On Coupling Robust Estimation with Regularization for High-Dimensional Data

Jan Kalina and Jaroslav Hlinka

Abstract Standard data mining procedures are sensitive to the presence of outlying measurements in the data. Therefore, robust data mining procedures are highly desirable, which are resistant to outliers. This work has the aim to propose new robust classification procedures for high-dimensional data and algorithms for their efficient computation. Particularly, we use the idea of implicit weights assigned to individual observation to propose several robust regularized versions of linear discriminant analysis (LDA), suitable for data with the number of variables exceeding the number of observations. The approach is based on a regularized version of the minimum weighted covariance determinant (MWCD) estimator and represents a unique attempt to combine regularization and high robustness, allowing to down-weight outlying observations. Classification performance of new methods is illustrated on real fMRI data acquired in neuroscience research.

1 Robustness and Regularization of Classification Methods

Classification methods (classifiers) have the aim to automatically assign new data to one of K groups ($K \geq 2$) based on decision rules constructed over a training data set. Sensitivity (non-robustness) of standard classifiers to the presence of outlying measurements (outliers) in the data has been repeatedly reported as a serious problem [3] and robust classification methods have been proposed as alternatives, which are resistant to the presence of outliers [8].

Linear discriminant analysis (LDA) as a standard (supervised) classification method assumes the data in each group to come from a Gaussian distribution, while the covariance matrix Σ is the same across groups. Its pooled estimator denoted by S is singular for high-dimensional data with $n < p$ or even $n \ll p$. For such data, which commonly appear in a variety of applications (e.g., in medicine, molecular

J. Kalina (✉) • J. Hlinka
Institute of Computer Science of the Czech Academy of Sciences, Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic

National Institute of Mental Health, Topolová 748, 250 67 Klecany, Czech Republic
e-mail: kalina@cs.cas.cz; hlinka@cs.cas.cz

F. Palumbo et al. (eds.), *Data Science*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-319-55723-6_2

genetics, chemometrics, or econometrics), regularized versions of LDA have been proposed to avoid the curse of dimensionality. They have become popular tools with a clear comprehensibility.

One common approach to regularized LDA is known as shrunken centroid regularized discriminant analysis (SCRDA) [4]. In this context, regularization brings benefits from both the computational and statistical point of view [13], which is true for $n < p$, as well as for $n > p$ with a relatively small n [5]. Its results may be superior to approaches based on a prior dimensionality reduction performed by selection of the most relevant variables.

However, regularized versions of LDA are sensitive to the presence of outlying values in the data. Unfortunately, most robust versions of LDA, which have been proposed within the framework of robust statistics, are computationally infeasible for $n < p$ [3, 8]. Xanthopoulos et al. [19] estimated high-dimensional covariance matrices allowing for measurement errors in the observed data. The resulting estimates are robust (insensitive) only to noise, but not robust to the presence of outliers. Robust procedures for high-dimensional data have been considered for regression models, including the proposal of a canonical correlation analysis [18] or partial least squares [7]. In the context of estimating a covariance matrix Σ of multivariate data, nonparametric correlation coefficients have been investigated by Croux and Öllerer [2] under the assumption that Σ^{-1} is sparse, which allows interesting applications in the area of graphical modeling. None of these approaches however exploits the idea of coupling the robustness with regularizing the estimated covariance matrix.

This paper exploits principles of robust statistics with the aim to propose new robust classification methods for high-dimensional data. We work with methods which are robust in terms of the breakdown point, which can be characterized as a global measure of robustness of an estimator against severe outliers in the data [9]. Methods with a high breakdown point are commonly denoted as highly robust. We presented a detailed overview of regularized versions of LDA in [11], however without considering robustness aspects. On the other hand, our previous work [10] on robust classification methods cannot be applied to high-dimensional data. Only the current paper exploits a unique coupling of regularization for $n \ll p$ and statistical robustness, which is based on implicit weighting, and thus ensures a high breakdown point.

