Robustness of High-Dimensional Data Mining

Jan Kalina, Jurjen Duintjer Tebbens, and Anna Schlenker

Institute of Computer Science AS CR, Praha, Czech Republic
kalina@cs.cas.cz

Abstract: Standard data mining procedures are sensitive to the presence of outlying measurements in the data. This work has the aim to propose robust versions of some existing data mining procedures, i.e. methods resistant to outliers. In the area of classification analysis, we propose a new robust method based on a regularized version of the minimum weighted covariance determinant estimator. The method is suitable for data with the number of variables exceeding the number of observations. The method is based on implicit weights assigned to individual observations. Our approach is a unique attempt to combine regularization and high robustness, allowing to downweight outlying high-dimensional observations. Classification performance of new methods and some ideas concerning classification analysis of high-dimensional data are illustrated on real raw data as well as on data contaminated by severe outliers.

1 Robustness in Data Mining

Numerous data mining procedures are commonly based on distance measures between observations or their groups, clusters, etc. Most measures for continuous data are very sensitive to the presence of outlying measurements in the data. This is true for Euclidean, Mahalanobis, and Manhattan distances, the cosine similarity, and many others. This paper has the aim to propose and study new robust data mining methods for continous data by means of robustifying the Mahalanobis distance.

Sensitivity of standard data mining as well statistical methods to the presence of outlying measurements in the data has been repeatedly reported as a serious problem [5]. This is true in various areas of applications, including classification analysis, clustering, dimensionality reduction, prediction models, etc. Robust methods in statistics and data mining are those resistant to the influence of noise and to the presence of outliers. Xanthopoulos et al. [27] obtained various robust data mining procedures for continuous data as a solution of optimization tasks taking into account uncertainty of the observed values. Robustness aspects of neural networks and support vector machines were overviewed in [15].

Another important problem in data mining commonly occurs if the number of variables p exceeds the number of observations n (i.e. $n < p$ or even $n \ll p$). In this context, we speak of high-dimensional data and their analysis is described as the large p/small n problem. Some standard methods suffer from the curse of dimensionality, which is manifested through numerical unstability or computational infeasibility [19]. Two most common solutions are suitable regularization and dimensionality reduction (variable selection). The concept of regularization encompasses a variety of approaches allowing to solve ill-posed or insoluble high-dimensional problems by means of additional information, assumptions, or penalization [7]. So far, regularized and at the same time robust data mining procedures for $n < p$ have been rarely discussed [14]. In this paper, we present a unique attempt to combine principles of regularization and robust statistics for $n \ll p$.

This paper has the following structure. Section 2 discusses various existing approaches to linear discriminant analysis for $n \ll p$. Following sections have the aim to combine regularization principles with ideas of robust statistics to propose new robust methods for classification analysis of high-dimensional data. In Section 3, a new regularized robust classification method based on M-estimation is proposed. In Section 4, a regularized highly robust estimator of multivariate scatter based on the minimum weighted covariance determinant estimator is proposed, which exploits the idea of implicit weights assigned to individual observations. The core Section 5 exploits previous sections to propose a regularized highly robust classification method, based on down-weighting less reliable high-dimensional observations. Following examples in Sections 6 and 7 illustrate important ideas concerning the high dimensionality in the classification task and the classification performance of individual methods. Finally, Section 8 concludes the paper.

2 Linear Discriminant Analysis

In this section, we recall the linear discriminant analysis (LDA) as a standard classification analysis procedure and its modifications suitable for data in the $n \ll p$ situation.

Classification analysis has the aim to construct (learn) a decision rule based on a training data set, which is able to automatically assign new data to one of K groups. It assumes n observations with p variables, observed in K different samples (groups) with $p > K \geq 2$,

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

where $n = \sum_{k=1}^{K} n_k$. LDA 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 will be denoted by S. In its standard form, LDA assumes $n > p$ and is based on computing the Mahalanobis distance

between a new observation Z and the mean of each of the K groups.

For $n < p$, the matrix S is singular and computing its inverse is not possible. The most important approaches include regularization, i.e. replacing the computation of S^{-1} by an appropriate alternative, performing a dimensionality reduction prior to LDA, computing a pseudoinverse matrix, which is however unstable due to a small n, generalized SVD decomposition, or elimination of the common null space of the between-group and within-group covariance matrices [4].

