Robust Regularized Discriminant Analysis Based on Implicit Weighting

Kalina, Jan
2016

Institute of Computer Science
The Czech Academy of Sciences

Robust Regularized Discriminant Analysis Based on Implicit Weighting

Jan Kalina and Jaroslav Hlinka

Technical report No. V-1241

December 2016

Pod Vodárenskou věží 2, 182 07 Prague 8, phone: +420 266 053 099, fax: +420 286 585 789,
e-mail: kalina@cs.cas.cz, hlinka@cs.cas.cz

Institute of Computer Science
The Czech Academy of Sciences

Robust Regularized Discriminant Analysis Based on Implicit Weighting

Jan Kalina and Jaroslav Hlinka[1]

Technical report No. V-1241

December 2016

Abstract:

In bioinformatics, regularized linear discriminant analysis is commonly used as a tool for supervised classification problems tailormade for high-dimensional data with the number of variables exceeding the number of observations. However, its various available versions are too vulnerable to the presence of outlying measurements in the data. In this paper, we exploit principles of robust statistics to propose new versions of regularized linear discriminant analysis suitable for highdimensional data contaminated by (more or less) severe outliers. The work exploits a regularized version of the minimum weighted covariance determinant estimator, which is one of highly robust estimators of multivariate location and scatter. The performance of the novel classification methods is illustrated on real data sets with a detailed analysis of data from brain activity research.

Keywords:
high-dimensional data, classification analysis, robustness, outliers, regularization

[1]Institute of Computer Science, The Czech Academy of Sciences, Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic, E-mail: kalina@cs.cas.cz, hlinka@cs.cas.cz.

Robust Regularized Discriminant Analysis Based on Implicit Weighting

Jan Kalina[1,2] and Jaroslav Hlinka[1,2]

[1] Institute of Computer Science of the Czech Academy of Sciences, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic
[2] National Institute of Mental Health, Klecany, Czech Republic

Abstract. In bioinformatics, regularized linear discriminant analysis is commonly used as a tool for supervised classification problems tailor-made for high-dimensional data with the number of variables exceeding the number of observations. However, its various available versions are too vulnerable to the presence of outlying measurements in the data. In this paper, we exploit principles of robust statistics to propose new versions of regularized linear discriminant analysis suitable for high-dimensional data contaminated by (more or less) severe outliers. The work exploits a regularized version of the minimum weighted covariance determinant estimator, which is one of highly robust estimators of multivariate location and scatter. The performance of the novel classification methods is illustrated on real data sets with a detailed analysis of data from brain activity research.

Keywords: high-dimensional data, classification analysis, robustness, outliers, regularization

1 Introduction

In bioinformatics, a common data analysis task is to learn a classification rule over high-dimensional data, i.e. data with the number of variables p exceeding the number of observations n ($n < p$ or even $n \ll p$) [8, 2]. Thus, supervised classification methods (classifiers) represent important tools for the analysis of data observed in K different samples (groups) as

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

while we assume $p > K \geq 2$ and denote $n = \sum_{k=1}^{K} n_k$. 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 [9, 33].

The linear discriminant analysis (LDA) is well known to be too sensitive to the presence of outlying values in the data because it exploits the classical (non-robust) estimates in the form of means and empirical covariance matrix. As an alternative, robust classification methods have been proposed which are

resistant to the presence of outliers [4, 16, 35]. Highly robust methods are defined as methods with a high breakdown point, which measures the sensitivity of an estimator 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 of the common scatter matrix are not shifted to infinity [6]. Nevertheless, robust classification methods are computationally feasible only for $n > p$ with a sufficiently small p.

In this paper, we propose new classification methods for high-dimensional data exploiting principles of robust statistics in a unique combination with the (Tikhonov) regularization. Section 2 recalls various existing approaches to regularized linear discriminant analysis for $n \ll p$. Section 3 recalls the minimum weighted covariance determinant estimator, which is one of highly robust estimators of multivariate location and scatter, and proposes its regularized version. Section 4 proposes four new robust regularized classification methods for high-dimensional data, exploiting the tools of Section 3. The following Section 5 illustrates the performance of the novel methods on several real data sets, while the largest attention is paid to data from brain activity research. A discussion follows in Section 6, where also the good comprehensibility of the newly proposed procedures is brought to attention, and finally Section 7 concludes the paper.

2 Regularized Linear Discriminant Analysis

Various available versions of a regularized LDA can be characterized as modifications of the standard LDA for the context of high-dimensional data. While these methods have found their applications in bioinformatics (e.g. [10, 34, 23]), they remain to be vulnerable to outliers because of the non-robustness of the empirical covariance matrix as well as means of each group.

Regularized LDA assumes a common covariance matrix Σ for each group. If $n < p$ or even $n \ll p$, the pooled estimator of Σ denoted by S is singular. Let us denote the mean of the observed values in the k-th group ($k = 1, \ldots, K$) by $\bar{X}_k$. Perhaps the simplest and most habitually used version 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

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

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

$$S^* = (1 - \lambda) S + \lambda T \tag{3}$$

for $\lambda \in (0, 1)$ denotes a regularized estimator of Σ and the target matrix T is a given symmetric positive definite matrix of size $p \times p$. Its most common choices include the identity matrix $\mathcal{I}_p$ or a diagonal (non-identity) matrix $T = \bar{s}\,\mathcal{I}_p$,

where $\bar{s} = \sum_{i=1}^{p} S_{ii}/p$. The regularized matrix (3) is guaranteed to be regular and positive definite even for $n \ll p$. A suitable value of λ is usually found by a cross validation.

Another important example of a regularized LDA is the shrunken centroid regularized discriminant analysis (SCRDA) [10], which performs also a regularization on the mean of each group, namely by shrinking each of the means towards the pooled mean in the L_1-norm.

Concerning the properties of regularized versions of LDA, it would be very difficult to assess them rigorously. Instead, they were rather investigated only by means of numerical simulations [34, 10]. Basically, regularized LDA methods remain to be non-robust to the presence of severe outliers, although the regularization leads to their local insensitivity (robustness) to small measurement errors as advocated within the framework of robust optimization [1].