In Sect. 2 of this paper, several new robust regularized methods for high-dimensional data are proposed based on down-weighting less reliable observations. The following Sects. 3 and 4 illustrate various methods on two real data sets and bring a detailed discussion of the results. Finally, Sect. 5 concludes the paper.

2 Classification Analysis Based on the Regularized Minimum Weighted Covariance Determinant Estimator

In this section, we propose several different robust versions of regularized LDA together with a discussion of their efficient computation. First, the robust regularized estimates of the covariance matrix and the means will be defined.

2.1 Estimation of the Covariance Matrix

In the whole paper, we assume n observations with p variables observed in K different groups

$$X_{11}, \ldots, X_{1n_1}, \ldots, X_{K1}, \ldots, X_{Kn_K}, \tag{1}$$

where $p > K \geq 2$ and $n = \sum_{k=1}^{K} n_k$.

Chen et al. [1] proposed regularized M-estimation of the population mean and covariance matrix of multivariate data based on a popular M-estimator of Tyler [16] and applied it to the task of mining wireless sensor data. While M-estimation represents a popular approach to robust estimation of parameters, it does not possess a high breakdown point in the multivariate model [9].

The minimum weighted covariance determinant (MWCD) estimator is one of highly robust estimators of the mean and at the same time of the covariance matrix Σ of multivariate data [14]. The estimate of the mean has the form of a weighted mean and the estimate of Σ has the form of a weighted covariance matrix. Prior to the computation, the user must specify magnitudes of weights, while the weights themselves are assigned to individual observations after an optimal permutation. Linearly decreasing weights in the form

$$w_i^* = 1 - \frac{i-1}{n}, \quad i = 1, \ldots, n, \tag{2}$$

if standardized to have the sum equal to 1, represents a simple and reasonable choice and will be considered also in the example of Sects. 3 and 4. The estimator remains to be reliable for data containing a large percentage of outliers [14].

While robust LDA based on the MWCD estimator was proposed in [10], the next sections propose classification methods based on the regularized MWCD estimator of the covariance matrix Σ in the form

$$\tilde{S}_{MWCD} = \lambda S_{MWCD} + (1-\lambda)T, \quad \lambda \in (0,1), \tag{3}$$

where a given target matrix T is symmetric positive definite of size $p \times p$. Such regularization ensures $\tilde{S}_{MWCD}$ to be regular and positive definite even for $n \ll p$. The

simplest choices for T are the identity matrix $T = \mathscr{I}_p$ or a diagonal (nonidentity) matrix

$$T = \bar{s}\,\mathscr{I}_p, \tag{4}$$

where $\bar{s} = \sum_{i=1}^{p} S_{ii}/p$. Within the classification procedures defined below, a suitable value of λ may be found by a cross validation in the form of a grid search over all possible values of $\lambda \in (0, 1)$.

2.2 Estimation of the Means

Based on general principles of regularization [4, 5], we propose to consider also the regularization of the means to improve the classification performance of the robust regularized LDA. While all of the classification methods, which will be newly proposed in this paper, consider the pooled covariance matrix to be estimated by the same regularized MWCD estimator $\tilde{S}_{MWCD}$, we will consider different ways for estimating the means of each of the K groups. The MWCD estimator of each of the mean will be denoted by $\bar{X}_{k,MWCD}$ for $k = 1, \ldots, K$.

The regularized MWCD-means will be now defined for the k-th group for $k = 1, \ldots, K$. They will be defined as shrinkage estimators in various norms including the L_2, L_1, and L_0 norm using a fixed value of the regularization parameter. We use the notation $\bar{X}^{MWCD}$ for the overall MWCD-mean across groups, $(x)_+$ for the positive part of $x \in \mathbb{R}^p$, and $\mathbb{1}(B)$ for the indicator function of a random event B.

