Sparse and robust PLS for binary classification

Irene Hoffmann[a,*], Peter Filzmoser[a], Sven Serneels[b], Kurt Varmuza[a]

[a] *Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna 1040, Austria*
[b] *BASF Corp., Tarrytown NY 10591, USA*

Abstract

Partial robust M regression (PRM), as well as its sparse counterpart sparse partial robust M regression (SPRM), have been reported to be regression methods that foster a partial least squares alike interpretation, while having good robustness and efficiency properties, as well as a low computational cost. In this paper, the partial robust M discriminant analysis classifier (PRM-DA) is introduced, which consists of dimension reduction through an algorithm closely related to PRM and a consecutive robust discriminant analysis in the latent variable space. The method is further generalized to sparse PRM-DA (SPRM-DA) by introducing a sparsity penalty on the estimated direction vectors. Thereby, an intrinsic variable selection is achieved, which yields a better graphical interpretation of the results, as well as more precise coefficient estimates, in case the data contain uninformative variables. Both methods are robust against leverage points within each class, as well as against *adherence outliers* (points that have been assigned a wrong class label). A simulation study investigates the effect of outliers, wrong class labels and uninformative variables on the proposed methods and its classical PLS counterparts, and corroborates the robustness and sparsity claims. The utility of the methods is demonstrated on data from mass spectrometry analysis (TOF-SIMS) of meteorite samples.

Keywords: Discriminant analysis, Partial least squares, Robustness,

*Correspondence to I. Hoffmann, TU Wien, Wiedner Hauptstrasse 8-10/105, A-1040 Wien, Austria.

Email addresses: irene.hoffmann@tuwien.ac.at (Irene Hoffmann), p.filzmoser@tuwien.ac.at (Peter Filzmoser), sven.serneels@basf.com (Sven Serneels), kurt.varmuza@tuwien.ac.at (Kurt Varmuza)

Supervised classification, Variable selection

1. Introduction

Partial Least Squares (PLS) [1], is a powerful and popular method for compressing high dimensional data sets. Commonly it is applied to two data blocks (predictors and response) and projects the data onto a latent structure such that the squared covariance between the blocks is maximized. PLS can deal with a high number of variables p and small sample size n and it is not affected by multicollinearity. Furthermore, it is popular in applied sciences because of the relative ease with which results can be visualised and interpreted. The latent components and scores can be displayed in biplots, which support the interpretation of the model and the understanding of the multivariate data structure.

Many classification methods can only be applied to data with more observations than variables. PLS is a well established tool for effective dimension reduction in the classification setting. Nguyen and Rocke [2] proposed a two step approach for binary classification based on PLS. First the class memberships are modelled as binary variables and are treated for the projection on the latent structure as if they were a continuous response. In the second step a standard classifier, e.g. Fisher's LDA, is applied to the transformed data in the low dimensional space. This method is here referred to as PLS-DA.

The feasibility of such approaches has been discussed by Kemsley [3] and in more detail by Barker [4] and Barker and Rayens [5]. They established the theoretical connection between PLS on binary response and classification and showed that PLS directions maximise the between group variance. PLS classification methods have been applied with considerable success in various scientific research areas, as well as in industry and production. They were used to analyse food quality with conventional sensory profiling data [6], to classify waste water pollution [7] and infrared spectra of olive oils and plant seeds [3]. They have been used for tumor classification with micro-array data [8] and for fault diagnosis in chemical processes [9].

Nevertheless, these methods have their drawbacks [10]. For experimental data two challenges arise frequently which will be addressed here, namely outliers, i.e. samples which are not coherent with the general trend of the data, and uninformative variables, which contain no explanatory power for the response and which commonly appear in large quantity in high dimensional data sets.

Contamination, defect of instruments or wrong assumptions about the distribution of the data, may lead to apparently unreasonable measurements in the samples. In classical PLS, outliers have a much higher influence on the model estimation than ordinary observations and thereby, they distort the model. To avoid this problem in the regression framework, various robust PLS methods have been developed (for an overview, see Filzmoser et al. [11]). Partial Robust M regression (PRM) [12] is among the most popular of these methods, for its trade-off between robustness and (statistical and computational) efficiency. It is robust with respect to leverage points (outliers in the predictor space) and vertical outliers (outliers in the response).

PRM-DA is presented here as a robust alternative to PLS-DA. It inherits the advantages of a PLS method, such as the ability to deal with high dimensional data, multicollinearity and the possibility to illustrate the model in biplots for interpretation. Furthermore, PRM-DA is closely related to PRM regression and as such, has good robustness properties, a high statistical efficiency, and is computationally fast. Due to the data structure of classification problems, the PRM algorithm for regression cannot be directly applied to a binary response but needs specific modifications for the detection of outliers. This aspect is presented in Section 3.

Another problem of increasing importance is the extraction of relevant variables from the data set. Variables which do not provide information about the class membership add unnecessary uncertainty to the model. For data with a high percentage of uninformative variables, biplots become overloaded and a sound interpretation of the model becomes tedious or even impossible [10]. These issues can be countered by sparse modelling. An overview of existing sparse methods in Chemometrics is given in Filzmoser et al. [13]. A sparse coefficient estimate is obtained by imposing a penalty term (e.g. the L_1 norm of the coefficient vector), thanks to which uninformative variables are excluded from the model. In case the data contain uninformative variables, sparsity improves model precision and a parsimonious model is easier to interpret. Chun and Keleş [14] introduced a sparse PLS regression method, which was adapted by Chung and Keleş [15] to the classification setting. Following this approach, the sparse and robust classifier SPRM-DA is introduced in Section 4, which performs intrinsic variable selection and is related to SPRM regression [16].

For the selection of the optimal model, the number of PLS components and a sparsity parameter are determined through K-fold cross validation. The procedure is described in Section 5. To demonstrate the performance of

the methods, simulation studies are conducted in Section 6 and data examples from mass spectrometry are given in Section 7.

2. Projection onto latent structure for discriminant analysis

In the binary classification problem, the data consist of observations from two different populations, henceforth referred to as *group A* and *group B* . Let n_A and n_B denote the number of observations of groups A and B, respectively, and $n = n_A + n_B$. The data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ with p variables can be divided into two subsets $\boldsymbol{X}_A \in \mathbb{R}^{n_A \times p}$ and $\boldsymbol{X}_B \in \mathbb{R}^{n_B \times p}$ containing the observations of groups A and B. In order to disencumber notation, we assume without loss of generality that $\boldsymbol{X}_A$ form the first n_A rows of $\boldsymbol{X}$, followed by the observations $\boldsymbol{X}_B$ of group B.

The first step of PLS-DA, the projection onto latent structure, is methodically equivalent to that in PLS regression. The class memberships of the data are coded in the vector $\tilde{\boldsymbol{y}}$ with 1 for group A and -1 for group B. It is centred and scaled and further treated as if it were a continuous response, denoted by $\boldsymbol{y}$. Furthermore, assume that $\boldsymbol{X}$ is column-wise centred.

