Data Mining and Knowledge Discovery (2020) 34:723–741
https://doi.org/10.1007/s10618-019-00666-8

Robust and sparse multigroup classification by the optimal scoring approach

Irene Ortner[1] · Peter Filzmoser[2] · Christophe Croux[3]

Received: 12 June 2017 / Accepted: 28 November 2019 / Published online: 20 February 2020

Abstract

We propose a robust and sparse classification method based on the optimal scoring approach. It is also applicable if the number of variables exceeds the number of observations. The data are first projected into a low dimensional subspace according to an optimal scoring criterion. The projection only includes a subset of the original variables (sparse modeling) and is not distorted by outliers (robust modeling). In this low dimensional subspace classification is performed by minimizing a robust Mahalanobis distance to the group centers. The low dimensional representation of the data is also useful for visualization purposes. We discuss the algorithm for the proposed method in detail. A simulation study illustrates the properties of robust and sparse classification by optimal scoring compared to the non-robust and/or non-sparse alternative methods. Three real data applications are given.

Keywords High dimensional data · Linear discriminant analysis · Penalization · Robustness · Supervised classification · Variable selection

1 Introduction

In linear discriminant analysis (LDA) the data originate from K different populations. The aim is to find linear decision boundaries to separate the observations from the K groups as good as possible and to predict the class membership of new, unlabeled observations. Several formulations for LDA exist. Fisher's approach to LDA searches for directions that maximize the between group variance given the within group variance.

Responsible editor: Jieping Ye.

✉ Irene Ortner
irene.ortner@applied-statistics.at

[1] Applied Statistics GmbH, Taubstummengasse 4/10, 1040 Wien, Austria

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

[3] Economics and Finance Faculty, Edhec Business School, Lille, France

Equivalently, one can take the conditional class densities as multivariate normal with the same covariance matrix, and apply the Bayes classification rule. The formulation for LDA considered in this paper is optimal scoring (Hastie et al. 1994). It recasts the classification problem into a regression framework and models the class-membership with a quantitative parameter for each class.

While these different approaches to LDA yield the same classification results (Johnson et al. 2002; Witten and Tibshirani 2011) they are all limited to settings with more observations n than variables p. Optimal scoring enables us to transfer new developments in high dimensional regression analysis to the classification context. In regression analysis the problem of high dimensional data, in particular data with more variables than observations, attracts a lot of attention. A variety of so-called sparse methods have been developed. The best known is the Lasso regression estimate (Tibshirani 2011). For a response $\boldsymbol{y} \in \mathbb{R}^n$ and a column-wise centered and scaled predictor matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, it is defined as

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1$$

for regression coefficients $\boldsymbol{\beta}$, where $\|\boldsymbol{a}\|^2 = \sum_{i=1}^p a_i^2$ is the squred Euclidean norm and $\|\boldsymbol{a}\|_1 = \sum_{i=1}^p |a_i|$ the L_1 norm, for a vector $\boldsymbol{a} = (a_1, \ldots, a_p)^T$. Fast algorithms have been developed for Lasso regression (Efron et al. 2004; Wu and Lange 2008). The Lasso shrinks several of the estimated regression coefficients to zero, and is therefore said to be *sparse*. The zero coefficients correspond to the variables that are not selected into the model. Hence, the Lasso performs simultaneous model estimation and variable selection. The sparsity tuning parameter is λ, and increasing values of λ will favor more coefficients equal to zero and sparser models. This is especially useful for data sets including uninformative variables which do not contribute information to predicting the response. When uninformative variables are excluded, the precision of the estimation increases and the models are easier to interpret. Recently, Clemmensen et al. (2012) proposed a sparse version of multigroup LDA, by adding an L_1 penalty to the objective function of the optimal scoring problem. This leads to a sparse discriminant analysis method applicable for $n < p$ as well.

In this paper we propose a *robust* version of sparse optimal scoring. It is robust because it is resistant to outliers. In linear discriminant analysis an observation is considered an outlier if it belongs to group k but differs from the majority of observations in group k, for $k = 1, \ldots, K$. It is well known, that outliers may render a statistical method completely unreliable. The sample covariance matrix and the group average can be heavily distorted by single extreme observations and classification rules based on them will be unreliable. This will not happen if a robust method is used.

A variety of robust classification methods have been proposed (Hubert and Van Driessen 2004; Todorov and Pires 2007) but generally they are not applicable for data with $n < p$. Vanden Branden and Hubert (2005) proposed a robust classifier for high dimensions based on SIMCA, but it does not use sparse modeling, so all variables are included in the model. A sparse and robust classification method based on partial least squares was proposed by Hoffmann et al. (2016) but only for binary classification problems. Robust optimal scoring, even the non-sparse case, was not considered before in the literature.

The paper is structured as follows. In Sect. 2, we review the optimal scoring approach to linear discriminant analysis. In Sect. 3, we introduce the proposed method and present the algorithm in detail. In Sect. 4, a strategy is outlined to select the sparsity tuning parameter. A simulation study competing with existing alternative methods is presented in Sect. 5. Illustrations on real world examples are given in Sect. 6.

2 Optimal scoring for multigroup classification

We follow the notation of Clemmensen et al. (2012) to outline the optimal scoring method. Let $\boldsymbol{X}$ be the $n \times p$ data matrix with the observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ in its rows and $\boldsymbol{Y}$ an $n \times K$ matrix of dummy variables coding the class membership of the observations, i.e. $y_{ik} = 1$ if and only if observation $\boldsymbol{x}_i$ belongs to group k, and zero otherwise. The rows of $\boldsymbol{Y}$ are denoted by $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. The columns of $\boldsymbol{X}$ are centered to have mean zero and scaled to have unit variance. The aim of optimal scoring is to find projection vectors $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_H$, such that each $\boldsymbol{X}\hat{\boldsymbol{\beta}}_h$ is a good prediction of the corresponding vector $\boldsymbol{Y}\hat{\boldsymbol{\theta}}_h$, for $h = 1, \ldots, H$, where the vector $\boldsymbol{Y}\hat{\boldsymbol{\theta}}_h$ contains the scores of the group each observation belongs to. The K components of the score vector $\hat{\boldsymbol{\theta}}_h$ are the numeric scores assigned to each of the groups. One takes H smaller than the number of groups K, commonly $H = K - 1$.

The projection vectors $\hat{\boldsymbol{\beta}}_h$ and the score vectors $\hat{\boldsymbol{\theta}}_h$ are obtained sequentially. Let $\boldsymbol{D} = \frac{1}{n}\boldsymbol{Y}^T\boldsymbol{Y}$ be a $K \times K$ diagonal matrix of class proportions. Set $\hat{\boldsymbol{\theta}}_0 = \boldsymbol{1}_K$, the K-vector of ones. Then solve for $h = 1, \ldots, H$

$$\min_{\boldsymbol{\beta}_h, \boldsymbol{\theta}_h} \frac{1}{n}\|\boldsymbol{Y}\boldsymbol{\theta}_h - \boldsymbol{X}\boldsymbol{\beta}_h\|^2 \quad \text{s.t.} \quad \boldsymbol{\theta}_h^T\boldsymbol{D}\boldsymbol{\theta}_h = 1, \quad \boldsymbol{Q}_h^T\boldsymbol{D}\boldsymbol{\theta}_h = \boldsymbol{0},$$

where $\boldsymbol{Q}_h = [\boldsymbol{Q}_{h-1}, \hat{\boldsymbol{\theta}}_{h-1}]$ is a $K \times h$ matrix.

