Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 51 (2007) 3718–3736

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

DALASS: Variable selection in discriminant analysis via the LASSO

Nickolay T. Trendafilov[a,*], Ian T. Jolliffe[b]

[a]*Department of Statistics, The Open University, Milton Keynes MK7 6AA, UK*
[b]*Department of Meteorology, University of Reading, Reading RG6 6BB, UK*

Received 25 January 2006; received in revised form 8 November 2006; accepted 13 December 2006
Available online 12 January 2007

Abstract

The objective of DALASS is to simplify the interpretation of Fisher's discriminant function coefficients. The DALASS problem—discriminant analysis (DA) modified so that the canonical variates satisfy the LASSO constraint—is formulated as a dynamical system on the unit sphere. Both standard and orthogonal canonical variates are considered. The globally convergent continuous-time algorithms are illustrated numerically and applied to some well-known data sets.

Keywords: Canonical variates; Orthogonal canonical variates; LASSO constraint; Penalty function; Continuous-time constrained optimization; Steepest ascent vector flows on manifolds

1. Introduction

Discriminant analysis (DA) is a descriptive multivariate technique for analyzing grouped data, i.e. the rows of the data matrix are divided into a number of groups that usually represent samples from different populations (Krzanowski, 2003; McLachlan, 1992). Recently DA has also been viewed as a promising dimensionality reduction technique (Dhillon and Modha, 2001; Hastie et al., 2001). Indeed, the presence of group structure in the data additionally facilitates dimensionality reduction. The best known variety of DA is linear discriminant analysis (LDA), whose central goal is to describe the differences between the groups in terms of discriminant functions defined as linear combinations of the original variables (Fisher, 1936). The same (linear) discriminant functions can be obtained if the different populations are assumed to be multivariate Gaussian with a common covariance matrix and probabilities of misclassification are minimized (Hastie et al., 2001; McLachlan, 1992). Under the same assumptions, discriminant functions appear in one-way MANOVA for best separation of the group means (Rencher, 2002). For example, canonical discriminant analysis by MATLAB (MATLAB, 2002) can be performed using the function for one-way MANOVA. The mathematical equivalent of DA is the generalized eigenvalue problem (Golub and Van Loan, 1991; Parlett, 1980).

The interpretation of the discriminant functions is based on the coefficients of the original variables in the linear combinations. The problem is similar to interpretation of principal components (Jolliffe, 2002): the interpretation can be clear and obvious if there are only few large coefficients and the rest are all close to or exactly zero. Unfortunately,

[*] Corresponding author. Tel.: +44 190 8652030; fax: +44 190 8655515.
E-mail addresses: N.Trendafilov@open.ac.uk (N.T. Trendafilov), ian@sandloch.fsnet.co.uk (I.T. Jolliffe).

0167-9473/$ - see front matter Crown Copyright © 2007 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.12.046

in many applications this is not the case. There are several approaches to the interpretation of the discriminant functions, each of which has disadvantages (Pedhazur, 1982; Rencher, 2002). These will be discussed in Section 2 below. Section 3 describes a modification of LDA in which vectors of coefficients are constrained to be orthogonal, and interpretation of its coefficients is compared to that of LDA for a well known example.

In Section 4 we consider the classical DA problem subject to additional LASSO constraints (Hastie et al., 2001). We call this technique DALASS. A similar idea has already been successfully applied to principal component analysis (Trendafilov and Jolliffe, 2006). The LASSO inequality constraint requires that the sum of the absolute values of the coefficients of a unit length vector **a** be less than some pre-specified threshold t, i.e.:

$$\sum_{i=1}^{p} a_i^2 = \|\mathbf{a}\|_2^2 = 1 \quad \text{and} \quad \sum_{i=1}^{p} |a_i| = \|\mathbf{a}\|_1 \leqslant t, \quad t \in [1, \sqrt{p}]. \tag{1}$$

The idea is very simple: as t decreases, an increasing number of coefficients are driven to zero, or close to it. Section 4 also describes an algorithm for implementing DALASS and hence producing discriminant functions with a only few of the original variables contributing non-trivially to each of them. The example of Section 3 is revisited. The corresponding graphical representations in DA will also be clearer in DALASS, as they are based on canonical variates composed almost entirely from only a few of the original variables. This can make a considerable difference when the number of original variables is large.

Two further examples are included in Section 5 and some concluding remarks are made in Section 6.

2. Canonical variates

In general the data for DA are as follows: p variables are measured and collected on a $(1 \times p)$ vector $\mathbf{x}$; the measurements are made on n individuals (cases) which are *a priori* divided into g groups; let n_i be the number of individuals in the ith group, i.e. $n_1 + n_2 + \cdots + n_g = n$. It is assumed that $n > p$. Then the $(1 \times p)$ vector $\mathbf{x}_{ij}$ denotes the measurements made on the jth individual belonging to the ith group. The $(n \times p)$ data matrix $\mathbf{X}$ collects the measurements of all individuals.

