Highly Robust Classification: A Regularized Approach for Omics Data

Jan Kalina[1,2] and Jaroslav Hlinka[1,2]
[1]*Institute of Computer Science CAS, Prague, Czech Republic*
[2]*National Institute of Mental Health, Klecany, Czech Republic*

Keywords: High-dimensional Data, Classification Analysis, Robustness, Outliers, Regularization.

Abstract: Various regularized approaches to linear discriminant analysis suffer from sensitivity to the presence of outlying measurements in the data. This work has the aim to propose new versions of regularized linear discriminant analysis suitable for high-dimensional data contaminated by outliers. We use principles of robust statistics to propose classification methods suitable for data with the number of variables exceeding the number of observations. Particularly, we propose two robust regularized versions of linear discriminant analysis, which have a high breakdown point. For this purpose, we propose a regularized version of the minimum weighted covariance determinant estimator, which is one of highly robust estimators of multivariate location and scatter. It assigns implicit weights to individual observations and represents a unique attempt to combine regularization and high robustness. Algorithms for the efficient computation of the new classification methods are proposed and the performance of these methods is illustrated on real data sets.

1 LINEAR DISCRIMINANT ANALYSIS AND ITS MODIFICATIONS

Classification methods (classifiers) in bioinformatics commonly have the aim to learn a classification rule over data with the number of variables p exceeding the number of observations n. Let us consider the total number n of p-dimensional observations

$$X_{11},\ldots,X_{1n_1},\ldots,X_{K1},\ldots,X_{Kn_K}, \qquad (1)$$

which are observed in K ($K \geq 2$) different samples (groups) with $p > K \geq 2$, where $n = \sum_{k=1}^{K} n_k$. Sensitivity of various standard classification procedures to the presence of outlying measurements (outliers) in such high-dimensional data has been repeatedly reported as a serious problem in data mining as well as multivariate statistics (Christmann and van Messem, 2008).

Linear discriminant analysis (LDA) as a standard (supervised) classification method assumes normally distributed data in each group, while the covariance matrix Σ is the same across groups. Let us denote the mean of the observed values in the k-th group ($k = 1,\ldots,K$) by $\bar{X}_k$. If $n < p$ or even $n \ll p$, the pooled estimator of the covariance matrix denoted by S is singular. One of the habitually used versions of regularized LDA, which we denote by LDA* to avoid confusion, assigns a new observation $Z = (Z_1,\ldots,Z_p)^T$ to group k, if $l_k^* > l_j^*$ for every $j \neq k$, where the regularized linear discriminant score for the k-th group ($k = 1,\ldots,K$) has the form

¿Cambia Z y SI falta ¿x barra? creo que con esa restricción de menor a... se van outliers

$$l_k^* = \bar{X}_k^T (S^*)^{-1} Z - \frac{1}{2} \bar{X}_k^T (S^*)^{-1} \bar{X}_k + \log \pi_k. \qquad (2)$$

Here, π_k is a prior probability of observing an observation from the k-th group,

Mismo estimador de la matriz de covarianza regularizado :)

$$S^* = \lambda S + (1-\lambda) T \qquad (3)$$

for $\lambda \in (0,1)$ denotes a regularized estimator of the covariance matrix across groups and T is a given symmetric positive definite matrix of size $p \times p$.

In addition, the standard LDA is too sensitive to the presence of outlying values in the data because of its construction using the maximum likelihood estimates of the means and covariance matrix (Filzmoser and Todorov, 2011). As an alternative, robust classification methods have been proposed which are resistant to the presence of outliers (Croux and Dehon, 2001; Hubert et al., 2008; Todorov and Filzmoser, 2009) in terms of a high breakdown point, which measures the sensitivity against noise or outliers in the data. Particularly, the finite-sample definition of the breakdown point corresponds to the maximal percentage of extremely severe outliers present in the data set, which still does not lead the method to a collapse, i.e. the estimators of the means and covariance matrix

Kalina, J. and Hlinka, J.
Highly Robust Classification: A Regularized Approach for Omics Data.
DOI: 10.5220/0005623500170026
In *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016) - Volume 3: BIOINFORMATICS*, pages 17-26
ISBN: 978-989-758-170-0

are not shifted to infinity (Davies and Gather, 2005; Huber and Ronchetti, 2009). Nevertheless, available robust classification procedures require $n > p$.

Numerous available versions of the regularized LDA, as described e.g. by (Guo et al., 2007; Tibshirani and Narasimhan, 2003; Krzanowski et al., 1995; Kindermans et al., 2014) are also vulnerable to outliers because of the non-robustness of S as well as means of each group. We will document this on examples in Section 3. While LDA* is considered to be robust from the point of view of the so-called robust data mining (Xanthopoulos et al., 2013), such robustness is defined only as insensitivity to measurement errors but not to the presence of severe outliers. In this paper, we perform a unique approach to combine the Tikhonov regularization for $n \ll p$ with statistical robustness.