Definition 1 (Robust Regularized Means)

1.

$$\bar{X}^{(2)}_{k,MWCD} = \delta^{(2)}\bar{X}_{k,MWCD} + (1 - \delta^{(2)})\bar{X}_{MWCD}, \quad \delta^{(2)} \in \mathbb{R} \tag{5}$$

2.

$$\begin{aligned}\bar{X}^{(1)}_{k,MWCD} &= \mathsf{sgn}(\bar{X}_{k,MWCD})\left(|\bar{X}_{k,MWCD}| - \delta^{(1)}\right)_+ \\ &= \mathsf{sgn}(\bar{X}_{k,MWCD})\max\left\{|\bar{X}_{k,MWCD}| - \delta^{(1)}, 0\right\}, \quad \delta^{(1)} \in \mathbb{R}\end{aligned} \tag{6}$$

3.

$$\bar{X}^{(0)}_{k,MWCD} = \bar{X}_{k,MWCD} \cdot \mathbb{1}\left[|\bar{X}_{k,MWCD} > \delta^{(0)}|\right], \quad \delta^{(0)} \in \mathbb{R} \tag{7}$$

All the estimators of Definition 1 can be interpreted as biased (Stein's shrinkage) versions of the MWCD-mean, while the biasedness allows to improve the mean square error [5]. The shrinkage within the estimator (6) is known as soft

thresholding, while (7) is known as hard thresholding, where the latter corresponds to the solution of L_0 regularization [6].

2.3 MWCD-RLDA

The first of the novel methods proposed in this paper, which is denoted as MWCD-RLDA, assigns an observation $Z = (Z_1, \ldots, Z_p)^T$ to group j if

$$(\bar{X}_{k,MWCD} - Z)^T (\tilde{S}_{MWCD})^{-1} (\bar{X}_{k,MWCD} - Z) - 2 \log \pi_k \tag{8}$$

over $k =, 1, \ldots, K$ is minimal exactly for j, where π_k denotes the prior probability of observing an observation from the k-th group for $k = 1, \ldots, K$. Equivalently, the classification rule can be also expressed by means of the robust and regularized linear discriminant score

$$\ell_k^* = (\bar{X}_{k,MWCD})^T (\tilde{S}_{MWCD})^{-1} Z - \frac{1}{2} (\bar{X}_{k,MWCD})^T (\tilde{S}_{MWCD})^{-1} \bar{X}_{k,MWCD} + \log \pi_k \tag{9}$$

and an observation Z is assigned to group j if $\ell_j^* > \ell_k^*$ for every $k \neq j$.

The situation with equal regularized linear discriminant scores $\ell_k^* = \ell_{k'}^*$ for $k' \neq k$ does not deserve a separate treatment, because it occurs with a zero probability for data coming from a continuous distribution. We can say that the method is based on a deformed (regularized) Mahalanobis distance between a new observation Z and the mean of each group. Because $\tilde{S}_{MWCD}$ depends on $\lambda \in (0, 1)$, its suitable value should be found by cross validation.

Because both (9) and (8) are rather obscure from the computational point of view, we propose to avoid computing the inverse matrix by solving a set of linear equations within the following algorithm based on eigendecomposition of the robust regularized covariance matrix.

Algorithm 1 avoids computing the inverse of $\tilde{S}_{MWCD}$. Instead, the group assignment in (8) is done in a more efficient way, which easily follows from

$$\begin{aligned} & (\bar{X}_{k,MWCD} - Z)^T (\tilde{S}_{MWCD})^{-1} (\bar{X}_{k,MWCD} - Z) - 2 \log \pi_k \\ = \; & (\bar{X}_{k,MWCD} - Z)^T \tilde{Q} \tilde{D}^{-1} \tilde{Q}^T (\bar{X}_{k,MWCD} - Z) - 2 \log \pi_k \\ = \; & \| \tilde{D}^{-1/2} \tilde{Q}^T (\bar{X}_k^{k,MWCD} - Z) \|^2 - 2 \log \pi_k . \end{aligned} \tag{13}$$

Possible improvements of Algorithm 1 in terms of computational stability include:

1. A possible tailor-made approach for the specific choice $T = \mathscr{I}_p$.
2. Replacing the eigendecomposition by the Cholesky decomposition of $\tilde{S}_{MWCD}$ in the form $\tilde{S}_{MWCD} = L L^T$, where L is a nonsingular lower triangular matrix. Then,

Algorithm 1 MWCD-RLDA for a general T based on eigendecomposition.