The sparse optimal scoring method of Clemmensen et al. (2012) simply adds an L_1 penalty to the objective function.

$$\min_{\boldsymbol{\beta}_h, \boldsymbol{\theta}_h} \frac{1}{n}\|\boldsymbol{Y}\boldsymbol{\theta}_h - \boldsymbol{X}\boldsymbol{\beta}_h\|^2 + \lambda\|\boldsymbol{\beta}_h\|_1 \quad \text{s.t.} \quad \boldsymbol{\theta}_h^T\boldsymbol{D}\boldsymbol{\theta}_h = 1, \quad \boldsymbol{Q}_h^T\boldsymbol{D}\boldsymbol{\theta}_h = 0. \tag{1}$$

Estimators $\hat{\boldsymbol{\beta}}_h$ and $\hat{\boldsymbol{\theta}}_h$ that solve (1) can be obtained iteratively. Starting with a random vector for $\boldsymbol{\theta}_h$ one computes the Lasso for $\boldsymbol{\beta}_h$. For a given $\boldsymbol{\beta}_h$ there exists an explicit solution of (1) for $\boldsymbol{\theta}_h$. One iterates further until convergence. For details, see Clemmensen et al. (2012).

Once the projection vectors are obtained, a standard LDA is performed in a low dimensional space of dimension H. Let denote by $z_1, \ldots, z_n$ the projected observations in the rows of $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B}$, $\boldsymbol{B} = [\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_H]$. Denote the group averages of the projected observations by $\boldsymbol{m}_k = \frac{1}{n_k}\sum_{i \in C_k} z_i$, where C_k denotes the index set for observations from class k, and n_k is the number of observations in class k, for $k = 1, \ldots, K$. The within group covariance matrix is

$$\boldsymbol{S} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i \in C_k} (\boldsymbol{z}_i - \boldsymbol{m}_k)(\boldsymbol{z}_i - \boldsymbol{m}_k)^T.$$

The Mahalanobis distance of an observation $\boldsymbol{z}$ to the center $\boldsymbol{m}_k$ is given by

$$MD(\boldsymbol{z}; \boldsymbol{m}_k, \boldsymbol{S}) = \left((\boldsymbol{z} - \boldsymbol{m}_k)^T \boldsymbol{S}^{-1} (\boldsymbol{z} - \boldsymbol{m}_k)\right)^{1/2}.$$

An observation $\boldsymbol{x}$, transformed to $\boldsymbol{z} = \boldsymbol{x}^T \boldsymbol{B}$, is then assigned to the class k with smallest value of

$$MD(\boldsymbol{z}; \boldsymbol{m}_k, \boldsymbol{S})^2 - 2\log(\pi_k).$$

Here, π_k is the prior probability belonging to group k, with $\pi_1 + \cdots + \pi_K = 1$. In the following, π_k is set to the class proportion of group k, so $\pi_k = n_k/n$.

3 Robust and sparse optimal scoring

We now propose an optimal scoring algorithm for data containing outliers and possibly more variables than observations. Furthermore, not all variables contribute information about the class membership of the observations, in the following referred to as uninformative variables, and therefore we aim at sparse estimation. To the best of our knowledge, sparse and robust classification methods for multiple groups have not been considered in literature so far. Even the non-sparse case, robust optimal scoring is a new approach to robust classification problems.

The aim of the algorithm is to reduce the influence of outlying observations on the model estimation. A common and powerful approach to achieve this in a regression model is the iteratively re-weighted least squares algorithm (Wolke and Schwetlick 1988). Given a robust initial estimator, the influence of observations with large residuals is down-weighted by case weights. The coefficient estimates and the case weights are iteratively re-estimated. Here we will take a related approach.

The data matrix $\boldsymbol{X}$ is robustly centered by the coordinate-wise median and each column is scaled by the median absolute deviation (MAD) (Hampel 1974). The MAD is defined by $\text{MAD}(a_1, \ldots, a_n) = 1.48 \,\text{med}_i\, |a_i - \text{med}_j\, a_j|$ where $\text{med}_i\, a_i$ denotes the median of $a_1, \ldots, a_n$ and 1.48 is a factor to get consistency at normal distribution.

3.1 Initial estimation

The vectors $\hat{\boldsymbol{\beta}}_h$ and $\hat{\boldsymbol{\theta}}_h$ are estimated sequentially for $h = 1, \ldots, H$. As before, $\hat{\boldsymbol{\theta}}_0 = \mathbf{1}_K$. First we obtain initial estimates for $\hat{\boldsymbol{\beta}}_h$ and $\hat{\boldsymbol{\theta}}_h$. It is important that they are robust to outliers and can be computed in high dimensions. These initial estimates

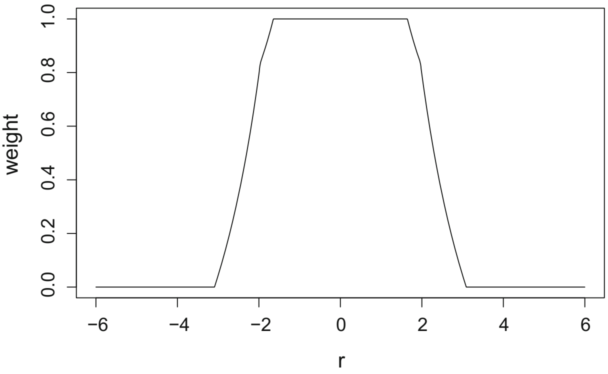

Fig. 1 Hampel's re-descending weighting function

start up the iterative procedure to get the final $\hat{\boldsymbol{\beta}}_h$ and $\hat{\boldsymbol{\theta}}_h$. "Appendix" provides full details for their computation.

3.2 Outlier weights

Residuals are computed as

$$r_i = \boldsymbol{y}_i^T \hat{\boldsymbol{\theta}}_h - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_h \quad \text{for } i = 1, \ldots, n.$$

The observations will be weighted so that potential outliers receive less weight. The weights are computed from the residuals. Weights are calculated separately for each group. Denote $r_i^{(s)}$ the robustly standardized residuals where we center by the median and scale by the MAD.

Hampel's re-descending weighting function (Hampel et al. 1986) is applied to the standardized residuals to obtain weights for each observation. This weighting function (plotted in Fig. 1) is given as

$$\omega_H(r) = \begin{cases} 1 & |r| \leq q_1 \\ \frac{q_1}{|r|} & q_1 < |r| \leq q_2 \\ \frac{q_3 - r}{q_3 - q_2} \frac{q_1}{|r|} & \text{if } q_2 < |r| \leq q_3 \\ 0 & q_3 < |r| \end{cases}$$

where the parameters q_1, q_2 and q_3 are set to the 0.95, 0.975 and 0.999 quantiles of the standard normal distribution, respectively, i.e. $q_1 = 1.64$, $q_2 = 1.96$ and $q_3 = 3.09$. The case weights are then $\omega_i = \omega_H(r_i^{(s)})$ for $i = 1, \ldots, n$. Under the assumption of normality of the residuals, 90% of the observations will receive weight $\omega_i = 1$ and 0.2% receive weight $\omega_i = 0$.