Our aim is to propose new classification methods for high-dimensional data exploiting principles of robust statistics. Section 2 proposes two new robust regularized methods for high-dimensional data, exploiting the idea of down-weighting less reliable observations. They are based on a regularized version of the minimum weighted covariance determinant estimator, which possesses a high breakdown point. The following Section 3 illustrates various methods on several real data sets. The results are presented in Section 4 and discussed in Section 5. Finally, Section 6 concludes the paper and discusses also the good comprehensibility of the newly proposed approach.

2 CLASSIFICATION ANALYSIS BASED ON THE REGULARIZED MWCD ESTIMATOR

We recall the regularized M-estimator of covariance matrix (Chen et al., 2011) in Section 2.1. Further, we propose a regularized version of the highly robust minimum weighted covariance determinant estimator in Section 2.2. For the iterative computation of this covariance matrix estimator, we recommend to use the Chen's estimator as an initial estimator of the covariance matrix. Finally, we propose two robust versions of regularized LDA in Sections 2.3 and 2.4.

2.1 Regularized M-estimator of Covariance Matrix

A regularized M-estimator of the population covariance matrix of multivariate data was proposed by (Chen et al., 2011). Assuming a single group of independent identically distributed (i.i.d.) data ($K = 1$), Chen's estimator can be characterized as a regularized M-estimator of (Tyler, 1987), in other words a regularized Huber-type estimator for multivariate data. As it depends on a regularization parameter $\rho \in (0,1)$, we will denote it as $S^*_{M,\rho}$.

Although M-estimation is the most common robust statistical approach (Huber and Ronchetti, 2009), M-estimators of parameters in the multivariate model do not possess a high breakdown point (Tyler, 2014). This is true also for Chen's estimator. Therefore, we consider more robust methods in Section 2.2, using Chen's estimator as an initial estimator of the covariance matrix.

2.2 Regularized MWCD Estimator

We recall the minimum weighted covariance determinant (MWCD) estimator, which is one of highly robust estimators of parameters of multivariate data (Roelant et al., 2009). Its appealing properties deserve to be overviewed and we propose its regularized version suitable for $n < p$.

Considering a single group of i.i.d. data ($K = 1$), the MWCD estimates the mean in the form of a weighted mean and at the same time estimates the covariance matrix Σ. Prior to the computation, the user must specify magnitudes of weights, while the weights themselves are assigned to individual observations after an optimal permutation. The idea is to assign small weights to outliers and larger weights to reliable data points.

The MWCD estimator can be interpreted as a direct generalization of the minimum covariance determinant (MCD) estimator of (Rousseeuw and van Driessen, 1999), allowing more general weight functions. In addition, the hard rejection rule of the MCD may increase its local sensitivity, while the MWCD does not suffer from such local non-robustness due to the weighting scheme; this is analogous to the least weighted squares regression reducing the local sensitivity of the least trimmed squares (Víšek, 2006).

The MWCD estimator is highly robust in terms of the breakdown point (Roelant et al., 2009) and is equal to the maximal breakdown point attainable for affine-equivariant estimators of Σ (Lopuhaä and Rousseeuw, 1991); this is true if the outliers obtain weights exactly equal to 0. The MWCD estimator has the largest efficiency for elliptically symmetric unimodal distributions. Also the Fisher consistency and influence function are known (Roelant et al., 2009). An approximative algorithm for computing the MWCD estimator may be obtained as a gener-

alization of the MCD algorithm (Rousseeuw and van Driessen, 1999).

Because the MWCD estimator cannot be computed for $n < p$, we define its regularized version computationally feasible for $n \ll p$. Let T denote a given symmetric positive definite matrix of size $p \times p$.

Algorithm 1: *Regularized MWCD estimator.*

Step 1. Initialize the value of the loss function as $+\infty$.

Step 2. Compute $S^*_{M,\rho}$ for a given $\rho \in (0,1)$. Compute Huber's M-estimator of μ and denote it as $\bar{X}_M$. Denote $B = \bar{X}_M$ and $C = S^*_{M,\rho}$. regularización

Step 3. Compute the regularized M-Mahalanobis distance

$$d(i;B,C) = \left[(X_i - B)^T C^{-1} (X_i - B)\right]^{1/2} \quad (4)$$

for each observation X_i. Sort these distances in ascending order. This determines a permutation $\pi(1),\ldots,\pi_n$ of the indexes $1,2,...,n$, which fulfills

$$d(\pi(1);B,C) \leq \cdots \leq d(\pi(n);B,C). \quad (5)$$

Sumatoria de las cov de todos los grupos con medida de robustes

Assign the weights to invididual observations according to the ranks of the Mahalanobis distances. Thus, e.g. the observation $X_{\pi(1)}$ obtains the weight w_1.

Step 4. Evaluate

$$S_w = \sum_{i=1}^{n} w_i (X_i - \bar{X}_w)(X_i - \bar{X}_w)^T \quad (6)$$

with the weights from Step 3. If the loss function evaluated as the determinant of the matrix

$$\det\left(\lambda S_w + (1-\lambda)T\right). \quad (7)$$

is smaller than the previously obtained value, continue with step 5. Otherwise go to step 6. cuál?