1. For a given $\delta \in (0,1)$, compute the matrix
$$A = [\bar{X}_{1,MWCD} - Z, \ldots, \bar{X}_{K,MWCD} - Z] \tag{10}$$
of size $p \times K$.
2. Compute $\tilde{S}_{MWCD}$ with a fixed $\lambda \in (0,1)$.
3. Compute and store the eigenvalues of $\tilde{S}_{MWCD}$ in the diagonal matrix $\tilde{D}$, and compute and store the corresponding eigenvectors of $\tilde{S}_{MWCD}$ in the orthogonal matrix $\tilde{Q}$.
4. Compute the matrix
$$B = \tilde{D}^{-1/2}\tilde{Q}^T A \tag{11}$$
and assign Z to group k, if
$$k = \underset{j=1,\ldots,K}{\operatorname{argmax}} \left\{ ||B_j||^2 - 2\log \pi_j \right\}, \tag{12}$$
where $||B_j||^2$ is the Euclidean norm of the j-th column of B.
5. Repeat steps 1 to 4 with different values of λ and find the classification rule with the best classification performance.

an efficient computation may exploit that

$$\begin{aligned} & (\bar{X}_{k,MWCD} - Z)^T (\tilde{S}_{MWCD})^{-1} (\bar{X}_{k,MWCD} - Z) - 2\log \pi_k \\ = & (\bar{X}_{k,MWCD} - Z)^T L^{-T} L^{-1} (\bar{X}_{k,MWCD} - Z) - 2\log \pi_k \\ = & \|L^{-1}(\bar{X}_k - Z)\|^2 - 2\log \pi_k. \end{aligned} \tag{14}$$

3. Using the truncated eigendecomposition instead of the (standard) eigendecomposition. Let us recall the latter in the form
$$\tilde{S}_{MWCD} = \sum_{i=1}^{r} d_i q_i q_i^T, \tag{15}$$
where r is rank of $\tilde{S}_{MWCD}$, $d_1, \ldots, d_r$ are nonzero eigenvalues, and $q_1, \ldots, q_r$ corresponding eigenvectors. The truncated eigendecomposition replaces (i.e., approximates) the expression (15) by
$$\tilde{S}_{MWCD} \approx \sum_{i=1}^{s} d_i q_i q_i^T = \tilde{Q}_* \tilde{D}_* \tilde{Q}_*^T, \tag{16}$$
where $\tilde{Q}_*$ has size only $p \times s$ and $\tilde{D}_*$ only $s \times s$ for a specified $s < r$.

2.4 Other Classification Methods

We propose several other classification methods which, in contrary to MWCD-RLDA, consider also regularizing the means of each of the groups of the data. They are denoted as MWCD-RLDA2, MWCD-RLDA1, or MWCD-RLDA0, which correspond to regularizing the means in the L_2, L_1, or L_0 norm, respectively. Within each classification method, suitable values of the regularization parameters λ and of (as the case may be) $\delta^{(l)}$ for $l \in \{0, 1, 2\}$ can be found by leave-one-out cross validation. We use the notation $\mathsf{diag}(A)$ to denote the diagonal matrix containing diagonal elements of A. The linear discriminant rules of the novel methods are defined as modifications of (9).

Definition 2 (MWCD-RLDA2)

$$\tilde{\ell}_k^{(2)} = (\bar{X}_{k,MWCD}^{(2)})^T (\tilde{S}_{MWCD})^{-1} Z - \frac{1}{2} (\bar{X}_{k,MWCD}^{(2)})^T (\tilde{S}_{MWCD})^{-1} \bar{X}_{k,MWCD}^{(2)} + \log \pi_k. \tag{17}$$

Definition 3 (MWCD-RLDA1)

$$\tilde{\ell}_k^{(1)} = (\bar{X}_{k,MWCD}^{(1)})^T (\tilde{S}_{MWCD})^{-1} Z - \frac{1}{2} (\bar{X}_{k,MWCD}^{(1)})^T (\tilde{S}_{MWCD})^{-1} \bar{X}_{k,MWCD}^{(1)} + \log \pi_k. \tag{18}$$