Dimension reduction is achieved by projection of the original variables onto a latent structure, such that the covariance between the projection of the predictors and the response is maximized. In detail, the *direction vector* $\boldsymbol{w}_h$ of a PLS model (also known as *weighting vector*) maximizes

$$\boldsymbol{w}_h = \underset{\boldsymbol{w}}{\operatorname{argmax}} \operatorname{cov}^2 (\boldsymbol{X}\boldsymbol{w}, \boldsymbol{y}) , \tag{1a}$$

for $h \in \{1, ..., H\}$ subject to

$$\| \boldsymbol{w}_h \| = 1 \quad \text{and} \quad \boldsymbol{w}_h^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w}_i = 0 \text{ for } 1 \leq i < h. \tag{1b}$$

The direction vectors form the columns of $\boldsymbol{W} \in \mathbb{R}^{p \times H}$ and define the *latent components* or scores $\boldsymbol{T} \in \mathbb{R}^{n \times H}$ as linear combinations of the data, i.e. $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W}$. Since $\boldsymbol{y}$ and $\boldsymbol{X}$ are centred, an estimate of the (squared) covariance in (1a) is

$$\operatorname{cov}^2 (\boldsymbol{X}\boldsymbol{w}, \boldsymbol{y}) = \left(\frac{1}{n-1} \boldsymbol{y}^T \boldsymbol{X} \boldsymbol{w} \right)^2 . \tag{2}$$

Many algorithms exist to solve the maximization problem (1a) with this standard estimator. One of the most prominent is the NIPALS algorithm [17], which will be used in what follows.

After the dimension reduction of the data to dimensionality $H < p$, a standard classifier can be applied to the scores $\boldsymbol{t}_i$, which are the rows of $\boldsymbol{T}$. Here a simple linear classification rule is used, Fisher's Linear Discriminant Analysis (LDA). It assumes equal covariance structure of both groups. The classical pooled within-groups covariance estimate is defined by

$$\begin{aligned} \hat{\boldsymbol{\Sigma}} &= \frac{1}{n-2} \sum_{k \in \{A,B\}} \sum_{i \in C_k} (\boldsymbol{t}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{t}_i - \hat{\boldsymbol{\mu}}_k)^T \qquad (3) \\ \text{with} \quad \hat{\boldsymbol{\mu}}_k &= \frac{1}{n_k} \sum_{i \in C_k} \boldsymbol{t}_i \quad \text{for} \quad k \in \{A, B\} \end{aligned}$$

where $C_A = \{1, ..., n_A\}$ is the index set for group A and $C_B = \{n_A + 1, ..., n\}$ for group B. Following the LDA decision rule, a new observation $\boldsymbol{x}$ is then assigned to that group $k \in \{A, B\}$ that has the largest value of the discriminant score

$$\delta_k(\boldsymbol{x}) = (\boldsymbol{x}^T \boldsymbol{W})^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_k - \frac{1}{2} \hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_k + \log(\pi_k), \qquad (4)$$

wherein π_k are the prior probabilities of group adherence, and $\pi_A + \pi_B = 1$.

Kemsley [3] has shown that the first PLS direction maximizes the univariate between-groups variance. Hence, good group separation can be expected from PLS dimension reduction, which facilitates classification in the score space. A more detailed discussion on the relationship between LDA and PLS is given in Barker and Rayens [5]. It establishes the theoretical foundation to use PLS dimension reduction for classification methods.

3. Robust discriminant analysis with PRM

Outliers in the data distort model estimation and therefore, predictions. Hence, it is essential to verify whether outliers are present in the data and to control their influence. In regression analysis, two types of outliers are generally distinguished: leverage points (outliers in the $\boldsymbol{X}$ space) and vertical outliers (in the $\boldsymbol{y}$ space). Within the framework of discriminant analysis, we need to deal with leverage points separately for each group, since each population has its own data structure. For each group those samples are identified which have values beyond a certain threshold, given the (robustly estimated) covariance structure of the data. The concept of vertical outliers in regression cannot be directly translated to discriminant analysis, since the response is a categorical variable. However, in practice, errors may occur

in the encoding of the group membership. These cases are *adherence outliers*. We label observations as adherence ($\boldsymbol{y}$) outliers, if the supervised group membership, i.e. the class coded in $\boldsymbol{y}$, is intrinsically wrong. This can be assessed by evaluating its position in the estimated score space.

A powerful tool in robust statistics to identify and diminish the influence of outliers, is the concept of M-estimation. Weights between zero and one are assigned to each sample to regularize its influence on the model estimation, whereas weights smaller than one reduce the contribution of an observation to the estimation of the model parameters (and eventually, a zero weight excludes it). In Serneels et al. [12] and Hoffmann et al. [16], it has been described how the concept of M regression can be translated to the PLS regression problem. PLS regression, as well as PLS classification, consists of two steps, which both need to be robustified against the influence of outliers. The first step of PLS-DA is the dimension reduction by projection onto the latent structure. The direction of that projection may be distorted by outliers. In order to construct a robust method, case weights are used in the covariance maximization step (Eq. (2)). The data are then iteratively reweighted to find optimal weights. These weights are then also used to perform weighted, robust LDA in the score space. An overview of the algorithm is presented in Algorithm 1.

The initial weights are derived from the position of an observation within its group. In high dimensions, the distances have less informative value since they get more and more similar with increasing dimensionality [18]. Therefore, the weights are not directly obtained from the distances in the original space. Instead, group-wise PCA is used for dimension reduction, as it has been similarly applied in Filzmoser et al. [19]: The data are split into the two groups $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$, and each column is robustly scaled. Then a classical PCA model is estimated for each group, where the number of components H_k for $k \in \{A, B\}$ can be determined by, e.g., the broken stick rule [1], i.e. to retain the hth component if its eigenvalue is larger than $1/p \sum_{i=h}^{p} 1/i$. Since the data is scaled robustly, the classical variance is dominated by the outliers. Therefore, the first PCA components will highlight variables, which are important for the detection of outliers.

The squared Mahalanobis distance of an observation $\boldsymbol{x}$ with respect to a

[1] Note that for these purposes, several alternative criteria could be applied. A good overview is given in [20].

center $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is defined as

$$\mathrm{MD}^2(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}). \tag{5}$$

Outliers can be detected by the robust squared Mahalanobis distance in the PCA score space, here defined for group A,

$$\tilde{d}_i = \mathrm{MD}^2(\boldsymbol{t}_i^{PCA};\hat{\boldsymbol{\mu}}_A^{PCA},\hat{\boldsymbol{\Sigma}}_A^{PCA}) \quad \text{for } i = 1,...,n_A. \tag{6}$$

It is the distance of the i-th PCA score vector $\boldsymbol{t}_i^{PCA}$ to $\hat{\boldsymbol{\mu}}_A^{PCA}$, the robust centre in the PCA score space of group A, given $\hat{\boldsymbol{\Sigma}}_A^{PCA}$, the robust covariance estimate of the PCA scores. Robust centre and covariance are determined by a fast, high breakdown joint estimate of location and covariance. Estimators suitable for these purposes are e.g. the MM estimator [21] or the Fast MCD algorithm [22]. The results shown in this article are computed using robust starting values based on Fast MCD. In the same way, distances for observations from group B are obtained. Then the distances used for outlier detection are

$$d_i = \frac{\tilde{d}_i}{\mathrm{med}_{j\in C_k}\,\tilde{d}_j}\chi^2_{H_k}(0.5) \quad \text{for } i \in C_k \text{ and } k \in \{A,B\}. \tag{7}$$

The robust squared Mahalanobis distance is approximately $\chi^2_{H_k}$ distributed with the dimension of the data as degrees of freedom if the majority of the data is normally distributed. By the transformation in (7), the median of d_i equals $\chi^2_{H_k}(0.5)$, the 0.5 quantile of the chi-squared distribution with H_k degrees of freedom equal to the number of principal components of the model of group k.