Step 5. Store the values of the weights. Compute the weighted mean and weighted covariance matrix using these weights. Continue with steps 2, 3, and 4. This is repeated as long as the value of the loss decreases.

Step 6. Repeatedly (10 000 times) perform the steps 1 to 5. The optimal weights are those which yield the minimal value of the loss function over all repetitions of steps 1 to 5.

In step 2, choosing robust rather than standard initial estimators is a common approach in a variety of iterative robust estimators (Huber and Ronchetti, 2009).

2.3 MWCD-LDA*

We propose a novel classification method denoted as MWCD-LDA*. The data are assumed to be observed in K different groups as in (1), while each of the groups has a Gaussian distribution with covariance matrix Σ. The method is based on estimating Σ by the regularized MWCD estimator. However, Σ does not play the role of the covariance matrix over all data and we will need to adapt Algorithm 1 for the situation with K groups.

In order to simplify the notation, let us denote the p-dimensional measurements (1) as $Y_1,\ldots,Y_n$. We will distinguish between Huber's M-estimator computed for all observations

$$\bar{Y}_M = (\bar{Y}_{M,1},\ldots,\bar{Y}_{M,p})^T \quad (8)$$

and Huber's estimator for the k-th group

$$\bar{Y}^k_M = (\bar{Y}^k_{M,1},\ldots,\bar{Y}^k_{M,p})^T. \quad (9)$$

The formula (6) will be replaced by

$$\tilde{S}_{MWCD} = \left(\tilde{S}_{ij}\right)_{i,j=1}^{p}, \quad (10)$$

Estimador de Huber

where

$$\tilde{S}_{ij} = \sum_{k=1}^{K} \sum_{l \in group\ k} w_l (Y_{li} - \bar{Y}^k_{M,i})(Y_{lj} - \bar{Y}^k_{M,j}). \quad (11)$$

Here, the summation over l runs over all observations $l = 1,\ldots,n$, which belong to the k-th group. The result of Algorithm 1 with such modification for the K groups are the optimal weights denoted as $\tilde{w}_1,\ldots,\tilde{w}_n$. The regularized MWCD estimator will be computed as the matrix (11) with weights equal to $\tilde{w}_1,\ldots,\tilde{w}_n$ and will be denoted as S^*_{MWCD}.

Se hizo el algoritmo 2 para cada w_i.

At the same time, the means of each of the K groups will be estimated by the MWCD estimator and we will distinguish between

$$\bar{Y}_{MWCD} = \sum_{l=1}^{n} \tilde{w}_l Y_l \quad (12)$$

and

$$\bar{Y}^k_{MWCD} = \sum_{l \in group\ k} \tilde{w}_l Y_l, \quad k = 1,\ldots,K. \quad (13)$$

Within a classification procedure, a suitable value of λ will be found by a cross-validation in the form of a grid search over all possible values of $\lambda \in (0,1)$. Formally, MWCD-LDA* will assign a new observation $Z = (Z_1,\ldots,Z_p)^T$ to group k, if $\ell^*_k > \ell^*_j$ for every $j \neq k$, where

$$\ell^*_k = (\bar{Y}_{k,MWCD})^T (S^*_{MWCD})^{-1} Z - \quad (14)$$
$$-\frac{1}{2}(\bar{Y}_{k,MWCD})^T (S^*_{MWCD})^{-1} \bar{Y}_{k,MWCD} + \log \pi_k.$$

Equivalently, the classification rule can be also expressed as follows. An observation Z is assigned to group k if

$$(\bar{Y}_{j,MWCD} - Z)^T (S^*_{MWCD})^{-1} (\bar{Y}_{j,MWCD} - Z) + \log \pi_k \quad (15)$$

over $j = ,1,\ldots,K$ is minimal exactly for k.

Nevertheless, both (14) and (15) 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 regularized covariance matrix.

Algorithm 2: *MWCD-LDA* for a general T based on eigendecomposition.*

1. For a fixed value of $\lambda \in (0,1)$, compute S^*_{MWCD} using a given T using Algorithm 1, replacing (6) by (10) with (11).
2. Denote the weights determined by the computation of S^*_{MWCD} by $\tilde{w}_1,\ldots,\tilde{w}_n$ and use them to compute $\bar{Y}^k_{MWCD}$ using (13).
3. For a given $\delta \in (0,1)$, compute the matrix
$$A = [\bar{Y}_{1,MWCD} - Z, \ldots, \bar{Y}_{K,MWCD} - Z] \quad (16)$$
of size $p \times K$.
4. Compute and store the eigenvalues of S^*_{MWCD} in the diagonal matrix $\tilde{D}$, and compute and store the corresponding eigenvectors of S^*_{MWCD} in the orthogonal matrix Q.
5. Compute the matrix
$$B = D^{-1/2} Q^T A \quad (17)$$
and assign Z to group k, if
$$k = \underset{j=1,\ldots,K}{\arg\max} \left\{ ||B_j||^2 + \log \pi_k \right\}, \quad (18)$$
where $||B_j||^2$ is the Euclidean norm of the j-th column of B.
6. Repeat steps 1 to 4 with different values of λ and find the classification rule with the best classification performance.