3.3 Solving the weighted sparse optimal scoring problem

Let $\boldsymbol{\Omega}$ be a diagonal matrix with the case weights $\omega_1, \ldots, \omega_n$ on the diagonal. Then define $\tilde{\boldsymbol{Y}} = \boldsymbol{\Omega}^{1/2}\boldsymbol{Y}$ and $\tilde{\boldsymbol{X}} = \boldsymbol{\Omega}^{1/2}\boldsymbol{X}$ the weighted data matrices. The diagonal matrix $\tilde{\boldsymbol{D}} = \frac{1}{\sum \omega_i}\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}}$ contains on its diagonal the share of the total weight coming from each group's observations. The weighted sparse optimal scoring problem is defined as

$$\min_{\boldsymbol{\beta}_h, \boldsymbol{\theta}_h} \frac{1}{\sum \omega_i} \|\tilde{\boldsymbol{Y}}\boldsymbol{\theta}_h - \tilde{\boldsymbol{X}}\boldsymbol{\beta}_h\|^2 + \lambda\|\boldsymbol{\beta}_h\|_1 \quad \text{s.t.} \quad \boldsymbol{\theta}_h^T\tilde{\boldsymbol{D}}\boldsymbol{\theta}_h = 1, \quad \boldsymbol{Q}_h^T\tilde{\boldsymbol{D}}\boldsymbol{\theta}_l = \boldsymbol{0}. \tag{2}$$

If no outliers are detected, all weights are one, $\sum \omega_i = n$, $\boldsymbol{\Omega}$ is the identity matrix and Eq. (2) is the standard optimal scoring problem Eq. (1).

Equation (2) is solved by an alternating iterative scheme. For given $\hat{\boldsymbol{\theta}}_h$ it reduces to the weighted Lasso regression problem

$$\hat{\boldsymbol{\beta}}_h = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{\sum \omega_i} \sum_{i=1}^{n} (\boldsymbol{y}_i^T\hat{\boldsymbol{\theta}}_h - \boldsymbol{x}_i^T\boldsymbol{\beta})^2\omega_i + \lambda\|\boldsymbol{\beta}\|_1. \tag{3}$$

which is equivalent to solving the Lasso for the weighted data, with sparsity parameter given by $\lambda \sum w_i/n$. For a given $\hat{\boldsymbol{\beta}}_h$, the optimization problem Eq. (2) is solved by

$$\hat{\boldsymbol{\theta}}_h = c\left\{\boldsymbol{I} - \boldsymbol{Q}_h(\boldsymbol{Q}_h^T\tilde{\boldsymbol{D}}\boldsymbol{Q}_h)^{-1}\boldsymbol{Q}_h^T\tilde{\boldsymbol{D}}\right\}(\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}_h) \tag{4}$$

where c is a scalar so that $\hat{\boldsymbol{\theta}}_h$ fulfills the side constraint $\hat{\boldsymbol{\theta}}_h^T\tilde{\boldsymbol{D}}\hat{\boldsymbol{\theta}}_h = 1$. The derivation of Eq. (4) is given in the "Appendix". Notice that the last part in parentheses in Eq. (4) is proportional to $(\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}})^{-1}\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}_h$, the OLS estimate of $\boldsymbol{\theta}_h$ when regressing $\tilde{\boldsymbol{Y}}$ on $\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}_h$ without side constraints.

After computing $\hat{\boldsymbol{\beta}}_h$ and $\hat{\boldsymbol{\theta}}_h$, new residuals r_i and case weights ω_i, for $i = 1, \ldots, n$, are calculated as described previously. New estimates of coefficient and score vectors are computed on the re-weighted data as in Eqs. (3) and (4).

The notation in Eq. (3) highlights the role of the weights ω_i: for $\omega_i = 0$ the term $(\boldsymbol{y}_i^T\hat{\boldsymbol{\theta}}_h - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$ does not contribute to the parameter estimation, for $\omega_i = 1$ the squared residual is fully considered in the objective function of Eq. (3). Intermediate values of the weights correspond to a reduced, but non-zero, influence of the observation on the estimators.

3.4 Convergence criterion

Let $\omega_1^j, \ldots, \omega_n^j$ denote the case weights and $\hat{\boldsymbol{\beta}}_h^j$ and $\hat{\boldsymbol{\theta}}_h^j$ the estimates in the jth iteration step. Then the weighted mean residual sum of squares with Lasso penalty in the jth iteration step is

$$L_h^j = \sum_{i=1}^{n} (\boldsymbol{y}_i^T \hat{\boldsymbol{\theta}}_h^j - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_h^j)^2 \omega_i^j + \sum_{i=1}^{n} \omega_i^j \lambda \|\hat{\boldsymbol{\beta}}_h^j\|_1.$$

The convergence criterion for stopping the iterative procedure is chosen as $|L_h^j - L_h^{j-1}|/L_h^j < 10^{-4}$.

3.5 Classification rule

The iterative procedure outlined in Sects. 3.1–3.4 provides a projection matrix $\boldsymbol{B} = [\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_H]$. We project the data into an H dimensional subspace, i.e. $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B}$, with the rows $z_1, \ldots, z_n$. We observed that for a large sparsity parameter λ, the last column(s) of $\boldsymbol{B}$ may consist of only zeros. Then the dimension of the classification problem on the projected data is reduced automatically.

Instead of computing sample averages and covariance matrices of the projected data, we compute a robust location and covariance matrix estimator. For this, we take the minimum covariance determinant (MCD) described in Rousseeuw and Van Driessen (1999). The robust group centers $\boldsymbol{m}_k$, for $k = 1, \ldots, K$, are the MCD location estimates of the projected observations from the kth group, i.e. of z_i, $i \in C_k$. Then the projected observations are group-wise centered, $\tilde{z}_i = z_i - \boldsymbol{m}_k$ for $i \in C_k$ and $k = 1, \ldots, K$. A robust covariance estimate $\boldsymbol{S}$ from these pooled centered observations is obtained by the MCD covariance matrix estimate (Rousseeuw and Van Driessen 1999). The decision rule for a new observation $\boldsymbol{x}$ is as follows. Project $\boldsymbol{x}$ to the subspace, $z = \boldsymbol{x}^T\boldsymbol{B}$ and compute the Mahalanobis distances to the group centers $\boldsymbol{m}_k$ with respect to $\boldsymbol{S}$. Assign $\boldsymbol{x}$ to group

$$\underset{k=1,\ldots,K}{\operatorname{argmin}} \, (z - \boldsymbol{m}_k)^T \boldsymbol{S}^{-1} (z - \boldsymbol{m}_k) - 2\log(\pi_k).$$

4 Model selection and evaluation

Two steps are necessary for the proper evaluation of the proposed method. First, a strategy to select an optimal sparsity parameter is needed, second, the prediction performance for new observations is evaluated. We split the data randomly into calibration data and test data.

