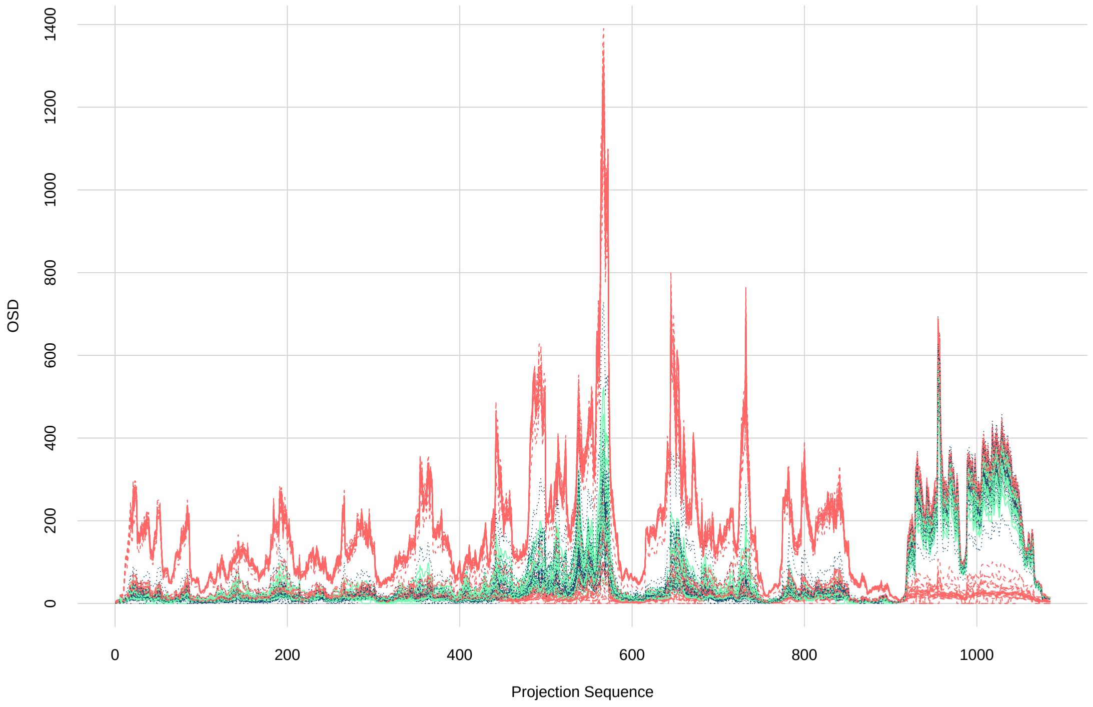


Figure 2.10: Diagnostic plot for the full fruit data set. Three groups are present. Additional group structure can be adumbrated. Especially the presence of outliers is evident. The observed group structure reflects the changes in the illumination system while collecting data from melon growth as described in various publications (e.g. Hubert and Van Driessen, 2004).

To show that the identification of additional group structures and outliers can be achieved, utilizing diagnostic plots for guided projections we further introduce the *glass vessels* (e.g. Filzmoser et al., 2008) dataset. Archaeological glass vessels from the 16^{th} and 17^{th} century were investigated by an electron-probe X-ray micro-analysis. In total, 1920 characteristics are used to describe each vessel. The presence of outliers, especially in one out of the four glass groups has been shown in previous studies (Serneels et al., 2005). We use the algorithm *pcout* (Filzmoser et al., 2008) to identify outliers in this group of observations. The diagnostic plot based on guided projections is visualized in Figure 2.11. We can see that the outliers from *pcout*, drawn in red/light dotted, correspond to two clear subgroups, being well separated from the majority of blue/dark observations for most projections. We further note that the outliers consist of at least two groups and are able to identify at least two groups of main observations and some additional candidates for outliers in the final projections of the series. It is not clear, what underlying nature this group structure is identified from and it seems to be undocumented so far by statistical publications working on the very same glass vessels data set. This information will be valuable for the analyst, because it can refer to problems in the measurement process, or to inconsistencies in the observations which are initially assumed to belong to one group.

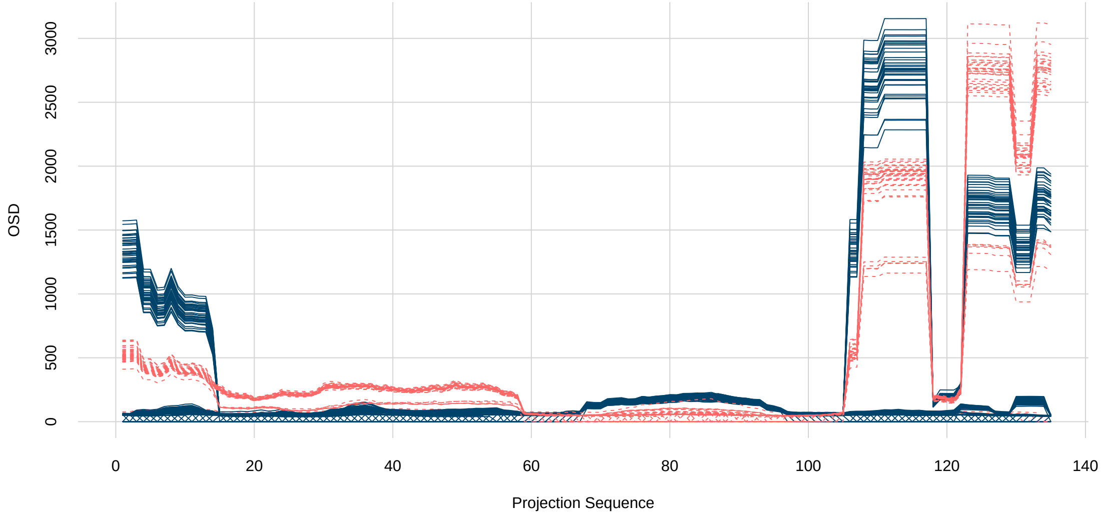


Figure 2.11: Diagnostic plot for the glass vessel data set. Only the main group of glass vessels is considered. Red/light lines correspond to identified outliers by the *pcout* algorithm from Filzmoser et al. (2008).

2.5 Conclusions and outlook

We have proposed guided projections as an alternative to existing data transformations which are applied prior to data structure evaluation methods. We project all observations on the space spanned by a small number of q observations which are selected in a way such that they are likely to come from the same group. We then exchange observations in this selection one by one and therefore create a series of projections. Each projection can then be treated as a new variable, but only the complete series is used for investigating the grouping structure contained in the data. Note that this approach differs conceptually from projection pursuit approaches, where the focus is on identifying one (or several) low-dimensional projections of the data that reveal the group structure.

While guided projections is motivated by the separation of groups using the full available information, its application can be extended onto all types of data structure analysis which is affected by high-dimensionality like outlier detection, cluster analysis, or discriminant analysis. Furthermore, a way for identifying the existence of group structure is provided by the introduced visualization of guided projections. This concept can be further extended to new diagnostic plots for identifying outliers and group structures in the data.

The results based on simulated data show the advantages and limitations of guided projections in comparison to other data transformation methods. Given favourable conditions in the data structure, namely informative variables in different subspaces, guided projections can vastly improve the degree of separation between existing groups in the data. Furthermore, guided projections turned out to be a lot more robust against additional non-informative variables. The results based on the real world data sets also prove the practical importance of guided projections.

There are multiple ways to further improve the concepts of guided projections. First, we can remove the restriction of considering only the projections $\mathcal{I}_L^s$ and $\mathcal{I}_R^s$ for each step. Instead, we can consider every projection of previous steps. Removing this limitation allows a more complex network of projections instead of an ordered series of projections. The setup requires additional research. The second adjustment is the implementation of different distance measures in the projection space. While PCA-based transformations create an orthogonal basis in the projection space, guided projections are highly correlated. Only few projections often provide enough information for a perfect separation. Identifying these projections is a task of its own.

Furthermore, a detailed evaluation of possible measures for OSD needs to be performed to allow a proper evaluation of the limitations and possibilities of guided projections.

Appendix

Equation (2.31) and (2.32) can be proven using the decomposition $\boldsymbol{x} = \boldsymbol{z}_1 + \boldsymbol{z}_2$, where $\boldsymbol{z}_1 \in span(\{\boldsymbol{x}_i : i \in \mathcal{I}^s\})$ and $\boldsymbol{z}_2 \in span^{\perp}(\{\boldsymbol{x}_i : i \in \mathcal{I}^s\})$. *span* represents all possible linear combinations of its observations and $span^{\perp}$ its orthogonal complement. Specifically, we write $\boldsymbol{z}_1 = \sum\limits_{i \in \mathcal{I}^s} a_i \boldsymbol{x}_i$. For the equality of Equation (2.31) it is important to note that also $\hat{\boldsymbol{\mu}}$ is a linear combination of $\boldsymbol{x}_i, i \in \mathcal{I}^s$, with constant coefficients $\frac{1}{q}$. Thus, we can use the property $\boldsymbol{x}_i = \boldsymbol{V}_{\mathcal{I}^s}\boldsymbol{D}_{\mathcal{I}^s}\boldsymbol{u}_i$, which holds for all $i \in \mathcal{I}^s$ where $\boldsymbol{u}_i$ represents the respective right singular vector:

$$\begin{aligned} OD_{\mathcal{I}^s}(\boldsymbol{z}_1) &= ||\boldsymbol{z}_1 - \hat{\boldsymbol{\mu}} - \boldsymbol{V}_{\mathcal{I}^s}\boldsymbol{V}'_{\mathcal{I}^s}(\sum_{i\in\mathcal{I}^s} a_i\boldsymbol{x}_i - \sum_{i\in\mathcal{I}^s}\frac{1}{q}\boldsymbol{x}_i)||, \quad a_i \in \mathbb{R}\ \forall i \in \mathcal{I}^s \\ &= ||\boldsymbol{z}_1 - \hat{\boldsymbol{\mu}} - (\sum_{i=1}^{q} a_i\boldsymbol{V}_{\mathcal{I}^s}\boldsymbol{V}'_{\mathcal{I}^s}\boldsymbol{V}_{\mathcal{I}^s}\boldsymbol{D}_{\mathcal{I}^s}\boldsymbol{u}_i - \sum_{i=1}^{q}\frac{1}{q}\boldsymbol{V}_{\mathcal{I}^s}\boldsymbol{V}'_{\mathcal{I}^s}\boldsymbol{V}_{\mathcal{I}^s}\boldsymbol{D}_{\mathcal{I}^s}\boldsymbol{u}_i)|| \end{aligned} \tag{2.44}$$

Since $\boldsymbol{V}'_{\mathcal{I}^s}\boldsymbol{V}_{\mathcal{I}^s} = \boldsymbol{I}$, one can see that the two linear combinations in Equation (2.44) sum up to $\boldsymbol{z}_1$ and $\hat{\boldsymbol{\mu}}$ respectively. Therefore, Equation (2.44) can be simplified to

$$OD_{\mathcal{I}^s}(\boldsymbol{z}_1) = ||\boldsymbol{z}_1 - \hat{\boldsymbol{\mu}} - (\boldsymbol{z}_1 - \hat{\boldsymbol{\mu}})|| = 0, \tag{2.45}$$

which proves Equation (2.31). To show Equation (2.32) we first note that $\hat{\boldsymbol{\Sigma}}_{\mathcal{I}^s}$ can be written as $\frac{1}{q-1}\boldsymbol{D}^2_{\mathcal{I}^s}$ and due to Equation (2.9) $\boldsymbol{V}'_{\mathcal{I}^s}\boldsymbol{x}_i = \boldsymbol{D}_{\mathcal{I}^s}\boldsymbol{u}_i$ holds. Therefore, we can rewrite the squared score distances for $\boldsymbol{x}_i$ for all $i \in \mathcal{I}^s$ as:

$$\begin{aligned} SD^2(\boldsymbol{x}_i) &= \left(\boldsymbol{V}'_{\mathcal{I}^s}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})\right)' \hat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{I}^s} \left(\boldsymbol{V}'_{\mathcal{I}^s}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})\right) \qquad (2.46) \\ &= \left(\boldsymbol{D}_{\mathcal{I}^s}\boldsymbol{u}_i - \frac{1}{q}\sum_{j\in\mathcal{I}^s}\boldsymbol{D}_{\mathcal{I}^s}\boldsymbol{u}_j\right)' (q-1)\boldsymbol{D}^{-2}_{\mathcal{I}^s} \left(\boldsymbol{D}_{\mathcal{I}^s}\boldsymbol{u}_i - \frac{1}{q}\sum_{l\in\mathcal{I}^s}\boldsymbol{D}_{\mathcal{I}^s}\boldsymbol{u}_l\right) \\ &= (q-1)\left(\boldsymbol{u}'_i\boldsymbol{u}_i - \frac{1}{q}\sum_{j\in\mathcal{I}^s}\boldsymbol{u}'_j\boldsymbol{u}_i - \frac{1}{q}\boldsymbol{u}'_i\sum_{l\in\mathcal{I}^s}\boldsymbol{u}_l + \frac{1}{q^2}\left(\sum_{j\in\mathcal{I}^s}\boldsymbol{u}'_j\right)\left(\sum_{l\in\mathcal{I}^s}\boldsymbol{u}_l\right)\right). \end{aligned}$$

Due to $\boldsymbol{U}_{\mathcal{I}^s}$ being a unitary matrix and therefore $\boldsymbol{u}'_i\boldsymbol{u}_j = \delta_{ij}$, δ_{ij} denoting Kronecker's delta, this expression can be simplified.

$$SD^2(\boldsymbol{x}_i) = (q-1)\left(1 - \frac{1}{q} - \frac{1}{q} + \frac{q}{q^2}\right) = \frac{(q-1)^2}{q} \tag{2.47}$$

which proves Equation (2.32).