The expensive and numerically unstable computation of the Mahalanobis distance is avoided by replacing the inversion of S^*_{MWCD} by an efficient group assignment in (15) based on

$$\begin{aligned} & (\bar{Y}_{k,MWCD} - Z)^T (S^*_{MWCD})^{-1} (\bar{Y}_{k,MWCD} - Z) \\ = \; & (\bar{Y}_{k,MWCD} - Z)^T Q D^{-1} Q^T (\bar{Y}_{k,MWCD} - Z) \\ = \; & \|D^{-1/2} Q^T (\bar{Y}^{k,MWCD}_k - Z)\|^2. \end{aligned} \quad (19)$$

Alternatively, the method can be computed using Cholesky decomposition. Besides, if a specific choice $T = I_p$ is considered, the computation of MWCD-LDA* can be performed by means of more efficient algorithms, which exceed the scope of this paper.

2.4 MWCD-LDA**

The second novel regularized robust version of LDA denoted as MWCD-LDA** combines robust covariance matrix estimation of Section 2.3 with shrinking the means towards the pooled mean (across groups).

The regularized estimator of the joint covariance matrix (across groups) is obtained as in Algorithm 2. Further, let us use the notation $\bar{X}^{MWCD}$ for the overall MWCD-mean across groups. The MWCD-means of individual groups is be shrunken towards $\bar{X}^{MWCD}$ replacing the classical mean of the k-th group by

$$\bar{Y}^{**}_{k,MWCD} = \delta \bar{Y}_{k,MWCD} + (1-\delta) \bar{Y}_{MWCD} \quad (20)$$

for $k = 1,\ldots,K$ and a fixed $\delta \in (0,1)$.

The new method MWCD-LDA** is defined as follows. An observation $Z = (Z_1,\ldots,Z_p)^T$ will be assigned to group k, if $\ell^{**}_k > \ell^{**}_j$ for every $j \neq k$, where

$$\begin{aligned} \ell^{**}_k \;=\; & (\bar{Y}^{**}_{k,MWCD})^T (S^*_{MWCD})^{-1} Z - \\ - \; & \frac{1}{2} (\bar{Y}^{**}_{k,MWCD})^T (S^*_{MWCD})^{-1} \bar{Y}^{**}_{k,MWCD} + \log \pi_k. \end{aligned} \quad (21)$$

The shrinkage in (20) can be interpreted as shrinkage in the L_2-norm as in ridge regression. Although this does not perform variable selection, which is obtained by the L_1-regularization, it is more suitable for such data that do not contain a small number of variables dominant for the classification task. Such shrinking the means towards the pooled mean is known to bring benefits e.g. in Bayesian hierarchical models (Mallick et al., 2009). Thus, MWCD-LDA** can be interpreted as a method based on an L_2-regularized Mahalanobis distance.

Suitable values of parameters λ and δ can be found by cross validation within algorithms analogous to those given above. The method is preferable 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 EXAMPLES

To illustrate the performance of the robust regularized versions of LDA, we analyze four different real data sets. Each data set fulfils $n < p$ and three of them can be described as omics data with $n \ll p$. The aim of each example is to distinguish between two groups of observations. Table 1 overviews the classification performance of 5-fold cross validation in the form of the Youden's index I, which is defined as

$$I = \text{sensitivity} + \text{specificity} - 1 \quad (22)$$

and fulfils $I \in [-1,1]$.

Table 1: Youden's index (22) as a classification performance measure computed for a 5-fold cross validation study on various data sets of Sections 3.1 to 3.4. Data from Section 3.2 are considered as raw as well as after a contamination by normally distributed outliers $\mathsf{N}(0,\sigma^2)$ for different values of σ.

		Section 3.2					
			Contam. for $\sigma =$				
	Section 3.1	Raw	0.1	0.2	0.3	Section 3.3	Section 3.4
n	48	168				42	32
p	38 614	4005				518	15
Regularized versions of LDA							
PAM	0.85	0.88	0.81	0.75	0.68	0.86	0.51
LDA*	1.00	1.00	0.95	0.94	0.92	0.89	0.71
SCRDA	1.00	1.00	1.00	1.00	0.99	0.91	0.80
MWCD-LDA*	1.00	1.00	1.00	1.00	1.00	0.91	0.79
MWCD-LDA**	1.00	1.00	1.00	1.00	1.00	0.92	0.80
Other classification methods							
SVM	1.00	1.00	0.99	0.98	0.96	0.92	0.85
Classification tree	0.94	0.96	0.95	0.91	0.92	0.84	0.11
Lasso-LR	0.97	0.99	1.00	0.97	0.94	0.87	0.82
Number of principal components	10	10	10			20	4
PCA $\Longrightarrow$ LDA	0.15	1.00	0.94	0.93	0.88	0.70	0.59
PCA $\Longrightarrow$ LDA*	0.51	1.00	0.95	0.94	0.89	0.62	0.59
PCA $\Longrightarrow$ SCRDA	0.62	1.00	0.95	0.94	0.89	0.72	0.59
Number of selected genes	10	10	10			20	4
MRMR $\Longrightarrow$ LDA	0.90	1.00	0.94	0.93	0.89	0.88	0.72
MRMR $\Longrightarrow$ LDA*	0.96	1.00	0.96	0.93	0.89	0.88	0.76
MRMR $\Longrightarrow$ SCRDA	1.00	1.00	0.96	0.93	0.89	0.90	0.76