To select the optimal sparsity parameter λ^*, five-fold cross validation is performed on the calibration set. We split the calibration data randomly into $J = 5$ blocks of approximately equal size such that the observations from each class are evenly spread across the blocks. Each of the five blocks is used in turn as validation set and the rest as training set. For a sequence of values for the sparsity parameter $\lambda_1, \ldots, \lambda_L$ (covering the range between the full and the empty model) classification models are estimated on the training data and evaluated on the validation data. Since the decision for the optimal sparsity parameter λ^* should not be influenced by outliers, we propose to use a weighted misclassification rate (wmcr) for evaluation. For the jth validation set, which consists of n_j observations $\boldsymbol{x}_1^j, \ldots, \boldsymbol{x}_{n_j}^j$ define

$$\mathrm{wmcr}(\boldsymbol{x}_1^j, \ldots, \boldsymbol{x}_{n_j}^j, \lambda) = \frac{1}{K} \sum_{k=1,\ldots,K} \frac{\sum_{i \in M_k^j(\lambda)} w_i^j(\lambda)}{\sum_{i \in C_k^j} w_i^j(\lambda)}, \tag{5}$$

where C_k^j is the index set of all observations from the validation set belonging to group k, and $M_k^j(\lambda)$ is the subset of C_k^j containing the indices of misclassified observations (for the model estimated with sparsity parameter λ). The weight $w_i^j(\lambda)$ of an observation $\boldsymbol{x}_i^j$ is derived from the Mahalanobis distance to its closest center in the projected subspace, i.e.

$$MD_i^j(\lambda) = \min_{k=1,\ldots,K} MD\left(\boldsymbol{x}_i^{jT} \boldsymbol{B}; \boldsymbol{m}_k, \boldsymbol{S}\right),$$

where $\boldsymbol{B}$, $\boldsymbol{m}_k$ and $\boldsymbol{S}$ are estimated on the jth training set with sparsity parameter λ. Then the weight is defined as

$$w_i^j(\lambda) = \begin{cases} 1 & MD_i^j(\lambda)^2 \leq \chi_H^2(0.975) \\ 1/MD_i^j(\lambda) & \text{else} \end{cases},$$

where $\chi_H^2(0.975)$ denotes the 97.5% quantile of the χ^2 distribution with H degrees of freedom. When all weights are equal to one, the wmcr is equivalent to the misclassification rate (mcr), the mean of the proportion of misclassified observations from each group.

The tuning parameter can now be selected such that the average wmcr over the $J = 5$ validation sets is minimized, i.e.

$$\tilde{\lambda} = \operatorname*{argmin}_{\lambda \in \{\lambda_1, \ldots, \lambda_L\}} \frac{1}{J} \sum_j l^j(\lambda),$$

where, for easier notation, $l^j(\lambda) = \mathrm{wmcr}(\boldsymbol{x}_1^j, \ldots, \boldsymbol{x}_{n_j}^j, \lambda)$, for $j = 1, \ldots, J$.

We then use the one-standard-error rule (Hastie et al. 2015): choose the model with largest sparsity parameter such that its average wmcr is still within one standard error of the minimum average wmcr. Thus, the optimal sparsity parameter with the one-standard-error rule is

$$\lambda^* = \max\left\{\lambda \in \{\lambda, \ldots, \lambda_L\} \mid \frac{1}{J}\sum_{j=1}^{J} l^j(\lambda) < \frac{1}{J}\sum_{j=1}^{J} l^j(\tilde{\lambda}) + \mathrm{se}(l^1(\tilde{\lambda}), \ldots, l^J(\tilde{\lambda}))\right\},$$

where $\mathrm{se}(a_1, \ldots, a_J) = \sqrt{\mathrm{var}(a_1, \ldots, a_J)/J}$ denotes the standard error. This strategy favors more parsimonious models and is a safeguard against over-fitting. With the optimal sparsity parameter λ^* the final model is estimated on the whole calibration data and we obtain $\boldsymbol{B}$, $\boldsymbol{m}_k$ and $\boldsymbol{S}$.

For the evaluation of the model, the class memberships of test data are predicted. Since the evaluation should not be distorted by outliers in the test data, we use the weighted misclassification rate Eq. (5). In the simulation study, since the test data are generated without outliers, we set the weights equal to one. In Sects. 5 and 6 we compare robust sparse optimal scoring (rSOS) with classical sparse optimal scoring (cSOS). The sparsity parameter for both methods is selected in the same way, but for cSOS the standard unweighted mcr is minimized.

5 Simulation study

Simulation schemes: The data are generated from $K = 3$ different p-dimensional normal distributions representing three groups. The distributions have equal covariance structure, but different mean vectors. For group k ($k = 1, 2, 3$) let the mean be a vector of length p with value 2 for the kth variable and zeros elsewhere. So the number of informative variables is $q = 3$. The diagonal of the covariance matrix is a vector of ones. The covariance between the informative variables is 0.1 and zero between all others. The number of observations is $n = 120$, where each group consists of $n_k = 40$ observations.

In the first scenario of this simulation study the effect of increasing $p \in \{3, 13, 23, 53, 103, 203\}$ is illustrated, i.e. of increasing the number of uninformative variables while the number of informative variables $q = 3$ is fixed. The second scenario shows the effect of outliers on the methods, also for increasing $p \in \{3, 13, 23, 53, 103, 203\}$. Outliers are included in the calibration data by taking 10% of the observations of the first group and replacing their values for the first variable by random values from $N(-10, 1)$. Hence there are still two uncontaminated informative variables. In a third scenario the number of uninformative variables is fixed at 50 and the third informative variable is removed, i.e. $p = 52$. Outliers are again only generated in the first group by replacing the values of the first variable by random values from $N(-20, 1)$. This setting is more challenging because only one uncontaminated informative variable remains, and because the outliers take more extreme values. The percentage of outliers in the first group ranges from 0 to 45% by steps of 5%, allowing to observe the influence of increasing levels of contamination. Finally, simulation scenario four is used to study the effect of the sample size n on the performance of the algorithms. For this purpose, the second scenario is modified as follows. The number of variables is fixed with $p = 53$ for an increasing number of observations $n \in \{120, 600, 1200, 120{,}000\}$.

Methods and evaluation: The results from robust sparse optimal scoring (rSOS) and classical sparse optimal scoring (cSOS) are compared. For settings where non-sparse classification methods can be applied (i.e. $n > p$), models are estimated with LDA and robust LDA (rLDA). The latter method uses the MCD of the pooled centered data as robust covariance matrix estimator, where the centers of each group are estimated by the location MCD estimator, see Hubert and Van Driessen (2004). Recall that LDA is equivalent to classical optimal scoring. As a benchmark, we first remove all uninformative variables and outliers from the calibration data and then apply LDA. This benchmark method cannot be applied in practice, since one does not know which variables are informative and which observations are outliers. We refer to this method

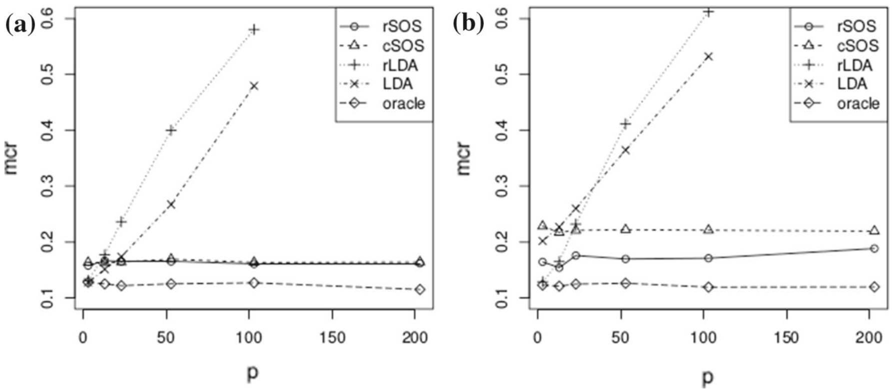

Fig. 2 Misclassification rate (mcr) averaged over 100 simulation runs as a function of p, the number of variables. **a** Scenario 1: models estimated on clean calibration data; **b** scenario 2: models estimated on calibration data with 10% outliers in one group

as oracle; it gives an estimate of the lower bound for the best misclassification rate we can achieve with linear boundaries.