Definition 4 (MWCD-RLDA0)

$$\tilde{\ell}_k^{(0)} = (\bar{X}_{k,MWCD}^{(0)})^T (\tilde{S}_{MWCD})^{-1} Z - \frac{1}{2} (\bar{X}_{k,MWCD}^{(0)})^T (\tilde{S}_{MWCD})^{-1} \bar{X}_{k,MWCD}^{(0)} + \log \pi_k. \tag{19}$$

Definition 5 (MWCD-PAM)

$$\begin{aligned} \ell_k^{PAM} = {} & (\bar{X}_{k,MWCD}^{(1)})^T (\mathsf{diag}\{\tilde{S}_{MWCD}\})^{-1} Z - \\ & - \frac{1}{2} (\bar{X}_{k,MWCD}^{(1)})^T (\mathsf{diag}\{\tilde{S}_{MWCD}\})^{-1} \bar{X}_{k,MWCD}^{(1)} + \log \pi_k. \end{aligned} \tag{20}$$

An efficient computation of the new methods can be performed by an analogy of Algorithm 1. If the classification rule based on (20) is formulated by means of the Mahalanobis distances, the formula (13) reduces to a simple form

$$\sum_{i=1}^{p} \frac{(\bar{X}_{ki} - Z_i)^2}{\tilde{S}_{i,MWCD}^2}, \tag{21}$$

where $\bar{X}_k = (\bar{X}_{k1}, \ldots, \bar{X}_{kp})^T$ and $\tilde{S}_{i,MWCD}^2$ denotes the i-th diagonal element of $\tilde{S}_{MWCD}$. MWCD-PAM represents a robust counterpart of the Prediction Analysis

of Microarrays [15], where the latter is nothing else than a diagonalized LDA with means regularized in the L_1 norm.

Here, MWCD-RLDA1 can be interpreted as a robust counterpart of SCRDA [4]. Because MWCD-RLDA1 contains an intrinsic variable selection in (6), it is especially suitable if the data set contains a small set of dominant (very relevant) variables. On the other hand, MWCD-RLDA2 can be recommended if the data contain a large number of variables with a small effect on the classification, but without any clearly dominant small subset of variables.

3 Example: Brain Activity Data

A data set on the spontaneous activity of various parts of the brain will be now analyzed, which has been captured by means of fMRI neuroimaging. We have participated on a neuroscience research of the spontaneous brain activity in the resting state (i.e., resting-state brain networks). Our aim now is to illustrate the behavior of the newly proposed classification methods.

The brain activity of $n = 24$ probands is measured by means of fMRI under seven different situations. One of them can be characterized as a resting state, i.e., rest without any stimulus. Besides, the probands were watching each of six different movies while the brain activity was measured. The fMRI divides the brain into 90 regions and we are interested only in values of correlation coefficients between a pair of brain regions. In this context, the correlation coefficient evaluates a (functional) connectivity between the two regions. Thus, we consider $p = 90 * 89/2 = 4005$ variables containing values of correlation coefficients for each of the 24 probands.

The task is to learn a classification rule allowing to discriminate between two groups (resting state and movie) over 24 individuals, i.e., all movies together are considered to be one class. This is a classification to two groups with $p = 4005$ variables. The resting-state group contains 24 observations, but the group corresponding to any movie contains $6 * 24 = 144$ observations. In common applications, fMRI measurements are known to be usually contaminated by noise as well as outliers. It is also true with our data and therefore robust methods are highly desirable for their analysis.

We performed the computations in *R* software. Standard machine learning methods are used with default settings of their parameters. For various regularized versions of LDA, we choose the target matrix T as either $T = \mathcal{I}_p$ or as (4). The results of leave-one-out cross classification are overviewed in Table 1. Performance of classifiers is measured by means of their accuracy, i.e., number of correctly classified cases divided by the total number of cases.