The initial weights are calculated from the distances d_i using Hampel's redescending weighting function

$$\omega_1(d) = \begin{cases} 1 & |d| \le Q_1 \\ \frac{Q_1}{|d|} & Q_1 < |d| \le Q_2 \\ \frac{Q_3-d}{Q_3-Q_2}\frac{Q_1}{|d|} & \text{if } \; Q_2 < |d| \le Q_3 \\ 0 & Q_3 < |d|. \end{cases} \tag{8}$$

A sensible choice for the parameters Q_1, Q_2 and Q_3 are the 0.95, 0.975 and 0.99 quantiles of the chi-squared distribution with as degrees of freedom the number of components used in the PCA model. Then the initial weights are

$\omega_i = \omega_1(d_i)$ for $i = 1, ..., n$. A diagonal matrix $\boldsymbol{\Omega} = \text{diag}(\omega_1, ..., \omega_n)$ is used to downweight the observations. Let $\boldsymbol{X}_\Omega = \boldsymbol{\Omega}\boldsymbol{X}$ be the weighted data matrix, where every observation has been multiplied by its case weight. The weighted response is denoted by $\boldsymbol{y}_\Omega = \boldsymbol{\Omega}\boldsymbol{y}$. Then the PLS maximization criterion is solved for the weighted data,

$$\hat{\boldsymbol{w}}_h = \underset{\boldsymbol{w}}{\text{argmax}}\, \text{cov}^2(\boldsymbol{X}_\Omega \boldsymbol{w}, \boldsymbol{y}_\Omega), \tag{9a}$$

subject to

$$\| \hat{\boldsymbol{w}}_h \| = 1 \quad \text{and} \quad \hat{\boldsymbol{w}}_h^T \boldsymbol{X}_\Omega^T \boldsymbol{X}_\Omega \hat{\boldsymbol{w}}_i = 0 \text{ for } 1 \leq i < h. \tag{9b}$$

The actual maximum is found by applying the NIPALS algorithm to the weighted data.

Starting with the PLS model estimated from data weighted with the initial weights, the case weights are updated iteratively. The score matrix $\boldsymbol{T} = \boldsymbol{X}\hat{\boldsymbol{W}}$ is divided into $\boldsymbol{T}_A$ and $\boldsymbol{T}_B$ with scores which belong to group A and B, respectively. From these matrices, robust Mahalanobis distances are calculated with the fast MCD estimator and then transformed as in (7). As before, the weighting function is applied to these distances d_i and the weights obtained, are $w_i^t = \omega_1(d_i)$. In the algorithm for regression, the calculation of these weights is simplified, because the side constraint = 0 leads to uncorrelated scores. For the classification setting, it is important to consider the covariance structure of the groups.

To identify observations which have probably the wrong coding of the class membership, we use an LDA related approach. Barker and Rayens [5] showed that for a classification problem with two groups, the first PLS direction is the direction that maximizes between group variance. Considering this property, we assume that projection on the first PLS component will lead to good group separation. Let $\boldsymbol{t}_1^{(s)}$ denote the vector of the group wise scaled (not centred) scores of the first component and let m denote the midpoint between the two robust group centers of $\boldsymbol{t}_1^{(s)}$. We use m as the point of separation. For each group the observations with values of $\boldsymbol{t}_1^{(s)}$ on the wrong side of m will be down-weighted. We define

$$\boldsymbol{v} = (\boldsymbol{t}_1^{(s)} - m\boldsymbol{1}_n)^T \tilde{\boldsymbol{y}}, \tag{10}$$

where $\tilde{\boldsymbol{y}}$ is the vector with the class memberships coded as 1 and -1, and $\boldsymbol{1}_n$ is a vector of ones. The entries v_i of $\boldsymbol{v}$ are negative for those observations

for which the corresponding value of the vector $\boldsymbol{t}_1^{(s)}$ does not accord with the given class membership in $\tilde{\boldsymbol{y}}$. Values smaller than a negative threshold should be excluded from the model estimation, since the label in $\tilde{\boldsymbol{y}}$ may be incorrect. For this purpose we use a modified Tukey's Biweight function

$$\omega_2(v) = \begin{cases} 0, & v \leq c \\ \left(1 - \left(\frac{v}{c}\right)^2\right)^2 & \text{if} \quad c < v \leq 0 \\ 1, & v > 0 \end{cases} \tag{11}$$

$$c := \begin{cases} N^{-1}(0.01) & \text{if } N^{-1}(0.01) < 0 \\ 0 & \text{else} \end{cases} \tag{12}$$

with the 0.01 quantile $N^{-1}(0.01)$ of the normal distribution $N(\text{med}(\boldsymbol{v}), 1)$. The weights are denoted by $w_i^y = \omega_2(v_i)$ for $i = 1, ..., n$.

The final case weights for the reweighting of the data are defined as

$$\omega_i = \sqrt{\omega_1\left(d_i\right)\omega_2\left(v_i\right)}. \tag{13}$$

For some situations the weights $\omega_2\left(v_i\right)$ are not reasonable, e.g. when the known class membership of the observations is reliable. Then only the robust distances within each group should be considered, i.e. $\omega_i = \omega_1\left(d_i\right)$. The data are weighted with the updated case weights $\boldsymbol{\Omega} = \text{diag}(\omega_1, ..., \omega_n)$. The reweighting of $\boldsymbol{X}$ by $\boldsymbol{X}_\Omega = \boldsymbol{\Omega}\boldsymbol{X}$ is repeated till convergence of the case weights ω_i.

In the second step of PRM-DA, a robust linear classifier is applied to the scores $\boldsymbol{T} = \boldsymbol{X}\hat{\boldsymbol{W}}$. Robust estimates are plugged into the LDA decision rule described in (4) using the weights derived in the first step. They are defined by

$$\begin{aligned} \hat{\boldsymbol{\mu}}_k &= \frac{\sum_{i \in C_k} \omega_i \boldsymbol{t}_i}{\sum_{i \in C_k} \omega_i} \quad \text{for} \quad k \in \{A, B\} \quad \text{and} \\ \hat{\boldsymbol{\Sigma}} &= \frac{1}{(\sum_{i=1}^n \omega_i) - 2} \sum_{k \in \{A,B\}} \sum_{i \in C_k} \omega_i (\boldsymbol{t}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{t}_i - \hat{\boldsymbol{\mu}}_k)^T. \end{aligned} \tag{14}$$

as in Todorov and Pires [23].