CHAPTER 3

Local projections for high-dimensional outlier detection

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.

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 observation. These novel interpretations lead to a univariate measure of outlyingness based 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.

Keywords: dimension reduction, data transformation, diagnostic plots, informative variables

3.1 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, 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).

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.

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., 2017a). 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 3.2 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 3.3. Section 3.4 describes all methods used in the comparison, which are then applied in two simulated settings in Section 3.5. Finally, in Section 3.6, 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 3.7. We conclude with a discussion in Section 3.8.

3.2 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 $F_{X_1}, \ldots, F_{X_q}$, 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.

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 $i = 1, \ldots, n$, not by all available observations but by a small subset, which is located close to $\boldsymbol{x}_i$.

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.

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.

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 between $\boldsymbol{y}$ and $\boldsymbol{x}_i$, denoted by $d(\boldsymbol{y}, \boldsymbol{x}_i)$ for all $i = 1, \ldots, n$:

$$knn(\boldsymbol{y}) = \{\boldsymbol{x}_i : d(\boldsymbol{y}, \boldsymbol{x}_i) \leq d_k\}, \tag{3.1}$$

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 parameter describing the 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 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

$$\boldsymbol{x}_0 = arg \min_{\boldsymbol{x}_i \in knn(\boldsymbol{y})} \{d_{(\lceil \alpha \cdot k \rceil)}(\boldsymbol{x}_i)\}, \tag{3.2}$$

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 $\boldsymbol{y}$:

$$\begin{aligned} core(\boldsymbol{y}) = \{\boldsymbol{x}_i : & d(\boldsymbol{x}_0, \boldsymbol{x}_i) < d_{(\lceil \alpha \cdot k \rceil)}(\boldsymbol{x}_0) \wedge \\ & \boldsymbol{x}_i \in knn(\boldsymbol{y}) \wedge \boldsymbol{x}_i \neq \boldsymbol{y}\} \end{aligned} \tag{3.3}$$

In order to provide an 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.

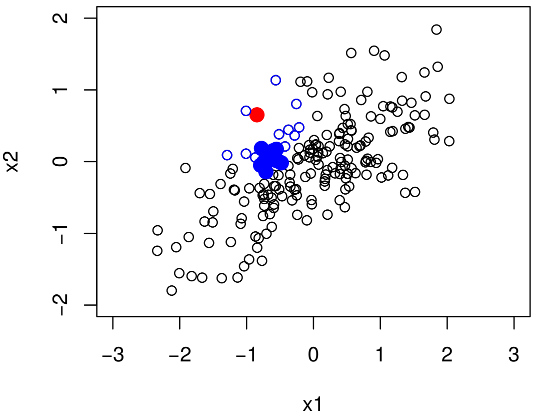


Figure 3.1: Visualization of the *core*-selection process. The red observation represents the initiating observation $\boldsymbol{y}$. The blue observations represent $knn(\boldsymbol{y})$ and the filled blue observations represent $core(\boldsymbol{y})$. x_0 itself is not visualized but it is known to be an element of $core(\boldsymbol{y})$.

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 for scaling the data, we can apply classical estimators on the core preserving robustness properties, since the observations have been included into the core in a robust way.

$$\boldsymbol{X}_{core(\boldsymbol{y})} = (\boldsymbol{x}_{\boldsymbol{y},1}, \dots, \boldsymbol{x}_{\boldsymbol{y},\lceil \alpha \cdot k \rceil})' \tag{3.4}$$

$$\boldsymbol{x}_{\boldsymbol{y},j} \in core(\boldsymbol{y}) \qquad \forall j \in \{1, \dots, \lceil \alpha \cdot k \rceil\}$$

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{y}} = \frac{1}{\lceil \alpha \cdot k \rceil} \sum_{\boldsymbol{x}_i \in core(\boldsymbol{y})} \boldsymbol{x}_i \tag{3.5}$$

$$\begin{aligned} \hat{\boldsymbol{\sigma}}_{\boldsymbol{y}} = &\Big(\sqrt{Var(x_{\boldsymbol{y},11}, \dots, x_{\boldsymbol{y},\lceil \alpha \cdot k \rceil 1})}, \dots, \\ &\sqrt{Var(x_{\boldsymbol{y},1p}, \dots, x_{\boldsymbol{y},\lceil \alpha \cdot k \rceil p})} \Big)' \\ = &(\hat{\sigma}_{\boldsymbol{y},1}, \dots, \hat{\sigma}_{\boldsymbol{y},p})', \end{aligned} \tag{3.6}$$

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

$$\boldsymbol{x}_{\boldsymbol{y}}^c = (x_{\boldsymbol{y},1}^c, \dots x_{\boldsymbol{y},p}^c)' = \boldsymbol{x}_{\boldsymbol{y}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{y}}, \tag{3.7}$$

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

$$\tilde{\boldsymbol{X}}_{\boldsymbol{y}} = \Bigg(\left(\frac{x^c_{\boldsymbol{y},11}}{\hat{\sigma}_{\boldsymbol{y},1}}, \ldots, \frac{x^c_{\boldsymbol{y},1p}}{\hat{\sigma}_{\boldsymbol{y},p}} \right)', \ldots, \left(\frac{x^c_{\boldsymbol{y}\lceil \alpha \cdot k \rceil 1}}{\hat{\sigma}_{\boldsymbol{y},1}}, \ldots, \frac{x^c_{\boldsymbol{y},\lceil \alpha \cdot k \rceil p}}{\hat{\sigma}_{\boldsymbol{y},p}} \right)' \Bigg)' \tag{3.8}$$

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 $\tilde{\boldsymbol{X}}_{\boldsymbol{y}}$,

$$\tilde{\boldsymbol{X}}_{\boldsymbol{y}} = \boldsymbol{U}_{\boldsymbol{y}} \boldsymbol{D}_{\boldsymbol{y}} \boldsymbol{V}'_{\boldsymbol{y}}. \tag{3.9}$$

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,

$$\boldsymbol{x}_{core(\boldsymbol{y})} = \boldsymbol{V}'_{\boldsymbol{y}} \left(\frac{x^c_1}{\hat{\sigma}_{\boldsymbol{y},1}}, \ldots, \frac{x^c_p}{\hat{\sigma}_{\boldsymbol{y},p}} \right)', \tag{3.10}$$

where $\boldsymbol{x}^c = (x^c_1, \ldots, x^c_p)'$ is computed as defined in Equation (3.5) and the orthogonal representation $\boldsymbol{x}_{orth(\boldsymbol{y})}$ given the orthogonal complement of the core space,

$$\boldsymbol{x}_{orth(\boldsymbol{y})} = \boldsymbol{x}^c - \boldsymbol{V}_{\boldsymbol{y}} \boldsymbol{x}_{core(\boldsymbol{y})}. \tag{3.11}$$

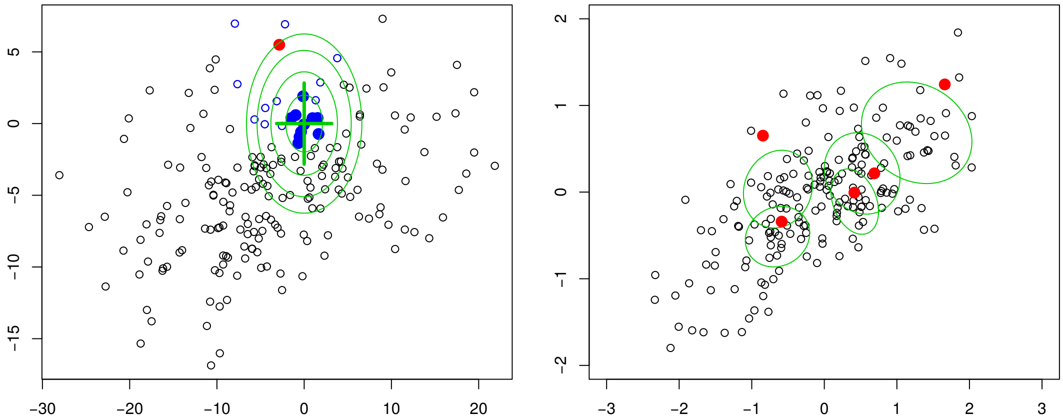
Figure 3.2a shows the representation of our 200 simulated observations in the core space. Note that in this special case, the orthogonal representation is constantly $\boldsymbol{0}$ due to the non-flat data structure of the core observations ($p < k$). We further see that the center of the core is now located in the center of the coordinate system.

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 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 (3.9), while the singular values given by the diagonal of the matrix $\boldsymbol{D_y}$ 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:

$$CD_{\boldsymbol{y}}(\boldsymbol{x}) = \sqrt{\frac{\boldsymbol{x}'_{core(\boldsymbol{y})} \boldsymbol{D_y}^{-1} \boldsymbol{x}_{core(\boldsymbol{y})}}{min(\lceil \alpha \cdot k \rceil - 1, p)}} \tag{3.12}$$

.



(a) (b)

Figure 3.2: Plot (a) provides a visualization of the transformed observations from Figure 3.1. 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 3.1 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 computation of core distances can be derived from Figure 3.2a. 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 3.2 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.

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 $\mathbf{0}$:

$$\boldsymbol{x}_{orth(\boldsymbol{y})} = \mathbf{0} \quad \forall \boldsymbol{x} \in core(\boldsymbol{y}) \tag{3.13}$$

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).

$$OD_{\boldsymbol{y}}(\boldsymbol{x}) = ||\boldsymbol{x}_{orth(\boldsymbol{y})}|| \tag{3.14}$$

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 3.2b 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.

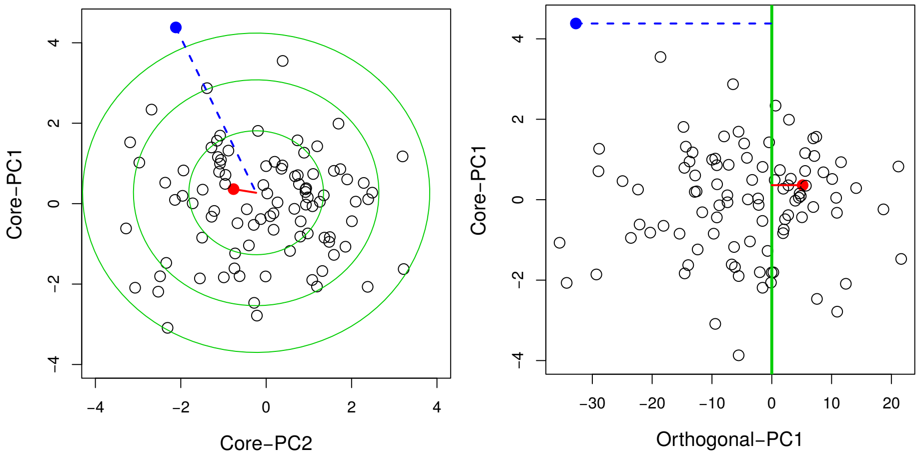
3.3 Interpretation and utilization of local projections

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 3.3 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



(a) (b)

Figure 3.3: 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 core space. Therefore, the probability of the core distance of a core observation being larger than the core distance of any other observation from the same group converges to 1 with increasing p:

$$\lim_{p\to\infty} P(CD_{\boldsymbol{y}}(\boldsymbol{x}_i) > CD_{\boldsymbol{y}}(\boldsymbol{z})) = 1, \quad \boldsymbol{x}_i \in core(\boldsymbol{y}), \tag{3.15}$$
$$\boldsymbol{z} \notin core(\boldsymbol{y})$$

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

$\boldsymbol{y}$:

$$w_{\boldsymbol{y}}(\boldsymbol{x}) = \begin{cases} 0, & \boldsymbol{x} \in core(\boldsymbol{y}) \\ \dfrac{\frac{1}{CD_{\boldsymbol{y}}(\boldsymbol{x})} - \min\limits_{\tilde{\boldsymbol{z}} \in \mathcal{X}} \left(\frac{1}{CD_{\tilde{\boldsymbol{z}}}(\boldsymbol{x})} \right)}{\sum\limits_{\boldsymbol{z} \in \mathcal{X}} \left(\frac{1}{CD_{\boldsymbol{z}}(\boldsymbol{x})} - \min\limits_{\tilde{\boldsymbol{z}} \in \mathcal{X}} \left(\frac{1}{CD_{\tilde{\boldsymbol{z}}}(\boldsymbol{x})} \right) \right)}, & else \end{cases} \tag{3.16}$$

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),

$$LocOut(\boldsymbol{x}) = \sum_{\boldsymbol{y} \in \mathcal{X}} \left(w_{\boldsymbol{y}}(\boldsymbol{x}) \cdot OD_{\boldsymbol{y}}(\boldsymbol{x}) \right), \tag{3.17}$$

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.

3.4 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).

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. Zimek et al., 2012; Campos et al., 2016). 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., 2011) 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 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 et al., 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).

---

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

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 3.6.