Suitable regularized estimators of the covariance matrix (e.g. [6]) are guaranteed to be regular and positive definite even for $n \ll p$. Habitually used regularized versions of LDA for $n \ll p$ have the form

$$S^* = \lambda S + (1-\lambda)T, \quad \lambda \in (0,1), \tag{2}$$

using a given target matrix T, which must be a regular symmetric positive definite matrix of size $p \times p$. Its most common choices include the identity matrix $\mathscr{I}_p$ or a diagonal (non-identity) matrix. A suitable value of λ can be found by cross-validation. Nevertheless, their fast computation and numerical stability remains to be an important issue [14].

3 Regularized M-LDA (M-RLDA)

M-estimators are the most common robust statistical estimators applicable to a variety of tasks [12]. They originated in the seminal paper of Huber [9], who investigated robust estimation of a location parameter. In this section, we propose a new method denoted as M-RLDA, which abbreviates the regularized version of robust linear discriminant analysis based on M-estimation. However, the disadvantage of M-estimators is their low robustness in terms of the breakdown point. Therefore, the approach of this section is rather intended to yield an initial regularized robust estimator for a highly robust approach in Section 5.

We consider the multivariate model with p-dimensional data $X_1, \ldots, X_n$ in the form $X_i = \mu + e_i$, where the noise random vectors $e_1, \ldots, e_n$ are independent following the normal distribution $N(0, \Sigma)$. Unfortunately, M-estimation does not allow to jointly estimate the mean and covariance matrix in a simple way [13]. Let ψ denote the Huber's function. The expectation (population mean) μ will be estimated by $\bar{X}_M = (\bar{X}_1^M, \ldots, \bar{X}_p^M)^T$, where the coordinate μ_i will be estimated by the Huber's estimator as the solution of

$$\sum_{j=1}^{n} \psi(X_{ij} - \bar{X}_i^M) = 0. \tag{3}$$

To estimate Σ for a given estimator $t \in \mathbb{R}^p$ of μ, Tyler [25] proposed an M-estimator which is computed iteratively as the solution of

$$p \cdot ave \left\{ \frac{(X_i - t)(X_i - t)^T}{(X_i - t)V_n^{-1}(X_i - t)^T} \right\} = V_n \tag{4}$$

where *ave* denotes the average of the given values over $i = 1, \ldots, n$. We will use its regularized version proposed by Chen et al. [2] for $n \ll p$. This estimator is based on a ridge regularization of the estimator of Σ in each iteration of the computation and will be denoted as S_M^*.

We assume the data (1) to observed in K groups. Let $\bar{X}_{k,M}$ denote the M-estimator of the mean in the k-th group. Let

$$\bar{X}_{k,M}^* = \delta \bar{X}_{k,M} + (1-\delta)\bar{X}_M, \quad k = 1, \ldots, K, \tag{5}$$

for $\delta \in (0,1)$ and let $\bar{X}^M$ denote the overall M-estimator across groups. Now we define a modified version of Mahalanobis distance based on M-estimation, denoted as M-Mahalanobis distance, and a corresponding version of regularized LDA, denoted as M-RLDA.

Definition 1. *The regularized M-Mahalanobis distance between an observation Z and the k-th group is defined as*

$$(\bar{X}_{k,M}^* - Z)^T S_M^{*-1} (\bar{X}_{k,M}^* - Z). \tag{6}$$

Algorithm 1. *M-RLDA.*

Step 1 *For a given $\delta \in (0,1)$, compute the matrix*

$$A = [\bar{X}_{1,M}^* - Z, \ldots, \bar{X}_{K,M}^* - Z] \tag{7}$$

of size $p \times K$ whose k-th column is $\bar{X}_{k,M}^* - Z$.

Step 2 *Compute S_M^* as*

$$S_M^* = \lambda S_M + (1-\lambda)T \tag{8}$$

with a fixed $\lambda \in (0,1)$ and a given target matrix T.

Step 3 *Compute and store the eigenvalues of S_M^* in the diagonal matrix D_*, and compute and store the corresponding eigenvectors of S^* in the orthogonal matrix Q_*.*

Step 4 *Compute the matrix*

$$B = D_*^{-1/2} Q_*^T A \tag{9}$$

and assign Z to group k if the column of B with largest Euclidean norm is the k-th column.

Step 5 *Repeat steps 1 to 4 with different values of δ and λ and find the classification rule with the best classification performance.*

For the special case $T = \mathscr{I}_p$, which is commonly denoted as Tikhonov or ridge regularization of S, a more efficient computation can be performed as follows.

