Computation of Regularized Linear Discriminant Analysis

Jan Kalina, *Institute of Computer Science AS CR*, kalina@cs.cas.cz
Zdeněk Valenta, *Institute of Computer Science AS CR*, valenta@cs.cas.cz
Jurjen Duintjer Tebbens, *Institute of Computer Science AS CR*, duintjertebbens@cs.cas.cz

Abstract. This paper is focused on regularized versions of classification analysis and their computation for high-dimensional data. A variety of regularized classification methods has been proposed and we critically discuss their computational aspects. We formulate several new algorithms for shrinkage linear discriminant analysis, which exploits a shrinkage covariance matrix estimator towards a regular target matrix. Numerical linear algebra considerations are used to propose tailor-made algorithms for specific choices of the target matrix. Further, we arrive at proposing a new classification method based on L_2-regularization of group means and the pooled covariance matrix and accompany it by an efficient algorithm for its computation.

Keywords. Classification analysis, Regularization, Matrix decomposition, Shrinkage eigenvalues, High-dimensional data

1 Introduction

Classification analysis methods have 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. Linear discriminant analysis (LDA) is a standard statistical classification method. In the whole paper, we consider n observations with p variables, observed in K different samples (groups) with $p > K \geq 2$,

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

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. LDA in its standard form assumes $n > p$ and is unsuitable for high-dimensional data with a number of variables exceeding the number of observations (large p/small n problem). In case where $n < p$, the matrix S of size p is singular and computing its inverse must be replaced by an appropriate alternative. Available approaches in this context are based e.g. on pseudoinverse matrices, which are however unstable due to a small n [4]. Other proposals are based on the generalized SVD decomposition or on elimination of the common null space of the between-group and within-group covariance matrices [2].

Various authors suggested to use a regularized version of LDA for $n \ll p$ [3, 2, 4, 5]. Suitable regularized estimators of the covariance matrix are guaranteed to be regular and positive definite even for $n \ll p$. They have become established e.g. in image analysis, chemometrics, molecular genetics, or econometrics, while their fast computation and numerical stability remains to be an important issue [4, 7]. We will describe the most important approaches and critically discuss their possible computation.

The first approach to a regularized discriminant analysis by [3] is based on a shrinkage covariance matrix with two parameters, which are searched for in a grid search minimizing the classification error. Later, the computation was criticized as computationally intensive in [8], where a linear shrinkage estimator of the covariance matrix was proposed and the asymptotically optimal value of the regularization parameter was derived. The method is implemented in the *corpcor* package of *R* software; however, its computation for a large p is very slow.

Habitually used regularized versions of LDA are based either on regularizing only Σ using one of approaches of [8] or on a double shrinkage applied on the covariance matrix as well as means of each group. The latter approach was proposed by Guo et al. [4], who performed shrinking of the covariance matrix towards an identity matrix and at the same time shrinking of the mean of each group to zero. The method is implemented in the *rda* package of *R* software. For specific values of the parameters, the computation is based on the SVD algorithm, without applying methods of numerical linear algebra to decrease computational costs. The optimal values of shrinkage parameters are optimized in a cross-validation over a 2-dimensional grid, which has been described as tedious [4]. Moreover, there are many possible tuning parameters giving the same cross-validation error rate. The computational effectivity and stability of habitually used algorithms is not investigated even in the recent monograph [7] on covariance matrix estimation for high-dimensional data.

This paper studies efficient algorithms for computing various regularized versions of LDA. Section 2 of this paper formulates several algorithms for shrinkage LDA, which exploits a shrinkage covariance matrix estimator towards a regular target matrix. The computational effectivity of the algorithms is inspected using arguments of numerical linear algebra. For a specific choice of the target matrix, we are able to propose a tailor-made algorithm with a lower computational cost compared to algorithms which are formulated for a general context. Besides, we arrive at proposing new versions of classification methods and accompany them by efficient algorithms for their computation in Section 3. The classification performance of the methods is illustrated on real data in Section 4.

2 Algorithms for Regularized Linear Discriminant Analysis

This section is devoted to proposing and comparing new algorithms for a habitually used version of the regularized LDA [4]. We use suitable matrix decompositions to propose efficient algorithms either for a general choice of T or for its specific choices. To the best of our knowledge, tailor-made algorithms for a specific T have not been described. We compare the new algorithms in terms of their computational costs as well as numerical stability.

We will describe one of habitually used regularized versions of LDA. This will be denoted as LDA* to avoid confusion, because the concept of regularized discriminant analysis encompasses several different methods [4]. A given target matrix T will be used, which must be a regular symmetric positive definite matrix of size $p \times p$. Its most common choices include the identity matrix I_p or a diagonal (non-identity) matrix; other target matrices have been considered by [8].