SCRDA as one of available regularized LDA versions turns out to perform reliably, while its classification rule is based only on 81 variables. Also the newly proposed robust LDA versions yield a very good performance. We do not find major differences in the classification performance of robust and non-robust various

Table 1 Results of the examples of Sects. 3 and 4

	Classification accuracy	
Classification method	Brain data	AIM data
SCRDA	1.00	0.86
MWCD-RLDA	1.00	0.86
MWCD-RLDA1	1.00	0.86
MWCD-PAM	0.98	0.77
SVM (Gaussian kernel)	1.00	0.85
Multilayer perceptron	Infeasible	Infeasible
Number of principal components	10	20
PCA $\Longrightarrow$ LDA	1.00	0.83
PCA $\Longrightarrow$ SCRDA	1.00	0.83
PCA $\Longrightarrow$ MWCD-RLDA with $T = \mathscr{I}_p$	1.00	0.84
PCA $\Longrightarrow$ MWCD-RLDA with (4)	1.00	0.84
PCA $\Longrightarrow$ MWCD-RLDA2	1.00	0.84
PCA $\Longrightarrow$ MWCD-RLDA1	1.00	0.84
PCA $\Longrightarrow$ MWCD-RLDA0	1.00	0.84
PCA $\Longrightarrow$ MWCD-PAM	0.96	0.75

Various classification methods are compared, while their classification accuracy is evaluated in a leave-one-out cross validation study

regularized versions of LDA. This can be explained by the fact that the data do not contain a remarkable percentage of outliers. Also the SVM method gives a perfect classification rule, while a multilayer perceptron with one hidden layer is computationally infeasible due to $n \ll p$ in the implementation in R software.

Additionally, we investigated the effect of dimensionality reduction by means of principal component analysis (PCA) on the classification performance. There seems no remarkable small group of genes responsible for a large portion of variability of the data and the first few principal components seem rather arbitrary. All the novel robust methods have a good classification ability if applied on principal components. Thus, the classification results after reducing the dimensionality bring other arguments in favor of the regularization approaches used in this paper.

In order to investigate the performance of various classification methods on data contaminated by noise, we generated proband-independent noise generated from normal distribution $\mathsf{N}(0, \sigma^2)$ for various values of σ. The noise was added to all measurements for each proband and classification rules are learned over this contaminated data set. Such contamination was repeated 100 times and the classification performance of various methods was evaluated for each case. We give the averaged values of the classification accuracy computed over the 100 cases in Table 2 only for selected classifiers, because their computation is rather demanding.

The results of the classification performance of various methods on data artificially contaminated by noise show an evidence of robustness of SCRDA. The larger the value of σ, the more influential outliers are present in the contaminated data set. Indeed, the reduction of the classification performance of the standard data mining

Table 2 Results of the brain activity analysis on data artificially contaminated by normally distributed outliers $\mathsf{N}(0, \sigma^2)$ for different values of σ

Classification method	Classification accuracy		
	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.3$
SVM (Gaussian kernel)	1.00	0.99	0.98
Classification tree	0.99	0.98	0.98
SCRDA	1.00	1.00	1.00
MWCD-RLDA1 with $T = \mathscr{I}_p$	1.00	1.00	1.00
PCA $\Longrightarrow$ LDA	0.99	0.99	0.99

The classification accuracy is evaluated by a leave-one-out cross validation. PCA is used with a fixed number of ten principal components

methods is not caused by the noise itself, but rather by severe outliers. SCRDA and the novel robust versions of LDA turn out to yield reliable results. The robustness of SCRDA to noise has not however been systematically investigated although it has been recommended as a promising alternative to the SVM [4].

Further, the classification rule distinguishing between the resting state and a particular movie is constructed, which is again a classification to 2 groups with $p = 4005$ variables. This time, each of the groups contains 24 observations. We computed SVM, SCRDA, and PCA $\Longrightarrow$ LDA for 6 different tasks, namely classification between the resting state and movie 1; between the resting state and movie 2, etc. In a leave-one-out cross validation study, every method yields a 100% classification accuracy in all the seven classification tasks. For the sake of comprehensibility, it is important that MWCD-RLDA1 turns out to be based only on a small number of variables, namely 1, 1, 2, 3, 3, and 7 variables. These are the most relevant sets of variables for the particular classification task, while SVM and PCA exploit observed values from the whole set of p variables. If PCA is performed keeping ten principal components, each of the considered classifiers keeps the 100% classification accuracy for each of the six classification tasks.