Algorithm 2. *M-RLDA for the ridge regularization.*

Step 1 *Compute the matrix (7) of size $p \times K$ whose k-th column is $\bar{X}_k - Z$ and compute the matrix Y as*

$$\begin{aligned} Y = \quad & [X_{11} - \bar{X}_M, \ldots, X_{1n_1} - \bar{X}_M, \quad \ldots, \\ & X_{K1} - \bar{X}_M, \ldots, X_{Kn_K} - \bar{X}_M]^T. \end{aligned} \tag{10}$$

Step 2 *Compute the singular value decomposition of Y as*

$$Y = P\Sigma Q^T, \tag{11}$$

with singular values $\{\sigma_1,\dots,\sigma_n\}$ and complement these singular values with $p-n$ zero values $\sigma_{n+1} = \cdots = \sigma_p = 0$.

Step 3 *For a fixed $\lambda \in (0,1)$, compute $D_* =$*

$$= diag\{\lambda\sigma_1^2 + (1-\lambda),\dots,\lambda\sigma_p^2 + (1-\lambda)\}. \tag{12}$$

Step 4 *Compute the matrix*

$$B = D_*^{-1/2} Q^T A \tag{13}$$

and assign Z to group k if the column of B with largest Euclidean norm is the k-th column.

Step 5 *Repeat steps 2 to 4 with different values of λ and find the classification rule with the best classification performance.*

4 Implicitly Weighted Robust Methods

Our aim is a regularized highly robust classification analysis procedure. Before we develop such method in Section 5, we will devote the present section to robust estimation of parameters of high-dimensional data. This section starts by recalling the least weighted squares regression estimator [26] and the minimum weighted covariance determinant (MWCD) estimator for multivariate data [16]. Both methods are highly robust estimation procedures based on assigning implicitly given weights to individual observations. However, they are computationally infeasible for $n < p$. As a new result, a regularized version of the MWCD estimator is proposed, which exploits the tools of Section 3.

Linear regression remains to be the most commonly investigated statistical model in the context of robust statistics [12]. Therefore, we will explain some important principles on the standard linear regression model

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i, \quad i = 1,\dots,n. \tag{14}$$

with independent identically distributed random errors $e_1,\dots,e_n$, without assuming their Gaussian distribution.

We will need the following notation. Let us consider (any) estimate $b = (b_1,\dots,b_p)^T \in \mathbb{R}^p$ of the parameter $\beta = (\beta_1,\dots,\beta_p)^T$. We denote the residual corresponding to the i-th observation by

$$u_i(b) = y_i - b_1 X_{i1} - \cdots - b_p X_{ip}, \quad i = 1,\dots,n \tag{15}$$

and let us order the squared residuals

$$u^2_{(1)}(b) \le u^2_{(2)}(b) \le \cdots \le u^2_{(n)}(b). \tag{16}$$

The idea of the highly robust LWS estimator is to down-weight less reliable observations, which are likely to be outliers. The weights are assigned to individual data during the computation of the estimator based on residuals. One possible choice of weights is based on an implicit permutation of given (fixed) magnitudes of the weights $w_1,\dots,w_n$ fulfilling $\sum_{i=1}^n w_i = 1$ to the data. Data-dependent adaptive weights of [3] are another choice.

The LWS estimator of $(\beta_1,\dots,\beta_p)^T$ in the model (14) is defined as

$$b^{LWS} = (b_1^{LWS},\dots,b_p^{LWS})^T = \arg\min_{b\in\mathbb{R}^p} \sum_{i=1}^n w_i u^2_{(i)}(b). \tag{17}$$

The computation of the LWS estimator with adaptive weights begins with an initial highly robust estimator and proceeds to proposing values of the weights based on comparing the empirical distribution function of squared residuals with the theoretical counterpart assuming normality.

Statistical methods based on ranks of observations have appealing properties [11] in a variety of situations. The LWS estimator has a high breakdown point, which is a statistical measure of sensitivity against outliers in the data [12]. The estimator has asymptotically a 100 % efficiency of the least squares estimator under Gaussian errors and its relative efficiency has been numerically evaluated as high (over 85 %) compared to maximum likelihood estimators under various distributional models [3]. Moreover, we accompanied the LWS estimator by a robust coefficient of determination and asymptotic hypothesis tests in [17].