3 Robust Estimation of Multivariate Location and Scatter

This section devoted to robust estimation of multivariate location and scatter starts by recalling the highly robust minimum weighted covariance determinant estimator in Section 3.1. We will use the regularized M-estimator of multivariate data proposed in [3] to obtain a reliable initial estimator of the scatter matrix. This allows to define a regularized version of the minimum weighted covariance determinant estimator in Section 3.2. For the iterative computation of this scatter matrix estimator, an initial estimate will be necessary and we recommend to use Chen's estimator [3] for this purpose.

3.1 Minimum Weighted Covariance Determinant Estimator

The Minimum Covariance Determinant (MCD) and the Minimum Weighted Covariance Determinant (MWCD) estimators are highly robust affine-equivariant estimators of multivariate location and scatter. Let us consider a single sample of independent identically distributed p-variate random variables $X_1, \ldots, X_n$ (i.e. $K = 1$). The estimators are formulated for multivariate data coming from a unimodal elliptically symmetric distribution with a location parameter $\mu \in \mathbb{R}^p$ and a scatter matrix $\Sigma \in \mathbb{R}^{p \times p}$ (cf. [17]). The scatter matrix is a more general concept compared to the covariance matrix, which must not necessarily exist. Nevertheless, the two concepts are identical for Gaussian data. Standard estimators of μ and Σ are highly vulnerable to the presence of outliers in the data. On the other hand, the MCD and MWCD estimators, which will be now recalled, are more suitable for severely contaminated data compared to classical estimates and also compared to multivariate M-estimators [26, 36].

The MCD estimator [31] is computed as the classical mean and the empirical covariance matrix taking into account however only the optimal subset of h observations, which yields the minimum determinant of the empirical covariance matrix over all such possible subsets of h observations. This corresponds

to assigning weights equal to 1 or 0 to the observations, while the number of ones is a fixed value equal to h which must be specified by the user prior to the computations. Properties of the estimator were overviewed in [17].

The MWCD estimator [29] represents a generalization of MCD allowing to consider non-negative (possibly continuous) weights $w_1, \ldots, w_n$ to be assigned to the observations. The user specifies magnitudes of weights prior to computing the estimator but the weights themselves are assigned to individual observations only after a permutation, which is determined only during the computation of the estimator. Such implicitly given weights are based on the idea to down-weight outliers and to increase the influence of the majority of "good data".

Properties of the MWCD estimator were derived in [29] including the efficiency, Fisher consistency or influence function. The method attains the maximal breakdown point which is possible for an affine-equivariant estimator [25]; this is true if the outliers obtain weights exactly equal to 0 and the data are assumed in general position [30]. An approximative algorithm for computing the MWCD estimator may be obtained as a generalization of the MCD algorithm [31]. The advantage of the weighting scheme is its ability to reduce the local sensitivity compared to the MCD estimator; this is analogous to the experience with implicitly weighted methods in robust regression [19].

3.2 Regularized MWCD Estimator

We consider again the data in one group as in Section 3.1. Because the MWCD estimator cannot be computed for $n < p$, we define its regularized version which is computationally feasible also for $n \ll p$.

Let us first recall the work of Chen et al. [3], who proposed a regularized M-estimator of the scatter matrix of multivariate data based on a popular (Huber-type [15]) M-estimator of [36]. However, M-estimators of parameters in the multivariate model do not possess a high breakdown point [37]. In addition, the estimation does not yield any corresponding estimator of the mean.

Our proposal of a regularized MWCD estimator presented as Algorithm 1 exploits Chen's regularized M-estimator as an initial estimator of the scatter matrix. This depends on the value of a regularization parameter $\rho \in (0, 1)$. If there is no prior idea how to choose its suitable value, a reasonable recommendation is to choose a very small ρ. Nevertheless, its suitable value may be easily found by cross validation in specific tasks. This will be the case of classification problems in Section 4. In step 2, choosing robust rather than standard initial estimators is a common approach in a variety of iterative robust estimators [18].

4 Robust Classification

Four novel robust versions of regularized LDA will be proposed in this section, together with algorithms for their efficient computation. The approach is suitable for multivariate data coming from a unimodal elliptically symmetric distribution as explained in Section 3.1.

Algorithm 1 Regularized MWCD estimator.

Input: p-dimensional observations $X_1, \ldots, X_n$, weights $w_1, \ldots, w_n$, positive definite symmetric matrix $T \in \mathbb{R}^{p \times p}$.

Output: $\bar{X}_{MWCD}$, S_{MWCD}.

1: **for** $i = 1$ to 10 000 **do**
2: Randomly select an initial set of $n/2$ observations. Compute Chen's estimator, which will be denoted by B (for the mean) and C (for the scatter matrix).
3: $j := 1$
4: $L_{ij} := +\infty$
5: **repeat**
6: Compute

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

Sort these values in ascending order and assign corresponding ranks to individual observations. This determines a permutation $\pi(1), \ldots, \pi(n)$ of the indexes $1, 2, ..., n$, which fulfills

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

7: Assign the weights to each observation according to its rank evaluated in the previous step. In this way, e.g. the observation $X_{\pi(1)}$ obtains the weight w_1.
8: Compute the weighted mean and weighted empirical covariance matrix S_w using these weights.
9: $j := j + 1$
10: $C := (1 - \lambda) S_w + \lambda T$
11: $L_{ij} := \det(C)$
12: **until** $L_{ij} \geq L_{i,j-1}$
13: **end for**
14: Determine the set of weights $\tilde{w}_1, \ldots, \tilde{w}_n$ minimizing L_{ij} over all considered i and j.
15: $\bar{X}_{MWCD} := \sum_{i=1}^{n} \tilde{w}_i X_i$
16:

$$S_{MWCD} := \sum_{i=1}^{n} \tilde{w}_i (X_i - \bar{X}_{MWCD})(X_i - \bar{X}_{MWCD})^T \tag{6}$$

4.1 MWCD-LDA*

We will propose a novel classification method denoted as MWCD-LDA*. We assume data (1) in K groups, while all the groups have the common scatter matrix $\Sigma \in \mathbb{R}^{p \times p}$. Our approach is based on estimating Σ as well as the means of the groups by the regularized MWCD estimator of Section 3.2.