We computed various classifiers in *R* software. All regularized versions of LDA use the unit matrix as the target matrix T and the MWCD-LDA* uses linearly decreasing weights in the form

$$w_i = \frac{\tilde{w}_i}{\sum_{j=1}^{n} \tilde{w}_j}, \quad i = 1,\ldots,n, \tag{23}$$

where

$$\tilde{w}_i = 1 - \frac{i-1}{n}, \quad i = 1,\ldots,n. \tag{24}$$

With such simple choice of decreasing weights, the outliers will obtain very small weights and their effect will be reduced considerably.

Besides the methods already described in this paper, we used also several standard machine learning methods, e.g. support vector machines (SVM) with a radial basis function kernel or logistic regression using the lasso regularization (lasso-LR) (Friedman et al., 2015).

To compare the results with the effect of dimensionality reduction, we also use the principal component analysis (PCA) and the Minimum Redundancy Maximum Relevance (MRMR), where the latter is a supervised variable selection with Pearson's correlation coefficient as measure of relevance and redundancy (Peng et al., 2005). The number of principal components and the number of selected variables by the MRMR procedure for the four particular data sets is given in Table 1 together with the results, which will be discussed later in Section 4.

3.1 Cardiovascular Genetic Study

We participated on a cardiovascular genetic study with the to identify a small set of genes associated with excess genetic risk for the incidence of a cardiovascular disease among $p = 38\,590$ gene transcripts (the link must stay hidden according to submission guidelines). The gene expressions are measured on $n = 48$ individuals, namely on 24 patients having a cerebrovascular stroke and 24 control persons.

3.2 Brain Activity Data

Further, we analyze a real data set from neuroscience research investigating the spontaneous activity of various parts of the brain by means of neuroimaging methods. Specific functions of individual parts of the brain have been already described (Duffau, 2011), but spontaneous brain activity and especially connection between pairs of brain parts in the resting state (i.e.

resting-state brain networks) has been a hot topic in current neuroscience (Hlinka et al., 2011).

We participated on a study of the brain activity of $n = 24$ probands, which was measured by means of fMRI under 7 different situations. One of them can be characterized as a resting state, i.e. rest without any stimulus. Besides, the probands were observing each of 6 different movies while measuring the brain activity in the same way. The fMRI divides the brain to 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 only $p = 90 * 89/2 = 4005$ variables containing values of correlation coefficients for each of the 24 probands. The basic task is to classify the resting state from (any) movie, i.e. all movies together are considered to be one class. In general, fMRI measurements are known to be contaminated by noise as well as outliers (Wager et al., 2005). It is also true with out data and therefore robust methods are highly desirable for their analysis.

The task is to learn a classification rule allowing to discriminate between two groups (resting state, movie). This is a classification to 2 groups with $p = 4005$ variables over $n = 168$ individuals, while the resting state group contains 24 observations and the group corresponding to any movie contains $6 * 24 = 144$ observations.

In addition, we investigate the performance of various classification methods on data contaminated by noise. For this purpose, 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 for the 5-fold cross valiation was evaluated for each case for various methods. We consider the noise with $\sigma = 0.1$ to be slight and with $\sigma = 0.3$ to be moderate, revealing already the advantage of robust methods compared to non-robust ones.

3.3 Metabolomic Profiles Data

We analyze a publicly available data set of prostate cancer metabolomic data (Sreekumar et al., 2009) of $p = 518$ metabolites measured over two groups of $n = 42$ patients, who are either those with a benign prostate cancer (16 patients) or with other cancer types (26 patients). The task in both examples is to learn a classification rule allowing to discriminate between the two classes of individuals.

3.4 Keystroke Dynamics Data

The last data set contains data from biometric authentication by means of keystroke dynamics. We participate on a study aiming at proposing and implementing a biometric authentication system for medical reports within a hospital based on keystroke dynamics measurements. Our detailed analysis goes beyond the results of (Kalina and Schlenker, 2015).

The training data set contains keystroke durations and keystroke latencies measured in milliseconds on $n = 32$ probands, who typed 10-times in their habitual speed a short sequence of 8 characters. This sequence (password) was the same for each proband. In spite of a small value of $p = 15$ variables, p exceeds the number of measurements for each individual.