Further, we consider the multivariate model with p-dimensional data $X_1,\dots,X_n$ in the form $X_i = \mu + e_i$ with noise modeled by independent random vectors $e_1,\dots,e_n$ following the normal distribution $N(0,\Sigma)$. The MWCD estimator, which estimates the parameters μ and Σ jointly under the assumption $n > p$, will be now recalled. Other available robust estimators of parameters of multivariate data were studied e.g. by [10, 5].

The MWCD-estimator of the mean of the data has the form of a weighted mean. At the same time, the MWCD-estimator of Σ has the form of a weighted covariance matrix. Both these estimators are computed with such weights, which correspond to the optimal permutation of given values $w_1,\dots,w_n$ with $\sum_{i=1}^n w_i = 1$. We do not assume the data to come from the normal distribution. However, the data are assumed to be in general position, i.e. any p observations among the total number of n observations are assumed to give a unique determination of Σ.

Definition 2. *The MWCD estimator of μ denoted as $\bar{X}_{MWCD}$ is equal to the weighted mean of $X_1,\dots,X_n$ in the form $\bar{X}_w = \sum_{i=1}^n w_i X_i$ with such permutation of $w_1,\dots,w_n$, for which the determinant of*

$$S_w = \sum_{i=1}^n w_i (X_i - \bar{X}_w)(X_i - \bar{X}_w)^T \tag{18}$$

is minimal. The MWCD estimator of Σ denoted as S_{MWCD} is equal to S_w with this optimal permutation of $w_1,\dots,w_n$.

The MWCD with adaptive weights attains the finite-sample breakdown point

$$\left\{ \left\lfloor \frac{n-p+1}{2} \right\rfloor \right\} / n \tag{19}$$

for any p-dimensional data $X_1, \ldots, X_n$ in general position, where $\lfloor a \rfloor$ stands for the integer part of a. At the same time, (19) is the maximal breakdown point of affine-equivariant estimators of Σ [18].

The estimator cannot be computed for $n < p$. Therefore, we define its regularized version, which exploits the regularized robust Mahalanobis distance based on M-estimation (6) and is computationally feasible even for $n \ll p$.

Algorithm 3. *Regularized MWCD estimator.*

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

Step 2 *Randomly select an initial set of* $n/2$ *observations. Compute* $\bar{X}^*_M$ *and* S^*_M *based on these observations. Denote* $T = \bar{X}^*_M$ *and* $C = S^*_M$.

Step 3 *Compute the regularized M-Mahalanobis distance*

$$d(i;T,C) = \left[(X_i - T)^T C^{-1} (X_i - T) \right]^{1/2} \tag{20}$$

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*

$$d(\pi(1);T,C) \leq d(\pi(2);T,C) \leq \cdots \leq d(\pi(n);T,C). \tag{21}$$

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 *The loss function is evaluated as the determinant of the matrix*

$$\det\left(\lambda S_w + (1-\lambda)T\right), \tag{22}$$

where S_w *is evaluated as (18) with the weights from Step 3. If the loss is smaller than the previously obtained value, continue with step 5. Otherwise go to step 6.*

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.*

5 Regularized Robust Classification Analysis

In this section, we propose a regularized robust version of the Mahalanobis distance together with a regularized robust version of LDA, exploiting the tools of Sections 3 and 4. The new method MWCD-RLDA represents an MWCD-based regularized linear discriminant analysis, computed using a deformed Mahalanobis distance in the multivariate space. We also present an efficient algorithm for its computation. The MWCD estimator yields an estimator of the expectation and covariance matrix of multivariate data jointly. While implicitly weighted robust methods are known to be computationally infeasible for high-dimensional data [1], our work overcomes the high dimensionality by a sophisticated regularization. To the best of our knowledge, this is a first attempt to consider a highly robust estimator for high-dimensional data which is based on implicit weights assigned to individual observations.

We assume p-dimensional data $X_1, \ldots, X_n$. The Mahalanobis distance will be formulated as a distance of an observation $Z = (Z_1, \ldots, Z_p)^T$ from a group of p-dimensional observations $X_1, \ldots, X_n$. Let $\bar{X}_{MWCD}$ and S_{MWCD} denote the estimators of μ and Σ obtained by the regularized MWCD estimator of Section 4.

We denote the regularized MWCD-estimator of the covariance matrix Σ as

$$S^*_{MWCD} = \lambda S_{MWCD} + (1-\lambda)T, \quad \lambda \in (0,1). \tag{23}$$

The parameter $\lambda \in (0,1)$ denotes a shrinkage estimator of the covariance matrix across groups. A suitable value of λ is found by a cross-validation in the form of a grid search over all possible values of $\lambda \in (0,1)$.