Concerning the estimator of the scatter matrix of data (1), we must be aware that Σ does not play the role of the covariance (nor scatter) matrix over all data but rather to the scatter matrix common for each of the groups. Therefore, we need to adapt Algorithm 1 to the situation with K groups. Algorithm 1 in step 8 considers the empirical weighted covariance matrix, which has to be replaced

for the data in K groups by

$$S^*_{MWCD} = \left(S^*_{ij}\right)_{i,j=1}^{p}, \tag{7}$$

where

$$S^*_{ij} = \sum_{k=1}^{K} \sum_{l=1}^{n_k} w_{kl} \left(X_{kli} - \bar{X}^k_{iw}\right) \left(X_{klj} - \bar{X}^k_{jw}\right). \tag{8}$$

Here, the summation over l runs over all observations $l = 1, \ldots, n$, which belong to the k-th group,

$$X_{kl} = (X_{kl1}, \ldots, X_{klp})^T \quad \text{for } k = 1, \ldots, K,\ l = 1, \ldots, n_k, \tag{9}$$

the weights are denoted as $w_{11}, \ldots, w_{1n_1}, \ldots, w_{K1}, \ldots, w_{Kn_K}$ and $\bar{X}^k_{iw}$ denotes the weighted mean of the k-th group with these weights. With this difference, Algorithm 1 yields S^*_{MWCD} as the estimator of Σ, which exploits the optimal weights denoted as $\tilde{w}_1, \ldots, \tilde{w}_n$.

The resulting weights will be now used also to define regularized MWCD-means of each group, which have the form

$$\bar{X}_{MWCD} = \sum_{l=1}^{n} \tilde{w}_l X_l \tag{10}$$

and

$$\bar{X}^k_{MWCD} = \sum_{l \in group\ k} \tilde{w}_l X_l, \quad k = 1, \ldots, K. \tag{11}$$

Now we come back to the original classification problem. Formally, MWCD-LDA* will assign a new observation $Z = (Z_1, \ldots, Z_p)^T$ to group k, if $\tilde{\ell}_k > \tilde{\ell}_j$ for every $j \neq k$, where

$$\begin{aligned} \tilde{\ell}_k &= (\bar{X}_{k,MWCD})^T \left(S^*_{MWCD}\right)^{-1} Z - \\ &- \frac{1}{2} (\bar{X}_{k,MWCD})^T \left(S^*_{MWCD}\right)^{-1} \bar{X}_{k,MWCD} + \log \pi_k. \end{aligned} \tag{12}$$

Equivalently, the classification rule can be also expressed exploiting a robust regularized Mahalanobis distance. In this respect, an observation Z is assigned to group k if

$$(\bar{X}_{j,MWCD} - Z)^T \left(S^*_{MWCD}\right)^{-1} (\bar{X}_{j,MWCD} - Z) + \log \pi_j \tag{13}$$

reaches its minimum over all $j = 1, \ldots, K$ exactly for k.

As both (12) and the group assignment (13) are rather obscure from the computational point of view, we recommend to avoid the expensive and numerically unstable computation of the Mahalanobis distance by solving a set of linear equations within Algorithm 2 for the task to classify an observation $Z = (Z_1, \ldots, Z_p)^T$.

The approach of Algorithm 2 is based on the eigendecomposition of S^*_{MWCD}. Its inversion is replaced is replaced by (16), which is based on expressing

$$\begin{aligned}&(\bar{X}_{k,MWCD} - Z)^T (S^*_{MWCD})^{-1} (\bar{X}_{k,MWCD} - Z)\\ &= (\bar{X}_{k,MWCD} - Z)^T Q D^{-1} Q^T (\bar{X}_{k,MWCD} - Z)\\ &= \|D^{-1/2} Q^T (\bar{X}_k^{k,MWCD} - Z)\|^2. \end{aligned} \tag{14}$$

While S^*_{MWCD} depends on the parameter λ, its suitable value will be found by a cross validation in the form of a grid search over all possible values of $\lambda \in (0, 1)$.

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

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

Input: Data (1), $Z \in \mathbb{R}^p$, weights $w_1, \ldots, w_n$, positive definite symmetric matrix $T \in \mathbb{R}^{p \times p}$.

Output: Assignment of Z to one of the groups $1, \ldots, K$.

1: **for** $i = 1$ to 100 **do**
2: $\lambda := i/100$
3: Compute S^*_{MWCD} and $\bar{X}^k_{MWCD}$ for $k = 1, \ldots, K$ by a modification of Algorithm 1 using the given T and λ, replacing S_w by (7) with (8).
4: Compute the matrix

$$A = \left(\bar{X}_{1,MWCD} - Z, \ \ldots \ , \bar{X}_{K,MWCD} - Z\right) \tag{15}$$

of size $p \times K$.
5: Compute the eigendecomposition $S^*_{MWCD} = QDQ^T$.
6: Compute $B = D^{-1/2} Q^T A$.
7: Assign Z to group k, if

$$k = \underset{l=1,\ldots,K}{\arg\max} \left\{ ||B_l||^2 + \log \pi_l \right\}, \tag{16}$$

where $||B_l||^2$ is the Euclidean norm of the l-th column of B.
8: **end for**
9: Determine the value of λ yielding the best classification performance and carry out steps 3 to 7 with it to find the final classification decision.

4.2 L_1-SCRRDA

Further, we propose to accompany regularizing the scatter matrix of Section 4.1 by regularizing the mean of each of the groups. The novel method represents a robustification of the SCRDA of [10] and will be denoted as L_1-SCRRDA, which abbreviates a shrunken centroid robust regularized discriminant analysis

with the means regularized in the L_1-norm. We can also perceive the method to be based on an L_1-regularized robust Mahalanobis distance.