Consider the following linear combinations $\mathbf{Y} = \mathbf{XA}$ also called discriminant scores (Pedhazur, 1982). This is a linear transformation of the original data $\mathbf{X}$ into another vector space. It is interesting to find a $(p \times s)$ transformation matrix $\mathbf{A}$ of the original data $\mathbf{X}$ such that the *a priori* groups are better separated in the dimensions of the transformed data $\mathbf{Y}$ than with respect to any of the original variables. The number of transformed dimensions s is typically much smaller than the original p. Fisher's LDA achieves both goals by finding a transformation $\mathbf{A}$ which produces the "best" discrimination of the groups by simultaneous maximization of the between-groups variance and minimization of the within-groups variance of $\mathbf{Y}$ (Fisher, 1936). The procedure of finding $\mathbf{A}$ is sequential: suppose $\mathbf{a}$ is the first column of $\mathbf{A}$. Then one can show (Fisher, 1936; Krzanowski, 2003) that formally the problem is

$$\max_{\mathbf{a}} \frac{\mathbf{a}^{\mathrm{T}}\mathbf{C_B}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{C_W}\mathbf{a}}, \tag{2}$$

where $\mathbf{C_B}$ and $\mathbf{C_W}$ are the between-groups and within-groups covariance matrices:

$$\mathbf{C_B} = \frac{\mathbf{B}}{g-1} \quad \text{and} \quad \mathbf{C_W} = \frac{\mathbf{W}}{n-g}, \tag{3}$$

with

$$\mathbf{W} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^{\mathrm{T}}, \tag{4}$$

$$\mathbf{B} = \sum_{i=1}^{g} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^{\mathrm{T}}, \tag{5}$$

and the within-group means and total mean are

$$\bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij} \quad \text{and} \quad \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{g} n_i\bar{\mathbf{x}}_i. \tag{6}$$

If **x** has a multivariate normal distribution, then Fisher's LDA objective function

$$F(\mathbf{a}) = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{C_B}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{C_W}\mathbf{a}} \tag{7}$$

has an F distribution with $g-1$ and $n-g$ degrees of freedom under the null hypothesis that there is no difference among the g group means, and thus $F(\mathbf{a})$ can be used to test that hypothesis. The larger the F value, the greater the divergence among the group means. The transformation **A** will successively produce the maximum possible divergence among the group means in its first few dimensions.

The problem (2) is equivalent to the following generalized eigenvalue problem (Krzanowski, 2003):

$$(\mathbf{C_B} - \lambda\mathbf{C_W})\mathbf{a} = 0, \tag{8}$$

which can also be written as

$$(\mathbf{C_W^{-1}}\mathbf{C_B} - \lambda\mathbf{I}_p)\mathbf{a} = 0. \tag{9}$$

Thus the maximum of the objective function F in (2) is the largest eigenvalue of $\mathbf{C_W^{-1}C_B}$ and is achieved at the corresponding eigenvector **a**. The problem looks quite similar to that of principal component analysis (PCA) but $\mathbf{C_W^{-1}C_B}$ is not symmetric. Moreover the rank of this matrix is $r \leqslant \max\,(p, g-1)$ and all the remaining eigenvalues are 0s. The number r is called dimension of the canonical variate representation. The number of useful dimensions for discriminating between groups, s, is smaller than r, and the transformation **A** is formed by the eigenvectors corresponding to the s largest eigenvalues ordered in decreasing order. Clearly the $(p \times s)$ transformation **A** determined by Fisher's LDA maximizes the discrimination among the groups and represents the transformed data in a lower s-dimensional space.

The solution of the eigenvalue problem (9) in matrix terms is $\mathbf{C_B A} = \mathbf{C_W A \Lambda}$, where $\mathbf{\Lambda}$ is the $(s \times s)$ diagonal matrix of the s largest eigenvalues of $\mathbf{C_W^{-1}C_B}$ ordered in decreasing order. This is not a symmetric eigenvalue problem and **A** is an oblique matrix (not orthogonal as in the symmetric case of PCA), i.e. **A** is of full column-rank s and $\mathrm{diag}(\mathbf{A}^{\mathrm{T}}\mathbf{A}) = \mathbf{I}_s$. The matrix $\mathbf{A}^{\mathrm{T}}\mathbf{C_W A}$ is diagonal and it is usual to normalize **A** such that $\mathbf{A}^{\mathrm{T}}\mathbf{C_W A} = \mathbf{I}_s$, i.e.:

$$\mathbf{A}_{\mathrm{raw}} = \mathbf{A}\,\mathrm{diag}(\mathbf{A}^{\mathrm{T}}\mathbf{C_W A})^{-1/2}, \tag{10}$$

where $\mathrm{diag}(\mathbf{A}^{\mathrm{T}}\mathbf{C_W A})^{-1/2}$ is the diagonal matrix containing the elementwise square roots of main diagonal of $\mathbf{A}^{\mathrm{T}}\mathbf{C_W A}$. The elements of the normalized matrix $\mathbf{A}_{\mathrm{raw}}$ are called *raw coefficients* (Pedhazur, 1982). They are, in fact, the coefficients of the variables in the discriminant functions. This normalization makes the within-groups variance of the discriminant scores **Y** equal to 1.