3.5 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),

$$\boldsymbol{\Sigma}^i = \begin{pmatrix} \boldsymbol{\Sigma}^i_{inf} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{noise} \end{pmatrix} \qquad \boldsymbol{\Sigma}^i_{inf} = \boldsymbol{\Omega}_i \begin{pmatrix} 1 & \rho^i & \dots & \rho^i \\ \rho^i & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho^i \\ \rho^i & \dots & \rho^i & 1 \end{pmatrix} \boldsymbol{\Omega}_i', \tag{3.18}$$

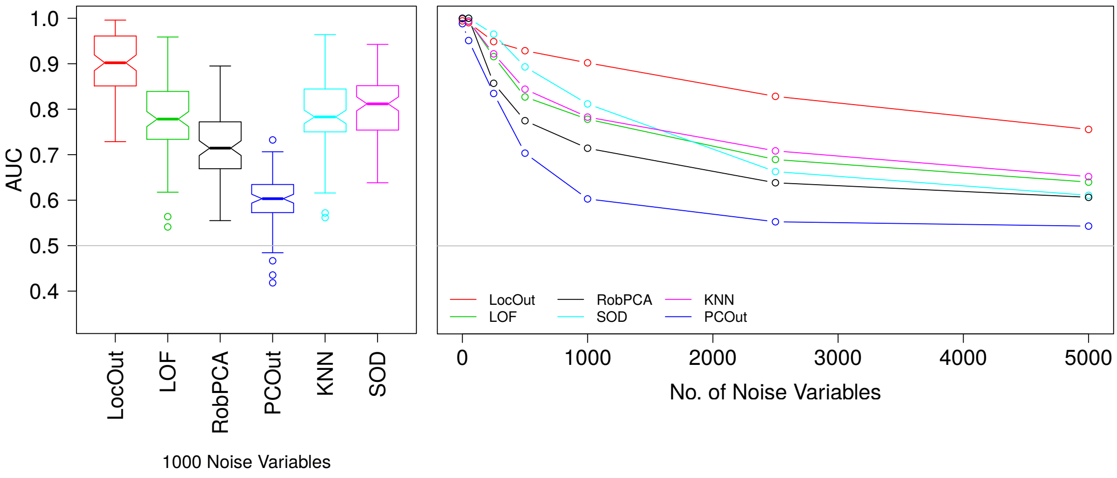
for $i = 1, 2, 3$, where $\boldsymbol{I}_{noise}$ is an identity matrix describing the covariance of uncorrelated noise variables and $\boldsymbol{\Sigma}^i_{inf}$ the covariance matrix of informative variables, which are variables containing information about the separation of present groups where ρ^i 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:

$$(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3) = \begin{pmatrix} \mu & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & \mu \\ \mu & 0 & 0 \\ 0 & \mu & 0 \\ \vdots & \vdots & \vdots \end{pmatrix}. \tag{3.19}$$

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.

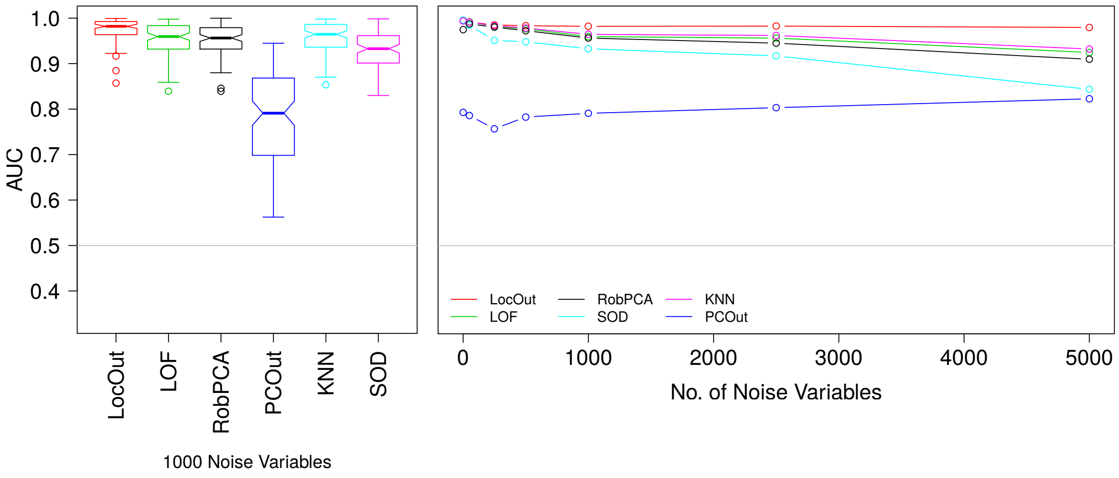
The outcome of the first simulation setup based on multivariate normal distribution is visualized in Figure 3.4. figure 3.4a shows the performance for 100 repetitions with 1000 noise variables as boxplots measured by the AUC value. We note that local projections (LocOut) outperform all other methods, 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 3.4b, 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 3.4: 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 3.5, 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.



(a) (b)

Figure 3.5: 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.

3.6 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.

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, 2014) 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 3.6. 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.

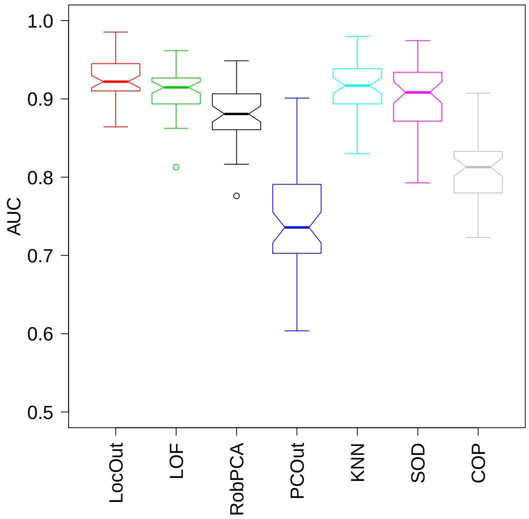


Figure 3.6: Performance of different outlier detection methods for the 25 dimensional olive oil dataset measured by the area under the ROC curve (AUC). For each method the configuration parameters are optimized based on the ground truth.

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 3.7 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.

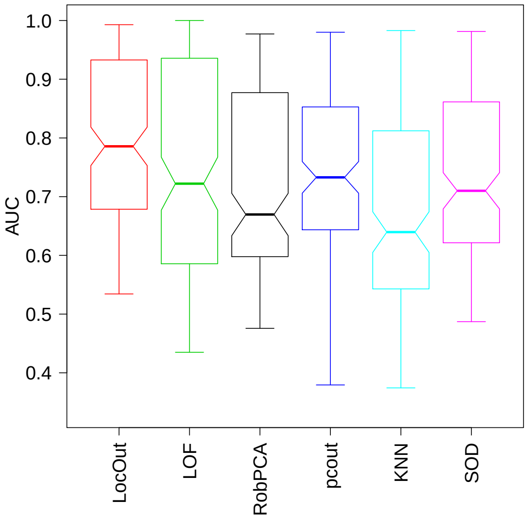


Figure 3.7: Evaluation of the performance of the outlier detection algorithms on the fruit data set, showing boxplots of the performance of 150 repetitions of outlier detection measured by the AUC.

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 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 3.8a 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 3.8b. Again, LocOut outperforms all compared methods, while LOF and PCOut have problems to deal with this data setup.

3.7 Discussion of runtime

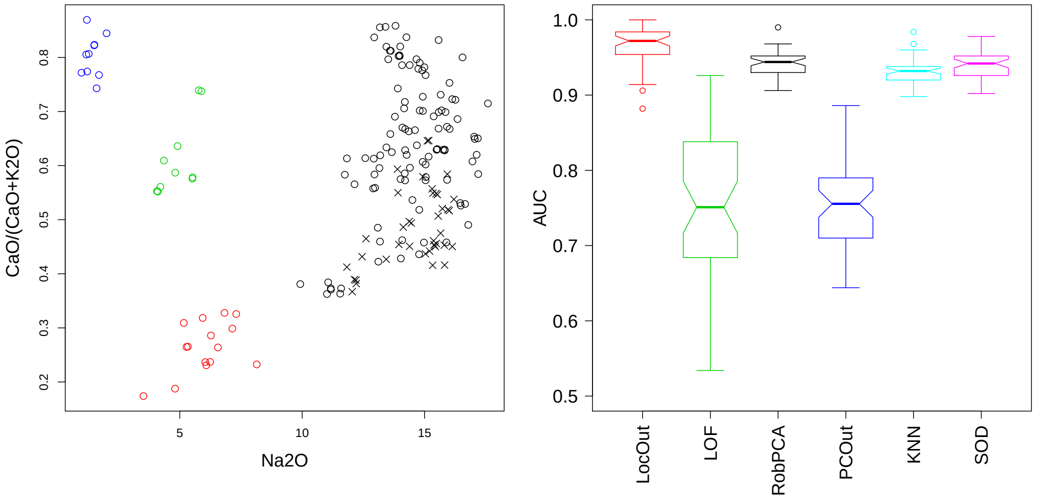
The algorithm for local projections was implemented in an R-package which is publicly available[2]. The package further includes the glass vessels data set used in Section 3.6. Based on this R-package, we performed simulations to test the required computational effort for the proposed algorithm and the 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 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

---

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



(a)

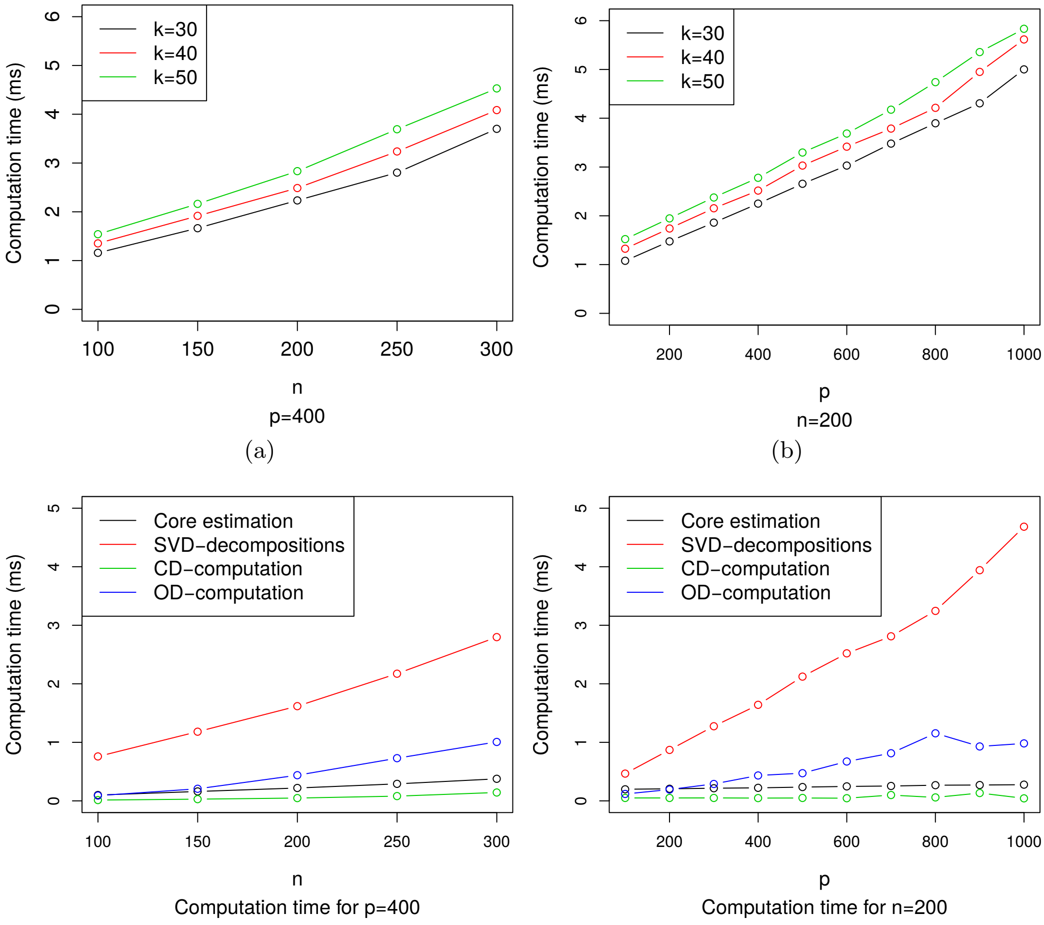
(b)

Figure 3.8: 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.

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 3.9 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 3.9 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 3.9: 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.

3.8 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* 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, 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.

CHAPTER 4

Multigroup discrimination based on weighted local projections

Abstract

Local Discriminant analysis (LDA) model is computed using the information within the projection space as well as the distance to the projection space. The models provide information about the quality of separation for each class combination. Based on this information, weights are defined for aggregating the LDA-based posterior probabilities of each subspace to a new overall probability. The same weights are used for classifying new observations.

In addition to the provided methodology, implemented in the R-package *lop*, a method of visualizing the connectivity of groups in high-dimensional spaces is proposed at the basis of the posterior probabilities. A deep evaluation is performed using three different real-world datasets, underlining the strenghts of local projection based classification and the provided visualization methodology.

4.1 Introduction

Supervised classification methods are widely used in research and industry, including tasks like tumor classification, speech recognition, or the classification of food quality. Observations are gathered from G distinct groups and for each observation the group membership is known. Decision boundaries are then estimated in the sample space, such that a new observation can be assigned to one of the G groups. The aim of discrimination methods is to find classification boundaries, which result in low misclassification rates for new observations, i.e. new observations are assigned to the correct class with high accuracy.