Algorithm 1: PRM-DA

1. Calculate initial case weights:
   - Estimate for each group a PCA model with groupwise robustly scaled data.
   - Choose the number of components of the PCA models by the broken stick rule.
   - Calculate the robust MD^2 (5) of the PCA scores for group A

$$\tilde{d}_i = \mathrm{MD}^2(\boldsymbol{t}_i^{PCA}; \hat{\boldsymbol{\mu}}_A^{PCA}, \hat{\boldsymbol{\Sigma}}_A^{PCA}) \quad \text{for } i = 1, ..., n_A \tag{15}$$

     and analogous for group B.
   - Distances are transformed to d_i as described in (7) and the initial case weights are defined by $\omega_i = \omega_1(d_i)$.
2. Centre data robustly about the column wise median: $\boldsymbol{X}$
   Centre and scale response with mean and standard deviation: $\boldsymbol{y}$
3. Reweighting process: Repeat until convergence of the case weights.
   - Weight data:

$$\begin{aligned} \boldsymbol{X}_\Omega &= \mathrm{diag}(\omega_1, ..., \omega_n)\boldsymbol{X} \\ \boldsymbol{y}_\Omega &= \mathrm{diag}(\omega_1, ..., \omega_n)\boldsymbol{y} \end{aligned}$$

   - Apply NIPALS algorithm for H components to $\boldsymbol{X}_\Omega$ and $\boldsymbol{y}_\Omega$ and obtain robust direction matrix $\boldsymbol{W}_\Omega$. Define scores $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W}_\Omega$.
   - Calculate weights for outliers in the predictor space.
     - Split scores $\boldsymbol{T}$ into group A and B, denoted by $\boldsymbol{T}_A$ and $\boldsymbol{T}_B$.
     - Calculate the robust MD^2

$$\tilde{d}_i = \mathrm{MD}^2(\boldsymbol{t}_i; \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k) \quad \text{for } i \in C_k \text{ and } k \in \{A, B\}. \tag{16}$$

     - $\omega_i^t = \omega_1(d_i)$ with

$$d_i = \frac{\tilde{d}_i}{\mathrm{med}_{j \in C_k} \tilde{d}_j} \chi^2_H(0.5) \quad \text{for } i \in C_k \text{ and } k \in \{A, B\}, \tag{17}$$

       where $\chi^2_H(0.5)$ is the 0.5 quantile of the chi-square distribution with H degrees of freedom.
   - Calculate weights for potentially wrong class labels.

- Robustly scale the first column of $\boldsymbol{T}$ group wise: $\boldsymbol{t}_1^{(s)}$
- Let m be the midpoint between robust group centres of $\boldsymbol{t}_1^{(s)}$.
- Define measure of group coherence

$$\boldsymbol{v} = (\boldsymbol{t}_1^{(s)} - m\mathbf{1}_n)\tilde{\boldsymbol{y}}$$

- Define $\omega_i^y = \omega_2(v_i)$.

- Update case weights $\omega_i = \sqrt{\omega_i^t \omega_i^y}$.

4. Classify with LDA decision rule (4) in the score space based on robust estimates described in (14).

4. Sparse robust discriminant analysis with SPRM

Sparse models are constructed such that only certain variables contribute to the prediction. In PLS based models, sparsity can be achieved when complete rows of $\boldsymbol{W}$ are zero. Then the corresponding variables have no influence on the scores $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W}$.

Chun and Keleş [14] introduced a sparse PLS regression method, which was extended by Chung and Keleş [15] to the classification setting. The central idea is to penalize the estimation of the direction vector $\boldsymbol{w}_h$ by an L_1 norm penalty. To gain more flexibility in the estimation and therefore more sparsity in the model, a surrogate direction vector $\boldsymbol{c}$ is introduced and the PLS criterion (9a) for downweighted data, with the standard covariance estimator (2) as plug-in, is modified to:

$$\min_{\boldsymbol{c},\boldsymbol{w}} -\frac{1}{2}\left(\boldsymbol{y}_\Omega^T \boldsymbol{X}_\Omega \boldsymbol{w}\right)^2 + \frac{1}{2}\left(\boldsymbol{y}_\Omega^T \boldsymbol{X}_\Omega (\boldsymbol{c} - \boldsymbol{w})\right)^2 + \lambda_1 \|\boldsymbol{c}\|_1 \tag{18a}$$

subject to

$$\|\hat{\boldsymbol{w}}\| = 1 \quad \text{and} \quad \hat{\boldsymbol{w}}^T \boldsymbol{X}_\Omega^T \boldsymbol{X}_\Omega \boldsymbol{w}_i = 0 \quad \text{for } 1 \le i < h. \tag{18b}$$

with $\hat{\boldsymbol{c}}$ and $\hat{\boldsymbol{w}}$ are the vectors minimizing (18a). The final estimate of the direction vector is

$$\hat{\boldsymbol{w}}_h = \frac{\hat{\boldsymbol{c}}}{\|\hat{\boldsymbol{c}}\|} \quad \text{for } h = 1, ..., H. \tag{18c}$$

The parameter λ_1 is the sparsity parameter, which controls for the amount of zeros in $\hat{\boldsymbol{c}}$.

Algorithm 2: sparse NIPALS
Let $\boldsymbol{X}$ denote a column wise centred matrix and $\boldsymbol{y}$ the centred response.
Define $\boldsymbol{E}_1 = \boldsymbol{X}$. For $h = 1, ..., H$:

- $\boldsymbol{z}_h = \boldsymbol{E}_h^T \boldsymbol{y} / \|\boldsymbol{E}_h^T \boldsymbol{y}\|$
- $\boldsymbol{v}_h = (|\boldsymbol{z}_h| - \eta \max_i |z_{ih}|) \odot \boldsymbol{I}\left(|\boldsymbol{z}_h| - \eta \max_i |z_{ih}| > 0\right) \odot \operatorname{sgn}(\boldsymbol{z}_h)$
- $\boldsymbol{t}_h = \boldsymbol{E}_h \boldsymbol{v}_h$
- $\boldsymbol{E}_{h+1} = \boldsymbol{E}_h - \boldsymbol{t}_h \boldsymbol{t}_h^T \boldsymbol{E}_h / \|\boldsymbol{t}_h\|^2$

where $\odot$ is the Hadamard (or element wise) matrix product. The weighting vectors $\boldsymbol{v}_h$ of the deflated matrix $\boldsymbol{E}_h$ form the columns of $\boldsymbol{V}$. Then the sparse PLS direction vectors for the transformation of $\boldsymbol{X}$ are defined by $\boldsymbol{W} = \boldsymbol{V}(\boldsymbol{V}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{V})^{-1}$ and the scores are $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W}$.

The minimization problem (18) has an exact solution [14], thanks to which a sparse NIPALS algorithm can be constructed. In Algorithm 2 weighting vectors are penalized by a fraction $\eta \in [0, 1)$ of its largest entry. The expresion $\eta \max_i |z_{ih}|$ replaces λ_1 to facilitate the parameter selection as described in Section 5 since the range of η is known. So the complexity of the models can be varied from the full model to a nearly empty model.