Definition 3. *The regularized MWCD-Mahalanobis distance between an observation* Z *and the data* $X_1, \ldots, X_n$ *is defined as*

$$(\bar{X}_{MWCD} - Z)^T S^{*-1}_{MWCD} (\bar{X}_{MWCD} - Z). \tag{24}$$

Further, we assume the data to be observed in K different groups as in (1). In this context, we replace (24) by such version, where the mean of each group if shrunken towards the overall MWCD-mean across groups. This can be interpreted as a regularized (biased) version of the MWCD-mean or Stein's shrinkage estimator, which improves the mean square error of the (unbiased) mean. It can be alternatively derived in a Bayesian setting.

Tibshirani et al. [24] applied such shrinkage on regular means of each group towards the overall mean, which improved the classification performance, and claimed to improve the robustness of their method PAM. We consider their argument as misleading, because the finite-sample breakdown point of PAM can be easily evaluated as only $1/n$.

The overall MWCD-mean across groups will be denoted as $\bar{X}^{MWCD}$ and the MWCD-mean of individual groups as $\bar{X}_1^{MWCD},\ldots,\bar{X}_K^{MWCD}$. We denote

$$\bar{X}^*_{k,MWCD} = \delta\bar{X}_{k,MWCD} + (1-\delta)\bar{X}_{MWCD} \qquad (25)$$

for $k = 1,\ldots,K$ and $\delta \in (0,1)$ to obtain the following form of the Mahalanobis distance.

Definition 4. *The regularized MWCD-Mahalanobis distance between an observation Z and the k-th group of the data is defined as*

$$(\bar{X}^*_{k,MWCD} - Z)^T (S^*_{MWCD})^{-1} (\bar{X}^*_{k,MWCD} - Z). \qquad (26)$$

The values of δ and λ can be found by cross-validation. In an analogous way, the Mahalanobis distance can be formulated as a distance between two groups etc.

Let us now consider data (1) observed in K groups. Let a new regularized robust version of LDA denoted as MWCD-RLDA be defined to assign an observation $Z = (Z_1,\ldots,Z_p)^T$ to group k, if

$$\arg\min_{j=1,\ldots,K} (\bar{X}^*_{j,MWCD} - Z)^T (S^*_{MWCD})^{-1} (\bar{X}^*_{j,MWCD} - Z) \quad (27)$$

is attained for $j = k$.

From the computational point of view, (27) should be always replaced by a rule based on regularized linear discriminant scores. Thus, Z is classified to the group k, if $l_k^* > l_j^*$ for every $j \neq k$, where

$$\begin{aligned} l_k^* &= (\bar{X}^*_{k,MWCD})^T (S^*_{MWCD})^{-1} Z + \log p_k - \\ &- \frac{1}{2} (\bar{X}^*_{k,MWCD})^T (S^*_{MWCD})^{-1} \bar{X}^*_{k,MWCD}, \end{aligned} \qquad (28)$$

and p_k is a prior probability of observing an observation from the k-th group for $k = 1,\ldots,K$. The situation with $l_k^* = l_{k'}^*$ for $k' \neq k$ does not need a separate treatment, because it occurs with a zero probability for data coming from a continuous distribution.

Algorithm 4. *MWCD-RLDA.*

Step 1 *For a given $\delta \in (0,1)$, compute the matrix*

$$A = [\bar{X}^*_{1,MWCD} - Z, \ldots, \bar{X}^*_{K,MWCD} - Z] \qquad (29)$$

of size $p \times K$.

Step 2 *Compute S^*_{MWCD} according to (23) with a fixed $\lambda \in (0,1)$.*

Step 3 *Compute and store the eigenvalues of S^*_{MWCD} in the diagonal matrix D_*, and compute and store the corresponding eigenvectors of S^*_{MWCD} in the orthogonal matrix Q_*.*

Step 4 *Compute the matrix*

$$B = D_*^{-1/2} Q_*^T A \qquad (30)$$

and assign Z to group k if the column of B with largest Euclidean norm is the k-th column.