Linear discriminant analysis (LDA) is a popular tool for classification. It estimates linear decision boundaries by maximizing the between-group to within-group variance and assumes equal covariance structure of the groups. LDA often gives surprisingly good results in low-dimensional settings, however, it cannot be directly applied if the number of variables exceeds the number of observations since then the within-group covariance estimate becomes singular and its inverse cannot be calculated. With restrictions on the covariance estimation the problem of singularity can be mended, but asymptotically (with increasing number of variables) the performance of LDA is not better than random guessing (Bickel and Levina, 2004; Shao et al., 2011).

In many classification tasks it is commonly the case that the underlying data has a flat structure, i.e. there are more variables than observations. Therefore, a great variety of alternative classification methods and extensions of LDA have been developed to overcome this limitations. Several proposed approaches consider projection of the data onto a lower dimensional subspace (Barker and Rayens, 2003; Chen et al., 2013b) or reducing the dimensionality by model-based variable selection (Witten and Tibshirani, 2011). Other methods are not based on covariance estimation and so they are not restricted to low-dimensional (non-flat) settings, e.g. k-nearest neighbour (KNN) classification, support vector machines (SVM) or random forests (RF). Nevertheless, the noise accumulation due to a large number of variables, which are not informative for the class separations, affects these methods as well.

We propose a new approach for supervised classification based on a series of projections into low-dimensional subspaces, referred to as local projections. In each subspace, we calculate an LDA model. The posterior probabilities of each LDA model are aggregated (weighted by a class-specific quality measure of the projection space) to obtain a final classification. The idea of aggregating posterior probabilities in the context of random forests has been proposed by Bosch et al. (2007) taking the average over the posterior probabilities from all trees.

The remainder of the paper is structured as follows. Section 4.2 presents the proposed method. First, local projections based on the k-nearest class neighbors of an observation and distances within and to the projection space are introduced in Section 4.2, resulting in the local discrimination space where an LDA model is estimated. Next, in Section 4.2, we introduce weights used for aggregating the posterior probabilities from the individual LDA models leading to a final classification rule. In Section 4.2, the range of the tuning parameter k, associated with the dimensionality of the local discrimination space, is discussed and a strategy to select the tuning parameter is presented. Section 4.3 introduces a way of visualizing the data structure and the degree of separation. In Section 4.4, three real-world datasets are used to evaluate the performance of our approach in comparison to other related and popular classification methods. The datasets cover settings with only 25 and up to almost 10.000 variables, including multigoup and binary classification problems, and a dataset where subgroups are known to exist. The effect of imbalanced group sizes in the training data is investigated and results are visualized by the techniques introduced in the previous section. Section 4.5 concludes the paper.

4.2 Methodology

Let $\boldsymbol{X}$ denote a data matrix of n observations $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)'$ in a p-dimensional space, $\boldsymbol{x}_i \in \mathbb{R}^p$, $i = 1, \dots, n$. We further assume the presence of G classes where the class memberships of the observations are stored in a categorical vector $\boldsymbol{y}$ with $y_i = g$ iff $\boldsymbol{x}_i$ comes from group g, for $g \in \{1, \dots, G\}$. The number of observations in group g is denoted by n_g with $n = n_1 + \cdots + n_G$. For all observations we assume that they have been drawn from G different continuous probability distributions.

For our methodology it is important that each subset of k observations which spans a space having a dimension of at least $G - 1$, and that there are no ties present in our data. These requirements automatically imply high dimensional dataspaces as the area of application. Both assumptions can be met by a preprocessing step, removing duplicate or linearly dependent observations. Note that these restrictions only apply for the training data, but not for new observations.

Previous research (Ortner et al., 2017a,b) shows the effectiveness of using series of projections to overcome the limitations dictated by a flat data structure. In this section, the local discrimination method is introduced which allows for the number of variables p to exceed the number of observations n. The idea is as follows. For a fixed observation $\boldsymbol{x}_i$, its k nearest neighbours are identified, called the *core* of $\boldsymbol{x}_i$, which are used to define a $k - 1$ dimensional hyperplane, the core space. The Euclidean distance to this hyperplane, called *orthogonal distance*, is calculated for each observations. The hyperplane and the orthogonal distance together define a k-dimensional subspace, the local discrimination space, where an LDA model is estimated. This approach is performed for each observation resulting in n LDA models. To assign the class membership to an observation, its posterior probabilities of all models are aggregated.

Local discrimination space

Let $d_k^g(\boldsymbol{x})$ denote the kth-smallest distance from $\boldsymbol{x}$ to any observation from class g, for $g \in \{1, \ldots, G\}$. According to Ortner et al. (2017a) and Ortner et al. (2017b), we define the *core* of $\boldsymbol{x}_i$ as the k-nearest class neighbors of $\boldsymbol{x}_i$,

$$core(\boldsymbol{x}_i) = \{\boldsymbol{x}_j : d(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq d_k^g(\boldsymbol{x}_i) \wedge y_i = y_j = g\} = \{\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_k}\}, \tag{4.1}$$

where $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ denotes the Euclidean distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and $i_1, \ldots, i_k$ are the indices of the core observations within $\boldsymbol{X}$. In contrast to Ortner et al. (2017b), we use all k-nearest class neighbours as we can use the group membership in order to guarantee a *clean* core, i.e. no observations from other groups within the core.

Any of the n available cores $core(\boldsymbol{x}_1), \ldots, core(\boldsymbol{x}_n)$ can be used to unambiguously define an affine subspace spanned by the core observations. In order to determine the projection onto this subspace, we center and scale the data with respect to the k core observations $\boldsymbol{x}_{i_j}$, $j = 1, \ldots, k$.

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{k} \sum_{j=1}^{k} \boldsymbol{x}_{i_j} \tag{4.2}$$

$$\begin{aligned} \hat{\boldsymbol{\sigma}}_i &= \left(\sqrt{\hat{Var}(x_{i_1 1}, \ldots, x_{i_k 1})}, \ldots, \sqrt{\hat{Var}(x_{i_1 p}, \ldots, x_{i_k p})} \right)' \\ &= (\hat{\sigma}_{i1}, \ldots, \hat{\sigma}_{ip})', \end{aligned} \tag{4.3}$$

where $\hat{Var}$ denotes the sample variance. For the ongoing work, we denote $\tilde{\boldsymbol{X}}^i = (\tilde{\boldsymbol{x}}_1^i, \ldots, \tilde{\boldsymbol{x}}_n^i)'$ as the data matrix of centered and scaled observations based on the location and scale estimators $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\sigma}}_i$ of the core of $\boldsymbol{x}_i$. A projection onto the subspace spanned by the core of $\boldsymbol{x}_i$ is defined by $\boldsymbol{V}_i$ from the singular value decomposition (SVD) of the centered and scaled core observations $(\tilde{\boldsymbol{x}}_{i_1}^i, \ldots, \tilde{\boldsymbol{x}}_{i_k}^i)' = \boldsymbol{U}_i \boldsymbol{D}_i \boldsymbol{V}_i'$. Since the core of $\boldsymbol{x}_i$ consists of exactly k linearly independent observations, $\boldsymbol{D}_i$ is a $k-1$ dimensional diagonal matrix, with non-zero singular values in the diagonal.

Since the idea of no ties being present in the data and each core consisting of linearly independent observations may appear like a strong limitation, an adjustment of the definitions can help in order to avoid a preprocessing step. If we interpret the core of $\boldsymbol{x}_i$ as a set of observations, where iteratively the observation from the same class, closest to $\boldsymbol{x}_i$ is added until a $k-1$ dimensional subspace is spanned, we only need to guarantee the existence of such subsets of observations, which is a much weaker assumption.

Given the projection matrix $\boldsymbol{V}_i$ from the decomposition, a representation of the data $\boldsymbol{X}$ in the core space is defined by down-projecting the centered and scaled data matrix, $\boldsymbol{Z}^i = \tilde{\boldsymbol{X}}^i \boldsymbol{V}_i$. The core representation consists of $k-1$ orthogonal variables, while the $p-k$ dimensional complement of $\boldsymbol{Z}^i$ defines the orthogonal complement of the core space. In contrast to commonly used procedures of first reducing the dimensionality using PCA and then performing a discrimination method like LDA, we acknowledge the fact that the last principal components might contain an important part of the information like exploited by modern outlier detection algorithms (e.g. Hubert et al., 2005; Kriegel et al., 2012). Since the reduction of dimensionality remains vital, we aggregate the information from the orthogonal complement by considering the Euclidean distance to the core space,

$$OD^i(\boldsymbol{x}_j) = ||\tilde{\boldsymbol{x}}_j^i - \boldsymbol{V}_i \boldsymbol{z}_j^i||, \tag{4.4}$$

where $\boldsymbol{z}_j^i = \boldsymbol{V}_i' \tilde{\boldsymbol{x}}_j^i$ denotes the core representation of $\boldsymbol{x}_j$ given $core(\boldsymbol{x}_i)$.

The combination of the core representation and the orthogonal distance OD in a matrix, $[\boldsymbol{Z}^i, OD^i]$, provides a k-dimensional representation for all observations of $\boldsymbol{X}$. This k-dimensional space is the *local discrimination space.* The reduction of the sample space to the local discrimination space results in a good description of the neighbourhood of an observation $\boldsymbol{x}_i$ and also includes grouping structure which is not described in the core space by the orthogonal distances.

An LDA model is estimated in the local discrimination space, excluding the observations from the core of $\boldsymbol{x}_i$. It is necesarry to exclude the core observations, because they have very specific properties in the local discrimination space and this would distort the within-group covariance estimation. In Ortner et al. (2017a) it is shown that

$$OD^i(\boldsymbol{x}_j) = 0 \qquad \forall \boldsymbol{x}_j \in core(\boldsymbol{x}_i) \tag{4.5}$$

$$SD^i(\boldsymbol{x}_j) \equiv const. \qquad \forall \boldsymbol{x}_j \in core(\boldsymbol{x}_i), \tag{4.6}$$

where SD represents the score distance, defined as the Euclidean distance within the core space. These properties hold because for $\boldsymbol{x}_j \in core(\boldsymbol{x}_i)$ the full information is located in the core space, so the orthogonal distances are zero. The scaling applied to the data based on the covariance estimation of the core observations leads to constant score distances for $\boldsymbol{x}_j \in core(\boldsymbol{x}_i)$. So the core observations must not be included in the computation of the LDA model. The model estimated on the remaining observations in the local discrimination space is denoted by LDA_i.

For the model LDA_i the posterior probability of group g given an observation $\boldsymbol{x}$ is defined by

$$P_{LDA_i}(g \mid \boldsymbol{x}) = \frac{h_g(\boldsymbol{x})}{\sum_{j=1}^{G} h_l(\boldsymbol{x})}, \tag{4.7}$$

where $h_g(\boldsymbol{x})$ denotes the estimated density of a multivariate normal distribution with the group mean of class g as center and the pooled within-group covariance matrix as covariance estimate.

Weighting/aggregating local projections

We now have a set of n local discrimination spaces and their respective LDA models. In order to receive an overall classification rule for a new observation $\boldsymbol{x}$, we need to aggregate the n available models from the core spaces. We accomplish such an aggregation by using the posterior probabilities defined in Equation (4.7). First we consider the mean over all n posterior probabilities of $\boldsymbol{x}$ belonging to group g, for $g \in \{1, \dots, G\}$,

$$\tilde{P}_{LP_1^k}(g \mid \boldsymbol{x}) = \frac{1}{n} \sum_{j=1}^{n} P_{LDA_j}(g \mid \boldsymbol{x}), \tag{4.8}$$

and we define the aggregated posterior probability of $\boldsymbol{x}$ belonging to group g, for $g \in \{1, \dots, G\}$, as

$$P_{LP_1^k}(g \mid \boldsymbol{x}) = \frac{\tilde{P}_{LP_1^k}(g \mid \boldsymbol{x})}{\sum_{j=1}^{G} \tilde{P}_{LP_1^k}(j \mid \boldsymbol{x})}. \tag{4.9}$$

These new aggregated posterior probabilities are based on a fixed number k describing the number of core observations as indicated by the index of LP_1^k.