There is another way to compute the raw coefficients using a symmetric eigenvalue problem as in PCA. To implement this, one needs to rewrite the basic LDA problem (2) in the following equivalent form:

$$\max_{\mathbf{a}} \mathbf{a}^{\mathrm{T}}\mathbf{C_B a} \quad \text{subject to } \mathbf{a}^{\mathrm{T}}\mathbf{C_W a} = 1. \tag{11}$$

Let $\mathbf{C_W} = \mathbf{U}^{\mathrm{T}}\mathbf{U}$ be the Cholesky factorization of $\mathbf{C_W}$ with **U** a positive definite upper triangular matrix. The substitution $\mathbf{a} := \mathbf{Ua}$ in (11) leads to the following symmetric eigenvalue problem:

$$\max_{\mathbf{a}} \mathbf{a}^{\mathrm{T}}\mathbf{U}^{-\mathrm{T}}\mathbf{C_B U}^{-1}\mathbf{a} \quad \text{subject to } \mathbf{a}^{\mathrm{T}}\mathbf{a} = 1, \tag{12}$$

whose (orthogonal) solution **A** is used in turn to find the raw coefficients $\mathbf{A}_{\mathrm{raw}} = \mathbf{U}^{-1}\mathbf{A}$. If $\mathbf{C_W}$ is ill-conditioned then the Cholesky factorization should be replaced by eigenvalue decomposition (Golub and Van Loan, 1991). One should also keep in mind that the LDA problem (8) can be solved using the original data **X** only without actually forming $\mathbf{C_B}$ and $\mathbf{C_W}$ (Golub and Van Loan, 1991).

Table 1
Tests of equality of group means

Vars.	Wilk's lambda	F	p
x_1	.651	16.072	.000
x_2	.998	.063	.803
x_3	.947	1.685	.204
x_4	.610	19.210	.000
x_5	.763	9.315	.005

Table 2
Canonical variates for Skull Data

Vars.	Raw coefficients	Standardized coefficients	Structure coefficients
x_1	.090	.367	.759
x_2	−.156	−.578	−.048
x_3	−.005	−.017	.246
x_4	.117	.405	.830
x_5	.117	.627	.578

The raw coefficients are considered difficult to interpret when one wants to evaluate the relative importance of the original variables. As in PCA, one can try to identify those raw coefficients that are large in magnitude in a particular discriminant function and conclude that the corresponding variables are important for discrimination between the groups. The problem is that such a conclusion can be misleading in LDA. The large magnitudes may indeed be caused by large between-groups variability, but also can be caused by small within-groups variability (Krzanowski, 2003). This problem with the interpretation of raw coefficients is overcome by an additional standardization of $\mathbf{A}_{\text{raw}}$ which makes all variables comparable:

$$A_{\text{std}} = \text{diag}(\mathbf{C_W})^{1/2}\mathbf{A}_{\text{raw}}, \tag{13}$$

and the new coefficients are called the *standardized coefficients* (Pedhazur, 1982).

Finally, the *structure coefficients* are defined as the correlation coefficients between the input variables and the discriminant scores (Pedhazur, 1982). They are considered by many authors as most appropriate for interpreting the importance of variables for discrimination. Their disadvantage is that the structure coefficients are univariate measures and do not represent the importance of a variable in the presence of other available variables (Rencher, 2002).

Example: Consider the data on 32 Tibetan skulls divided into two groups (1–17 and 18–32) and discussed and studied in Everitt and Dunn (2001). On each skull five measurements (in millimeters) were obtained: greatest length of skull (x_1), greatest horizontal breadth of skull (x_2), height of skull (x_3), upper face height (x_4), and face breadth, between outermost points of cheek bones (x_2).

The data are subject to LDA with SPSS (SPSS, 2001). There is only one canonical variate ($s = 1$) in this example. The value of Fisher's LDA objective function (7) (F value) is 28.012. The output provided in Table 1 gives the criterion generally used to assess each variable's contribution to Hotelling's T^2 and their importance for discrimination.

According to Table 1 one can conclude that the variable x_4 is the most important for discrimination, followed closely by x_1, and then by x_5 some distance behind. The other two variables do not seem interesting for the problem and can be dropped from further analysis.

No such clear decision can be made if one bases the interpretation on the raw and standardized coefficients—see Table 2. For the raw coefficients x_2, which was least important in Table 1, is now most important. This variable is also important for the standardized coefficients, but x_5 now has the largest coefficient. The structure coefficients imply interpretations similar to those deduced from Hotelling's T^2, but the dominance of x_1 and x_4 is less clear-cut. For two groups it can be argued that Hotelling's T^2 provides the best way to interpret the single discriminant function. However in case of several groups, Hotelling's T^2 measures the *overall* contribution of each variable to group separation, and thus is not helpful for interpretation of any particular discriminant function (Rencher, 2002).

For comparison, one can apply stepwise logistic regression and remove from analysis the predictor variable with the largest p-value at each step. For the Skull Data, only a single variable, upper face height (x_4), is selected as reported in Everitt and Dunn (2001). Surprisingly, MINITAB (MINITAB, 2000) logistic regression selects a different single variable, the greatest length of skull (x_1).

A promising new approach for variable selection, which can simplify the interpretation of a linear combination of variables, is to impose additional constraints of the LASSO (**L**east **A**bsolute **S**hrinkage and **S**election **O**perator) type (Hastie et al., 2001). The application to multiple regression was recently extended to PCA (Trendafilov and Jolliffe, 2006). In this paper the close relation between PCA and LDA is employed and the LASSO approach is applied to improve the interpretability of the canonical variates.