For rSOS the sparsity parameter λ is selected with five-fold cross validation on the calibration data ($n = 120$) from a grid of values between 0.1 and 2 with step size 0.05, as described in Sect. 4.

To evaluate the models, test data of size $n = 120$ are generated in the same way as the calibration data, but without outliers for all scenarios. The predicted class membership of the test data is compared to the known, true class membership and the misclassification rate (mcr) is reported. Other criteria of the quality of the model concern the number of correctly selected variables. The false negative rate (FNR) is the fraction of informative variables not included in the model, the false positive rate (FPR) refers to the fraction of uninformative variables included in the model.

Simulation results: The results from the first scenario demonstrate the advantage of sparse modeling when the number of uninformative variables increases. Figure 2a shows the misclassification rate for test data, averaged over 100 simulation runs. The benchmark mcr for this simulation design is about 12.5%, as can be seen from the results of the oracle. Hardly any difference between the performance of cSOS and rSOS is visible in this setting. The mcr of both methods remains stable with increasing number of uninformative variables. In very low dimensions, for instance $p = 3$, LDA and rLDA slightly outperform cSOS and rSOS, but with increasing p, LDA and rLDA quickly break down and give bad classification results, even for p still smaller than n. This shows that excluding uninformative variables is crucial for the quality of the prediction performance.

Table 1a shows the quality of the variable selection for cSOS and rSOS. The false negative rate is slightly higher for cSOS whereas the false positive rate is slightly higher for rSOS. Overall, both rates are low for both methods, which implies that the variable selection with the L_1 penalty achieves good results.

In the second scenario the effect of 10% outliers is investigated, see Fig. 2b. The benchmark given by the oracle is of course again about 12.5%. For $p = 3$ the robust

Table 1 Variable selection: the false negative rate (FNR) and the false positive rate (FPR) is averaged over 100 simulation runs for classical and for robust SOS for **a** scenario 1: models estimated on clean calibration data; **b** scenario 2: models estimated on calibration data with 10% outliers in one group

p	3	13	23	53	103	203
(a) Scenario 1						
FNR cSOS	0.02	0.02	0.02	0.04	0.01	0.04
FNR rSOS	0.01	0.02	0.02	0.02	0.00	0.02
FPR cSOS		0.02	0.02	0.01	0.01	0.00
FPR rSOS		0.04	0.03	0.02	0.01	0.02
(b) Scenario 2						
FNR cSOS	0.32	0.32	0.32	0.33	0.33	0.33
FNR rSOS	0.05	0.03	0.07	0.05	0.06	0.08
FPR cSOS		0.02	0.02	0.00	0.00	0.00
FPR rSOS		0.07	0.05	0.03	0.04	0.02

methods rLDA and rSOS outperform the classical methods LDA and cSOS. Increasing the number of variables heavily affects LDA, but also rLDA. The best performing method is rSOS, since it can cope with both increasing dimensions and outliers. Note that for cSOS the presence of outliers substantially increases the mcr but the number of uninformative variables has no further notable effect; for rSOS the mean mcr slightly increases when p tends to its highest value.

Table 1b shows that cSOS fails to identify the informative variables in presence of outliers. The FNR of cSOS is around 33% in this scenario, since the first of the three informative variables, the contaminated one, is not included anymore in the model. In this scenario the variables selected by cSOS do not contain any outliers, but since the information present in the first variable is lost, it still leads to an increased mcr. With rSOS this information can be recovered: rSOS down-weights the outliers and is able to reveal that this first variable contributes enough information to be selected. Comparing the FNR of rSOS in Table 1a and b shows an increase in the setting with outliers, but considerably smaller compared with cSOS. Finally, note that the FPR for rSOS is low, but slightly higher than for cSOS. In the second scenario, rSOS selects on average 4.7 variables, a bit more than the average of 2.2 variables for cSOS.

Scenario three illustrates how the percentage of outliers influences the classification performance of the different methods. Figure 3 pictures the mcr as a function of the proportion of outliers in the calibration data, for $p = 52$. The benchmark given by the oracle is about 22.2%, and indicates a lower bound for the mcr. When there are no outliers, cSOS and rSOS are close to the oracle. However, already for only 5% outliers the cSOS is strongly affected in its prediction performance, whereas rSOS remains to give reasonable results for larger percentages of outliers. As expected, the mcr of LDA and rLDA is inflated due to the $p - q = 50$ uninformative variables resulting in a high mean mcr, which increases slightly for higher percentages of contamination.

Figure 3b presents the results from scenario four and shows the best area of application as well as the limitations of the proposed algorithm. The number of variables is of moderate size and increasing the number of observations improves the performance of both non-sparse classifiers LDA and rLDA. Outliers in the simulation setting lead to heavy distortion of the classical methods LDA and cSOS. For small sample size

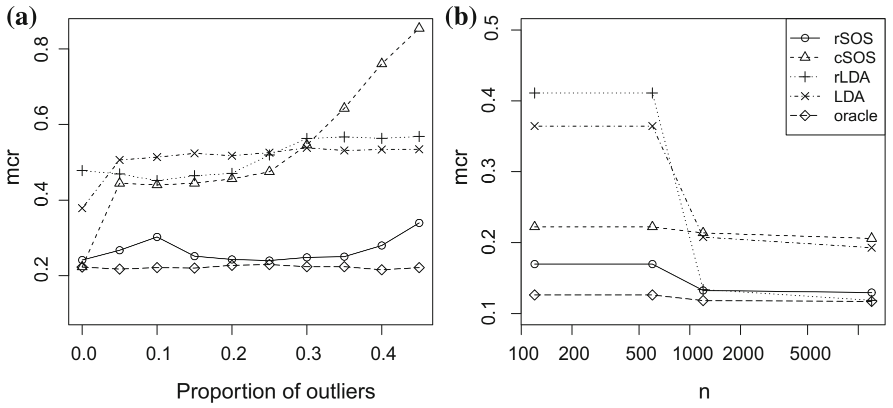

Fig. 3 Misclassification rate (mcr) averaged over 100 simulation runs, **a** scenario 3: as a function of increasing outlier proportion; $p = 52$, $n = 120$, **b** scenario 4: as a function of n, the number of observations (on log-scale); $p = 53$, outlier proportion 10%

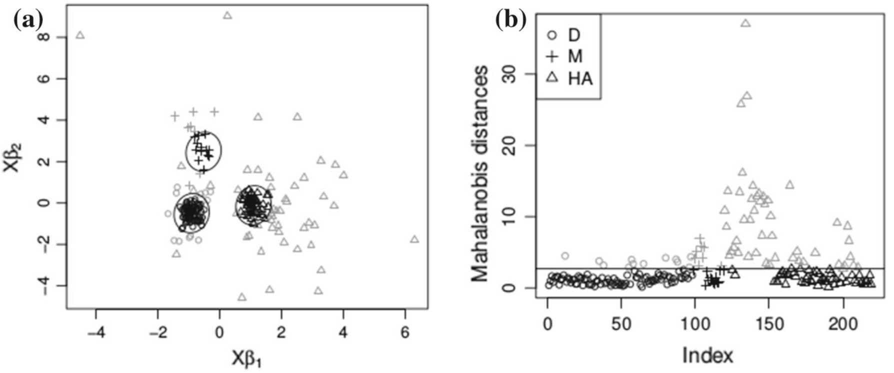

Fig. 4 Fruit data: **a** visualization of 219 test observations in the projected subspace, **b** Mahalanobis distance of each projected test observation to its group center. Observations with weight smaller than one are colored in gray

($n = 120, n = 600$) rSOS clearly outperforms all competitors whereas with sufficiently large sample size the robust covariance structure in rLDA can be properly estimated and slightly outperforms rSOS.