Additionally, classification between pairs of movies (e.g., classification between movie 1 and movie 2) yields results with 100% classification accuracy, while the number of variables contributing to the classification rule of MWCD-RLDA1 is between 2 and 30 for each of the tasks.

4 Example: Cardiovascular Genetic Study

Further, we illustrate the performance of the novel robust classifiers on the data from a cardiovascular genetic study performed in the Center of Biomedical Informatics in Prague in the years 2006–2011. From the point of view of analyzing the data, the very aim originally was to reduce the dimensionality [12], i.e., to find a small set of genes responsible for the development of the acute myocardial infarction (AMI). Gene expressions of $p = 38{,}590$ were measured across the whole genome. These

correspond to the activity of individual genes leading to synthesis of proteins and consequent biological processes. The BeadChip microarray technology was used to acquire the data over $n = 95$ individuals, including 46 AMI patients and 49 controls.

Robust versions of regularized LDA perform well. We find it a success that the method is computationally feasible for such p at all due to the high dimensionality, which severely complicates a potential identification of outliers. While the SVM classifier formally gives a perfect classification result, it suffers from a heavy overfitting, not only because its optimization of parameters tends to a very local optimum for $n < p$, but mainly because the SVM contains too many support vectors and does not capture the multivariate structure of the data. Although it is designed as a black box, we can say that it classifies each new observation with a too strong emphasis on its nearest neighbors.

If the classification rule is learned only over the set of the 20 principal components, robust versions of regularized LDA are able to slightly outperform available (non-robust) classifiers. There seems however no difference among the individual robust methods of Sect. 2, because only negligible values of the regularization parameter for the means are selected for each of the methods and the effect of this regularization is negligible itself.

5 Conclusions

Some of the standard methods of data mining or multivariate statistics are computationally infeasible for high-dimensional data, while others suffer from a numerical instability and lack of robustness to noise or outlying measurements. Therefore, this paper proposes new robust classification methods for high-dimensional observations, i.e., assuming the number of variables to exceed the number of observations. We combine robustness to the presence of outliers with regularized estimation of the population means and covariance matrix of the multivariate data in a unique way.

We propose several robust classifiers, which are based on implicit weighting of individual observations in Sect. 2. The methods are based on a regularized version of a robust covariance matrix, while also the mean of each group is computed by means of a robust regularized estimator. At the same time, implicit weights ensure a high breakdown point with respect to a larger percentage of outliers in a variety of other situations [17]. All the methods require intensive computations. Efficient algorithms allow the methods to be computed even for $n \ll p$.

The robust regularized versions of LDA can be interpreted as modifications of robust LDA corrected for small sample sizes. At the same time, we point out the connection to the shrinkage statistical approach, following Stein's result of estimating the mean of multivariate normal data [5].

We consider all of the newly proposed robust methods to be comprehensible. Particularly, let us discuss the classification rule of MWCD-RLDA1. It assigns an observation Z based on a deformed Mahalanobis distance between Z itself and a (robust) centroid of each of the K groups. Such variables contribute the most to the

classification rule, which are most relevant for the separation among groups. Also the implicit weights assigned to individual observations allow a clear interpretation. They deform the Mahalanobis distance, while less reliable observations (potential outliers) obtain small or negligible weights.

In addition, we analyzed two high-dimensional data sets. The fMRI data come from a brain research study, which has the aim to investigate connections among brain parts during a resting state. Results of various classification methods show distinct differences between the resting and non-resting state. At the same time, different movies shown to the set of 24 probands turn out to activate different connections between pairs of brain parts. Future neuroscience research is intended to search for a small set of variables allowing to distinguish schizophrenic patients from control individuals based only on the fMRI measurements of the brain in the resting state. The cardiovascular genetic data set with a dimensionality even larger ($p = 38{,}590$) compared to the fMRI data shows a slight advantage of the newly proposed methods compared to available classifiers. The analysis of this data set allows to detect a predisposition for infarction based only on gene expressions.