The posterior probabilities of the LDA models, $P_{LDA_i}(g \mid \boldsymbol{x})$, compared to the true class membership of $\boldsymbol{x}$ reflect the quality of separation in the respective local projection. We distinguish between two quality measures. Let q_i^{g+} denote the mean posterior probability of belonging to class g over all observations actually coming from class g, with respect to the model LDA_i, i.e.

$$q_i^{g+} = \frac{1}{n_g} \sum_{k:y_k=g} P_{LDA_i}(g \mid \boldsymbol{x}_k) \tag{4.10}$$

and q_i^{g-} the mean posterior probability of non-class-g observations being classified as class g observations given the model LDA_i, i.e.

$$q_i^{g-} = \frac{1}{n - n_g} \sum_{k:y_k \neq g} P_{LDA_i}(g \mid \boldsymbol{x}_k). \tag{4.11}$$

Based on q_i^{g+} and q_i^{g-}, we define weights w_i^g representing the quality of each local projection $i = 1, \dots, n$ for each group $g \in \{1, \dots, G\}$,

$$w_i^g = exp\left(q_i^{g+} - q_i^{g-}\right). \tag{4.12}$$

Based on these quality measures w_i^g, we redefine the overall posterior probabilities from Equation (4.9) by weighting each projection for each class with the respective weight. Note that these weights are class-specific and, therefore, a class-individual standardization of weights is required. In our notation, we remove the subscript 1 from Equation (4.8) and Equation (4.9), which represents constant weights of 1 for each local projection, resulting in:

$$\tilde{P}_{LP^k}(g \mid \boldsymbol{x}) = \frac{1}{\sum_{i=1}^{n} w_i^g} \sum_{i=1}^{n} w_i^g P_{LDA_i}(g \mid \boldsymbol{x}) \tag{4.13}$$

$$P_{LP^k}(g \mid \boldsymbol{x}) = \frac{\tilde{P}_{LP^k}(g \mid \boldsymbol{x})}{\sum_{j=1}^{G} \tilde{P}_{LP^k}(j \mid \boldsymbol{x})} \tag{4.14}$$

Equivalently to classical LDA, we use these posterior probabilities to assign an observation $\boldsymbol{x}$ to a class $\hat{y} = \underset{g \in \{1,\dots,G\}}{\operatorname{argmax}} P_{LP^k}(g \mid \boldsymbol{x})$. This decision rule defines the local discrimination model.

The choice of k

The computation of LDA models in the full dimensional space, given more variables than observations are available, demands for data preprocessing including dimension reduction (e.g. Barker and Rayens, 2003; Chen et al., 2013b) or the parallel performance of model estimation and variable selection (e.g. Witten and Tibshirani, 2011; Hoffmann et al., 2016). The concept of local projections allows us to compute an LDA model for each local projection due to the low dimensional core space. The parameter determining the dimensionality is k of the k-nearest class neighbours. It is important to properly tune k since it defines the degree of locality for each projection. Smaller k's are able to better describe a lower dimensional manifold on which groups might be located but increase the risk of not being able to properly describe the local data structure.

The number of classes G as well as the number of observations n_g for $g \in \{1, \dots, G\}$ provide a first limitation for the range of k. In order to compute an LDA model with G classes, a dimensionality equal to at least $G - 1$ is required. Therefore,

$$G - 1 \leq k \tag{4.15}$$

provides a lower boundary for k.

To identify an upper boundary for k, two properties of the core observations must be taken into account. Due to the specific properties of the core observations stated in Equations (4.5) and (4.6) they are not included in the computation of the LDA model. So an upper boundary for k is given by

$$k \leq n - (k + 1) \tag{4.16}$$

to guarantee a non-singular covariance estimation. It is useful to further reduce the upper boundary of k in order to allow for a reasonable covariance estimation. Here we take three times more observations than variables leading to the limitation

$$3k \leq n - k. \tag{4.17}$$

With these restrictions on k, LDA models in the core spaces can be computed, but for the evaluation of the models further limitations are necessary. To be able to evaluate the LDA models we depend on the posterior probabilities of observations for each class in order to determine the risks of misclassification. Since a core consists of observations from the same class only and the core observations are excluded from the LDA model, the size of the smallest class needs to exceed k.

$$k + 1 \leq \min_{g \in \{1,\dots,G\}} n_g - 1 \tag{4.18}$$

Due to the identified restrictions on k we optimize k within the following interval:

$$\left[G - 1, \min \left(\frac{n}{4}, \min_{g \in \{1,\dots,G\}} n_g - 1 \right) \right] \tag{4.19}$$

For a given k, the misclassification rate of the local discrimination model is calculated by summing up the number of misclassified observations (again excluding the core observations) divided by $n - k$, the total number of observations. The tuning parameter k is chosen from within the interval described in Equation (4.19) such that the misclassification rate is minimized.

4.3 Visualization of the discrimination

In linear discriminant analysis, the projection space is used for the visualization of the discrimination (e.g. Hair et al., 1998). The Mahalanobis distances of observations to the class centers refer to the posterior probabilities of the observations for the respective classes. This approach is not feasible for local discrimination, since each LDA model refers to a different subspace and the aggregated posterior probabilities do not refer to one specific low dimensional space, where the posterior probabilities could be visualized.

We therefore focus on visualizing the aggregated posterior probabilities and follow an approach for compositional data using ternary diagrams. We present the visualization technique on the four-group *Olitos* dataset which is used as a benchmark dataset for robust, high-dimensional data analysis. The dataset is publicly available in the R-package *rrcovHD* and was originally described by Armanino et al. (1989).

Hron and Filzmoser (2013) used ternary diagrams to visualize the outcome of (three-group) fuzzy clustering results, which can be interpreted the same way as posterior probabilities of discrimination models. The difficult aspect about ternary diagrams is the limitation to three variables. Therefore, we select two classes, use the respective posterior probabilities and as third composition the sum of posterior probabilities for all remaining classes. This new three-class composition is visualized in Figure 4.1.

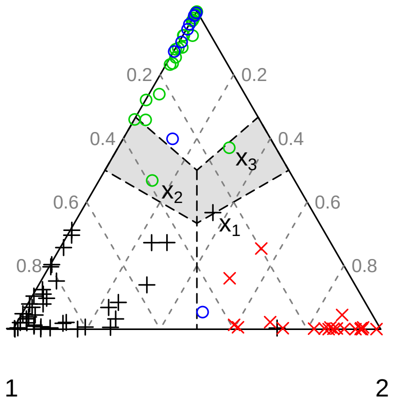
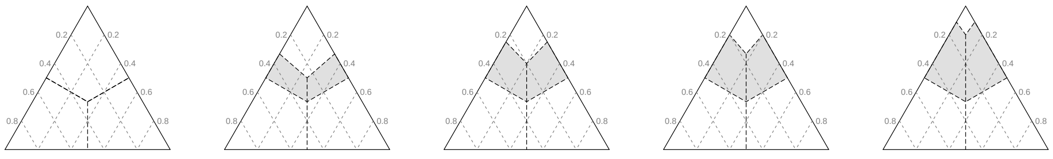


Figure 4.1: The aggregated posterior probabilities of two classes (class 1 and class 2) compared to the remaining classes is visualized as a ternary diagram. The dashed lines represent classification rules. Observations located in the white areas can be assigned to the respective group, while the grey area represents an uncertain area, where no certain statement can be made. The grey dashed lines refer to posterior probabilities of the selected classes.

Figure 4.1 shows the proposed representation for a sample of the Olitos dataset. The focus of this representation is the evaluation of the separation between the two selected classes 1 and 2. The gray dashed lines and the numbers on the left side show the posterior probability for an observation to belong to group 1. The two white areas at the bottom separated by a vertical dashed line represent the classification rules for the separation between class 1 and 2. Observations in the left area are assigned to group 1 and in the right area to group 2. In the bottom right area we can identify one outlier from the *blue* class and one from class 1 which are wrongly assigned to group 2. Besides these two false classifications, additional information can be gained from the diagram.

Firstly, the grey area represents the region, where no statement about classification can be made with certainty. Two observations x_1 and x_2 are highlighted there. While x_1 is located in the uncertainty area we can still tell that it will be misclassified since the posterior probability for class 2 is larger than for class 1. This decision is indicated by the vertical dashed line within the uncertainty area. But from this figure it is not possible to say whether it will be assigned to class 2 or to one of the other classes. The same holds for observation x_2. The posterior probability for class 1 is close to 0.4, for class 2 close to 0.15. Therefore, the posterior probabilities for classes 3 and 4 sum up to approximately 0.45. Depending on the class-specific allocation the maximal posterior probability for the classes 3 and 4 varies between 0.225 and 0.45 and the largest posterior probability for x_2 can originate from class 1, 3 or 4.

Secondly, the white classification area at the top of the triangle visualizes those observations which with certainty will not be assigned to class 1 or class 2. We note a minor risk of misclassifying observations in the direction of class 1. Note that the size of the uncertainty area and therefore the size of the third classification area highly depends on the number of groups to be aggregated. In a 3-group case, all posterior probabilities can be visualized and no area of uncertainty exists, as shown in Figure 4.2a. The remaining plots of Figure 4.2 show the impact of increasing numbers of groups on the area of uncertainty.



(a) (b) (c) (d) (e)

Figure 4.2: The effect of the number of aggregated groups is visualized. Figure (a) refers to a three-class case, (b) to a four-class case, (c) to a five-class case, (d) to a six-class case and (e) to a ten-class case.

Finally the positioning of the observations in the ternary diagram provides some insight on the connection between the groups. The red observations from class 2 in Figure 4.1 are mostly aligned along the axis from 1 to 2. Observations from the aggregated group are also mostly aligned between 1 and 3 while the black observations split up between class 2 and 3. Therefore, class 2 is strongly connected to class 1 but has no connection to the aggregated classes. Observations like x_3 strongly deviate from the typical class direction and should therefore be candidates for further investigation in the context of outlier analysis.

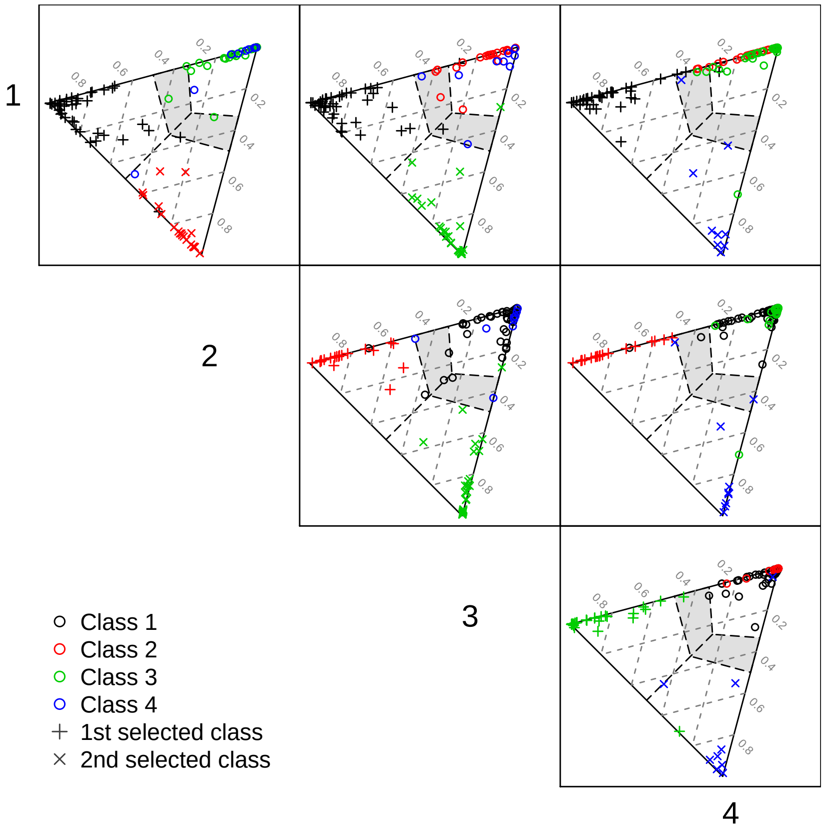


Figure 4.3: A set of ternary diagrams is used to visualize the classification performance for each possible combination of classes. While the color remains constant, we switch labels to emphasize the currently selected classes labeled on the bottom and left of the diagrams.

Since the representation in Figure 4.1 uses a three-components representation, which can not illustrate the overall discrimination result, we propose to use a combination of ternary diagrams in the form of a scatterplot matrix, as presented in Figure 4.3. Each combination of possible two-class plus one aggregated group classifications is presented as described for Figure 4.1. In order to align the groups and increase the readability, the diagrams have been rotated accordingly.

Besides providing information on the quality of discrimination and risks of misclassification, we can derive an overall picture about group connectivity. We already remarked on the location of class 2. The positioning of the observations of the green and blue observations in the second and third column of the matrix reveals that the green group has stronger ties to the black group, while the blue observations are equally drawn to its direct neighbors, the black and the blue group. Such insight on connectivity provides a feeling for the location of groups in high dimensional spaces which is in general a non-trivial task limited by the human spacial sense.

4.4 Evaluation

In order to evaluate the performance of our proposed local discrimination approach, abbreviated by LP for Local Projections, we use three real-world datasets which have previously been used as benchmark datasets for high dimensional data analysis. Based on those datasets, we compare with well established classification methods from the fields of computer science and statistics. While the visualization introduced in Section 4.3 provides interesting insights into each dataset and will be provided as well, we focus on comparing the used methods based on the misclassification rate.

For each dataset we split the available observations into training and test dataset. The same training dataset is used for each method to estimate the discrimination model and the same test dataset is used to evaluate the performance of all the models by reporting the misclassification rates.

The datasets under evaluation consist of groups of different numbers of observations. Since the outcome of a method can be strongly affected by the specific choice of the training and test set, we resample the observations 50 times per dataset, creating a series of training and test datasets resulting in a series of misclassification rates. The overall performance is then measured based on the median misclassification rate as well as on the deviation from the median misclassification rate.

Compared methods

The selection of classification methods is based on the popularity of the methods, the importance for our setups and the relevance for our proposed approach. The most important aspect is the applicability on the evaluated datasets. The crucial factor is the flat data structure (more variables than observations), especially in class-specific subsets of the overall dataset. In order to cover related classification methods, we include Linear Discrimination Analysis (LDA) as this is the classification method internally used for each local projection. We further include statistical advancements of LDA which try to deal with disadvantageous properties of our datasets of interest, namely penalized LDA and partial least squares for discriminant analysis. The most related method from the field of computer science is KNN-classification as our local projections are based on a knn-estimation. The final methods included in the evaluation are support vector machines and random forests to cover the most commonly used classification approaches from the field of computer science.