Computations are performed in R (R Core Team 2016). For classical sparse optimal scoring, R code is available in the package `sparseLDA` (Clemmensen and Kuhn 2012). The code for robust sparse optimal scoring is included in the package `rrcovHD` (Todorov 2016) in the function `SosDiscRobust`. The number of iteration steps for this algorithm depends on p with an average of 8.7 iterations for $p = 3$ and 11.7 for $p = 203$ in the simulation studies with $n = 120$. Increasing n leads to fewer iterations.

Table 2 Fruit data: average (w)mcr is the (weighted) mcr averaged over the five test data sets

	cSOS	rSOS
Average mcr	0.028 (0.0062)	0.041 (0.0068)
Average wmcr	0.016 (0.0116)	0.009 (0.0029)

6 Examples

Fruit data: This data set consist of $n = 1095$ measurements with $p = 256$ wavelengths from $K = 3$ different cultivars of the cantaloupe melon, named D, M and HA. We have 490 measurements from group D, 106 from group M and 499 from group HA. It is a well known benchmark data set to demonstrate the stability of robust classification methods (Hubert et al. 2008; Hubert and Van Driessen 2004; Vanden Branden and Hubert 2005). From former analyses it is known that the change of illumination led to outliers.

The data are 5 times split into calibration and test data (80% versus 20%), such that all observations are included in the test data once and the observations from each class are evenly spread across the test sets. For each calibration data set the optimal sparsity parameter λ^* is selected as described in Sect. 4 from a fine grid starting with 10^{-4} up to 10^{-1} with step size 0.002, covering model sizes from nearly full to empty.

The procedure is repeated for all calibration and test sets and results are summarized in Table 2. The weighted misclassification rate wmcr is calculated from the test data as in Eq. (5). The weights from the rSOS model are also used to calculate the wmcr for cSOS. Thereby observations which are detected as outliers by the rSOS model receive small influence on the wmcr of cSOS. Table 2 shows that the mcr of the cSOS is smaller than the mcr of the rSOS. On the other hand, the wmcr has a lower value for rSOS than for cSOS. The classical method tries to model the outliers and since outliers are present in the test data it achieves better results as well. The robust method, on the other hand, mainly models the non-outliers. So the weighted misclassification rate, which excludes the outliers, is lower for rSOS than for cSOS.

For visualization of the results we randomly select one of the five data splits and apply rSOS to the calibration data. Figure 4a shows the test observations projected into the subspace. The ellipses defined by the sets $\{z \in \mathbb{R}^2 | \mathrm{MD}\,(z; \boldsymbol{m}_k, \boldsymbol{S}) = \sqrt{\chi^2_2(0.975)}\}$, for $k = 1, 2, 3$ enclose those observations which are considered non-outliers and which did receive weight one in the wmcr. The observations outside of the ellipses are colored in gray. Most outliers are from group HA which is in line with previous analyses (Vanden Branden and Hubert 2005). In Fig. 4b the Mahalanobis distances of each test observation to its group center are shown. The horizontal line represents the cut-off value $\sqrt{\chi^2_2(0.975)}$. Again we see that many observations of HA have a large Mahalanobis distance in the projected space. Figure 4b pinpoints other anomalous observations in all three groups.

Olive oil data: The data set *olitos* (Armanino et al. 1989) available in the R package `rrcovHD` (Todorov 2016) contains $n = 120$ measurements on olive oil samples with $p = 25$ variables from fatty acids, sterols and triterpenic alcohols. The olive oils

Table 3 Olive oil data: average (w)mcr is the (weighted) mcr averaged over the five test data sets

	LDA	rLDA	cSOS	rSOS
Average mcr	0.178 (0.0645)	0.358 (0.1460)	0.175 (0.0308)	0.153 (0.0312)
Average wmcr	0.183 (0.0604)	0.353 (0.1201)	0.175 (0.0431)	0.133 (0.0282)

SE are reported in parentheses

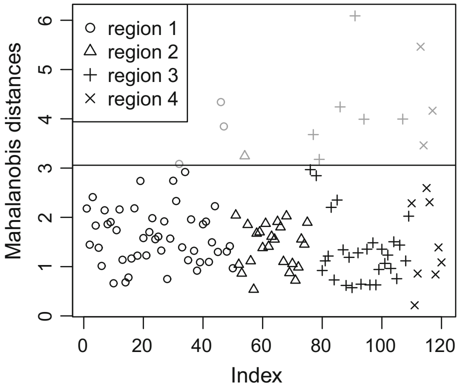

Fig. 5 Olive oil data: Mahalanobis distances from rSOS of each projected observation to its group center. Observations with weight smaller than one are colored in gray

originate from Tuscany in Italy and are grouped into $K = 4$ classes representing different regions of production with group size 50, 25, 34 and 11. In this example, the number of variables is quite low, but rSOS can still be an appropriate method. We will compare its results to cSOS as well as to LDA and rLDA. To estimate the models and to evaluate them, the same approach is taken as described previously for the fruit data. The optimal sparsity parameter is searched on a grid from 0.01 to 1 with step size 0.05, which covers various model sizes from the full model to the empty model.

Table 3 summarizes the quality of the resulting models. Our proposed method rSOS performs better on this data set than the other methods with an average wmcr of 13.3%. Interestingly, also the mcr is lowest for rSOS with 15.3%. This may happen if there is no pattern in the outlier configuration.

The classical sparse method cSOS outperforms LDA slightly, and robust LDA has a much lower prediction quality than all other methods. Figure 5 shows the Mahalanobis distances from the rSOS estimator of the projected test data. Especially region 3 and 4 have some observations with large distance to its group centers in the projected space. In Fig. 6 the projection of all observations into the subspace is visualized.

Audio data: This database consists of more than 4000 audio samples with 679 features. The audio samples come from 12 different groups such as recordings from classical and electronic music, cats and whale sounds, the noise of crowds, speech recordings from men and women and others (Brodinova et al. 2015).

We construct 10 experimental data sets to demonstrate the performance under varying data structure. For each data set we randomly choose $K = 4$ of the audio groups and randomly select 80, 70, 60 and 50 audio samples from the each of the groups, respectively. Further we add 23 observations from the remaining groups to the data set

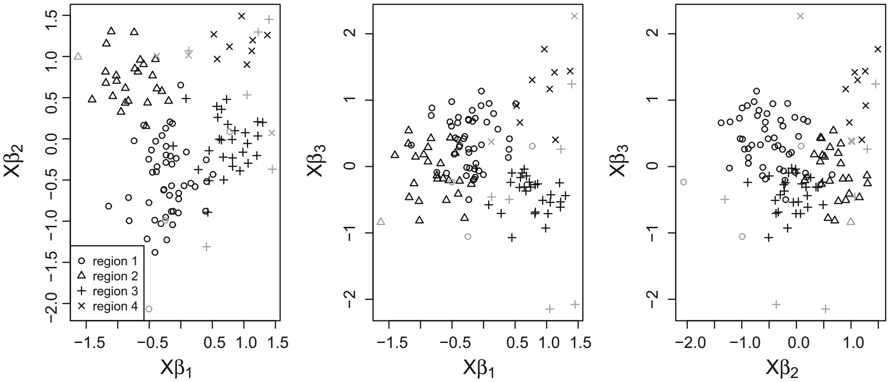

Fig. 6 Olive oil data: pairwise scatter plots of data projected into the 3-dimensional subspace derived from rSOS. Observations with weight smaller than one are colored in gray

Table 4 Audio data: average (w)mcr is the (weighted) mcr averaged over the 10 experimental settings

	cSOS	rSOS	Clean data
Average mcr	0.053 (0.0100)	0.057 (0.0128)	0.029 (0.0109)
Average wmcr	0.047 (0.0161)	0.020 (0.0089)	–

SE are reported in parentheses. Here *clean data* refers to results when the classical method is applied to the data set without artificial outliers

and assign them randomly to any of the four classes, thus generating outliers. Hence, each data set contains $n = 283$ observations, of which approximately 8% are artificial outliers, and $p = 679$ features.