Step 5 *Repeat steps 1 to 5 with different values of δ and λ and find the classification rule with the best classification performance.*

The main computational costs are in step 3; the eigendecomposition costs about $9 \cdot p^3$ floating point operations. Note that we need not (and should never) compute the inverse of S^*_{MWCD}, thus avoiding additional computations of the Mahalanobis distance, which is expensive of order p^3 and numerically rather unstable. The group assignment (27) is done by using

$$\begin{aligned} & (\bar{X}^*_{k,MWCD} - Z)^T (S^*_{MWCD})^{-1} (\bar{X}^*_{k,MWCD} - Z) \\ = \; & (\bar{X}^*_{k,MWCD} - Z)^T Q_* D_*^{-1} Q_*^T (\bar{X}^*_{k,MWCD} - Z) \\ = \; & \|D_*^{-1/2} Q_*^T (\bar{X}_k - Z)\|^2. \end{aligned} \qquad (31)$$

An algorithm for MWCD-RLDA using the ridge regularization with $T = \mathscr{I}_p$ can be obtained as a straightforward modification of Algorithm 2.

6 Metabolomic Profiles Example

We present an example on real molecular genetic data sets in order to illustrate the behavior of the newly proposed classification methods.

Classification method	Classification accuracy
LDA	Infeasible
PAM	0.98
SCRDA	1.00
MWCD-RLDA	1.00
SVM	1.00
Classification tree	0.97
Self-organizing map	0.93
PCA $\Longrightarrow$ LDA	0.90
PCA $\Longrightarrow$ SCRDA	0.92
PCA $\Longrightarrow$ PAM	0.81
PCA $\Longrightarrow$ MWCD-RLDA	0.92

Table 1: Metabolomic profiles example. Classification accuracy of various classification methods. PCA uses 20 principal components.

A prostate cancer metabolomic data set [23] is analyzed, which contains $p = 518$ metabolites measured over two groups of patients, namely those with a benign prostate cancer (16 patients) and 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.

Classification method	Data	
	Raw	Contam.
LDA	Infeasible	Infeasible
PAM [24]	0.96	0.91
SCRDA [6]	1.00	0.95
MWCD-RLDA	1.00	1.00
SVM	1.00	0.98
Classification tree	0.55	0.51
Self-organizing map	0.93	0.90

Table 2: Keystroke dynamics example. Classification accuracy of various classification methods for raw data and for data contaminated by outliers.

We use various classification methods with default settings in *R* software to distinguish between the 2 groups of observations. The new method MWCD-RLDA is used with linearly decreasing weights. The classification performance is measured by the accuracy, i.e. number of correctly classified cases divided by the total number of cases [22].

The results of various classification methods are overviewed in Table 1. The newly proposed method MWCD-RLDA turns out to perform reliably. We do not find major differences in the classification performance of various regularized versions of LDA. At the same time, MWCD-RLDA has a good classification ability if applied on principal components.

Our consequent analysis of principal components of the data brings other arguments in favor of the regularization appraoches used in this paper. 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. In such situations, if there seems no clear interpretation of the principal components, the preferable type of regularization seems to be the Tikhonov regularization, i.e. the regularization in the L_2 norm, which is used throughout this paper.

7 Keystroke Dynamics Example

We analyze a real data subset from a larger keystroke dynamics study aimed at proposing a fast online software system for person authentication. Here, we work with data measured on $K = 2$ probands, who were asked to write the word *kladruby* 5-times each. The authentication is a classification task to $K = 2$ groups with $n = 10$, where there are $p = 15$ ($p > n$) variables including 8 keystroke durations and 7 keystroke latencies measured in milliseconds. Our detailed analysis goes beyond the results of [21], where 32 probands were considered.

Table 2 gives classification accuracy of various methods obtained on raw data. The best results are obtained with MWCD-RLDA, SCRDA, and support vector machines (SVM).

Further, we investigate the classification performance of various methods on data artificially contaminated by severe outliers. Proband-independent noise is generated from normal distribution $\mathsf{N}(0, \sigma^2 = 225)$. Its absolute values are added to all measurements for each proband and classification rules are learned over this contaminated data set. Only MWCD-RLDA turns out to be resistant to outliers, while regularized versions of LDA seem not to suffer so strongly from their presence, compared to methods constructed without any regularization tools. It is remarkable that Prediction Analysis for Microarrays (PAM) turns out to be heavily influenced by outliers, because it has been interpreted as a denoised version of diagonalized LDA [24].

Finally, we investigate the structure of the covariance matrix S. The structure of S will be investigated by analyzing principal components of the data.