In Hoffmann et al. [16] this approach was robustified for regression analysis. Here the related SPRM-DA algorithm for classification is introduced, which follows the steps described in Algorithm 1 for PRM-DA with the sparse NIPALS (see Algorithm 2) instead of the NIPALS.

5. Parameter selection

For PRM-DA models, the number of latent components H needs to be determined and for the sparse methods additionally the sparsity parameter η has to be specified (see Algorithm 2). K-fold cross validation is a common tool to decide for the model parameters. Thereunto, the samples are divided randomly into K subsets. Each subset is used once as test data, while the rest of the samples are the training data. For a fixed parameter combination H and η, the model is estimated on the training data and the class membership

is predicted for the test data. To compare the predictions across different models, a robust cross validation criterion is introduced.

Since the predicted class membership of outliers is not reliable, its effect on the evaluation should be downweighted. Let $M_A := \{i : y_i = 1 \wedge \text{sign}(\hat{y}_i) = -1\}$ denote the set of indices of misclassified observations from group A within the test data and $C_A := \{i : y_i = 1\}$ all indices from group A test data (analogous for group B). Within each cross validation loop, weights are calculated for the test data according to their position in the estimated score space. Let $\omega_1, \ldots, \omega_n$ denote the resulting weights of all test observations. Then we define the robust misclassification rate as

$$rmcr = \frac{\left(\frac{\sum_{i \in M_A} \omega_i}{\sum_{i \in C_A} \omega_i} + \frac{\sum_{i \in M_B} \omega_i}{\sum_{i \in C_B} \omega_i}\right)}{2}. \tag{19}$$

The class membership of an observation with weight zero has no influence at all on this decision criterion, whereas the misclassification of an observation which is not considered as outlier has the largest influence. This reflects the idea that observations with increasing distance from the main data structure have diminishing influence on the choice of the model. The model with minimum robust misclassification rate is chosen as optimal model.

For data without outliers, i.e. weights equal to one, this criterion is the common misclassification rate, which gives equal importance to the correct classification of both groups, independent of their group size,

$$mcr = \frac{\left(\frac{c_A}{n_A} + \frac{c_B}{n_B}\right)}{2}, \tag{20}$$

where c_A and c_B denote the number of misclassified observations which belong to group A and B, respectively.

6. Simulation studies

We generate data coming from two groups under the assumption that the variables follow a latent structure. Therefore, let $\boldsymbol{D} \in \mathbb{R}^{n \times q}$ consist of a block $\boldsymbol{D}_A$ with $n_A = 60$ rows and a second block $\boldsymbol{D}_B$ with $n_B = 60$ rows coming from multivariate normal distributions with mean $(M, -M, ..., -M) \in \mathbb{R}^q$ and $(M, ..., M) \in \mathbb{R}^q$, respectively, and with equal covariance. The covariance matrix has a block structure with two uncorrelated blocks of equal size; the

covariance between the variables of each block is 0.7 and each variable has variance 1. The size of M determines whether or how much the groups are overlapping. We set $M = 1$ and $q = 10$. Then $\boldsymbol{y}$, which consists of the group memberships, is defined as $y_i = 1$ for $i = 1, ..., n_A$ and $y_i = -1$ for $i = n_A + 1, ..., n$.

We set $H = 2$ and apply the NIPALS algorithm to $\boldsymbol{D}$ and $\boldsymbol{y}$ in order to obtain a direction matrix $\boldsymbol{A}$ and loadings $\boldsymbol{P}$. The scores are $\boldsymbol{T} = (\boldsymbol{D} - \hat{\boldsymbol{\mu}})\boldsymbol{A}$, where $\hat{\boldsymbol{\mu}}$ is the column wise estimated centre of $\boldsymbol{D}$. Then the generated data is given by

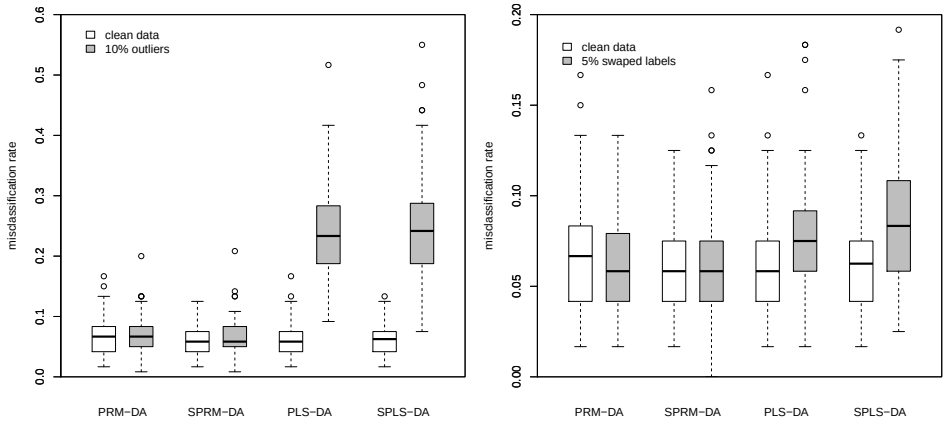
$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P} + \boldsymbol{E}, \tag{21}$$

where the values of $\boldsymbol{E} \in \mathbb{R}^{n \times p}$ come from the independent normal distribution $N(0, 0.2^2)$.

In this study, the data $\boldsymbol{X}$ is manipulated in three different ways to simulate common data problems: (i) Outliers in the predictor space: 10% outliers are generated by replacing the first $0.2n_A$ rows of $\boldsymbol{X}$ by independent values coming from a normal distribution with mean $(0, 10M, ..., 10M)$ and a diagonal covariance matrix with variance 0.1 for each variable. (ii) Wrong class labels: in group B, 10% of the group labels y_i are switched to 1. (iii) Uninformative variables: the rows of $\boldsymbol{T}\boldsymbol{P} \in \mathbb{R}^{n \times q}$ are extended by values from a $p - q = 500$ dimensional normal distribution with zero mean, variances of one and covariances of 0.1. These 500 variables give no information about the class membership.