Let us consider the mean of the k-th group to be estimated by

$$\begin{aligned}\bar{X}^{(1)}_{k,MWCD} &= \mathsf{sgn}(\bar{X}_{k,MWCD}) \left(|\bar{X}_{k,MWCD}| - \Delta\right)_+ \\ &= \mathsf{sgn}(\bar{X}_{k,MWCD}) \max\left\{|\bar{X}_{k,MWCD}| - \Delta, 0\right\}, \end{aligned} \tag{17}$$

where $\Delta \in \mathbb{R}^p$ and $(x)_+$ denotes the positive part of $x \in \mathbb{R}^p$. In other words, the MWCD-mean is shrunken towards zero in the L_1-norm, which can be interpreted as a regularized (biased) version of the MWCD-mean.

The method L_1-SCRRDA exploits the matrix S^*_{MWCD} as in Section 4.1. It assigns an observation $Z = (Z_1, \ldots, Z_p)^T$to group k, if $\ell^{(1)}_k > \ell^{(1)}_j$ for every $j \neq k$, where

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

Within the classification method, suitable values of both regularization parameters λ and Δ can be found by cross validation as in Algorithm 2, which can be adapted to the context of L_1-SCRRDA.

Remark 1. L_1-SCRRDA distinguished between two groups of variables as follows, using the notation $\bar{X}_{jk,MWCD}$ and $\bar{X}^{(1)}_{jk,MWCD}$ to evaluate the j-th coordinate of $\bar{X}_{k,MWCD}$ and $\bar{X}^{(1)}_{k,MWCD}$, respectively.

1. Major (more relevant) variables fulfilling $|\bar{X}_{jk,MWCD}| > \Delta$ for at least one k. Their values of $\bar{X}^{(1)}_{jk,MWCD}$ for these k are obtained by shrinking $\bar{X}_{jk,MWCD}$ towards zero by the amount of exactly Δ.
2. Minor (less relevant) variables fulfilling $|\bar{X}^{(1)}_{jk,MWCD}| \leq \Delta$ for each k. Their values of $\bar{X}^{(1)}_{jk,MWCD}$ are equal to 0.

Remark 2. L_1-regularization is generally understood to introduce sparseness and reduce the dimensionality. However, the universality of this property is rather a "golden legend" and holds e.g. in linear regression (lasso estimator) but not for (any) regularized LDA, although there have been misleading statements on variable selection and sparseness also in this context (cf. [34, 10]). In the light of Remark 1, we stress that L_1-SCRRDA remains to depend also on the major variables, because $\bar{X}_{k,MWCD} - Z$ is the same for all k, but the variable influences the linear discriminant score through the scatter matrix. Also the computational complexity of L_1-SCRRDA is not reduced compared to a method regularizing in the L_2-norm instead, which will be proposed in the next subsection.

4.3 L_2-SCRRDA

An alternative regularized robust version of LDA denoted as L_2-SCRRDA is proposed, which combines the scatter matrix estimation of Section 4.1 with

shrinking the means towards the pooled mean (across groups) in the L_2-norm. Thus, we denote this version of robust SCRDA as L_2-SCRRDA.

The pooled scatter matrix (across groups) is estimated again by S^*_{MWCD}. The classical mean of the k-th group is replaced by the MWCD-mean shrunken towards the overall MWCD-mean across groups $\bar{X}_{MWCD}$, i.e. we consider

$$\bar{X}^{(2)}_{k,MWCD} = \delta \bar{X}_{k,MWCD} + (1-\delta)\bar{X}_{MWCD} \tag{19}$$

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

The method L_2-SCRRDA assigns an observation Z to group k, if $\ell^{(2)}_k > \ell^{(2)}_j$ for every $j \neq k$, where

$$\begin{aligned} \ell^{(2)}_k = {} & (\bar{X}^{(2)}_{k,MWCD})^T (S^*_{MWCD})^{-1} Z - \\ & - \frac{1}{2} (\bar{X}^{(2)}_{k,MWCD})^T (S^*_{MWCD})^{-1} \bar{X}^{(2)}_{k,MWCD} + \log \pi_k. \end{aligned} \tag{20}$$

Suitable values of parameters λ and δ can be found again by cross validation. The method may be preferable to L_1-SCRRDA 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. The shrinkage in (19) performed in the L_2-norm is analogous to shrinking estimates of parameters in ridge regression.

Algorithm 3 M-LDA* based on the Cholesky decomposition.

Input: Data (1), $Z \in \mathbb{R}^p$.
Output: Assignment of Z to one of the groups $1, \ldots, K$.

1: Compute Huber's estimators $\bar{X}^1_M, \ldots, \bar{X}^K_M$.
2: Compute the matrix

$$A = \left(\bar{X}^1_M - Z, \ \ldots \ , \bar{X}^K_M - Z \right) \tag{21}$$

of size $p \times K$.
3: **for** $i = 1$ to 100 **do**
4: $\rho := i/100$
5: Compute $S^*_{M,\rho}$ (Section 4.4).
6: Compute the Cholesky decomposition $S^*_{M,\rho} = L_* L_*^T$, where L_* is a (regular) lower triangular matrix.
7: Compute $B = L_*^{-T} A$.
8: Assign Z to group k, if

$$k = \underset{l=1,\ldots,K}{\arg\max} \left\{ ||B_l||^2 + \log \pi_l \right\}, \tag{22}$$

where $||B_l||^2$ is the Euclidean norm of the l-th column of B.
9: **end for**
10: Determine the value of ρ yielding the best classification performance and carry out steps 4 to 8 with them to find the final classification decision.

4.4 M-LDA*

We can also define a version of robust LDA based on M-estimation. Assuming again the data (1) as denoted in Section 1, the M-estimator of the mean of the k-th group denoted as $\bar{X}_M^k$ for $k = 1, \ldots, K$ will be considered. As in Section 4.1, the Chen's regularized estimator is considered as the estimate of the scatter matrix Σ common for each of the groups. This matrix S_M^* will be rather denoted as $S_{M,\rho}^*$ to stress its dependence on the regularization parameter $\rho \in (0, 1)$.