3. Orthogonal canonical variates

LDA does not provide orthogonal projection of the data, as PCA does, because "the canonical variate space is derived by deforming the axes in the original data space" (Krzanowski, 2003). If orthogonal projections between the original data space are sought for maximal discrimination of the existing groups, then LDA needs to be modified in a PCA fashion (Jolliffe, 2002). For this reason the basic LDA problem (2):

$$\max_{\mathbf{a}_i} \frac{\mathbf{a}_i^{\mathrm{T}} \mathbf{C}_{\mathbf{B}} \mathbf{a}_i}{\mathbf{a}_i^{\mathrm{T}} \mathbf{C}_{\mathbf{W}} \mathbf{a}_i} \quad \text{subject to } \mathbf{a}_i^{\mathrm{T}} \mathbf{C}_{\mathbf{W}} \mathbf{a}_i = 1 \text{ and } \mathbf{a}_i^{\mathrm{T}} \mathbf{C}_{\mathbf{W}} \mathbf{a}_j = 0, \tag{14}$$

for $i = 1, 2, \ldots, s$; $i \neq j$, is replaced in Krzanowski (1995) by the following blend between PCA and LDA, namely a LDA objective function subject to PCA constraints:

$$\max_{\mathbf{a}_i} \frac{\mathbf{a}_i^{\mathrm{T}} \mathbf{C}_{\mathbf{B}} \mathbf{a}_i}{\mathbf{a}_i^{\mathrm{T}} \mathbf{C}_{\mathbf{W}} \mathbf{a}_i} \quad \text{subject to } \mathbf{a}_i^{\mathrm{T}} \mathbf{a}_i = 1 \text{ and } \mathbf{a}_i^{\mathrm{T}} \mathbf{A}_{i-1} = \mathbf{0}_{i-1}^{\mathrm{T}}, \tag{15}$$

where the matrix $\mathbf{A}_{i-1}$ is composed of all preceding vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{i-1}$, i.e. $\mathbf{A}_{i-1}$ is the $p \times (i-1)$ matrix defined as $\mathbf{A}_{i-1} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{i-1})$. The solutions $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_s)$ are called *orthogonal* canonical variates.

In these notations, the LDA problem (14) can be rewritten in a PCA-like form as suggested in (12):

$$\max_{\mathbf{a}_i} \mathbf{a}_i^{\mathrm{T}} \mathbf{U}^{-\mathrm{T}} \mathbf{C}_{\mathbf{B}} \mathbf{U}^{-1} \mathbf{a}_i \quad \text{subject to } \mathbf{a}_i^{\mathrm{T}} \mathbf{a}_i = 1 \text{ and } \mathbf{a}_i^{\mathrm{T}} \mathbf{A}_{i-1} = \mathbf{0}_{i-1}^{\mathrm{T}}, \tag{16}$$

where $\mathbf{U}$ is the positive definite upper triangular matrix from the Cholesky factorization of $\mathbf{C}_{\mathbf{W}}$, i.e. $\mathbf{C}_{\mathbf{W}} = \mathbf{U}^{\mathrm{T}}\mathbf{U}$. One keeps in mind that the (orthogonal) solution $\mathbf{A}$ of (16) is used in turn to find the raw coefficients $\mathbf{A}_{\mathrm{raw}} = \mathbf{U}^{-1}\mathbf{A}$. This reformulation of (14) in a PCA-like format will be used hereafter.

Thus, there are two very similar, PCA-like problems, to be solved. The difference is in the objective functions to be maximized.

The LDA problem (16) to find standard canonical variates can be solved by a number of well-known algorithms (Golub and Van Loan, 1991). The algorithm proposed by Krzanowski (1995) for finding the orthogonal canonical variates defined in (15) is a sequential one inspired by PCA and works in the same manner: find an unit vector $\mathbf{a}_1$ maximizing the objective function, then form the linear subspace orthogonal to $\mathbf{a}_1$ and find an unit vector $\mathbf{a}_2$ from this subspace which maximizes the objective function and so on.

Example: (continued) In Table 3 are given the orthogonal canonical variate (raw) coefficients, and the corresponding structure coefficients. The objective function (15) at this solution is, as before, 28.012. The raw coefficients are again

Table 3
Orthogonal canonical variates for Skull Data

Vars.	Raw coefficients	Structure coefficients
x_1	.290	.851
x_2	−.505	−.066
x_3	−.017	.332
x_4	.574	.900
x_5	.575	.701

difficult to interpret, while the structure coefficients are quite similar to those from Table 2. Again it is difficult to interpret the discrimination contribution of the original variables based on the raw coefficients' magnitudes.