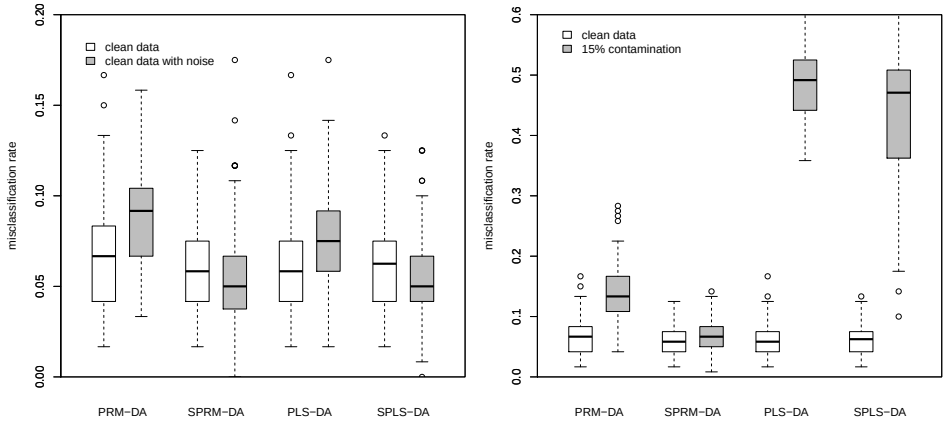
For the evaluation of the proposed methods, PRM-DA and SPRM-DA as well as their classical counterparts PLS-DA and SPLS-DA, training and test data are generated. The parameters are selected with 10-fold cross validation as described in Section 5, with choices of $H = 1, ..., 5$ and for the sparse methods $\eta = 0, 0.1, ..., 0.9$. For the selected parameters a model is estimated on the whole training data set. The model is then evaluated on the test data. Depending on the simulation setting the training data is contaminated with abnormal observations, i.e. outliers in the predictor space or wrong class labels. The test data is free from such contamination and so the accuracy of the classification model can be evaluated with the misclassification rate *mcr* defined in (20).

Figure 1 summarizes the results of the simulation study. The contamination in the predictor space leads to a heavy increase of the *mcr* for the classical methods (see Figure 1a). Also wrong class labels distort the classical methods, while no qualitative change in the *mcr* is visible for the robust



(a) Outliers in the predictor space

(b) Wrong class labels



(c) Uninformative variables

(d) All modifications

Figure 1: Misclassification rate of test data.

methods (see Figure 1b). The limitations of PRM-DA and PLS-DA get visible when uninformative variables are added to the predictors (see Figure 1c). The effect of the combination of all three data problems is presented in Figure 1d. PLS-DA and SPLS-DA fail completely with a median misclassification rate of approximately 50%, which could have been obtained with equal likelihood from random group assignment. The median misclassification rate of PRM-DA does not represent reasonable models either and shows that the method is no longer robust to outliers in the presence of these 500 noise variables. The best results are obtained by SPRM-DA. While the interquartile range increases by the modification of the data, the median misclassification rate of SPRM-DA remains nearly the same.

7. Mass Spectra of Extraterrestrial Material

COSIMA [24] is a TOF-SIMS (time-of-flight secondary ion mass spectrometry) instrument. It is on-board of ESA's Rosetta mission, where it collects dust particles of the comet Churyumov-Gerasimenko on gold or silver targets to study their chemical composition [25]. A twin laboratory instrument of COSIMA, located at Max Planck Institute for Solar System Research (Göttingen Germany), was used to analyze samples of meteorites from the Natural History Museum Vienna to support the analysis of the comet data.

One challenge is to identify the exact positions of comet dust particles on the target and to take measurements there. The spectra are typically obtained at rectangular grid positions located in the estimated area of the particles. The resulting data set consists of spectra taken on the grain as well as spectra from the background of the target.

We demonstrate the utility of the proposed methods for different research questions related to TOF-SIMS measurements on two meteorites (both prepared on the same target). One is meteorite Ochansk (observed fall 30 Aug 1887 near Perm, Russia), the other is meteorite Tieschitz (observed fall 15 Jul 1878 near Olomouc, Czech Republic); both are ordinary chondrites. The number of spectra used is 63 spectra from target background (gold), 155 spectra measured at or near an Ochansk particle, and 25 spectra measured at or near a Tieschitz particle. An original TOF-SIMS spectrum consists of the numbers of secondary ions in 30,222 flight-time bins for the mass range 0 to 400.52 mu. Preprocessing of the spectral data is briefly summarized as follows: mass range used 1 - 150 mu; only mass windows for inorganic ions are considered [26]; signals from the primary ions ($^{115}In^{+}$) excluded; resulting

in $p = 2612$ variables. Because qualitative aspects are of interest, the spectra were normalized to a constant sum (100) of the ion counts (rows in matrix $\boldsymbol{X}$).

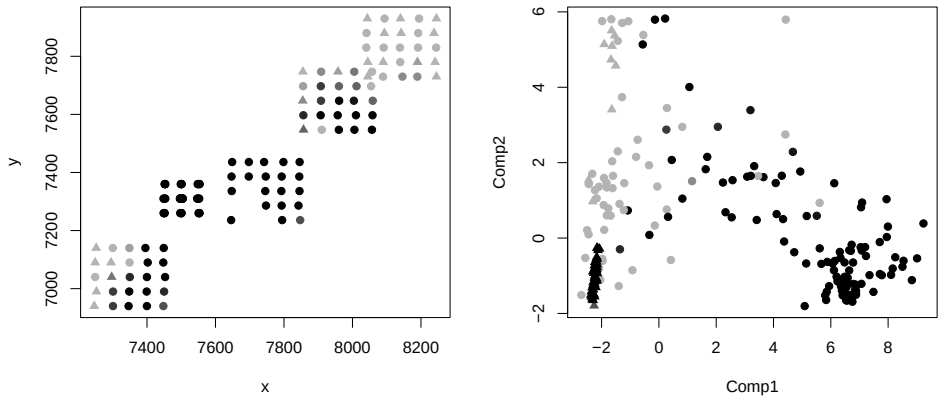
Mislabelled data: TOF-SIMS spectra are measured across grids in the area where the material of interest is suspected. For the Ochansk measurements, visual inspection is possible to locate the grain, but the meteorite material may be spread invisibly in a larger area. At the edge of the meteorite, the spectra consist of a mixture of background and meteorite. For TOF-SIMS measurements of comet grains, it is difficult to locate the dust particles precisely and the recognition of potentially relevant spectra becomes especially important.

We split the spectra measured on and in the neighbourhood of the meteorite (group A) and background spectra (group B) randomly into five subsets, such that the sizes of the groups across each split are approximately the same. Sequentially, each of the subsets is used as test data, while the rest is training data. An SPRM-DA model is estimated on the training data and the parameters of the model are chosen as described in Section 5 with 10 fold cross validation within the training set. Then the model is used to predict the class membership and to calculate weights of the test data.

To obtain a meaningful model, it is important for the estimation that spectra are used which were actually measured on the meteorite, i.e. that those spectra which come from the grain with a high probability have weight one. Class assignment of test samples to a group is only reliable for observations that are embedded in training data with weights equal to one. So one has to look jointly at weights and class prediction to gain more insight into the structure of the data and the meaning of the classification model.

The results for group A are shown in Figure 2a given the x- and y-coordinates of the measurements on the target. The weights for potentially mislabelled data ω_i^y are represented by the grey tone, black for weight one and continuously lighter grey for weights smaller than one. Small weights mean, that the corresponding sample is located close to the background samples. The area with black samples in Figure 2a coincides well with the area where the grain is visible on the target. All these spectra are predicted to belong to group A. It shows that this approach builds classification models based on the relevant data and by prediction of the weights also gives information about the applicability domain of the models.