For **Linear Discriminant Analysis** (LDA) it is assumed that the covariance structure is the same for each class and has elliptical shape. Under this assumption the optimal decision boundaries to separate the groups are linear. The separation of the classes is achieved by taking $G-1$ orthogonal directions which maximize the within-group variance to the between-group variance. In this $G-1$ dimensional space the Euclidean distance to the group centers is used to assign an observation to the group with the closest center.

For the calculations the function `lda` from the R-package `MASS` is used. This implementation can be applied to data with $p > n$ by performing singular value decomposition and reducing the dimensionality to the rank of the data.

Penalized LDA (PLDA) introduced by Witten and Tibshirani (2011) is a regularized version of Fisher's linear discriminant analysis. A penalty on the discriminant vectors favours zero entries, which leads to variable selection. The influence of the penalty is controlled by the sparsity parameter λ: larger values of λ lead to fewer variables in the model.

The sparsity parameter λ is selected from 10 values between 10^{-4} and 5 by 10-fold cross-validation on the training data using as selection criterion the minimum mean misclassification rate. The number of discriminating vectors is set to $G-1$. The functions for cross validation and model estimation are provided in the R-package `penalizedLDA` (Witten, 2011).

Partial least squares for discriminant analysis (PLSDA) was theoretically established by Barker and Rayens (2003), where its relationship to LDA and the application to flat data was discussed. PLSDA performs in a first step a projection onto K latent variables, which considers the grouping information of y. Then LDA is performed in the reduced space.

For the evaluation the R-package `DiscriMiner` (Sanchez and Determan, 2013) is used, which provides code for the selection of the number of components K by leave-one-out cross-validation.

Support Vector Machines (SVMs) are a popular machine learning method for classification. The margins between the groups of the training data are maximized in a data space induced by the selected kernel. While a variety of kernels is available (e.g. linear, polynomial, sigmoid, etc.) we limit the optimization procedure to the radial basis kernel, which is suggested as standard configuration.

We use an R-interface to *libsvm* (Chang and Lin, 2011) included in the R-package *e1071*. The internal optimization of SVM is based on a k-fold cross-validation on the training dataset, providing a range of values for the cost parameter and for γ. For multi-class-classification, libsvm internally trains K(K-1)/2 binary 'one-against-one' classifiers based on a sparse data representation matrix.

Random Forest (RF) is an ensemble-based learning method commonly used for classification and regression tasks. It builds a forest of decision trees using bootstrap samples of the training data and random feature selection for each tree. The final prediction is made as an average or majority vote of the predictions of the ensemble of all trees.

The RF implementation in the R-package *randomForest* uses Breiman's random forest algorithm (Breiman, 2001) for multigroup classification. In order to optimize the classification model we use the internal optimization procedure, starting with $\sqrt{p}$ randomly sampled variables as candidates for splits and increase this number with a factor of 1.5, in each optimization step.

In **KNN-Classification** (KNN), the class-membership of the k-nearest neighbors of an observation based on Euclidean distances is used for determining the class of the respective observation. For $k = 1$, the class of the nearest neighbor is used, for $k > 1$, the class with the highest frequency is used. In the case of ties, a random decision is performed. We use one-fold cross-validation in order to optimize k individually for each sampled dataset.

Olive oil

We want to consider datasets with various specific features. The first dataset under evaluation consists of 120 samples of 25 chemical compositions (fatty acids, sterols, triterpenic alcohols) of olive oils from Tuscany, Italy, and was first introduced by Armanino et al. (1989). The dataset is publicly available in the R-package *rrcovHD* (Todorov, 2014) where it is used as a reference dataset for robust high-dimensional data analysis.

The olive oils are separated in four classes of 50, 25, 34 and 11 observations. In order to have enough training observations from each group available we use 80% of observations for the training dataset and the remaining 20% as test observations. We repeatedly create such an evaluation setup 50 times. Each training dataset therefore consists of 96 observations, which yields the only setup where we have more observations than variables available. Therefore, classical LDA is expected to perform fairly well. Note that the smallest number of training observations per class is still much smaller than the number of overall variables. Therefore, class-specific covariance estimation as it is performed in quadratic discriminant analysis (Friedman, 1989) cannot be performed in this setup or on any other of our considered datasets.

LP and LDA perform exactly the same, which can be seen in Figure 4.4. PLDA slightly outperforms LP, while PLSDA, SVM, RF and especially KNN get outperformed. In most cases all variables are included in the PLDA model, but only a subset of variables contributes to each discriminant vector. This variable selection leads to a slight improvement over LDA and LP.

Arcene

The second real world dataset is part of the NIPS (Neural Information Processing Systems) 2003 feature selection challenge (Guyon et al., 2007). The task is to distinguish between cancer and non-cancer patterns from mass-spectrometric data with $p = 9961$ variables. Therefore, we deal with a two-class separation with continuous variables. The data was obtained from two different data sources, the National Cancer Institute (NCI) and the Eastern Virginia Medical School (EVMS). The observations represent patients with ovarian or prostate cancer and health or control patients. Very small and large masses have been removed from the spectrometric data in order to compress the data. In addition, a preprocessing step including baseline removal, smoothing and scaling was performed. All these details are described in Guyon et al. (2007).

The initial setup contained of 100 training and 100 validation observations, consisting

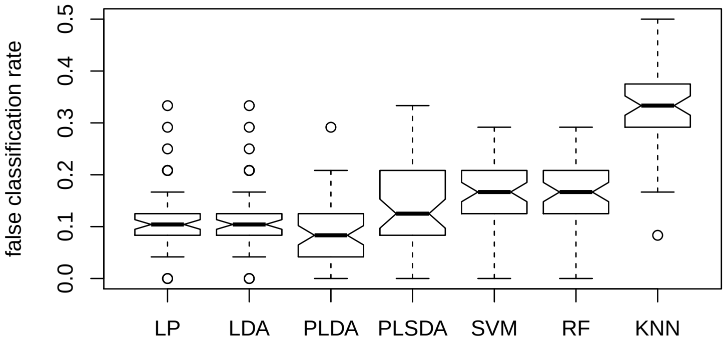


Figure 4.4: The performance in terms of false classification rates of all considered classification methods for 50 repetitions of the *Olitos* dataset is visualized by boxplots.

of a total of 112 non-cancer samples and 88 cancer samples. In order to have a non-equal ratio of observations to create again an imbalanced scenario, we merge both groups and resample 22 cancer training observations and 84 non-cancer trainining observations. The remaining observations are used as test observations. This procedure is repeated 50 times as for the other datasets under evaluation.

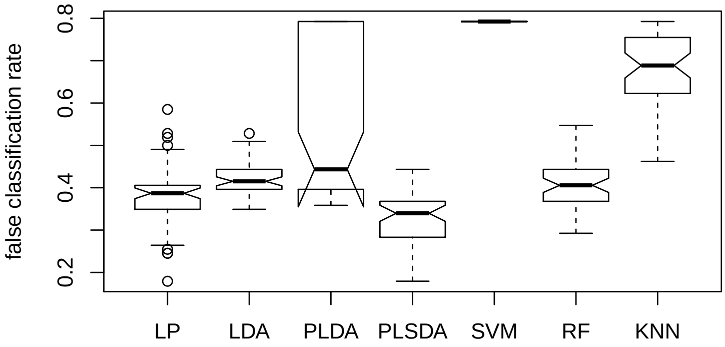


Figure 4.5: The performance in terms of false classification rates of all considered classification methods for 50 repetitions of the *Arcene* dataset is visualized by boxplots.

The performance of Arcene is evaluated in terms of boxplots in Figure 4.5. The classification for this dataset and the designed setup is more challenging than for the other real-world datasets. LP compares well compared to the other evaluated approaches being outperformed only by PLSDA. The performane of 80% false classification rate by SVM might be misleading as worse than random classification. All observations from the non-cancer samples are classified as cancer samples. This could be improved by strategies like oversampling or by changing the majority class assignment to a weighted class assignment. For LP it is not necessary to make adjustments for group imbalance.

Melon

Our final dataset consists of measurements of three types of melons based on spectra analyses of 256 frequencies. The fruits are pertain to three different melon cultivars, with group sizes of 490, 106 and 499, but additional subgroups are known to be present due to changes in the illumination system during the cultivation. The dataset is regularly used as a benchmark dataset for high dimensional and robust data analysis methods (e.g. Hubert and Van Driessen, 2004). Especially the subgroups usually affect non-robust analysis methods. Figure 4.6 provides some insight on the structure of the dataset.

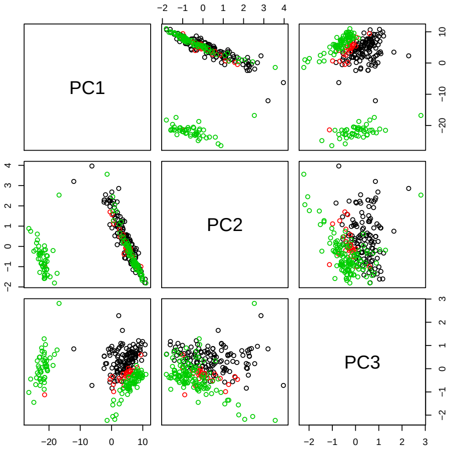


Figure 4.6: Visualization of the first three principal components of one sample of training observations of the *Melon* dataset. We see one subgroup of the green class, represented in the first principal components and a strong overlapping structure for the remaining observations.

We repeatedly sample 25% of the observations from each group as training observations, using the remaining 75% for testing the model performance. The smallest training class therefore consists of 26 observations, leading to a complex classification problem. The performance of the methods under evaluation is presented in Figure 4.7. LP can handle the challenges of the Melon dataset the best and significantly outperforms all compared methods. Especially PLDA results in a high false classification rate which is assumed to be connected to the subgroups and outliers affecting the variable selection.

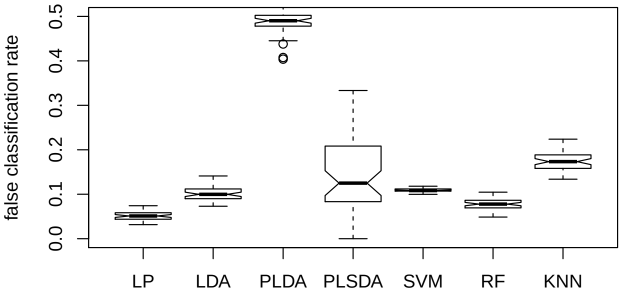
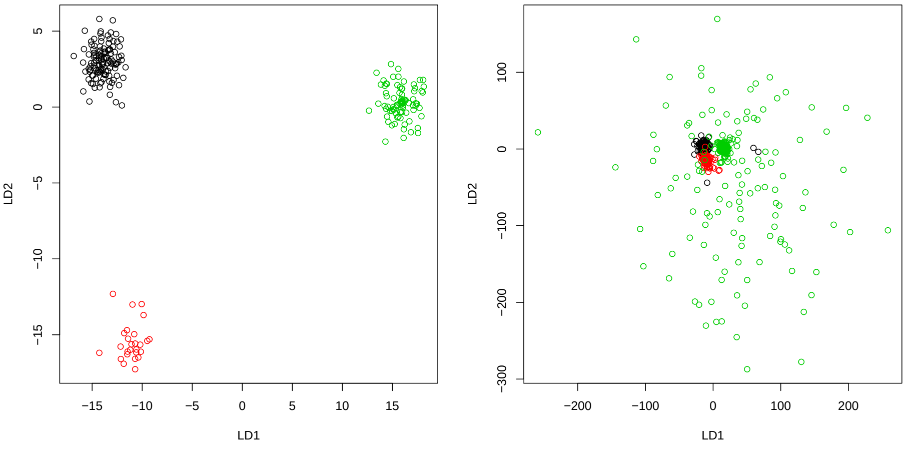


Figure 4.7: The performance in terms of false classification rates of all considered classification methods for 50 repetitions of the *Melon* dataset is visualized by boxplots.

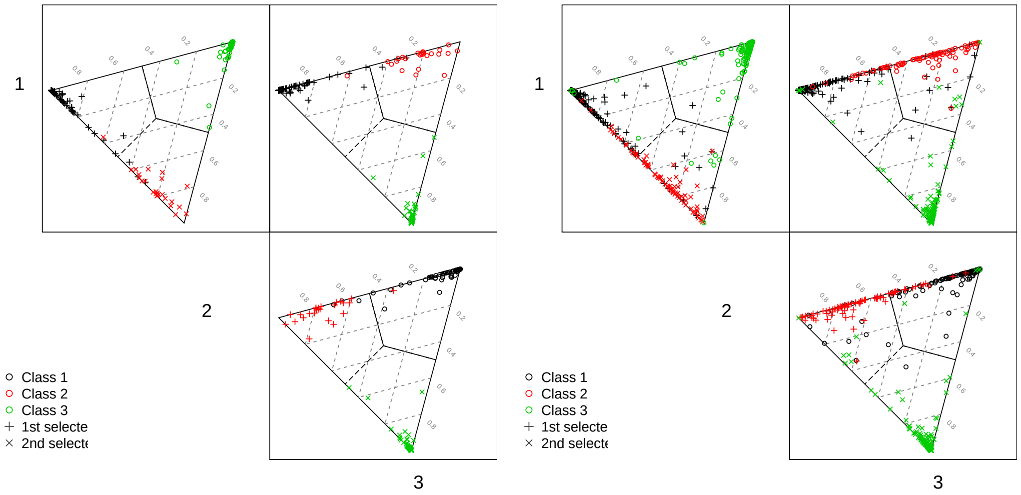
One problem during the visualization of LDA models is the property that in high dimensional spaces with more variables than observations, the training observations will almost always be well separated. Therefore, in a situation where we do not have enough observations to validate the model based on additional observations, a visualization of the discrimination space does not provide a lot of insight on the risks for misclassification of this model. These challenges are visualized in Figure 4.8a and Figure 4.8b.