In the practical application, one of the 32 individuals identifies himself/herself (say as XY) and types the password. The aim of the analysis is to verify if the individual typing on the keyboard is or is not the person XY. Thus, the authentication task is a classification problem to assign the individual to one of $K = 2$ groups.

4 RESULTS

4.1 Cardiovascular Genetic Study

This data set contains the largest number of variables among the data sets analyzed in the whole paper. Some methods reach the classification performance correct in 100 % of cases, including SVM and also some of the regularized versions of LDA. This is true also for MWCD-LDA* and MWCD-LDA**.

While dimensionality reduction by PCA has drastic consequences, it turns out that there is a small number of variables responsible for the separation between the two groups. The bad performance of PCA can be explained by its unsupervised nature, ignoring the grouping structure of the data. A supervised variable selection by MRMR yields much improved results and there are 10 genes selected which allow to separate both groups again with a 100 % correctness if SCRDA is used.

4.2 Brain Activity Data

Averaged values of the classification accuracy computed over the 100 cases are given in Table 1. Several classification methods including the MWCD-LDA* and MWCD-LDA** yield results correct in 100 % of cases. Because these raw data are not contaminated by severe outliers, there seems no advantage

of the robust regularized LDA over non-robust versions. Still, PAM turns out to be heavily influenced by them, although it was actually proposed as a denoised version of diagonalized LDA (Tibshirani and Narasimhan, 2003).

The results on the contaminated data reveal an evidence of robustness of the new approach. The classification performance of standard methods including SVM or the (non-robust) SCRDA is decreased compared to raw data, while MWCD-LDA* and MWCD-LDA** are able to outperform them. The MRMR variable selection allows to find a small set of variables with an ability to diagnose schizophrenic patients based only on the fMRI measurements of the brain in the resting state, which is an interesting result from the point of view of neuroscience research.

4.3 Metabolomic Profiles Data

There seem to be no severe outliers in the data. MWCD-LDA* and MWCD-LDA** are still able to slighly outperform other regularized versions of LDA. Their result are only slightly different and comparable to the SVM. Other classifiers yield inferior results. The MRMR variable selection performs better compared to the unsupervised dimensionality reduction by means of PCA, while there is to be no remarkable small group of variables responsible for a large portion of variability of the data and the first few principal components seem rather arbitrary for the classification task.

4.4 Keystroke Dynamics Data

The last column of Table 1 gives classification accuracy of various methods obtained on the keystroke dynamics data. This data set is used for comparison to reveal the behavior of regularized methods on a small number of variables ($p = 15$). The best results are obtained with SVM, which is again based on a large number of support vectors (≥ 90 % of observations), while MWCD-LDA* or SCRDA are slightly inferior. Dimensionality reduction leads to a loss of information compared to methods using all variables. The data contain approximately 10 % of rather severe outliers. MWCD-LDA* and MWCD-LDA** retain their performance if the outliers are ignored, while SVM and non-robust versions of LDA seem to be slightly affected by their presence.

5 DISCUSSION

Regularized LDA has been advocated for both computational and statistical benefits (Pourahmadi, 2013), which is true not only for $n < p$ but also for $n > p$ with a relatively small n (Hastie et al., 2008). The regularized covariance matrix can be theoretically justified as a Stein's estimator (Pourahmadi, 2013), in analogy to Stein's shrinkage estimator of the mean of multivariate normal data (Hastie et al., 2008; Hausser and Strimmer, 2009). Some authors claim that regularization of the covariance matrix ensures a robustness, although there is no theoretical justification for this belief. Actually, regularized LDA may be interpreted from the point of view of robust optimization (Xanthopoulos et al., 2013), which deals with small (local) changes of the measured data. Nevertheless, regularized LDA is not robust to more severe noise or outliers, as revealed in our examples.

SVM yields the best classification performance in some of the examples, especially those with a relatively smaller p. Nevertheless, we perceive its following drawbacks.

- It depends on too many support vectors. In the examples, more than 90 % of the observations play the role of support vectors.
- The necessity to optimize its parameters over a sufficiently large number of observations (Cai and Shen, 2010).
- A tendency to overfitting for $n < p$ (Han and Jiang, 2014).
- It works as a black box.
- Non-robustness to outliers.

Appealing properties of regularized LDA have lead us to the idea of joining principles of statistical robustness with a suitable regularization. Advantages of the two newly proposed methods denoted as MWCD-LDA* and MWCD-LDA** include:

- High robustness to outliers thanks to a high breakdown point of MWCD, which is ensured by the implicit weights, similarly to linear regression (Víšek, 2006; Kalina, 2012);
- No assumption on the distribution of the outliers;
- An efficient algorithm based on numerical linear algebra;
- No need for a prior dimensionality reduction;
- Comprehensibility.

5.1 Comprehensibility

Comprehensibility of the newly proposed methods, which represents an important requirement in a wide variety of classification tasks in bioinformatics, deserves to be discussed as a separate section.