4. Gradient ascent flows for canonical and orthogonal canonical variates

To achieve more easily interpretable canonical variates, additional LASSO constraints:

$$\|\mathbf{a}_i\|_1 \leqslant t_i \quad \text{for } i = 1, 2, \ldots, s \quad \text{with } t_i \in [1, \sqrt{p}], \tag{17}$$

can be imposed on the standard LDA problems from the previous section. In this way the loadings $\mathbf{a}_i$ are penalized to take only few non-zero values as the threshold parameters t_i are decreased from $\sqrt{p}$ to 1. Then the LDA problems become:

$$\max_{\mathbf{a}} \mathbf{a}^{\mathrm{T}}\mathbf{U}^{-\mathrm{T}}\mathbf{C}_{\mathbf{B}}\mathbf{U}^{-1}\mathbf{a} \tag{18}$$

and

$$\max_{\mathbf{a}} \frac{\mathbf{a}^{\mathrm{T}}\mathbf{C}_{\mathbf{B}}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{C}_{\mathbf{W}}\mathbf{a}} \tag{19}$$

both

$$\text{subject to } \|\mathbf{a}\|_1 \leqslant t,\ \|\mathbf{a}\|_2^2 = 1 \text{ and } \mathbf{a}^{\mathrm{T}}\mathbf{A}_{i-1} = \mathbf{0}_{i-1}^{\mathrm{T}}. \tag{20}$$

A standard way to eliminate the LASSO inequality constraint in (20) is by introducing an exterior penalty function P into the objective functions to be maximized. The idea is to penalize an unit vector $\mathbf{a}$ which does not satisfy the LASSO constraint by reducing the value of the new objective function. Thus, the LDA problems are modified as follows:

$$\max_{\mathbf{a}}[\mathbf{a}^{\mathrm{T}}\mathbf{U}^{-\mathrm{T}}\mathbf{C}_{\mathbf{B}}\mathbf{U}^{-1}\mathbf{a} - \mu P(\|\mathbf{a}\|_1 - t)] \tag{21}$$

and

$$\max_{\mathbf{a}} \left[\frac{\mathbf{a}^{\mathrm{T}}\mathbf{C}_{\mathbf{B}}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{C}_{\mathbf{W}}\mathbf{a}} - \mu P(\|\mathbf{a}\|_1 - t)\right] \tag{22}$$

both

$$\text{subject to } \|\mathbf{a}\|_2^2 = 1 \text{ and } \mathbf{a}^{\mathrm{T}}\mathbf{A}_{i-1} = \mathbf{0}_{i-1}^{\mathrm{T}}. \tag{23}$$

The exterior penalty function P is zero if the LASSO constraint is fulfilled. It "switches on" the penalty μ (a large positive number) if the LASSO constraint is violated. Moreover the more severe violations are penalized more heavily. A typical example of an exterior penalty function for inequality constraints is the Zangwill penalty function $P(x) = \max(0, x)$, which will be used hereafter.

The new LDA problems (21)–(23) and (22)–(23) require maximization of more complicated objective functions but subject to the well-known constraint from PCA. Because of the PCA-like nature of (23) it seems natural to attack the problems sequentially following the PCA tradition.

Such an approach was already proposed in Trendafilov and Jolliffe (2006) for solving a similar problem: PCA subject to additional LASSO constraint. This is a sequential algorithm based on the variational PCA formulation (Jolliffe, 2002) but translated into a sequence of dynamical systems. According to the variational definition of PCA the vector of loadings $\mathbf{a}_i$ for the ith principal component of a correlation matrix $\mathbf{R}$, is a vector that solves:

$$\text{Maximize} \quad \mathbf{a}^{\mathrm{T}}\mathbf{R}\mathbf{a}, \tag{24}$$

$$\text{subject to} \quad \|\mathbf{a}\|_2 = 1 \text{ and } \mathbf{a}^{\mathrm{T}}\mathbf{A}_{i-1} = \mathbf{0}_{i-1}^{\mathrm{T}}. \tag{25}$$

This maximizer can also be found as a solution of the initial value problem for the following vector ordinary differential equation:

$$\frac{\mathrm{d}\mathbf{a}_i}{\mathrm{d}t} = \mathbf{\Pi}_i \nabla_{\mathbf{a}^{\mathrm{T}}\mathbf{R}\mathbf{a}}(\mathbf{a}_i), \tag{26}$$

starting with an appropriate initial value $\mathbf{a}_{i,in}$ with $\|\mathbf{a}_{i,in}\|_2^2=1$. Note that $\nabla_{\mathbf{a}^{\mathrm{T}}\mathbf{Ra}}(\mathbf{a})$ is the gradient of the PCA function $\mathbf{a}^{\mathrm{T}}\mathbf{Ra}$ to be maximized with respect to the standard Frobenius (Euclidean) matrix norm (Golub and Van Loan, 1991). The projector $\mathbf{\Pi}_i$ in (26) is defined as follows:

$$\mathbf{\Pi}_i = \mathbf{I}_p - \mathbf{A}_i\mathbf{A}_i^{\mathrm{T}}. \tag{27}$$

Note that the orthogonality constraints are automatically fulfilled as the ith consequent ascent gradient vector flow (26) is defined on an unit sphere in $\mathbb{R}^p$ and orthogonal to all preceding $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{i-1}$. The reason for this reformulation is that the additional LASSO constraint can be easily incorporated. One needs to solve

$$\frac{\mathrm{d}\mathbf{a}_i}{\mathrm{d}t} = \mathbf{\Pi}_i \nabla_{\mathbf{a}^{\mathrm{T}}\mathbf{Ra}-\mu P(\|\mathbf{a}\|_1-t)}(\mathbf{a}_i), \tag{28}$$

instead of (26) as in the ordinary PCA case (Trendafilov and Jolliffe, 2006). The problem is that both the penalty function P and the LASSO constraint are not differentiable and thus the gradient ∇ in (28) cannot be computed. This is overcome in Trendafilov and Jolliffe (2006) by their smoothing