The robust regularized LDA based on M-estimation, which we denote as M-LDA*, may be performed by Algorithm 3, which is formulated for an observation $Z \in \mathbb{R}^p$ exploiting the Cholesky decomposition of the scatter matrix. Algorithm 3 can be also adapted to be suitable for previously mentioned classifiers (MWCD-LDA*, L_1-SCRRDA and L_2-SCRRDA) in a straightforward way.

5 Examples

5.1 Methods in the Computations

To illustrate the performance of the novel robust regularized versions of LDA, we analyze several different real data sets with $n < p$. Each of the examples learns a classification rule to two groups ($K = 2$).

We performed the computations in R software. Each classification task for each of the data sets is analyzed by means of a 5-fold cross validation. Within such approach, the data set is randomly divided into 5 subsamples of (approximately) equal sizes. Among all possible partitions, we select randomly 100 of them and compute the average Youden's index I as a classification performance measure over them. The averaged values are presented in Table 1. We recall Youden's index to be defined as

$$I = \text{sensitivity} + \text{specificity} - 1, \tag{23}$$

i.e. it fulfils $I \in [-1, 1]$.

Standard classifiers used in the examples include also the lasso-regularized logistic regression denoted as lasso-LR or a support vector machine (SVM) with a radial basis function kernel. Concerning the choice of parameters of individual classifiers, all regularized versions of LDA use $T = \mathcal{I}_p$. For standard methods, default settings of parameters were used whenever appropriate. Concerning the choice of the implicit weights, the MWCD-LDA* and L_2-SCRRDA use linearly decreasing weights in the following form. Starting with the simple choice

$$w_i = 1 - \frac{i-1}{n}, \quad i = 1, \ldots, n, \tag{24}$$

we standardize them to $\sum_{i=1}^n w_i = 1$ to obtain the final formula for the weights

$$\tilde{w}_i = \frac{2\,(n-i+1)}{n(n+1)}, \quad i = 1, \ldots, n, \tag{25}$$

which are very small for outliers reducing considerably their influence.

To investigate the effect of dimensionality reduction, we also use the principal component analysis (PCA) and a robust Minimum Redundancy Maximum Relevance (MRMR) of [22]. The latter is a robust supervised variable selection method selecting a small set of a fixed (given) number of the most relevant variables while penalizing for redundancy [27].

Table 1. Youden's index (23) as a classification performance measure computed for a 5-fold cross validation study on various real data sets of Section 5. Particularly, data from Section 5.3 are considered not only raw but also after a contamination by normally distributed outliers $\mathsf{N}(0, \sigma^2)$ for different values of σ.

	Section 5.3				Section		
		Contam. for $\sigma =$					
	Raw	0.1	0.2	0.3	5.5	5.6	5.7
n	168				48	42	32
p	4005				38 614	518	15
Regularized versions of LDA							
PAM	0.88	0.81	0.75	0.68	0.85	0.86	0.51
LDA*	1.00	0.95	0.94	0.77	1.00	0.89	0.71
SCRDA	1.00	1.00	1.00	0.99	1.00	0.91	0.80
MWCD-LDA*	1.00	1.00	1.00	1.00	1.00	0.91	0.79
L_2-SCRRDA	1.00	1.00	1.00	1.00	1.00	0.92	0.80
Other classification methods							
SVM	1.00	0.99	0.98	0.96	1.00	0.92	0.85
Classification tree	0.96	0.95	0.91	0.92	0.94	0.84	0.11
Lasso-LR	0.99	1.00	0.97	0.94	0.97	0.87	0.82
Number of principal components	10	10			10	20	4
PCA $\Longrightarrow$ LDA	1.00	0.94	0.93	0.88	0.15	0.70	0.59
PCA $\Longrightarrow$ LDA*	1.00	0.95	0.94	0.89	0.51	0.62	0.59
PCA $\Longrightarrow$ SCRDA	1.00	0.95	0.94	0.89	0.62	0.72	0.59
Number of selected genes	10	10			10	20	4
MRMR $\Longrightarrow$ LDA	1.00	0.94	0.93	0.89	0.90	0.88	0.72
MRMR $\Longrightarrow$ LDA*	1.00	0.96	0.93	0.89	0.96	0.88	0.76
MRMR $\Longrightarrow$ SCRDA	1.00	0.96	0.93	0.89	1.00	0.90	0.76

5.2 Description of the Brain Activity Study

We participate on a neuroscience research investigating the spontaneous activity of various parts of the brain by means of neuroimaging methods, motivated by a distant aim to investigate modifications of the resting-state brain networks in schizophrenic patients. Specific functions of individual parts of the brain have been already rigorously described [7], but spontaneous brain activity and especially connections between pairs of brain parts in the resting state (i.e.

resting-state brain networks) are acknowledged as a hot topic in current neuroscience [14]. We will now describe the whole study leading to acquiring the original real data set of brain scans by functional magnetic resonance imaging (fMRI), while the performance of various classification methods will be compared in the following sections on several different classification tasks.

Our data are measured on $n = 24$ healthy probands (i.e. without a manifested psychiatric disease) participating in the study, who were examined 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. For the sake of the fMRI imaging, the brain is divided to 90 regions and we are interested only in values of correlation coefficients between a pair of brain regions.

The most widely spread method for measuring the (functional) connectivity between a pair of brain regions is the correlation (i.e. Pearson's correlation coefficient) of activity time series derived from these regions by e.g. simple spatial averaging across all the voxels in the brain regions [14]. In this spirit, we consider $p = 90 * 89/2 = 4005$ 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 commonly contaminated by measurement errors as well as outliers [38]. It is also true with our data, which makes the newly proposed robust methods appealing for their analysis.

We consider the classification task to separate between the resting state and (any) movie for healthy individuals in Section 5.3, while a more specific task to classify between the resting state and one particular movie is investigated in Section 5.4.