Concerning the limitations of our analysis, both SCRDA and its robust counterparts are reliable under an implicit assumption that the variability is not substantially different across variables. Still, the methods seem to yield reliable results although this assumption is violated in the data.

To summarize practical recommendations based on the example, the new robust methods seem to perform reliably for high-dimensional data with a small number of observations. The level of noise in the original data seems to be moderate and the advantage of robust methods compared to non-robust ones is not revealed primarily after adding an artificial contamination to the data. SCRDA itself turns out to be reasonably robust, which can be explained as an effect of the regularization reducing the influence of noise in the data. The main result of the examples is however the reliability of the newly proposed methods for both original and contaminated data sets.

Acknowledgements The work is supported by the project "National Institute of Mental Health (NIMH-CZ)", grant number CZ.1.05/2.1.00/03.0078 of the European Regional Development Fund, Neuron Fund for Support of Science, and the Czech Science Foundation project No. 13-23940S.

References

1. Chen, Y., Wiesel, A., Hero, A.O.: Robust shrinkage estimation of high dimensional covariance matrices. IEEE Trans. Signal Process. **59**, 4097–4107 (2011)
2. Croux, C., Öllerer, V.: Robust and sparse estimation of the inverse covariance matrix using rank correlation measures. Technical Report, KU Leuven (2015)
3. Filzmoser, P., Todorov, V.: Review of robust multivariate statistical methods in high dimension. Anal. Chim. Acta **705**, 2–14 (2011)

4. Guo, Y., Hastie, T., Tibshirani, R.: Regularized discriminant analysis and its application in microarrays. Biostatistics **8**, 86–100 (2007)
5. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2nd edn. Springer, New York (2009)
6. Herlands, W., De-Arteaga, M., Neill, D., Dubrawski, A.: Lass-0: sparse non-convex regression by local search (2016, submitted)
7. Hoffmann, I., Serneels, S., Filzmoser, P., Croux, C.: Sparse partial robust M regression. Chemom. Intel. Lab. Syst. **149**, 50–59 (2015)
8. Hubert, M., Rousseeuw, P.J., Van Aelst, S.: High-breakdown robust multivariate methods. Stat. Sci. **23**, 92–119 (2008)
9. Jurečková, J., Sen, P.K., Picek, J.: Methodology in Robust and Nonparametric Statistics. CRC Press, Boca Raton (2012)
10. Kalina, J.: Highly robust statistical methods in medical image analysis. Biocybern. Biomed. Eng. **32**(2), 3–16 (2012)
11. Kalina, J.: Classification analysis methods for high-dimensional genetic data. Biocybern. Biomed. Eng. **34**, 10–18 (2014)
12. Kalina, J., Zvárová J.: Decision support systems in the process of improving patient safety. In: Bioinformatics: Concepts, Methodologies, Tools, and Applications, pp. 1113–1125. IGI Global, Hershey (2013)
13. Pourahmadi, M.: High-Dimensional Covariance Estimation. Wiley, Hoboken (2013)
14. Roelant, E., Van Aelst, S., Willems, G.: The minimum weighted covariance determinant estimator. Metrika **70**, 177–204 (2009)
15. Tibshirani, R., Hastie, T., Narasimhan, B.: Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Stat. Sci. **18**, 104–117 (2003)
16. Tyler, D.E.: A distribution-free M-estimator of multivariate scatter. Ann. Stat. **15**, 234–251 (1987)
17. Víšek, J.Á.: Consistency of the least weighted squares under heteroscedasticity. Kybernetika **47**, 179–206 (2011)
18. Wilms, I., Croux, C.: Robust sparse canonical correlation analysis. BMC Systems Biology **10**, 72 (2016)
19. Xanthopoulos, P., Pardalos, P.M., Trafalis, T.B.: Robust Data Mining. Springer, New York (2013)