$$\|\mathbf{a}\|_1 = \mathbf{a}^{\mathrm{T}}\,\mathrm{sign}(\mathbf{a}) \approx \mathbf{a}^{\mathrm{T}}\tanh(\gamma\mathbf{a}) \tag{29}$$

and

$$P(x) = \max\,(0, x) \approx \frac{x(1+\tanh(\gamma x))}{2}, \tag{30}$$

for some large γ, e.g. $\gamma = 1000$. The same continuous-time algorithm can be readily applied for solving the LDA problems (21)–(23) and (22)–(23). Let the functions to be maximized in (21) and (22) be denoted by $F_\mu(\mathbf{a})$:

$$F_\mu(\mathbf{a}) = \mathbf{a}^{\mathrm{T}}\mathbf{U}^{-\mathrm{T}}\mathbf{C}_{\mathbf{B}}\mathbf{U}^{-1}\mathbf{a} - \mu P(\mathbf{a}^{\mathrm{T}}\tanh(\gamma\mathbf{a}) - t) \tag{31}$$

and

$$F_\mu(\mathbf{a}) = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{C}_{\mathbf{B}}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{C}_{\mathbf{W}}\mathbf{a}} - \mu P(\mathbf{a}^{\mathrm{T}}\tanh(\gamma\mathbf{a}) - t). \tag{32}$$

Then both the standard and orthogonal canonical variates can be found as s consequent ascent gradient vector flows, each of them defined on an unit sphere in $\mathbb{R}^p$ and orthogonal to all preceding canonical variates. The loadings for the canonical variates $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_s$ can be computed as solutions of s consequent initial value problems for the following vector ordinary differential equations:

$$\frac{\mathrm{d}\mathbf{a}_i}{\mathrm{d}t} = \mathbf{\Pi}_i \nabla_{F_\mu(\mathbf{a}_i)}(\mathbf{a}_i), \tag{33}$$

starting with an appropriate initial value $\mathbf{a}_{i,in}$ with $\|\mathbf{a}_{i,in}\|_2^2=1$ for $i = 1, 2, \ldots, s$. The gradient of the penalty function P is the same as for PCA and is given in Trendafilov and Jolliffe (2006).

The values of the parameters μ and γ control the accuracy of preserving the LASSO constraint. The parameter μ prevents the cases $\|\mathbf{a}\|_1 > t$, while γ keeps $\|\mathbf{a}\|_1$ as close as possible to t from below and, thus, the best possible maximum within this particular constraint. In general, by increasing μ and γ one improves the solution, but also increases the CPU time required. In practice, one needs to find the "saturation" point for the problem, after which any further increase of μ and γ seems unreasonable with respect to the increase of the CPU time required.

The variation of the CPU time required for the data sets considered in this work is not considerable for different values of μ and γ. Nevertheless, one should clearly distinguish the roles of these parameters. The parameter γ controls how well the absolute value function in the LASSO constraints (17) can be approximated by tanh. It has nothing to do with the data of the particular discrimination problem and thus, some universal value can be adopted. A reasonable choice is $\gamma = 1000$. However, the parameter μ is data related as it controls the objective function of the particular discrimination problem. Its value should be evaluated individually for each data set.

In Fig. 1 there are four plots of the ℓ_1 norms of the DALASS canonical variates (left panel), and of the maxima achieved of the LDA function (2) as functions of μ and γ both in the range from 0 to 2000 with a step 200. The top two plots regard the Skull Data. One can see that changing μ does not really change the way the LASSO constraint