Let us denote the mean of the observed values in the k-th group ($k = 1, \ldots, K$) by $\bar{X}_k$. LDA* 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 has the form

$$l_k^* = \bar{X}_k^T (S^*)^{-1} Z - \frac{1}{2} \bar{X}_k^T (S^*)^{-1} \bar{X}_k + \log p_k, \quad k = 1, \ldots, K, \tag{2}$$

where p_k is a prior probability of observing an observation from the k-th group and

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

for $\lambda \in [0, 1]$ denotes a shrinkage estimator of the covariance matrix across groups. 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. Equivalently, LDA* assigns a new observation Z to group k, if

$$(\bar{X}_k - Z)^T S^{*-1} (\bar{X}_k - Z) = \min_{j=1,\ldots,K} \left\{ (\bar{X}_j - Z)^T S^{*-1} (\bar{X}_j - Z) \right\}. \tag{4}$$

First, the standard approach for computing LDA* may be improved by employing the eigendecomposition of S^* for a fixed λ. A suitable value of λ is found by a cross-validation in the form of a grid search over all possible values of $\lambda \in [0, 1]$.

Algorithm 2.1.
LDA* for the general regularization (3) based on eigendecomposition.

Step 1 *Compute the matrix*

$$A = [\bar{X}_1 - Z, \ldots, \bar{X}_K - Z] \tag{5}$$

of size $p \times K$ whose k-th column is $\bar{X}_k - Z$.

Step 2 *Compute S^* according to (3) with a fixed $\lambda \in [0, 1]$.*

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

Step 4 *Compute the matrix*

$$B = D_*^{-1/2} Q_*^T A \tag{6}$$

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.*

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^*, thus avoiding additional computations of the Mahalanobis distance, which is expensive of order p^3 and numerically rather unstable. The group assignment (4) is done by using

$$(\bar{X}_j - Z)^T S^{*-1}(\bar{X}_j - Z) = (\bar{X}_j - Z)^T Q_* D_*^{-1} Q_*^T (\bar{X}_j - Z) = \|D_*^{-1/2} Q_*^T (\bar{X}_j - Z)\|^2. \tag{7}$$

The algorithm can be made cheaper by replacing the eigendecomposition of S^* with its Cholesky decomposition

$$S^* = L_* L_*^T, \tag{8}$$

where L_* is a nonsingular lower triangular matrix. The costs of Cholesky decomposition are about $1/3 \cdot p^3$ floating point operations. On the other hand, Cholesky decomposition will suffer from instability when S^* is not positive definite.

Algorithm 2.2.
LDA* for the general regularization (3) based on Cholesky decomposition.

Step 1 *Compute the matrix*

$$A = [\bar{X}_1 - Z, \dots, \bar{X}_K - Z] \tag{9}$$

of size $p \times K$ whose k-th column is $\bar{X}_k - Z$.

Step 2 *Compute S^* according to (3) with a fixed $\lambda \in [0,1]$.*

Step 3 *Compute the Cholesky factor L_* of S^*.*

Step 4 *Compute the matrix*

$$B = L_*^T A \tag{10}$$

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.*

For specific target matrices, we can further reduce computational costs by using the following algorithm for LDA*. The pooled estimator S can be written in the form

$$S = Y^T Y, \qquad Y = [X_{11} - \bar{X}, \dots, X_{1n_1} - \bar{X}, \dots, X_{K1} - \bar{X}, \dots, X_{Kn_K} - \bar{X}]^T \tag{11}$$

where Y is of size $n \times p$. Then using the singular value decomposition (SVD) of Y in the form

$$Y = P \Sigma Q^T, \tag{12}$$

we can express the eigendecomposition of S as

$$S = Y^T Y = (P\Sigma Q^T)^T P \Sigma Q^T = Q \Sigma^2 Q^T. \tag{13}$$

The costs will be about $4 \cdot np^2$ floating point operations, thus with $p \gg n$ the gain is considerable. Moreover, if

$$S^* = \lambda S + (1-\lambda) I_p, \quad \lambda \in [0,1], \tag{14}$$

we immediately obtain the needed eigendecomposition of S^* as

$$S^* = \lambda S + (1-\lambda) I_p = Q\left(\lambda \Sigma^2 + (1-\lambda) I_p\right) Q^T. \tag{15}$$

The SVD can be computed in a backward stable way with all singular values accurate up to machine precision level [1]. For the special case (14), which is commonly denoted as Tikhonov or ridge regularization of S, a more efficient computation can be performed as follows.