Table 3 evaluates the contribution of the 9 principal components to the separation between both groups as the classification accuracy of LDA based on individual components and pairs of components. The first principal components turn out to be extremely influenced by outliers, while their discrimination ability is larger than that of the last components. The first two principal components are shown in Figure 1. The last components contribute a little to the variability of the data and their contribution to the separation of the groups is also negligible.

It is tempting to perform LDA or regularized LDA only on the major principal components. Actually, principal components corresponding to zero eigenvalues have been rigorously proven to have no discrimination effect for $n < p$ only recently [4]. Nevertheless, our example shows the destructive effect of outliers on S and therefore on principal component analysis (PCA). To explain the structure of S in more details, let us mention that the data matrix X has 10 singular values, the between-group covariance matrix

$$B = \frac{1}{K-1} \sum_{k=1}^{K} n_j (X_j - \bar{X})(X_j - \bar{X})^T \qquad (32)$$

has 1 positive eigenvalue and the within-group covariance matrix

$$W = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (X_{ki} - \bar{X_k})(X_{ki} - \bar{X_l})^T \qquad (33)$$

has 10 positive eigenvalues, while the covariance matrix S of size 15×15 has 9 positive eigenvalues and 6 eigenvalues exactly equal to 0. At any case, approaches avoiding PCA are preferable, because PCA suffers heavily from outliers and its regularized version using the Tikhonov regularization can be easily shown to be exactly equal to standard PCA.

8 Conclusions

Standard data mining procedures are very sensitive to the presence of outlying measurements in the data, which

Principal component	Classification accuracy Indiv.	Pair	Cumulative
1	0.5	0.9	0.6
2	0.9		0.9
3	1.0	0.8	1.0
4	0.9		0.9
5	0.9	0.7	1.0
6	1.0		1.0
7	1.0	0.5	1.0
8	1.0		1.0
9	0.6		1.0

Table 3: Keystroke dynamics example: study of principal components of the data. Classification accuracy of LDA as the percentage of correctly identified samples based on individual principal components and pairs of two principal components, together with the cumulative contribution of components $1,\ldots,r$ for $r = 1,\ldots,9$.

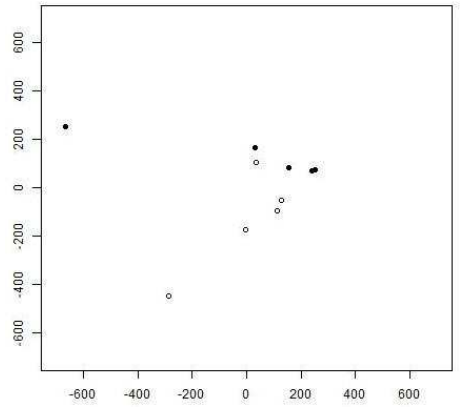

Figure 1: Keystroke dynamics example: study of principal components of the data. The 1st and 2nd principal component of the data in group 1 (bullets) and group 2 (empty circles).

makes robustness to their presence to be an important requirement. This paper proposes several robust versions of standard as well as regularized classification procedures and algorithms for their computation.

Mining high-dimensional data with a large number of variables p becomes an important task in a variety of applications [19]. Numerous new methods have been proposed in the literature for the analysis of big data, particularly with a large number of observations n. On the other hand, smaller attention attention has been paid to methods for data mining and multivariate statistics for this $n \ll p$ situation. Some of standard methods of data mining or multivariate statistics are computationally infeasible for high-dimensional data, others suffer from a numerical instability and lack of robustness to noise or outlying measurements. Sometimes it is claimed that some of the regularized methods have been empirically observed to possess reasonable robustness properties [14] although robustness properties of regularized methods have never been systematically investigated.

We propose new robust versions of the Mahalanobis distance between high-dimensional observations with the number of variables exceeding the number of variables. Particularly, we used M-estimation and the idea of implicit weighting to obtain new robust measures of distance between multivariate observations. This has allowed us to formulate new methods of classification analysis together with algorithms for their computation suitable for $n \ll p$. The new methods combine robustness with regularization in the multivariate model in a unique way. Because robust methods are computationally intensive themselves, which requires a sophisticated approach to regularization. Several new algorithms for shrinkage LDA are proposed, exploiting a shrinkage covariance matrix estimator towards a regular target matrix (unit matrix).