We see a perfect separation in Figure 4.8a and have no indication of risks of misclassification. The risk of misclassification can be evaluated using the aggregated posterior probabilities of LP, defined in Equation (4.14), which provide the advantage that each of the classification models is located in a low dimensional space. Figure 4.9 provides the visualization of the same data setup as used in Figure 4.8. The risk of misclassifying observations from class 1 as class 2 and vice verse becomes evident in Figure 4.9a and the realization of this risk becomes evident in Figure 4.9b. Note that this visualization can be adapted and used for posterior probabilities computed through cross-validation by any arbitrary classification method.



(a) (b)

Figure 4.8: Plot (a) shows the training observations of the LDA-projection space for one repetition of the *Melon* evaluation. Plot (b) shows the same projection for the respective setup.



(a) (b)

Figure 4.9: The same data setup as in Figure 4.8 is used. Plot (a) shows the proposed visualization of aggregated posterior probabilities from local projections for the training observations. Plot (b) visualizes the same aggregations for the respective test observations.

A further experiment is carried out with the Melon dataset: While in the previous experiment, the training datasets contained 25% of the observations from the original groups (with sizes 490, 106 and 499), we now investigate the effect of modifying the group sizes to be very imbalanced. Six different scenarios under investigation with changing group sizes, but the same overall sample size of $n = 250$ are given in Table 4.1. Figure 4.10 shows the mean misclassification rate over 50 repetitions. Scenario 1 and scenario 6, with the most extreme difference in the group sizes, result in the worst results for several methods. The LDA models are very stable but most of the time they are outperformed by LP. LP is only slightly affected by scenario 1, but otherwise it leads to similar results for the different settings, outperforming all other methods.

Table 4.1: Group sizes for simulation scenarios for the Melon dataset. We vary the numbers of observations per group in order to simulate highly imbalanced group sizes.

Scenario	1	2	3	4	5	6
Class 1	25	50	75	100	125	150
Class 2	75	75	75	75	75	75
Class 3	150	125	100	75	50	25

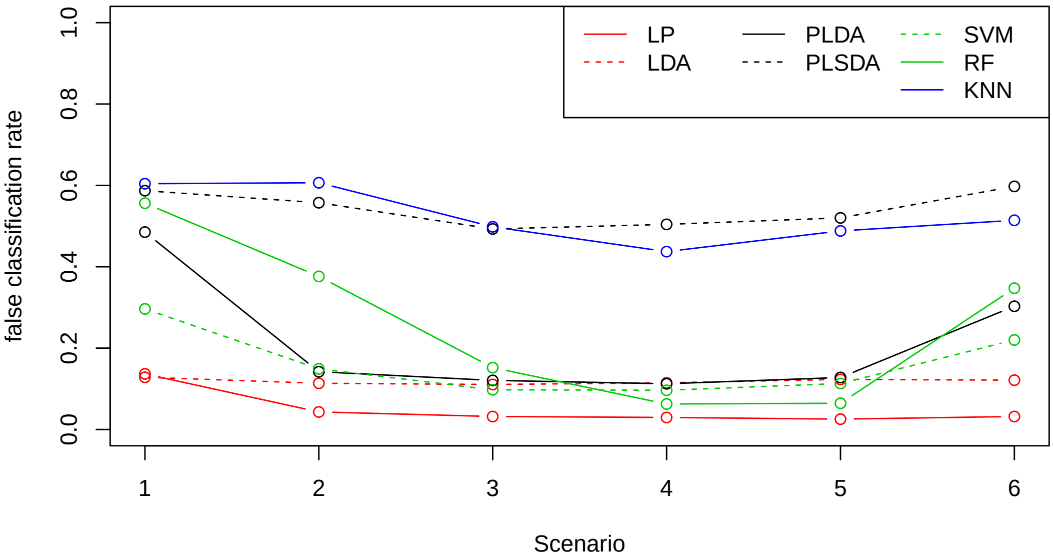


Figure 4.10: The performance of the evaluated classification methods for highly imbalanced datasets is evaluated. The 6 scenarios refer to the scenarios described in Table 4.1. Especially setup 1 and 6 cause a problem for most approaches while LP presents itself as mostly robust towards imbalanced group sizes.

4.5 Conclusions and outlook

We proposed a methodology for supervised classification combining aspects from the field of computer science as well as from the field of statistics. We use the concept of local projections to compute a set of linear discriminant models taking the information within each projection space into account as well as the distance to the projection spaces. The LDA models are then aggregated based on the projection based degree of separation. As shown in Ortner et al. (2017a), local projections can help identifying group structure in high dimensional spaces. Therefore, this way of computing aggregated probabilities for class-membership allows the utilization of LDA for high dimensional spaces while exploiting the advantages of identifying group structure by local projections.

A novel visualization based on ternary diagrams has been proposed which reveals the links between the groups in the high dimensional space. The visualization makes use of the posterior probabilities computed from the local projections, and it thus allows to draw conclusions about the uncertainty of the class assignment, supported by gray areas in the plot for uncertain assignment.

The performance of LP in comparison to related supervised classification methods (LDA, PLDA, PLSDA, SVM, RF and KNN) based on three different real-world datasets demonstrated the advantage of LP in various different settings: two- and multi-group classification tasks, higher number of observations than variables and vice versa, inhomogeneous groups caused by outliers, and imbalanced group sizes. The only tuning parameter for LP is the number k of nearest neighbors, for which a lower and upper boundary has been proposed.

While we utilize linear discriminant analysis performed on the projection space of each local projection there is no reason to limit ourselves to LDA. Depending on the data setup, other methods can be preferred over LDA and still benefit from the local projection based aggregation. A general combination of classification approaches with local projections is still to be evaluated in future work.

List of Figures

1.1 Figure (a) shows the convergence of the variance of $\frac{||X_p||}{E||X_p||}$ of a multivariate p-dimensional independent identically normally distributed variable. Figure (b) shows the relative difference between the maximum distance and the minimum distance between observations. Figure (c) shows the individual behaviour of d_{min} and d_{max}. Note that the behaviour in (b) and (c) is strongly influenced by the sample size n as well. 5

1.2 The expected distance of Y to X as well as the distribution of $||X||$ using various quantiles is visualized for varying p. 6

1.3 The development of Mahalanobis distances for 15 observations from $X \sim N(\mathbf{0}, \mathbf{1})$ (black) and one observation from $Y \sim N(\mathbf{0}, 1.25 \cdot \mathbf{1})$ (red) with increasing p are visualized. All Mahalanobis distances are computed based on the covariance structure of X. Therefore the observations of X converge against 1, while the observation of Y converges against a different constant. . 7

1.4 The degree of locality by using different k in the nearest neighbor based estimation of location and covariance is visualized. For $k = 3$ the data is highly overfitted, while it is underfitted for $k = 40$. The optimal choice of k depends on the data structure as well as on the research question. 10

2.1 This plot demonstrates the concept of guided projections. The figure shows two group structures and the corresponding true covariance structures described by solid ellipses. Each small subset of three observations, represented by x-symbols, will provide a local approximation of this group structure as visualized by the dashed ellipses. The aim of the proposed guided projections approach is to provide a series of such selections, offering a good overall description of the present group structures. Each blue/dark set of x-symbols represents selections from the same group, providing useful information about the group separation, the red/light group represents a mixed selection, where the group structure is masked, i.e. observations from both groups are present inside the ellipse. 19

2.2 Visualization of the selection procedure. To keep the observations in a constant location for each plot we use a two-dimensional space. The distances $OSD_{\mathcal{I}}$ to a selection of observations $\mathcal{I}$ are represented by dashed ellipses. The red/light ellipse represents the smaller distance and therefore the choice for the next observation to be selected. If an observation is part of $\mathcal{I}_L$ or $\mathcal{I}_R$, it is marked with an L or R respectively. Solid points represent observations which have not been selected so far, empty circles have been selected before or are part of a current selection. The next observation to be added to the sequence is marked by a red/light dot. 25

2.3 Subset of the series of projections for simulated data, consisting of two groups with 100 observations each, generated from two different fifty-dimensional normal distributions. The groups are visualized with red circles and blue plus symbols. Each plot represents one step of guided projections, where all observations are projected onto the space spanned by 10 selected observations. 27

2.4 Diagnostic plot utilizing guided projections for the simulated data from Figure 2.3. The colors represent the two clusters, originally located in a fifty-dimensional space. The projection index on the x-axis stands for the index j of $GP_j(\boldsymbol{x})$ of Equation (2.29). For each observation we can follow the change in OSD while slightly changing the projection direction. Similar observations are represented in parallel lines, close to each other. 30

2.5 The solid (black) line refers to the previously described setup one (informative variables only), the dashed (red) line to setup two (including non-informative variables). The transformed validity measures for both setups have been independently scaled to the interval $[0, 1]$ for a better visualization. Both lines are depending on the number of variables related to the respective setup. In total, 1000 observations are simulated for each simulation setup and group to evaluate the considered measures. 31
2.6 The optimization procedure for SPC is visualized. On the x-axis the sparsity parameter is presented, on the y-axis the number of sparse components. The quality of each parameter combination is presented by the color of the respective combination. Red/dark corresponds to a high value of the considered validity measure, grey/light to a low value. Figure (a) shows the optimization for the Gamma index, Figure (b) for the F-measure. For each index, the individual optimum is selected. The sparsity parameter for the F-measure is selected slightly larger than for the Gamma index. The optimal F-measure requires 20 sparse principal components while the Gamma index uses one. . . 33
2.7 For each selected validation measure, we show the mean performance of guided projections (solid blue/dark line) and one other transformation (dashed red/light line) as well as their respective standard errors (dotted lines). The performance of no transformation is shown by the *Raw* category. The start index of the informative variable on the x-axis refers to the parameter r of Equation (2.34) and (2.36). 35
2.8 For the selected validation indices, we analyze the impact of additional noise variables. The mean optimal performance and the respective standard error is visualized for an increasing number of noise variables for guided projections (solid blue/dark line) and one compared method (dashed red/light line). In general we expect a decrease in quality with increasing noise variables. 37
2.9 The performance of data transformations is measured by four different validation measures. 50 randomly selected subsets of the fruit data set are evaluated, based on the originally provided labels. 39
2.10 Diagnostic plot for the full fruit data set. Three groups are present. Additional group structure can be adumbrated. Especially the presence of outliers is evident. The observed group structure reflects the changes in the illumination system while collecting data from melon growth as described in various publications (e.g. Hubert and Van Driessen, 2004). 40

2.11 Diagnostic plot for the glass vessel data set. Only the main group of glass vessels is considered. Red/light lines correspond to identified outliers by the *pcout* algorithm from Filzmoser et al. (2008). 41

3.1 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 $core(\boldsymbol{y})$. 50

3.2 Plot (a) provides a visualization of the transformed observations from Figure 3.1. 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 3.1 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. . 52

3.3 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. 55

3.4 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. 60

3.5 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. 61

3.6 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. 62

3.7 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. 63

3.8 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. 65

3.9 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. 66

4.1 The aggregated posterior probabilities of two classes (class 1 and class 2) compared to the remaining classes is visualized as a ternary diagram. The dashed lines represent classification rules. Observations located in the white areas can be assigned to the respective group, while the grey area represents an uncertain area, where no certain statement can be made. The grey dashed lines refer to posterior probabilities of the selected classes. 77

4.2 The effect of the number of aggregated groups is visualized. Figure (a) refers to a three-class case, (b) to a four-class case, (c) to a five-class case, (d) to a six-class case and (e) to a ten-class case. 78

4.3 A set of ternary diagrams is used to visualize the classification performance for each possible combination of classes. While the color remains constant, we switch labels to emphasize the currently selected classes labeled on the bottom and left of the diagrams. 79

4.4 The performance in terms of false classification rates of all considered classification methods for 50 repetitions of the *Olitos* dataset is visualized by boxplots. 84

4.5 The performance in terms of false classification rates of all considered classification methods for 50 repetitions of the *Arcene* dataset is visualized by boxplots. 84

4.6 Visualization of the first three principal components of one sample of training observations of the *Melon* dataset. We see one subgroup of the green class, represented in the first principal components and a strong overlapping structure for the remaining observations. 85
4.7 The performance in terms of false classification rates of all considered classification methods for 50 repetitions of the *Melon* dataset is visualized by boxplots. 86
4.8 Plot (a) shows the training observations of the LDA-projection space for one repetition of the *Melon* evaluation. Plot (b) shows the same projection for the respective setup. 87
4.9 The same data setup as in Figure 4.8 is used. Plot (a) shows the proposed visualization of aggregated posterior probabilities from local projections for the training observations. Plot (b) visualizes the same aggregations for the respective test observations. 87
4.10 The performance of the evaluated classification methods for highly imbalanced datasets is evaluated. The 6 scenarios refer to the scenarios described in Table 4.1. Especially setup 1 and 6 cause a problem for most approaches while LP presents itself as mostly robust towards imbalanced group sizes. . . 88

List of Tables

4.1 Group sizes for simulation scenarios for the Melon dataset. We vary the numbers of observations per group in order to simulate highly imbalanced group sizes. 88

Bibliography

Abdi, H. and Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4):433–459.