5.3 Brain Activity: Resting State vs. Movie

Let us now consider the task to learn a classification rule over the training data from Section 5.2 to distinguish between the resting state and a movie. Considering thus all six movies together to belong to one class, this is a classification task to $K = 2$ groups with $p = 4005$ variables. The resting state group contains 24 observations while the group of movies consists of $6 * 24 = 144$ observations.

Several classification methods yield the best result ($I = 1.00$) as shown in Table 1. This is true for standard (non-robust) methods as well as for MWCD-LDA* or L_2-SCRRDA. While the standard LDA is computationally infeasible, SCRDA as one of its available regularized version turns out to perform reliably. There seems no advantage of the robust regularized LDA over non-robust versions, which may be explained by the fact that the raw data are not contaminated by a remarkable percentage of severe outliers. Only PAM turns out to be heavily influenced by them, although it was originally presented as a denoised version of diagonalized LDA [34]. The classification rule of L_1-SCRRDA distinguishes between 81 major variables and the remaining minor variables in this example. An SVM formally gives a perfect classification result, while our critical evaluation of SVM will be presented in Section 6.1.

Additionally, we investigated the effect of dimensionality reduction 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. L_2-SCRRDA has 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.

We additionally performed an artificial contamination of the original data in order to investigate the performance of the novel robust classification methods. Each single measurement for each proband was contaminated by noise, which was generated as proband-independent following normal distribution $\mathsf{N}(0, \sigma^2)$ for various values of σ. The noise was added to all measurements and classification rules are learned over this contaminated data set. 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. Such contamination was repeated 100-times and the classification performance of various methods was evaluated for each case and finally averaged.

The results of the classification performance of various methods on data artificially contaminated by noise, as presented again in Table 1, show an evidence of a reasonable robustness of SCRDA as well the novel methods. The larger 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 methods is not caused by the noise itself, but rather by severe outliers. The robustness of SCRDA to (small) measurement errors has not however been systematically investigated [21] and we think that its ability to outperform the SVM has not been documented sufficiently in the literature. Still, the robustness of the new methods MWCD-LDA* and L_2-SCRRDA is even able to outperform the relatively robust SCRDA.

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. Let us inspect the effect of dimensionality reduction performed by other approaches. If the variables are arranged according to values of the statistic of the two-sample t-test, the best performance with the Youden's index $I = 1.00$ can be obtained only if at least 21 variables are selected. If PAM is used to arrange the variable according to their contribution to separating the two groups, then at least 36 variables are need in order to reach $I = 1.00$.

5.4 Brain Activity: A Closer Look on Individual Movies

In addition, we solve more particular tasks to learn the classification rule allowing to distinguish between the resting state and only one given movie over the training data from Section 5.2. Such six tasks to classify between the resting state and the i-th movie ($i = 1, \ldots, 6$) always deal with $p = 4005$ variables

Table 2. Example of Section 5.4. Youden's index (23) as a classification performance measure computed for a 5-fold cross validation study. PCA is used with a fixed number of 10 principal components. The number of major variables within L_1-SCRRDA is also shown (see Remark 1 in Section 4.2).

	Resting state vs. movie					
Classification method	#1	#2	#3	#4	#5	#6
SVM	1.00	1.00	1.00	1.00	1.00	1.00
L_2-SCRRDA	1.00	1.00	1.00	1.00	1.00	1.00
L_1-SCRRDA	1.00	1.00	1.00	1.00	1.00	1.00
Number of major variables	3	7	2	3	1	1
PCA $\Longrightarrow$ LDA	1.00	1.00	1.00	1.00	1.00	1.00

and 24 observations for each of the two groups. A wide variety of classification procedures is able to reach $I = 1.00$ as shown in Table 2 for selected methods. By a robust variable selection method of [22], we additionally verified that small sets of variables can be found allowing to solve the classification tasks easily.

Finally, we considered other classification tasks with the aim to separate individuals pairs of movies, i.e. classifying between movie #1 and movie #2, between movie #1 and movie #3 etc. The results for each of these 15 tasks again show that $I = 1.00$ can be attained easily, even using a small number of variables. Particularly, we used again the robust variable selection of [22]. The minimal number of variables needed to obtain the $I = 1.00$ result turns out to be greater or equal to 2 and always less or equal to 30. Such small numbers can be explained by a small number of observations in each of the groups.

5.5 Cardiovascular Genetic Study

We illustrate the performance of the novel classifiers on data acquired within the cardiovascular genetic study of the Center of Biomedical Informatics in Prague. The data set was described and analyzed by standard methods in [20]. The aim of the study was 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 gene expressions were measured on $n = 48$ individuals, namely on 24 patients having a cerebrovascular stroke and 24 control persons.

Some methods reach the classification performance $I = 1.00$, which is true for some standard methods including SVM and LDA* and also for the novel methods MWCD-LDA* and L_2-SCRRDA. This can be explained by the very large p, compared to other data sets of this paper.

The dimensionality reduction by means of PCA has drastic consequences, which can be explained by its unsupervised nature ignoring the grouping structure of the data. Indeed, it is the MRMR variable selection which confirms this opinion. MRMR shows that there is a small number of variables responsible for the separation between the two groups and is able to yield much improved results, namely with only 10 most relevant genes allowing to separate both groups with $I = 1.00$.

5.6 Metabolomic Profiles Data

We analyze a publicly available benchmarking data set of prostate cancer metabolomic data [32] of $p = 518$ metabolites measured over two groups of patients, who are either those with a benign prostate cancer (16 patients) or with other cancer types (26 patients).

A detailed analysis of the data reveals that they are not contaminated by severe outliers. Still, MWCD-LDA* and L_2-SCRRDA are able to slightly outperform other regularized versions of LDA. Their result are comparable to the SVM classifier while other classifiers yield inferior results. The MRMR variable selection performs better compared to the unsupervised dimensionality reduction by means of PCA, while there is no 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.

5.7 Keystroke Dynamics Data

The last data set contains data from a biometric authentication study by means of keystroke dynamics, which was described and analyzed in [22]. Our aim is to illustrate the performance of the novel classifiers also on this data set with a small number of variables. A set of probands was asked to type the same short sequence (password) consisting of 8 characters repeatedly. The particular task now is to classify to 2 groups, i.e. to distinguish between two probands exploiting $p = 15$ variables (keystroke durations and latencies in milliseconds) available for each of them.