Algorithm 2.3.
LDA for the ridge regularization (14).*

Step 1 *Compute the matrix*

$$A = [\bar{X}_1 - Z, \ldots, \bar{X}_K - Z] \tag{16}$$

of size $p \times K$ whose k-th column is $\bar{X}_k - Z$ and compute the matrix Y in (11).

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

$$Y = P\Sigma Q^T, \tag{17}$$

with singular values $\{\sigma_1, \ldots, \sigma_n\}$ 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*

$$D_* = diag\{\lambda\sigma_1^2 + (1-\lambda), \ldots, \lambda\sigma_p^2 + (1-\lambda)\}. \tag{18}$$

Step 4 *Compute the matrix*

$$B = D_*^{-1/2} Q^T A \tag{19}$$

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.*

Eigenvalues of the regularized covariance matrix forming the matrix D^* in (18) can be interpreted as shrinkage eigenvalues.

In an analogous manner, algorithms for a regularized quadratic discriminant analysis (QDA) can be obtained, using a regularized estimator of the covariance matrix in each group separately.

3 L_2-regularized linear discriminant analysis

Disadvantages of SCRDA [4] include a computational intensity as well as an inconsistent approach to shrinkage. The means are namely modified by an L_1-norm regularization and the covariance matrix in the sense of the L_2-norm. As an alternative, this section proposes a new regularized version of LDA denoted as L_2-LDA together with an efficient algorithm for its computation. It employs a shrinkage estimator of Σ and shrunken means towards the overall mean across groups. As a unique feature, both shrinkage approaches have the form of an L_2-norm regularization.

The classification rule of L_2-LDA assigns a new observation Z to the k-th group, if $l_k^\dagger > l_j^\dagger$ for every $j \neq k$, where

$$l_k^\dagger = \bar{X}_k'^T (S^*)^{-1} Z - \frac{1}{2} \bar{X}_k'^T (S^*)^{-1} \bar{X}_k' + \log p_k \tag{20}$$

and $\bar{X}_k'$ denotes the shrunken mean of the k-th group towards the overall mean computed across groups. The method can be interpreted as based on a L_2 regularized Mahalanobis distance. As another contrast with the habitually used algorithm of SCRDA [4], we will estimate the parameter λ in a straightforward way using an asymptotically optimal value minimizing the mean square error [8]. To avoid confusion, the asymptotically optimal value of λ will be denoted by $\lambda^\dagger$ and the corresponding shrinkage covariance matrix by

$$S^\dagger = \lambda^\dagger S + (1 - \lambda^\dagger)T. \tag{21}$$

Algorithm 3.1.
L_2-LDA.

Step 1 *Compute $\lambda^\dagger$ as*

$$\lambda^\dagger = \frac{2\sum_{i=2}^{p}\sum_{j=1}^{i-1}\widehat{var}(S_{ij})}{2\sum_{i=2}^{p}\sum_{j=1}^{i-1}S_{ij}^2 + \sum_{i=1}^{p}(S_{ii}-1)^2}, \tag{22}$$

where $\widehat{var}(S_{ij})$ is the maximum likelihood estimator of the variance of values S_{ij} for a fixed i and j.

Step 2 *Compute and store the eigenvalues of $S^\dagger$ in the diagonal matrix D_*, and compute and store the corresponding eigenvectors of $S^\dagger$ in the orthogonal matrix Q_*.*

Step 3 *For a fixed $\delta \in [0, 1]$, compute $\bar{X}_k' = \delta\bar{X}_k + (1-\delta)\bar{X}, \qquad k = 1, \dots, K.$*

Step 4 *Assign Z to group k, if*

$$\|D_*^{-1/2}Q_*^T(\bar{X}_k' - Z)\| = \min_{j=1,\dots,K}\|D_*^{-1/2}Q_*^T(\bar{X}_j' - Z)\|. \tag{23}$$

Step 5 *Repeat steps 3 and 4 for various δ and find the optimal classification rule yielding the best classification performance.*

Algorithm 3.1 is formulated for a general target matrix T. For a specific choice of T, a computationally cheaper method can be obtained in an analogous way as Algorithms 2.2 and 2.3 from the general algorithm 2.1.

Another possibility is to regularize the within-group covariance matrix instead of regularizing S, which is however computationally more intensive.

4 Examples

We present two examples on real molecular genetic data sets in order to illustrate the behavior of the newly proposed L_2-LDA method.

Example 1 contains data from a cardiovascular genetic study of the Center of Biomedical Informatics in Prague performed in 2006–2011. The data contain expressions of $p = 38\,590$ gene transcripts measured on 24 patients having a cerebrovascular stroke and 24 control persons.

In Example 2, a prostate cancer metabolomic data set [9] 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.