Achlioptas, D. (2003). Database-friendly random projections: Johnson-lindenstrauss with binary coins. Journal of computer and System Sciences, 66(4):671–687.

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. (2005). On k-anonymity and the curse of dimensionality. In Proceedings of the 31st international conference on Very large data bases, pages 901–909. VLDB Endowment.

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

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3):175–185.

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.

Baker, F. B. and Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis. Journal of the American Statistical Association, 70(349):31–38.

Barker, M. and Rayens, W. (2003). Partial least squares for discrimination. Journal of chemometrics, 17(3):166–173.

Bellman, R. (1961). Adaptative control processes.

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is “nearest neighbor” meaningful? In International conference on database theory, pages 217–235. Springer.

Bickel, P. J. and Levina, E. (2004). Some theory for fisher’s linear discriminant function,’naive bayes’, and some alternatives when there are many more variables than observations. Bernoulli, pages 989–1010.

Bosch, A., Zisserman, A., and Munoz, X. (2007). Image classification using random forests and ferns. In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pages 1–8. IEEE.

Breiman, L. (2001). Random forests. Machine learning, 45(1):5–32.

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.

Bühlmann, P. and Van De Geer, S. (2011). Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media.

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.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3):27.

Chen, D., Cao, X., Wen, F., and Sun, J. (2013a). Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3025–3032.

Chen, L., Wang, Y., Liu, N., Lin, D., Weng, C., Zhang, J., Zhu, L., Chen, W., Chen, R., and Feng, S. (2013b). Near-infrared confocal micro-raman spectroscopy combined with pca–lda multivariate analysis for detection of esophageal cancer. Laser Physics, 23(6):065601.

Coifman, R. R. and Lafon, S. (2006). Diffusion maps. Applied and computational harmonic analysis, 21(1):5–30.

Cook, D., Buja, A., and Cabrera, J. (1993). Projection pursuit indexes based on orthonormal function expansions. Journal of Computational and Graphical Statistics, 2(3):225–250.

Cook, D., Buja, A., Cabrera, J., and Hurley, C. (1995). Grand tour and projection pursuit. Journal of Computational and Graphical Statistics, 4(3):155–172.

De Leeuw, J. (2011). History of nonlinear principal component analysis. Visualization and Verbalization of Data.

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

Desgraupes, B. (2013). Clustering indices. University of Paris Ouest-Lab Modal'X, 1:34.

Desgraupes, B. (2016). clustercrit: Compute clustering validation indices. R package version 1.2.7.

Donoho, D. L. et al. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. AMS Math Challenges Lecture, 1:32.

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 [software].

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

Friedman, J. H. (1989). Regularized discriminant analysis. Journal of the American statistical association, 84(405):165–175.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. IEEE Transactions on computers, 100(9):881–890.

Gattone, S. A. and Rocci, R. (2012). Clustering curves on a reduced subspace. Journal of Computational and Graphical Statistics, 21(2):361–379.

Gorban, A. N., Kégl, B., Wunsch, D. C., Zinovyev, A. Y., et al. (2008). Principal manifolds for data visualization and dimension reduction, volume 58. Springer.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar):1157–1182.

Guyon, I., Li, J., Mader, T., Pletscher, P. A., Schneider, G., and Uhr, M. (2007). Competitive baseline methods set new standards for the nips 2003 feature selection benchmark. Pattern recognition letters, 28(12):1438–1444.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L., et al. (1998). Multivariate data analysis, volume 5. Prentice hall Upper Saddle River, NJ.

Hall, P., Marron, J., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(3):427–444.

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.

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

Hoffmann, I., Filzmoser, P., Serneels, S., and Varmuza, K. (2016). Sparse and robust pls for binary classification. Journal of Chemometrics, 30(4):153–162.

Hron, K. and Filzmoser, P. (2013). Robust diagnostics of fuzzy clustering results using the compositional approach. Synergies of Soft Computing and Statistics for Intelligent Data Analysis, pages 245–253.

Hu, Y., Murray, W., and Shan, Y. (2011). Rlof: R parallel implementation of local outlier factor (lof). R package version, 1(0).

Hubert, L. and Schultz, J. (1976). Quadratic assignment as a general data analysis strategy. British journal of mathematical and statistical psychology, 29(2):190–241.

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.

Hung, Y.-C. and Tseng, N.-F. (2013). Extracting informative variables in the validation of two-group causal relationship. Computational Statistics, 28(3):1151–1167.

Ilies, I. and Wilhelm, A. (2010). Projection-based partitioning for large, high-dimensional datasets. Journal of Computational and Graphical Statistics, 19(2):474–492.

Indyk, P. and Motwani, R. (1998). Approximate nearest neighbors: towards removing the curse of dimensionality. In Proceedings of the thirtieth annual ACM symposium on Theory of computing, pages 604–613. ACM.

Janssens, K. H., Deraedt, I., Schalm, O., and Veeckman, J. (1998). Composition of 15–17th century archaeological glass vessels excavated in antwerp, belgium. In Modern Developments and Applications in Microbeam Analysis, pages 253–267. Springer.

Keogh, E. and Mueen, A. (2011). Curse of dimensionality. In Encyclopedia of Machine Learning, pages 257–258. Springer.

Korn, F., Pagel, B.-U., and Faloutsos, C. (2001). On the" dimensionality curse" and the" self-similarity blessing". IEEE Transactions on Knowledge and Data Engineering, 13(1):96–111.

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.

Larsen, B. and Aone, C. (1999). Fast and effective text mining using linear-time document clustering. In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 16–22. ACM.

Lee, E.-K. and Cook, D. (2010). A projection pursuit index for large p small n data. Statistics and Computing, 20(3):381–392.

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.

Li, P., Hastie, T. J., and Church, K. W. (2006). Very sparse random projections. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 287–296. ACM.

Olkin, I. (1992). Quadratic forms in random variables: Theory and applications. Journal of the American Statistical Association, 87(420):1244–1246.

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

Ortner, T., Filzmoser, P., Zaharieva, M., Brodinova, S., and Breiteneder, C. (2017b). Local projections for high-dimensional outlier detection. arXiv preprint arXiv:1708.01550.

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

Qiu, W. and Joe, H. (2015). clustergeneration: Random cluster generation. R package version, 1(4).

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.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. Mathematical statistics and applications, 8:283–297.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20:53–65.

Rousseeuw, P. J. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41(3):212–223.

Sanchez, G. and Determan, C. (2013). Discriminer: tools of the trade for discriminant analysis. R package version, pages 01–29.

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.

Shao, J., Wang, Y., Deng, X., Wang, S., et al. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. The Annals of statistics, 39(2):1241–1265.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288.

Todorov, V. (2014). rrcovhd: Robust multivariate methods for high dimensional data. R package version 0.2-3. Available at https://CRAN. R-project. org/package= rrcovHD, 426:429.

Todorov, V., Filzmoser, P., et al. (2009). An object-oriented framework for robust multivariate analysis. na.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American statistical association, 58(301):236–244.

Wickham, H., Cook, D., Hofmann, H., Buja, A., et al. (2011). tourr: An r package for exploring multivariate data with projections. Journal of Statistical Software, 40(2):1–18.

Witten, D. (2011). penalizedlda: Penalized classification using fisher's linear discriminant. R package version, 1.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using fisher's linear discriminant. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(5):753–772.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics, 10(3):515–534.

Zhang, H., Berg, A. C., Maire, M., and Malik, J. (2006). Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, volume 2, pages 2126–2136. IEEE.

Zhang, K., Hutter, M., and Jin, H. (2009). A new local distance-based outlier detection approach for scattered real-world data. Advances in knowledge discovery and data mining, pages 813–822.

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.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. Journal of computational and graphical statistics, 15(2):265–286.

Index

core, 8, 49, 72
core observations, 8
core space, 8, 51

data sets
  arcene, 83
  fruit/melon, 38, 85
  glass vessels, 40, 63
  melon/fruit, 63
  olive oil, 62, 83
diffusion maps, 17
distances
  core distance, 9, 52
  Mahalanobis distance, 6
  orthogonal distance, 9, 20, 53, 73
  orthogonbal distance, 73
  score distance, 9, 20, 73
  weighted orthogonal distance, 56

evaluation measures, 29

flat data, 1, 46, 70

guided projections, 15, 18
  algortihm, 22
  visualization, 28

high-dimensionality
  blessing, 5
  curse, 3

k-nearest neighbours, 58
  knn-classification, 82

linear discriminant analysis, 70, 81
local discrimination space, 72, 73
local outlier factor, 57
local projection, 8
local projections, 48
LocOut, 56

Outlier Detection in Arbitrary Subspaces (COP), 58

partial least squares for discriminant analysis, 81
PCOut, 57
penalized linear discriminant analysis, 81
posterior probability, 74
  aggregated, 75
principal component analysis, 17
projection pursuit, 17

R-package
  lop, 16

random forest, 82
random projections, 17
RobPCA, 57

sparse principal component analysis, 17
subspace-based outlier detection, 58
support vector machines, 82

variable selection, 17
variables
    informative, 4
    noise, 4, 36
visualization of discrimination, 76

Thomas Ortner

Curriculum Vitae
September 2017

Contact Address
Vienna University of Technology
Institute of Statistics and Mathematical Methods in Economics
Research Group Computational Statistics

Wiedner Hauptstraße 8-10
A-1040 Vienna, Austria

Email: thomas.ortner@tuwien.ac.at
Phone: +43 660 8171344

Career Summary

10/2014 - 10/2017	**TU Wien, Vienna (Austria)** Project assistant in K-Projekt DEXHELP Project assistant in FAMOUS (WWTF IKT12-010)
since 2014	**self-employed: Applied-Statistics OG** Statistical Consulting
2012 - 2014	**self-employed Consultant**
2008 - 2012	**Hauptverband der oesterr. Sozialversicherungsträger** Department of Health Economics

Education

10/2014 - 10/2017	**TU Wien, Vienna (Austria)** PhD training: supervisor Peter Filzmoser *Local Projections for high-dimensional data analysis*
2014	**TU Wien, Vienna (Austria)** Dipl-Ing (M.Sc.), Vienna University of Technology In Mathematics / Mathematical Economics

Publications

Ortner, T., Filzmoser, P., Zaharieva, M., Breiteneder, C., and Brodinova, S., Guided projections for analysing the structure of high-dimensional data. Submitted to Journal of Computational and Graphical Statistics.

Ortner, T., Filzmoser, P., Zaharieva, M., Breiteneder, C., and Brodinova, S., Guided projections for analysing the structure of high-dimensional data. Submitted to Statistical Analysis and Data Mining.

Ortner, T., Hoffmann, I., Filzmoser, P., Zaharieva, M., Breiteneder, C., and Brodinova, S., "Multigroup discrimination based on weighted local projections. Submitted to Journal of Computational and Graphical Statistics.

Brodinova, S., Filzmoser, P., Zaharieva, M., Ortner, T., and Breiteneder, C. Robust and sparse clustering for high-dimensional data. In Proceedings of the Classification and Data Analysis Group (CLADAG) (2017)

Ortner, T., Filzmoser, P., and Endel, G. Identifying Structural Changes in Austrian Social Insurance Data. In Proceedings of 8th Vienna International Conference on Mathematical Modelling (MATHMOD) (2015).

Ortner, T. Multivariate statistische Analyse von Gesundheitsdaten oesterreichischer Sozialversicherungstraeger. Springer-Verlag, 2014.

Conference Talks

Presentation at the *R School* in Grodno (Belarus), April 2017.

Ortner, T., Filzmoser, P., and Endel, G. Identifying Structural Changes in Austrian Social Insurance Data. In 8th Vienna International Conference on Mathematical Modelling (MATHMOD) (2015).

Ortner, T., Breiteneder, C., Brodinova, S., Filzmoser, P., and Zaharieva, M. Guided projections for analysing the structure of high-dimensional data. In International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics) (2016).

Ortner, T., Filzmoser, P., Brodinova, S., Zaharieva, M., and Breiteneder, C. Forward projection for high-dimensional data. In International Conference on Computer Data Analysis and Modeling (CDAM) (2016).

Ortner, T., Filzmoser, P., Brodinova, S., Zaharieva, M., and Breiteneder, C. Local projection for outlier detection. In Oloucian Days of Applied Mathematics (ODAM) (2017).

Selection of Professional Experiences

2016 -	**UNIDO:** Statistical Programmer for Business Intelligence
2016 -	**OEBB:** Data Scientist and Statistician
2016	**Safer Cities:** Prognostics for break-ins in private housing
2015 - 2016	**NOEGUS:** Consultancy in preclinical trials and analysis of medical data
2015 -	**Arsanis Bioscience:** Development of robust IC50 estimation
2015	**Hirslanden Clinic:** Assistance with statistical analysis plan in phase II study
2014 - 2015	**ASFINAG:** Statistical analysis of road conditions