As Table 1 indicates, the best results are obtained with L_2-SCRRDA, while other robust regularized LDA versions together with SCRDA remain slightly inferior. Again, the SVM classifier is based on a large number of support vectors (≥ 90 % of observations). Dimensionality reduction leads to a loss of information compared to methods using all variables. Our detailed analysis of the data reveals the percentage of severe outliers to be about 10 %. Indeed, if we additionally performed a manual outlier detection and then ignored the outliers from the data set, MWCD-LDA* and L_2-SCRRDA still retain their performance which was not affected by the outliers. On the other hand, the performance of SVM and non-robust versions of LDA is suddenly improved.

6 Discussion

6.1 Advantages of robust regularized classification

Regularized LDA has been advocated for its computational and statistical benefits, which may be revealed not only for $n < p$ but also for $n > p$ with a relatively small n [13]. Regularization is generally believed to ensures a robustness [34, 12], although this does not hold as a universal principle. In the context of regularized LDA, only a robustness with respect to small (local) changes of the measured

data is ensured as it has been observed empirically [1]. Nevertheless, regularized LDA is not robust to more severe noise or outliers, as revealed in our examples.

Appealing properties of regularized LDA have lead us to the idea of joining principles of suitable Tikhonov-type regularization with statistical robustness. Advantages of the newly proposed MWCD-LDA*, L_1-SCRRDA and L_2-SCRRDA include:

- High robustness to outliers thanks to a high breakdown point of MWCD (as a consequence of using the implicit weights similarly to the context of linear regression [19]);
- No assumption on the distribution of the outliers;
- Availability of efficient algorithms based on numerical linear algebra;
- No need for a prior dimensionality reduction;
- Comprehensibility.

While an SVM classifier yields the best classification performance in some of the examples, especially those with a relatively smaller p, we perceive also its drawbacks and try to summarize them.

- It depends on too many support vectors for $n < p$ (more than 90 % of the observations play the role of support vectors in the examples);
- The necessity to optimize its parameters over a sufficiently large number of observations;
- A tendency to overfitting for $n < p$ [11];
- Internal structure not supposed to be understood (black box);
- Non-robustness to outliers.

6.2 Comprehensibility

Comprehensibility represents an important requirement in a wide variety of classification tasks in bioinformatics. Therefore, the discussion of comprehensibility of the newly proposed methods deserves to be presented as a separate subsection.

We consider the classical LDA itself to be comprehensible in the sense that it is based on the Mahalanobis distance of a given (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. means of the corresponding groups and scatter matrix.

The classification rules of MWCD-LDA*, L_1-SCRRDA, L_2-SCRRDA and M-LDA* can be interpreted as based on a deformed (regularized) Mahalanobis distance between a new observation Z and the mean of each group. Let us discuss the particular situation of MWCD-LDA* and consider the singular value decomposition (SVD) of S^*_{MWCD} in the form $S^*_{MWCD} = Q\Lambda Q^T$. The aforementioned deformed Mahalanobis distance can be interpreted as the Euclidean distance applied on $\Lambda^{-1/2}Q^TZ$. More specifically, if we assume Z to come from one of the groups with the covariance matrix Σ, we obtain in a straightforward way

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

The deformed Mahalanobis distance of L_1-SCRRDA and L_2-SCRRDA takes additionally into account a regularization of the means.

Regularizing the means can be theoretically justified as exploiting Stein's statistical estimation [13, 28], extending Stein's shrinkage estimator originally proposed for the mean of multivariate normal data. The regularization of the means within L_1-SCRRDA and L_2-SCRRDA replaces (unbiased) arithmetic means by their (biased) shrunken counterparts, allowing to reduce their mean square error if the regularization parameters are sufficiently small.

Also the implicit weights assigned to individual observations in methods of Sections 4.1 to 4.3 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 empirical covariance matrix. The weights are used to compute the weighted mean and weighted empirical covariance matrix. In the numerical examples, we have verified that outlying measurements obtain small weights, which ensures the robustness of the method.

6.3 Limitations

Let us mention also the limitations of the newly proposed classification methods.

- Suitability of all the novel methods for data following an elliptically symmetric unimodal multivariate distribution.
- All the novel methods require an intensive computation.
- The implicit weights in methods of Sections 4.1 to 4.3 are assigned to individual observations (rather than perhaps to individual variables).
- The variability not substantially different across variables is unexpressedly assumed for all regularized LDA methods. Still, the novel methods seem to yield reliable results on the data sets of Section 5, although this implicit assumption of homogeneous variances of all variables is violated in them.
- L_1-SCRRDA and L_2-SCRRDA are more computationally demanding compared to MWCD-LDA*, but yield comparable results, i.e. there seems no major added value of regularizing the means in contrary to the experience of e.g. [10].

Finally, we need to recall that the regularization itself may be a too radical modification of the original problem of rank n, which is replaced by a problem of rank p, which may be much larger. Such increase of the dimensionality of the scatter 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 regularized problem was solved in an arbitrary-precision arithmetic [24, 5].

7 Conclusions and Future Work

The analysis of high-dimensional data with the number of variables p largely exceeding the number of observations n becomes an important task in numerous

tasks of bioinformatics. While numerous available algorithms for the regularized LDA are popular for the analysis of high-dimensional data [21], 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.

In this paper, we combine robustness to the presence of outliers with regularized estimation of the scatter matrix of the multivariate data in a unique way. As a result, four new robust classification methods for high-dimensional observations are proposed in Section 4. Three of the methods are based on implicit weighting of individual observations, while M-LDA* is based on M-estimation. In addition, newly proposed methods L_1-SCRRDA and L_2-SCRRDA replace also the sample mean of each group by a regularized (shrunken) robust estimator.

We analyzed several real data sets fulfilling $n < p$ in Section 5. These data sets coming from various problems of (bioinformatics) research can be characterized as high-dimensional in sense of $n < p$ or even $n \ll p$.