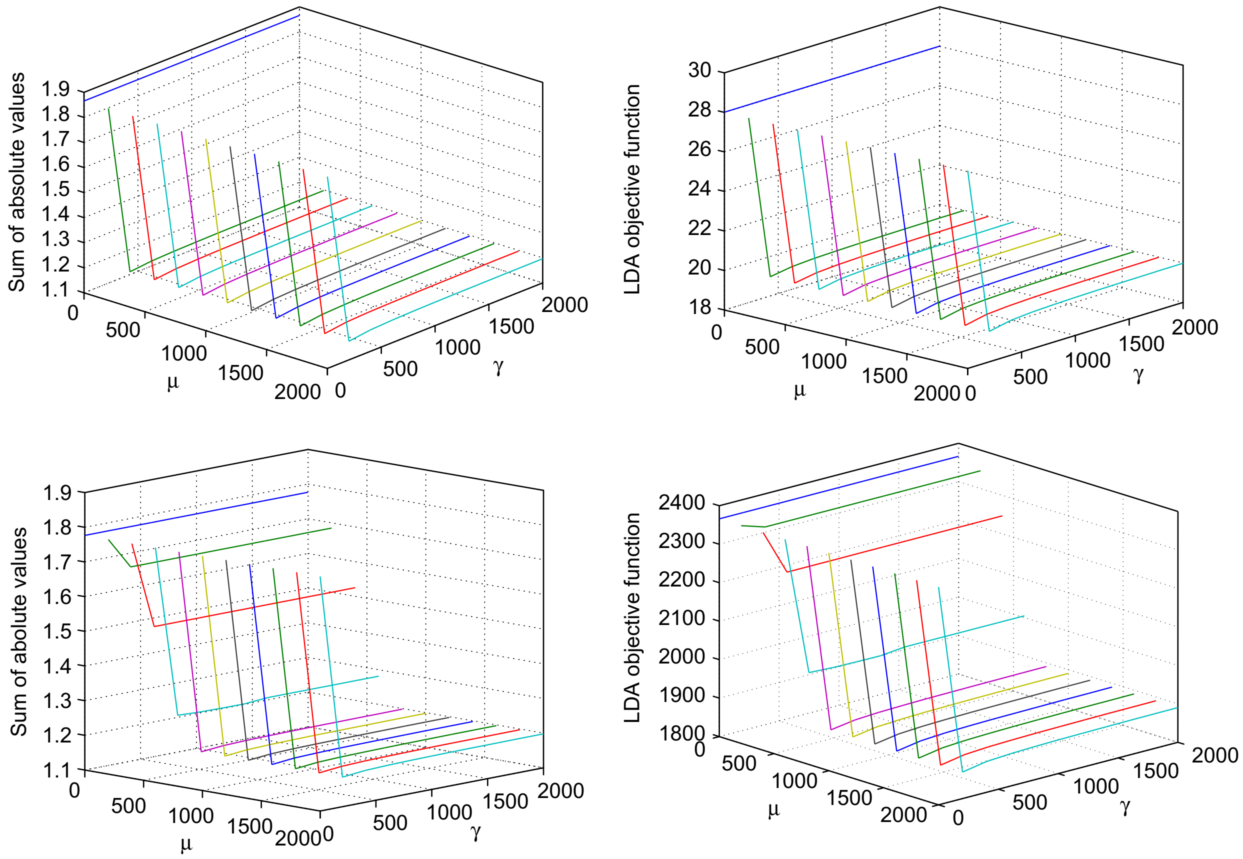

Fig. 1. The sum of the absolute values of the raw coefficients of the DALASS canonical variate (left), and the achieved maximum of the LDA function (2) as functions of μ and γ. The top two plots regard the Skull Data; the bottom two—the first canonical variate of the Fisher's Iris Data from Section 5.1.

Table 4
DALASS canonical variates for Skull Data ($t = 1.2$)

Vars.	Raw coefficients	Standardized coefficients	Structure coefficients
x_1	.108	.829	.987
x_2	−.005	−.032	.099
x_3	.002	.011	.422
x_4	.053	.228	.848
x_5	.006	.036	.612

is fulfilled; for small γ the LASSO constraint is severely violated, but after $\gamma = 500$ it is pretty well preserved. The achieved maxima of the LDA objective function exhibit the same behavior (right). These identical profiles suggest that a very low value of μ will suffice. Indeed, the DALASS canonical and orthogonal canonical variates reported later on in Tables 4 and 5 can be found with $\mu = 10$ (and $\gamma = 1000$). Note that $\mu = 0$ and/or $\gamma = 0$ correspond to the standard LDA solution.

As one can see from the bottom two plots of Fig. 1, the situation is quite different for the Fisher's Iris Data considered in Section 5.1. The DALASS canonical and orthogonal canonical variates are obtained with $\mu = 800$ and 1500, respectively. For the Vowel Data considered in Section 5.2 the corresponding parameters are $\mu = 50$ and 80.

Example: (continued) DALASS is applied to the Skull Data with tuning parameter $t = 1.2$ and the solution is given in Table 4. According to both raw and standardized coefficients one picks up a single variable, greatest length of skull

Table 5
DALASS orthogonal canonical variates for Skull Data ($t = 1.2$)

Vars.	Raw coefficients	Structure coefficients
x_1	.110	.822
x_2	−.053	.057
x_3	.000	.321
x_4	.992	.993
x_5	.038	.629

(x_1), as most important for the group discrimination in this data set. The structure coefficients are not helpful as they suggest several variables be taken into account and thus complicate the interpretation. This result is in concordance with the finding from MINITAB logistic regression. The objective function at this solution is 19.933.

The DALASS orthogonal canonical variates given in Table 5 provide even clearer interpretation: the raw coefficients suggest a single variable, the upper face height (x_4), as important for discriminating between the groups in these Skull Data. This result is in concordance with the logistic regression finding reported in Everitt and Dunn (2001). The objective function at this solution is 22.050—a smaller drop compared to 28.012. Thus the orthogonal canonical variates can be considered a better solution of the problem. As before the corresponding structure coefficients are clearly more difficult to interpret. DALASS works very well for the Skull Data and the interpretation of the results can be entirely based on the discriminant function coefficients.

5. Further examples

5.1. *Fisher's Iris Data (Fisher, 1936)*

Illustration of DALASS application on the famous Iris data set (Fisher, 1936) is considered in this section. This data set is four-dimensional and the four dimensions are: sepal length (x_1), sepal width (x_2), petal length (x_3) and petal width (x_4). It contains 50 observations in each of the three groups of plants: *Iris setosa*, *Iris versicolor and Iris virginica*.

In Tables 6 and 7 are given the standard and the orthogonal two-dimensional ($s = r = 2$) canonical variates solutions. Fisher's LDA objective function $F(\mathbf{a})$ (7) for the first standard CV is 2366.11. The value at the second CV is 20.98 (the third and fourth are 0), so the relative importance of the second CV is less than 1%. The objective values at the orthogonal CVs are 2366.11 and 705.50 (total 3071.61), i.e. the second CV is considerably more important for this analysis.