The classical LDA itself is considered to be comprehensible, because it is based on the Mahalanobis distance of a new measurement from each of the groups of data. The contribution of an individual observation to the final classification rule is only through the sufficient statistics, i.e. mean of the corresponding groups and covariance matrix. The classification rules of MWCD-LDA* and MWCD-LDA** can be interpreted as based on a deformed Mahalanobis distance. To explain this, let us consider the singular value decomposition (SVD) of S^*_{MWCD} in the form

$$S^*_{MWCD} = Q\Lambda Q^T. \tag{25}$$

We claim that the deformed Mahalanobis distance of MWCD-LDA* can be interpreted as the Euclidean distance applied on $\Lambda^{-1/2}Q^TZ$. To explain this, let us consider the aim to assign a new observation Z, which is not outlying, to one of the groups. In a straightforward way, we obtain

$$\begin{aligned} \mathsf{var}\,\Lambda^{-1/2}Q^TZ = \Lambda^{-1/2}Q^T \cdot \mathsf{var}\,Z \cdot Q\Lambda^{-1/2} \approx \\ \approx \Lambda^{-1/2}Q^T \cdot Q\Lambda Q^T Q\Lambda^{-1/2} = I. \end{aligned} \tag{26}$$

Also the implicit weights assigned to individual observations allow a clear interpretation. Less reliable observations (potential outliers) obtain small or negligible weights. Such permutation of the weights is used which minimizes the determinant of a weighted covariance matrix. The weights are used to compute the weighted mean and weighted covariance matrix. In the examples, we have verified that outlying measurements obtain small weights, which ensures the robustness of the method.

Particularly for the MWCD-LDA**, the means are replaced by shrunken means, while we use an original idea to shrink towards the pooled mean across groups. This is a difference from all available algorithms shrinking towards zero (Guo et al., 2007).

5.2 Limitations

Let us mention also the limitations of MWCD-LDA* and MWCD-LDA**.

- Suitability for data following a contaminated multivariate normal distribution.
- Intensive computations are required.
- The weights are assigned to individual observations rather than to individual variables.
- An implicit assumption that the variability is not substantially different across variables. This is the same as for SCRDA or other regularized LDA methods and the novel methods seem to yield reliable results although this implicit assumption of homogeneous variances of all variables is violated in all the data sets of Section 3.

Finally, we need to recall the regularization itself to be a rather drastic intervention to the original problem of rank n, which is replaced by a problem of a much larger rank p. Such increase of the dimensionality of the covariance matrix may cause the new problem to be very distant from the original problem even if an extremely small λ is used and if the new problem is solved with a perfect numerical precision (Kůrková and Sanguineti, 2005; Davies, 2014).

6 CONCLUSIONS AND FUTURE WORK

Numerous available algorithms for the regularized LDA are popular for the analysis of high-dimensional data (Kalina, 2014). However, regularized LDA turns out to be vulnerable to the presence of outliers, because it is based on the same maximum likelihood estimation principle as the standard LDA. It is the maximum likelihood estimation which causes the high sensitivity of the standard as well as of various regularized versions of LDA to outliers.

This paper proposes new robust classification methods for high-dimensional observations, i.e. assuming the number of variables n to exceed (perhaps largely) the number of variables p. We combine robustness to the presence of outliers with regularized estimation of the covariance matrix of the multivariate data in a unique way. Two robust classification methods denoted as MWCD-LDA* and MWCD-LDA** are proposed in Section 2, which are based on implicit weighting of individual observations and on a regularized version of a highly robust covariance matrix. MWCD-LDA* considers only a regularization of the covariance matrix, while MWCD-LDA** additionally replaces the mean of each group by a regularized robust estimator.

We analyzed four data sets fulfilling $n < p$ in Section 3, while three of them coming from bioinformatics research are high-dimensional in sense of $n \ll p$. On the whole, we can say that MWCD-LDA* performs very well for raw high-dimensional data as well as after contamination by noise. It is an artificial contamination of the data which reveals the robustness of MWCD-LDA* and MWCD-LDA** as a strong point of these methods. In the examples, various classification methods show distinct differences between the groups of observations.

In comparison to MWCD-LDA*, MWCD-LDA** allows to replace standard sample means of each group by shrinkage counterparts. This is possible only at a cost of a much higher increase of computational complexity. Nevertheless, the results of

MWCD-LDA** are only slightly improved compared to MWCD-LDA*.

Open problems concerning the newly proposed methods as well as more general ideas for a future research in the area of robust analysis of high-dimensional data contain the following tasks.

- Finding more efficient algorithms for specific choices of the target matrix T.
- Comparing various approaches to regularizing the means, mainly comparing the effect of L_2 and L_1 norm. In addition, comparing the effect of shrinking the means towards the common mean vs. towards zero.
- Comparing the performance and robustness of the new methods with approaches based on robust PCA.
- Investigating the non-robustness of other standard regularized classification methods.
- Applying the regularized robust Mahalanobis distance to modify other methods based on the Mahalanobis distance, such as classification trees, entropy estimators, k-means clustering, or dimensionality reduction.
- Combining regularization and robustness to other methods, including neural networks or SVM or even linear regression (Jurczyk, 2012).