The constructed data sets are randomly spit into calibration (80%) and test data (20%) such that the class proportions are preserved and the models are evaluated on the test data. Similar as in the simulation study the sparsity parameter is chosen from a grid of values between 0.1 and 2 with step size 0.05.

The average mcr and wmcr computed from the 10 different data sets are reported in Table 4. We also report the results when the artificial outliers are removed from the data sets and cSOS is applied. First, note that the average mcr is not much different between cSOS and rSOS. To reduce the influence of outliers on the evaluation criterion, the wmcr should be used. Here rSOS clearly outperforms cSOS. Note that the average mcr calculated with cSOS from the data set without artificial outliers is 0.029 and close to the average wmcr calculated with rSOS. This shows, that the wmcr gives an approximation of the mcr without outliers.

7 Conclusion

This paper introduced a robust and sparse optimal scoring method for multigroup classification. It yields a new supervised classification method, applicable if the number of variables is large with respect to the sample size *and* with possible presence of outliers in the data. Using an iterative algorithm, it searches for an optimal projection into a subspace using only a subset of the original variables; the most informative ones. Potential outliers are down-weighted, reducing their influence in the search for this optimal projection. The final classification is then carried out in this $(K-1)$-dimensional subspace. As shown in the examples in Sect. 6, the resulting low-dimensional representation of the data is also useful for visualization and interpretation.

The algorithm we developed, see Sect. 3, is implemented and publicly available in the R-package `rrcovHD` (Todorov 2016). This package contains outlier detection methods and robust statistical procedures for high dimensions. A call to the function `SosDiscrRobust`, with the data matrix and the class memberships as input, returns the estimated model.

Only few proposals exist so far for robust classification in high dimensions. Our proposal has the important feature of being sparse, simultaneously performing variable selection and model estimation, by using a (robust) Lasso-type approach. The simulation study has shown the importance to consider both sparse modeling *and* robust estimation. If either of them is missing, the prediction performance may decrease drastically.

Acknowledgements This work is supported by the Austrian Science Fund (FWF), Project P 26871-N20. We would like to thank the referees for useful comments.

Appendix

Derivation of expression (4) for the score vector estimates

Let $\omega_1, \ldots, \omega_n$ be case weights for each observation. $\boldsymbol{\Omega}$ is a diagonal matrix with these case weights in the diagonal. Then the weighted data matrices are $\tilde{\boldsymbol{Y}} = \boldsymbol{\Omega}^{1/2}\boldsymbol{Y}$ and $\tilde{\boldsymbol{X}} = \boldsymbol{\Omega}^{1/2}\boldsymbol{X}$. The diagonal matrix with weighted class proportions is $\tilde{\boldsymbol{D}} = \frac{1}{\sum \omega_i}\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}}$. The optimization problem (2) in step h for a given $\hat{\boldsymbol{\beta}}$ can be rewritten as

$$\min_{\boldsymbol{\theta}} \|\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{Y}}\boldsymbol{\theta}\|^2 \quad \text{s.t.} \quad \boldsymbol{\theta}^T\tilde{\boldsymbol{D}}\boldsymbol{\theta} = 1 \text{ and } \boldsymbol{C}\boldsymbol{\theta} = \boldsymbol{0} \in \mathbb{R}^h \tag{6}$$

with $\boldsymbol{C} = [\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_{h-1}]^T\tilde{\boldsymbol{D}}$, and we drop the depending on the index h for ease of notation.

We use the method of Lagrange multipliers. The Lagrangian associated to Eq. (6) is given by

$$L = (\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{Y}}\boldsymbol{\theta})^T(\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{Y}}\boldsymbol{\theta}) - \eta(\boldsymbol{\theta}^T\tilde{\boldsymbol{D}}\boldsymbol{\theta} - 1) - 2\boldsymbol{\gamma}^T\boldsymbol{C}\boldsymbol{\theta}.$$

The partial derivative set to zero gives

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = -2\tilde{\boldsymbol{Y}}^T(\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{Y}}\boldsymbol{\theta}) - 2\eta\tilde{\boldsymbol{D}}\boldsymbol{\theta} - 2\boldsymbol{C}^T\boldsymbol{\gamma} = \boldsymbol{0}.$$

Hence,

$$\boldsymbol{\theta} = (\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Y}} - \eta \tilde{\boldsymbol{D}})^{-1} (\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{X}} \hat{\boldsymbol{\beta}} + \boldsymbol{C}^T \boldsymbol{\gamma}).$$

To solve for the Lagrange multipliers η and $\boldsymbol{\gamma}$, the side constraints are used.

$$0 = \boldsymbol{C}\boldsymbol{\theta} = \boldsymbol{C}(\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Y}} - \eta \tilde{\boldsymbol{D}})^{-1} \tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{X}} \hat{\boldsymbol{\beta}} + \boldsymbol{C}(\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Y}} - \eta \tilde{\boldsymbol{D}})^{-1} \boldsymbol{C}^T \boldsymbol{\gamma}$$

So

$$\boldsymbol{\gamma} = -\left(\boldsymbol{C}(\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Y}} - \eta \tilde{\boldsymbol{D}})^{-1} \boldsymbol{C}^T\right)^{-1} \boldsymbol{C}(\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Y}} - \eta \tilde{\boldsymbol{D}})^{-1} \tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{X}} \hat{\boldsymbol{\beta}}.$$

We conclude

$$\begin{aligned}\boldsymbol{\theta} = {} & (\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Y}} - \eta \tilde{\boldsymbol{D}})^{-1} \\ & \left\{\boldsymbol{I} - \boldsymbol{C}^T (\boldsymbol{C}(\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Y}} - \eta \tilde{\boldsymbol{D}})^{-1} \boldsymbol{C}^T)^{-1} \boldsymbol{C}(\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Y}} - \eta \tilde{\boldsymbol{D}})^{-1}\right\} (\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{X}} \hat{\boldsymbol{\beta}}). \qquad (7)\end{aligned}$$

Since $\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Y}}$ is proportional to $\tilde{\boldsymbol{D}}$, there exists a scalar c such that

$$(\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Y}} - \eta \tilde{\boldsymbol{D}})^{-1} = c \tilde{\boldsymbol{D}}^{-1}.$$

Formula (7) can be simplified to

$$\boldsymbol{\theta} = c \left\{\boldsymbol{I} - \tilde{\boldsymbol{D}}^{-1} \boldsymbol{C}^T (\boldsymbol{C} \tilde{\boldsymbol{D}}^{-1} \boldsymbol{C}^T)^{-1} \boldsymbol{C}\right\} \tilde{\boldsymbol{D}}^{-1} \tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{X}} \hat{\boldsymbol{\beta}}.$$

Due to the symmetry of $\tilde{\boldsymbol{D}}$ and with the definition of $\boldsymbol{C} = \boldsymbol{Q}^T \tilde{\boldsymbol{D}}$ we obtain

$$\boldsymbol{\theta} = c \left\{\boldsymbol{I} - \boldsymbol{Q}(\boldsymbol{Q}^T \tilde{\boldsymbol{D}} \boldsymbol{Q})^{-1} \boldsymbol{Q}^T \tilde{\boldsymbol{D}}\right\} \tilde{\boldsymbol{D}}^{-1} \tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{X}} \hat{\boldsymbol{\beta}}.$$

The scalar c can then be scaled so that the side constraint $\boldsymbol{\theta}^T \tilde{\boldsymbol{D}} \boldsymbol{\theta} = 1$ is fulfilled.