Particularly, we pay the largest attention to the analysis of an original brain activity data set from a neuroscience research study investigating 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.

To investigate the performance of individual methods on data contaminated by noise, we also introduced an artificial contamination to the brain activity data. Indeed, robustness to moderate or severe outliers is an important requirement in the analysis of high-dimensional data, especially if the number of observations is small. The results reveal a regularized LDA in a standard form to be sensitive to outliers. SCRDA turns out to be moderately robust, which is an effect of the regularizing also the means, which yields denoised versions of standard means. The novel methods turn out to be even more robust also against severely outlying measurements. It is an artificial contamination of the data which reveals the robustness of the novel methods as their strength and the whole study with artificial contamination reveals the advantage of robust methods compared to non-robust ones. However, regularizing the means applied on the robust methods does not bring any major additional benefit compared to MWCD-LDA*, while it requires a high increase in computational complexity.

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.

- Formulating more efficient algorithms tailor-made for important specific choices of the target matrix T as alternatives to Algorithms 2 or 3.
- Comparing various approaches to regularizing the means (i.e. for various norms or various shrinkage targets) in a large simulation study.
- Comparing the performance and robustness of the new methods with approaches based on a robust PCA.

- Investigating the non-robustness of other standard regularized classification methods (e.g. of PAM).
- Extending the combination of regularization and robustness to 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 (e.g. robust lasso estimator).
- Developing other multivariate methods based on the regularized MWCD estimators, e.g. robust PAM or robust regularized PCA.

From the point of view of the neuroscience research, future investigations are planned 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 measured in the resting state.

Acknowledgments

Preliminary results were first presented at the BIOSTEC/BIOINFORMATICS 2016 conference (21-23 February 2016 in Rome), where they were published in the proceedings.

The work was supported by the project "National Institute of Mental Health (NIMH-CZ)", grant number ED2.1.00/03.0078, 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.

References

1. Ben-Tal, A., El Ghaoui, L. and Nemirovski, A. (2009). *Robust optimization.* Princeton University Press, Princeton.
2. Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data.* Springer, New York.
3. 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.
4. Croux, C. and Dehon, C. (2001). Robust linear discriminant analysis using S-estimators. *Canadian Journal of Statistics*, 29:473–493.
5. 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.
6. Davies, P. L. and Gather, U. (2005). Breakdown and groups. *Annals of Statistics*, 33:977–1035.
7. Duffau, H. (2011). *Brain mapping. From neural basis of cognition to surgical applications.* Springer, Vienna.
8. Dziuda, D.M. (2010). *Data mining for genomics and proteomics: Analysis of gene and protein expression data.* Wiley, New York.

9. Filzmoser, P. and Todorov, V. (2011). Review of robust multivariate statistical methods in high dimension. *Analytica Chinica Acta*, 705:2–14.
10. Guo, Y., Hastie, T., and Tibshirani, R. (2007). Regularized discriminant analysis and its application in microarrays. *Biostatistics*, 8:86–100.
11. Han, H. and Jiang, X. (2014). Overcome support vector machine diagnosis overfitting. *Cancer Informatics*, 13:145–148.
12. Hansen, P.C. (1998). *Rank-deficient and discrete ill-posed problems: Numerical aspects of linear inversion.* SIAM, Philadelpha.
13. Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The elements of statistical learning.* Springer, New York, 2nd edition.
14. 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.
15. Huber, P. J. and Ronchetti, E. M. (2009). *Robust statistics.* Wiley, New York, 2nd edition.
16. Hubert, M., Rousseeuw, P. J., and van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statistical Science*, 23:92–119.
17. Hubert, M. and Debruyne, M. (2010). Minimal covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:36–43.
18. Jurečková, J. and Portnoy, S. (1987). Asymptotics for one-step M-estimators in regression with application to combining efficiency and high breakdown point. *Communications in Statistics Theory and Methods*, 16:2187–2199.
19. Kalina, J. (2012). Implicitly weighted methods in robust image analysis. *Journal of Mathematical Imaging and Vision*, 44:449–462.
20. Kalina, J., Seidl, L., Zvára, K., Grünfeldová, H., Slovák, D. and Zvárová J. (2013). System for selecting relevant information for decision support. *Studies in Health Technology and Informatics*, 183:83–87.
21. Kalina, J. (2014). Classification analysis methods for high-dimensional genetic data. *Biocybernetics and Biomedical Engineering*, 34:10–18.
22. Kalina, J. and Schlenker, A. (2015). A robust and regularized supervised variable selection. *BioMed Research International*, Article 320385.
23. Kindermans, P.-J., Schreuder, M., Schrauwen, B., Müller, K.-R., and Tangermann, M. (2014). True zero-training brain-computer interfacing–An online study. *PLoS One*, 9, Article 102504.
24. Kůrková, V. and Sanguineti, M. (2005). Learning with generalization capability by kernel methods of bounded complexity. *Journal of Complexity*, 21:350–367.
25. 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.
26. Maronna, R.A., Martin, D.R. and Yohai, V.J. (2006). *Robust statistics: Theory and methods.* Wiley, New York.
27. 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.
28. Pourahmadi, M. (2013). *High-dimensional covariance estimation.* Wiley, New York.
29. Roelant, E., van Aelst, S., and Willems, G. (2009). The minimum weighted covariance determinant estimator. *Metrika*, 70:177–204.
30. Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust regression and outlier detection.* Wiley, New York.

31. Rousseeuw, P. J. and van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223.
32. Sreekumar, A. et al. (2009). Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, 457:910–914.
33. Steinwart, I. and Christmann, A. (2008) *Support vector machines.* Springer, New York.
34. Tibshirani, R. and Narasimhan, B. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, 18:104–117.
35. Todorov, V. and Filzmoser, P. (2009). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(3):1–47.
36. Tyler, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *Annals of Statistics*, 15:234–251.
37. Tyler, D. E. (2014). Breakdown properties of the M-estimators of multivariate scatter. http://arxiv.org/pdf/1406.4904v1.pdf.
38. 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.