For illustrative purposes, an SPRM-DA model is estimated for the com-



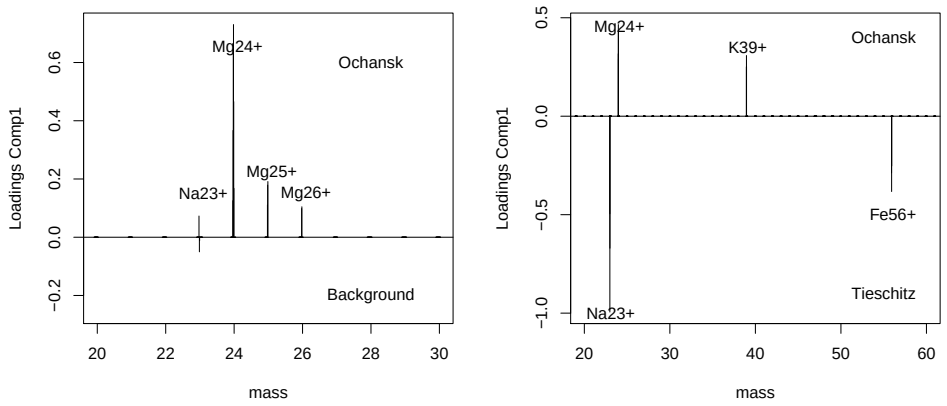
(a) Predicted weights for potentially mislabelled samples (black: weight one) and class prediction (circle for grain group) given their position (x,y in μm) on the target.

(b) Scores of the first two components of a SPRM-DA model for all Ochansk (circles) and background (triangles) data. The grey scores have case weights ω_i smaller than one.

Figure 2: SPRM-DA model for Ochansk and background spectra

plete data set. The parameters found with 10 fold cross validation are $H = 5$ and $\eta = 0.1$. The remaining number of variables (mass bins) in the model is 128. Figure 2b shows the scores of both groups for the first two components. From this two dimensional projection we can already see that samples from the meteorite group (circles) which are close to the background data (triangles) have small weights, i.e. are colored in light grey. Figure 3a shows that in the sparse loadings of the first component the magnesium isotopes are relevant for the separation between meteorite and background.

Outliers in the predictor space: Data from the meteorite Ochansk (group A) are compared to data of Tieschitz (group B). In this context the measurements on the two meteorite grains should form the two discriminant groups, while off grain measurements or other irregular data are considered as outliers. A pre-selection of grain spectra secures that the groups are not dominated by background spectra. Therefore, models of a meteorite grain and background data as described in the previous paragraphs are used to predict the class membership of the test data. Samples from the meteorite



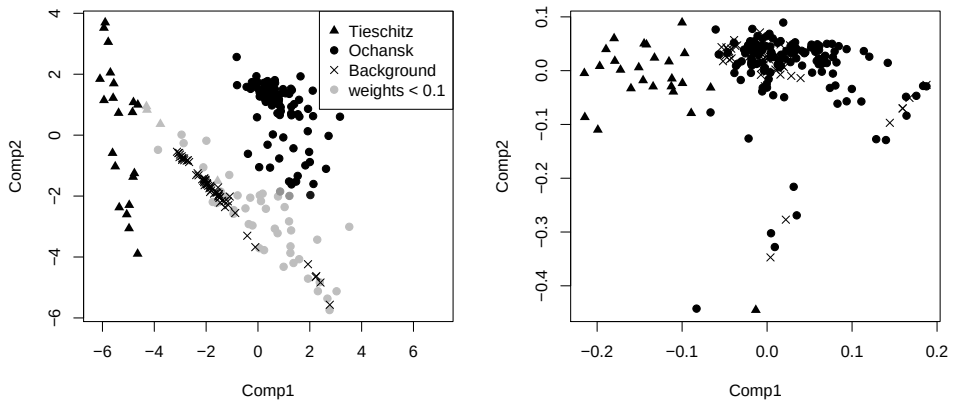
(a) Ochansk and background model (b) Ochansk and Tieschitz model

Figure 3: Selected range of the loadings of the first component for SPRM-DA model.

group that are assigned to the background group, are excluded. This leads to $n_A = 155$ and $n_B = 25$. Due to the small group size we use 5 fold cross validation to choose the model parameters. They are $H = 2$ and $\eta = 0.2$, which leads to a model with 30 mass bins.

In spite of the pre-selection, we expect the main source of outliers to come from background measurements considered as meteorite spectra. To validate the model, the group of background spectra is projected into the SPRM score space. Figure 4a shows, that several Ochansk spectra (and one Tieschitz spectra) are located in the same area as the background spectra. Since the scores in this area receive weights smaller than 0.1, they are reasonably identified as outliers and they are not relevant for the model estimation. In comparison, Figure 4b shows the score plot for an SPLS model with two components. Background spectra projected into the score space are spread over the whole area of the Ochansk spectra, so that in the group of Ochansk no distinction between spectra measured on grain or off grain is possible.

The SPRM model separates the two ordinary chondrite meteorites well and the first component gives insight into the different elemental compositions of the two meteorites (see Figure 3b). Tieschitz has higher counts for sodium and iron and Ochansk for magnesium and potassium. This is also visible in the mean spectra of the two groups in Figure 5.



(a) SPRM scores of Ochansk and Tieschitz - with background spectra projected into the score space.

(b) SPLS score plot of Ochansk and Tieschitz - with background spectra projected into the score space.

Figure 4: Score plots for models of Ochansk and Tieschitz.

8. Conclusion

In this paper, a novel methodology for robust and when necessary, sparse, classification has been outlined. Several methods exist to estimate robust or sparse classification models but to the best of our knowledge this is the first proposal of a sparse and robust method for binary classification. It inherits the visualisation and interpretation advantages that PLS-DA offers over many machine learning tools, the latter tendentionally yielding black box solutions. In contrast to classical PLS-DA, however, the new method is robust both to leverage points within each class, as well as to class adherence outliers. The method thanks its robustness essentially to a double pronged iterative reweighting scheme wrapped around the (sparse) NIPALS algorithm. Thereby, it is very germane to the earlier (sparse and non-sparse) partial robust M regression method for regression and has similar robustness and

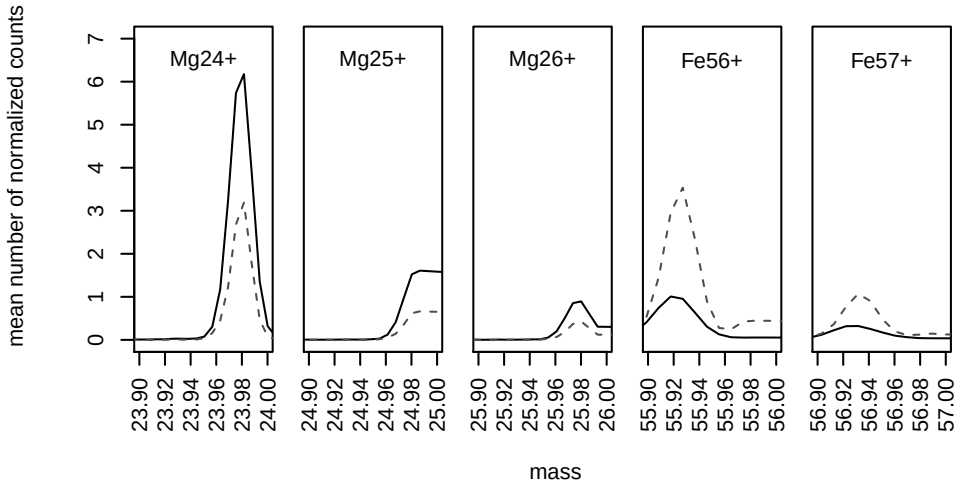


Figure 5: Selected mass ranges of mean spectra for Ochansk (black, solid line) and Tieschitz (grey, dashed line).

sparsity properties [2].