Algorithm for the computation of the initial estimates for $\boldsymbol{\beta}_h$ and $\boldsymbol{\theta}_h$

Input: h, $\boldsymbol{Q}_h$, $\boldsymbol{X}$, $\boldsymbol{Y}$, λ

(i) Compute $\boldsymbol{D} = \frac{1}{n} \boldsymbol{Y}^T \boldsymbol{Y}$.

(ii) Generate $\boldsymbol{\theta}_*$, a random vector from $N(0, 1)$ of length K.

(iii) Compute $\hat{\boldsymbol{\theta}}_h = c \left\{\boldsymbol{I} - \boldsymbol{Q}_h (\boldsymbol{Q}_h^T \boldsymbol{D} \boldsymbol{Q}_h)^{-1} \boldsymbol{Q}_h^T \boldsymbol{D}\right\} \boldsymbol{\theta}_*$, with c so that $\hat{\boldsymbol{\theta}}_h^T \boldsymbol{D} \hat{\boldsymbol{\theta}}_h = 1$.

Apply twice the following steps:

1. For fixed $\hat{\boldsymbol{\theta}}_h$ apply sparse least trimmed squares (sparse LTS) regression (Alfons et al. 2013) to the response $\boldsymbol{Y}\hat{\boldsymbol{\theta}}_h$ and predictors $\boldsymbol{X}$.
   Let $a = 0.5n$ and $\|\boldsymbol{r}\|_{1:a}^2 = \sum_{i=1}^a r_{(i)}^2$ denote the sum of the a smallest squared elements of the vector $\boldsymbol{r}$. The sparse LTS estimator is a robust version of the Lasso and defined as
   $$\min_{\boldsymbol{\beta}} \frac{1}{a} \|\boldsymbol{Y}\boldsymbol{\theta}_h - \boldsymbol{X}\boldsymbol{\beta}\|_{(1):(a)}^2 + \lambda \|\boldsymbol{\beta}\|_1.$$

As in Alfons et al. (2013), a re-weighting step is carried out afterwards yielding $\hat{\boldsymbol{\beta}}_h$.

2. For fixed $\hat{\boldsymbol{\beta}}_h$ apply least absolute deviation (LAD) regression with response $\boldsymbol{X}\hat{\boldsymbol{\beta}}_h$ and predictor matrix $\boldsymbol{Y}$:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \|\boldsymbol{Y}\boldsymbol{\theta} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_h\|_1.$$

The LDA estimator is robust to outliers in the dependent variable, but not to leverage points (i.e. outliers in the covariate space). Since the covariates are dummy variables here, leverage points cannot occur. Then we apply the transformation for satisfying the side constraints:

$$\hat{\boldsymbol{\theta}}_h = c\left\{\boldsymbol{I} - \boldsymbol{Q}_h(\boldsymbol{Q}_h^T\boldsymbol{D}\boldsymbol{Q}_h)^{-1}\boldsymbol{Q}_h^T\boldsymbol{D}\right\}\boldsymbol{\theta}^*.$$

Output: Initial estimators $\hat{\boldsymbol{\beta}}_h$ and $\hat{\boldsymbol{\theta}}_h$.

References

Alfons A, Croux C, Gelper S et al (2013) Sparse least trimmed squares regression for analyzing high-dimensional large data sets. Ann Appl Stat 7(1):226–248

Armanino C, Leardi R, Lanteri S, Modi G (1989) Chemometric analysis of tuscan olive oils. Chemom Intell Lab Syst 5(4):343–354

Brodinova S, Ortner T, Filzmoser P, Zaharieva M, Breiteneder C (2015) Evaluation of robust PCA for supervised audio outlier detection. In: Proceeding of 22nd international conference on computational statistics (COMPSTAT)

Clemmensen L, Kuhn M (2012) sparseLDA: sparse discriminant analysis. R package version 0.1-6. https://CRAN.R-project.org/package=sparseLDA. Accessed 21 Oct 2015

Clemmensen L, Hastie T, Witten D, Ersbøll B (2012) Sparse discriminant analysis. Technometrics 53(4):406–413

Efron B, Hastie T, Johnstone I, Tibshirani R et al (2004) Least angle regression. Ann Stat 32(2):407–499

Hampel F (1974) The influence curve and its role in robust estimation. J Am Stat Assoc 69(346):383–393

Hampel F, Ronchetti E, Rousseeuw P, Stahel W (1986) Robust statistics: the approach based on influence functions. Wiley, Hoboken

Hastie T, Tibshirani R, Buja A (1994) Flexible discriminant analysis by optimal scoring. J Am Stat Assoc 89(428):1255–1270

Hastie T, Tibshirani R, Wainwright M (2015) Statistical learning with sparsity: the lasso and generalizations. CRC Press, Boca Raton

Hoffmann I, Filzmoser P, Serneels S, Varmuza K (2016) Sparse and robust PLS for binary classification. J Chemom 30(4):153–162

Hubert M, Van Driessen K (2004) Fast and robust discriminant analysis. Comput Stat Data Anal 45(2):301–320

Hubert M, Rousseeuw P, Van Aelst S (2008) High-breakdown robust multivariate methods. Stat Sci 23(1):92–119

Johnson R, Wichern D et al (2002) Applied multivariate statistical analysis, vol 5. Prentice Hall, Upper Saddle River

R Core Team: R (2016) A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

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

Tibshirani R (2011) Regression shrinkage and selection via the lasso: a retrospective. J R Stat Soc B 73(3):273–282

Todorov V (2016) rrcovHD: robust multivariate methods for high dimensional data. R package version 0.2-4. https://CRAN.R-project.org/package=rrcovHD. Accessed 17 Feb 2016

Todorov V, Pires A (2007) Comparative performance of several robust linear discriminant analysis methods. REVSTAT Stat J 5(1):63–83

Vanden Branden K, Hubert M (2005) Robust classification in high dimensions based on the simca method. Chemom Intell Lab Syst 79(1):10–21

Witten D, Tibshirani R (2011) Penalized classification using fisher's linear discriminant. J R Stat Soc Ser B (Statistical Methodology) 73(5):753–772

Wolke R, Schwetlick H (1988) Iteratively reweighted least squares: algorithms, convergence analysis, and numerical comparisons. SIAM J Sci Stat Comput 9(5):907–921

Wu T, Lange K (2008) Coordinate descent algorithms for lasso penalized regression. Ann Appl Stat 2(1):224–244

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.