ACKNOWLEDGEMENTS

This publication was 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. The work of J. Kalina was financially supported by the Neuron Fund for Support of Science. The work of J. Hlinka was supported by the Czech Science Foundation project No. 13-23940S. The authors are thankful to Anna Schlenker for the data analyzed in Section 3.4.

REFERENCES

Cai, T. and Shen, X. (2010). *High-dimensional data analysis*. World Scientific, Singapore.

Chen, Y., Wiesel, A., and Hero, A. O. (2011). Robust shrinkage estimation of high dimensional covariance matrices. *IEEE Transactions on Signal Processing*, 59:4097–4107.

Christmann, A. and van Messem, A. (2008). Bouligand derivatives and robustness of support vector machines for regression. *Journal of Machine Learning Research*, 9:915–936.

Croux, C. and Dehon, C. (2001). Robust linear discriminant analysis using S-estimators. *Canadian Journal of Statistics*, 29:473–493.

Davies, P. (2014). *Data analysis and approximate models: Model choice, location-scale, analysis of variance, nonparametric regression and image analysis*. Chapman & Hall/CRC, Boca Raton.

Davies, P. L. and Gather, U. (2005). Breakdown and groups. *Annals of Statistics*, 33:977–1035.

Duffau, H. (2011). *Brain mapping. From neural basis of cognition to surgical applications.* Springer, Vienna.

Filzmoser, P. and Todorov, V. (2011). Review of robust multivariate statistical methods in high dimension. *Analytica Chinica Acta*, 705:2–14.

Friedman, J., Hastie, T., Simon, N., and Tibshirani, R. (2015). glmnet: Lasso and elastic-net regularized generalized linear models. http://cran.r-project.org/web/packages/glmnet/index.html.

Guo, Y., Hastie, T., and Tibshirani, R. (2007). Regularized discriminant analysis and its application in microarrays. *Biostatistics*, 8:86–100.

Han, H. and Jiang, X. (2014). Overcome support vector machine diagnosis overfitting. *Cancer Informatics*, 13:145–148.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The elements of statistical learning*. Springer, New York, 2nd edition.

Hausser, J. and Strimmer, K. (2009). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, 10:1469–1484.

Hlinka, J., Paluš, M., Vejmelka, M., Mantini, D., and Corbetta, M. (2011). Functional connectivity in resting-state fMRI: Is linear correlation sufficient? *NeuroImage*, 54:2218–2225.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust statistics*. Wiley, New York, 2nd edition.

Hubert, M., Rousseeuw, P. J., and van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statistical Science*, 23:92–119.

Jurczyk, T. (2012). Outlier detection under multicollinearity. *Journal of Statistical Computation and Simulation*, 82:261–278.

Kůrková, V. and Sanguineti, M. (2005). Learning with generalization capability by kernel methods of bounded complexity. *Journal of Complexity*, 21:350–367.

Kalina, J. (2012). Implicitly weighted methods in robust image analysis. *Journal of Mathematical Imaging and Vision*, 44:449–462.

Kalina, J. (2014). Classification analysis methods for high-dimensional genetic data. *Biocybernetics and Biomedical Engineering*, 34:10–18.

Kalina, J. and Schlenker, A. (2015). A robust and regularized supervised variable selection. *BioMed Research International*, 2015(320385).

Kindermans, P.-J., Schreuder, M., Schrauwen, B., Mller, K.-R., and Tangermann, M. (2014). True zero-training brain-computer interfacing–An online study. *PLoS One*, 9(102504).

Krzanowski, W. J., Jonathan, P., McCarthy, W. V., and Thomas, M. R. (1995). Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data. *Applications of Statistics*, 44:101–115.

Lopuhaä, H. P. and Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Annals of Statistics*, 19:229–248.

Mallick, B. K., Gold, D., and Baladandayuthapani, V. (2009). *Bayesian analysis of gene expression data*. Wiley, New York.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 27:1226–1238.

Pourahmadi, M. (2013). *High-dimensional covariance estimation*. Wiley, New York.

Roelant, E., van Aelst, S., and Willems, G. (2009). The minimum weighted covariance determinant estimator. *Metrika*, 70:177–204.

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

Sreekumar, A. et al. (2009). Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, 457:910–914.

Tibshirani, R. and Narasimhan, B. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, 18:104–117.

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

Tyler, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *Annals of Statistics*, 15:234–251.

Tyler, D. E. (2014). Breakdown properties of the M-estimators of multivariate scatter. http://arxiv.org/pdf/1406.4904v1.pdf.

Víšek, J. Á. (2006). The least trimmed squares. Part III: Asymptotic normality. *Kybernetika*, 42:203–224.

Wager, T. D., Keller, M. C., Lacey, S. C., and Jonides, J. (2005). Increased sensitivity in neuroimaging analyses using robust regression. *NeuroImage*, 26:99–113.

Xanthopoulos, P., Pardalos, P. M., and Trafalis, T. B. (2013). *Robust data mining*. Springer, New York.