The analysis of two real data sets reveals that the classification performance of the newly proposed method MWCD-RLDA seems to be comparable to available classification procedures, while it turns out to be superior for data contaminated by noise. Apart from a good classification performance, it is important to overview other suitable properties of the newly proposed approach. Its formula is elegant due to using the same kind of regularization in the L_2 norm for both means and covariance matrices. The implicit weights assigned to individual observations allow a clear interpretation. At the same time, implicit weights are known to ensure a high breakdown point with respect to a larger percentage of outliers in a variety of other situations [26, 17, 16]. MWCD-RLDA can be computed with an efficient algorithm even for $n \ll p$, while even faster algorithms can be proposed tailor-made for specific choices of the target matrix T. Finally, our methods search for optimal values of regularization parameters directly in the classification task, while existing approaches to regularized covariance matrix estimation exploit an analytical expression for the parameters, which are however not optimal for the classification task [20].

Alternative regularization approaches based on the L_1-regularization or variable selection were described by [7]. To give an example of a regularization approach combining various regularization types together, let us mention the ideas of [6] using L_2-regularization on covariance matrices and L_1-regularization on means. Such approaches may be less suitable for data without several clearly strong principal components contributing to explaining a large portion of the variability of the data.

Other possible extensions of our ideas, which exceed the scope of this paper, include the possibility to propose other regularized robust data mining procedures, e.g. classification trees, entropy estimation, k-means clustering, or dimensionality reduction by the minimum redundancy maximum relevance (MRMR) algorithm. Some of them may directly exploit the regularized robust Mahalanobis distance of Section 5. Our future work is intended to investigate theoretical connections between regularized ver-

sions of LDA and PCA and to study alternative regularization approaches in the context of LDA.

Acknowledgments

The work was financially supported by the Neuron Foundation for Supporting Science. The work of J. Kalina was supported by the grant GA13-17187S of the Czech Science Foundation. The work of J. Duintjer Tebbens was supported by the grant GA13-06684S of the Czech Science Foundation. The work of A. Schlenker was supported by the specific research projects 264513 of Charles University in Prague and 494/2013 of CESNET Development Fund.

References

[1] Alfons, A., Croux, C., Gelper, S.: Sparse least trimmed squares regression for analyzing high-dimensional large data sets. Annals of Applied Statistics **7** (2013) 226–248

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

[3] Čížek, P.: Semiparametrically weighted robust estimation of regression models. Computational Statistics & Data Analysis **55** (2011) 774–788

[4] Duintjer Tebbens, J., Schlesinger, P. Improving implementation of linear discriminant analysis for the high dimension/small sample size problem. Computational Statistics & Data Analysis **52** (2007) 423–437

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

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

[7] Hastie, T., Tibshirani, R., Friedman, J. The elements of statistical learning. Springer, New York, 2001

[8] Howland, P., Jeon, M., Park H.: Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. SIAM Journal on Matrix Analysis and Applications **25** (2003) 165–179

[9] Huber, P.: Robust estimation of a location parameter. Annals of Mathematical Statistics **35** (1964) 73–101

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

[11] Jurečková, J., Kalina, J.: Nonparametric multivariate rank tests and their unbiasedness. Bernoulli **18** (2012) 229–251

[12] Jurečková, J., Picek, J.: Robust statistical methods with R. Chapman & Hall/CRC, Boca Raton, 2006

[13] Jurečková, J., Sen, P.K.: Robust statistical procedures: Asymptotics and inter-relations. Wiley, New York, 1996

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

[15] Kalina, J.: Machine learning and robustness aspects. Serbian Journal of Management **9** (2014) 131–144

[16] Kalina, J.: Highly robust statistical methods in medical image analysis. Biocybernetics and Biomedical Engineering **32** (2012) 3–16

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

[18] Marazzi, A.: Algorithms, routines, and S-functions for robust statistics. Chapman & Hall/CRC, Wadsworth, 1993

[19] Martinez, W.L., Martinez, A.R., Solka, J.L.: Exploratory data analysis with MATLAB. 2nd edn. Chapman & Hall/CRC, London, 2011

[20] Pourahmadi, M.: High-dimensional covariance estimation. Wiley, Hoboken, 2013

[21] Schlenker, A., Bohunčák, A., Kalina J. (2014): Pilot study of authentication methods based on keystroke dynamics for multi-factor authentication in biomedicine. Submitted.

[22] Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing and Management **45** (2009) 427–437

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

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

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

[26] Víšek, J.Á.: Consistency of the least weighted squares under heteroscedasticity. Kybernetika **47** (2011) 179–206

[27] Xanthopoulos, P., Pardalos, P.M., Trafalis, T.B.: Robust data mining. Springer, New York, 2013