A simulation study has shown that outliers (leverage points and adherence outliers), as well as the presence of uninformative variables, can mislead PLS-DA and artificially inflate the misclassification rate. The new methods, on the contrary, still yield virtually unaffected misclassification performance in the presence of outliers. The sparse method (SPRM-DA) is the only method that also yields pristine performance when the data contain both outliers and a non-negligible number of uninformative variables, even though also in this setting, PRM-DA still outperforms both classical methods, showing that the impact of outliers is the more harsh type of contamination studied. The simulations have also shown that for data without outliers, the performance of (S)PLS-DA and (S)PRM-DA is very similar. One would usually expect a slight advantage of the classical over the robust methods, in particular for very low sample sizes, because under normality the classical methods are known to be statistically more efficient than robust methods [28]. In practice, however, only the robust method allows to verify if outliers are present or not

[2]Implementations of these two methods have been made publicly available through the R package `sprm`, which can be downloaded through the CRAN network since 2014. Both new classification methods, as well as cross-validation and visualisation tools, have been appended to the same package in the latest version update [27].

by investigating the case weights. The performance of SPRM-DA has been tested on a data set from meteorite samples, where it has largely managed to identify outliers and to classify samples according to the compositional classes they should belong to.

Acknowledgments

This work is supported by the Austrian Science Fund (FWF), project P 26871-N20. The authors thank F. Brandstätter, L. Ferrière, and C. Koeberl (Natural History Museum Vienna, Austria) for providing meteorite samples, C. Engrand (Centre de Sciences Nucléaires et de Sciences de la Matière, Orsay, France) for sample preparation, and M. Hilchenbach (Max Planck Institute for Solar System Research, Göttingen, Germany) for TOF-SIMS measurements.

References

[1] S. Wold, M. Sjöström, and L. Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58 (2):109–130, 2001.

[2] D.V. Nguyen and D.M. Rocke. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1): 39–50, 2002.

[3] E.K. Kemsley. Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. *Chemometrics and Intelligent Laboratory Systems*, 33(1):47–61, 1996.

[4] R. Barker. *Partial least squares for discrimination*. PhD thesis, University of Kentucky, 2000.

[5] M. Barker and W. Rayens. Partial least squares for discrimination. *Journal of Chemometrics*, 17(3):166–173, 2003.

[6] K. Rossini, S. Verdun, V. Cariou, E.M. Qannari, and F.S. Fogliatto. PLS discriminant analysis applied to conventional sensory profiling data. *Food Quality and Preference*, 23(1):18–24, 2012.

[7] E. Sääksjärvi, M. Khalighi, and P. Minkkinen. Waste water pollution modelling in the southern area of Lake Saimaa, Finland, by the SIMCA pattern recognition method. *Chemometrics and Intelligent Laboratory Systems*, 7(1):171–180, 1989.

[8] M. Pérez-Enciso and M. Tenenhaus. Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Human Genetics*, 112(5-6):581–592, 2003.

[9] L.H. Chiang, E.L. Russell, and R.D. Braatz. Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 50(2):243–252, 2000.

[10] N. Kettaneh, A. Berglund, and S. Wold. PCA and PLS with very large data sets. *Computational Statistics & Data Analysis*, 48(1):69–85, 2005.

[11] P. Filzmoser, S. Serneels, R. Maronna, and P.J. Van Espen. Robust multivariate methods in chemometrics. In S.D. Brown, R. Tauler, and B. Walczak, editors, *Comprehensive Chemometrics*, volume 3, pages 681–722. Elsevier, Oxford, 2009.

[12] S. Serneels, C. Croux, P. Filzmoser, and P.J. Van Espen. Partial robust M-regression. *Chemometrics and Intelligent Laboratory Systems*, 79(1-2):55 – 64, 2005.

[13] P. Filzmoser, M. Gschwandtner, and V. Todorov. Review of sparse methods in regression and classification with application to chemometrics. *Journal of Chemometrics*, 26(3-4):42–51, 2012.

[14] H. Chun and S. Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.

[15] D. Chung and S. Keleş. Sparse partial least squares classification for high dimensional data. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

[16] I. Hoffmann, S. Serneels, P. Filzmoser, and C. Croux. Sparse partial robust M regression. *Chemometrics and Intelligent Laboratory Systems*, 2015. In print.

[17] H. Wold. Soft Modeling by Latent Variables; the Nonlinear Iterative Partial Least Squares Approach. *Perspectives in Probability and Statistics. Papers in Honour of M. S. Bartlett*, 1975.

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

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

[20] D.A. Jackson. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, pages 2204–2214, 1993.

[21] S. Van Aelst M. Salibián-Barrera and G. Willems. Principal components analysis based on multivariate MM-estimators with fast and robust bootstrap. *Journal of the American Statistical Association*, 101: 11981211, 2006.

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

[23] V. Todorov and A.M. Pires. Comparative performance of several robust linear discriminant analysis methods. *REVSTAT Statistical Journal*, 5: 63–83, 2007.

[24] J. Kissel, K. Altwegg, B.C. Clark, L. Colangeli, H. Cottin, S. Czempiel, J. Eibl, C. Engrand, H.M. Fehringer, B. Feuerbacher, et al. COSIMA–high resolution time-of-flight secondary ion mass spectrometer for the analysis of cometary dust particles onboard Rosetta. *Space Science Reviews*, 128(1-4):823–867, 2007.

[25] R. Schulz, M. Hilchenbach, Y. Langevin, J. Kissel, J. Silen, C. Briois, C. Engrand, K. Hornung, D. Baklouti, A. Bardyn, et al. Comet

67P/Churyumov-Gerasimenko sheds dust coat accumulated over the past four years. *Nature*, 518(7538):216–218, 2015.

[26] K. Varmuza, C. Engrand, P. Filzmoser, M. Hilchenbach, J. Kissel, H. Krüger, J. Silén, and M. Trieloff. Random projection for dimensionality reduction – applied to time-of-flight secondary ion mass spectrometry data. *Analytica Chimica Acta*, 705(1):48–55, 2011.

[27] S. Serneels and I. Hoffmann. *sprm: Sparse and Non-Sparse Partial Robust M Regression and Classification*, 2015. R package version 1.2.

[28] R. Maronna, R.D. Martin, and V.J. Yohai. *Robust Statistics: Theory and Methods*. Wiley, New York, 2006.