In both examples, we computed the classification methods described in this paper using the algorithms of Sections 2 and 3. For comparison, we computed also other available classification

				Youden's index	
Method	S^*	*R* Package	Function	Example 1	Example 2
SVM	-	*e1071*	*svm*	1.00	1.00
Classification tree	-	*tree*	*tree*	0.94	0.97
Self-organizing map	-	*kohonen*	*som*	0.88	0.93
Multilayer percpetron	-	*nnet*	*nnet*	Infeasible	Infeasible
LDA	-	*MASS*	*lda*	Infeasible	Infeasible
SCRDA	(14)	*rda*	*rda*	1.00	1.00
LDA*	(14)	-	-	1.00	1.00
LDA*	(24)	-	-	1.00	1.00
L_2-LDA	(14)	-	-	1.00	1.00
L_2-LDA	(24)	-	-	1.00	1.00
PCA $\Longrightarrow$ LDA	-	-	-	0.54	0.90
PCA $\Longrightarrow$ SCRDA	(14)	-	-	0.71	0.92
PCA $\Longrightarrow$ LDA*	(14)	-	-	0.63	0.81
PCA $\Longrightarrow$ LDA*	(24)	-	-	0.63	0.81
PCA $\Longrightarrow$ L_2-LDA	(14)	-	-	0.71	0.92
PCA $\Longrightarrow$ L_2-LDA	(24)	-	-	0.71	0.92
PCA $\Longrightarrow$ MWCD-LDA	-	-	-	0.69	0.90

Table 1. Results of Example 1 and Example 2. LDA* was computed using Algorithm 2.3 for the choice (14) and Algorithm 2.2 for (24). L_2-LDA was computed using Algorithm 3.1. PCA uses 20 principal components.

methods, including the support vector machines (SVM), a classification tree, Kohonen's self-organizing map, a multilayer perceptron with 2 hidden layers, or the highly robust classification method MWCD-LDA of [6]. Various regularized versions of LDA include the most common choice $T = I_p$ or another choice

$$S^* = \lambda S + (1-\lambda) s I_p, \quad \lambda \in [0,1], \quad s = \sum_{i=1}^{p} S_{ii}/p. \tag{24}$$

We used the default settings to compute them in *R* software packages, which are listed also in Table 1. The classification performance is measured by means of the Youden's index, which is defined as sensitivity + specificity -1. The dimensionality reduction was performed by the principal component analysis (PCA) with 20 principal components.

The results performed on raw data as well as after a dimensionality reduction reveal that the regularized versions of LDA perform quite similarly. The newly proposed method L_2-LDA with an efficient algorithm seems to perform comparably with the available regularized methods with less efficient computation. Besides, the choice of the target matrix T does not seem to play an important role.

Further, we investigated the reduction in classification performance after reducing the dimensionality to 20 principal components in both examples. The approach of Algorithm 3.1 (PCA $\Longrightarrow$ L_2-LDA) yields improved results compared to its standard counterpart (PCA $\Longrightarrow$ LDA). The results of regularized methods do not greatly differ from the robust MWCD-LDA procedure, which indicates that regularizaed versions of LDA do not greatly suffer by the presence of outlying measurements in the data. Nevertheless, the robustness of regularized methods with respect to outliers has not been systematically investigated [5].

To conclude the paper, several new algorithms for shrinkage LDA are proposed, exploiting a shrinkage covariance matrix estimator towards a regular target matrix. Some algorithms are tailor-made for a specific choice of the target matrix and their computational costs are discussed. A new regularized classification method L_2-LDA is proposed and accompanied by an efficient algorithm. An analysis of two real data sets reveals its classification performance to be comparable to available regularized classification methods for high-dimensional data.

Acknowledgement

The work of J. Kalina was financially supported by the Neuron Foundation for Supporting Science. The work of Z. Valenta was supported by the institutional support RVO:67985807. The work of J. Duintjer Tebbens was supported by the grant GA13-06684S of the Czech Science Foundation.

Bibliography

[1] Barlow, J.L., Bosner, N. and Drmač, Z. (2005) *A new stable bidiagonal reduction algorithm.* Linear Algebra and its Applications, **397**, 35–84.

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

[3] Friedman, J.H. (1989) *Regularized discriminant analysis.* Journal of the American Statistical Assocation, **84**, 165–175.

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

[5] Kalina, J. (2014) *Classification methods for high-dimensional data.* Biocybernetics and Biomedical Engineering, **34** (1), 10–18.

[6] Kalina, J. (2012) *Highly robust statistical methods in medical image analysis.* Biocybernetics and Biomedical Engineering, **32** (2), 3–16.

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

[8] Schäfer, J. and Strimmer K. (2005) *A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics.* Statistical Applications in Genetics and Molecular Biology, **4**, Article 32.

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