These two solutions are depicted in Figs. 2 and 3, respectively. The three groups are well separated in both of the plots. The latter seems to be superior because the groups are more compact. More objective comparison of the quality of these two discriminations can be achieved by applying them to classify the original observations. The standard CV solution misclassifies three observations: 52 (3), 103 (2) and 104 (2) with error rate of 2%, while the orthogonal CV solution misclassifies four: 40 (2), 78 (3), 104 (2) and 121 (3) with error rate of 2.67%. Unfortunately, the discriminant functions coefficients for both of the solutions do not provide a basis for their unique and simple interpretation.

DALASS with tuning parameter $t = 1.2$ is applied to obtain both standard and orthogonal two-dimensional canonical variates solutions. They are given in Tables 8 and 9 and depicted in Figs. 4 and 5, respectively.

The standardized coefficients of the DALASS canonical variates suggest that they are both mainly composed of two of the original variables: the first CV is dominated by sepal length, x_1, and petal length, x_3, and the second CV by sepal length, x_1, and sepal width, x_2. One can conclude that the discrimination between the three groups in the *Iris* Data can be based on the length of the flowers, and the sepal size. The objective values at the CVs are 1888.10 and 255.60 (total 2143.71), i.e. the second CV is considerably more important (13.5%) for this analysis than for the standard CV from Table 6. The total value of the objective function is reasonably high compared to the original 2387.08, so this solution can be considered as quite successful.

The raw coefficients of the DALASS orthogonal canonical variates suggest even simpler interpretations. Both CVs are mainly composed of a single original variable: the first CV is dominated by petal width, x_4, and the second CV by petal length, x_3. According to this result one can discriminate between the three groups in the *Iris* Data based on the

Table 6
Canonical variates for Fisher's Iris Data

Vars.	Raw coefficients		Standardized coefficients		Structure coefficients	
x_1	−.08	−.00	−.43	−.01	.79	−.22
x_2	−.15	−.22	−.52	−.74	−.53	−.76
x_3	.22	.09	.95	.40	.98	−.05
x_4	.28	−.28	.58	−.58	.97	−.22

Table 7
Orthogonal canonical variates for Fisher's Iris Data

Vars.	Raw coefficients		Structure coefficients	
x_1	−.21	−.15	.79	.84
x_2	−.39	.04	−.53	−.48
x_3	.55	.76	.98	.99
x_4	.71	−.62	.97	.91

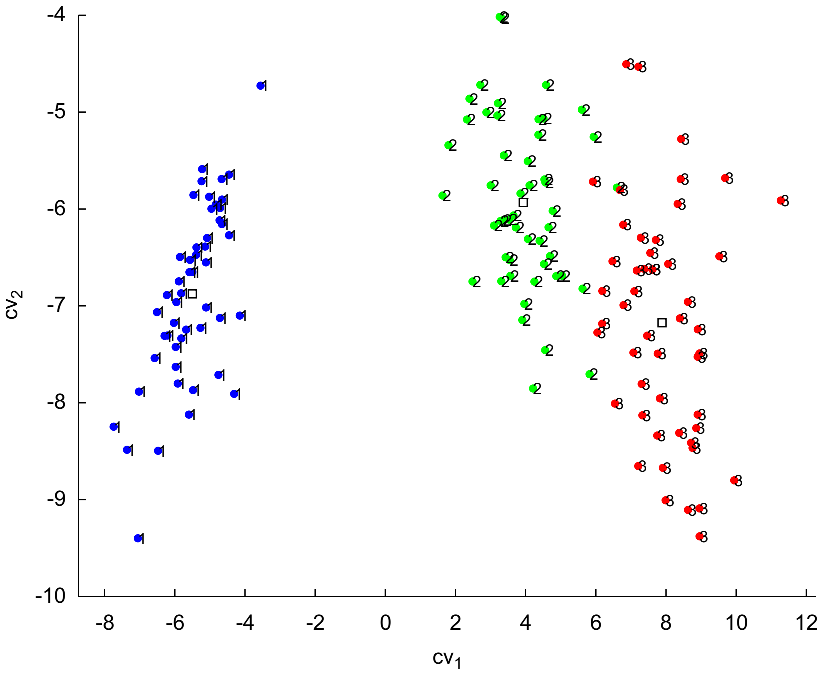

Fig. 2. *Iris* Data plotted against the first two canonical variates: 1 = *Iris setosa*; 2 = *Iris versicolor*; 3 = *Iris virginica*. Squares denote group means.

petal size only. This interpretation is surprisingly simple and also seems reliable. Indeed, the objective values at the DALASS orthogonal CVs are 1430.15 and 1348.91 (total 2779.06, which is only a 10% drop from the original total of 3071.61). Note that the orthogonal CVs are nearly equally important for this analysis. An appropriate orthogonal rotation of the CVs can make the discrimination problem essentially one-dimensional.

The quality of these DALASS discriminations is assessed by using them to classify the original observations. The DALASS CVs solution misclassifies four observations: 63 (2), 103 (2), 104 (2) and 121 (3) with error rate of 2.67%, while the DALASS orthogonal CVs solution misclassifies five: 9 (3), 50 (2), 78 (3), 104 (2) and 121 (3) with error rate of 3.33%, only marginally worse than for the standard analyzes.

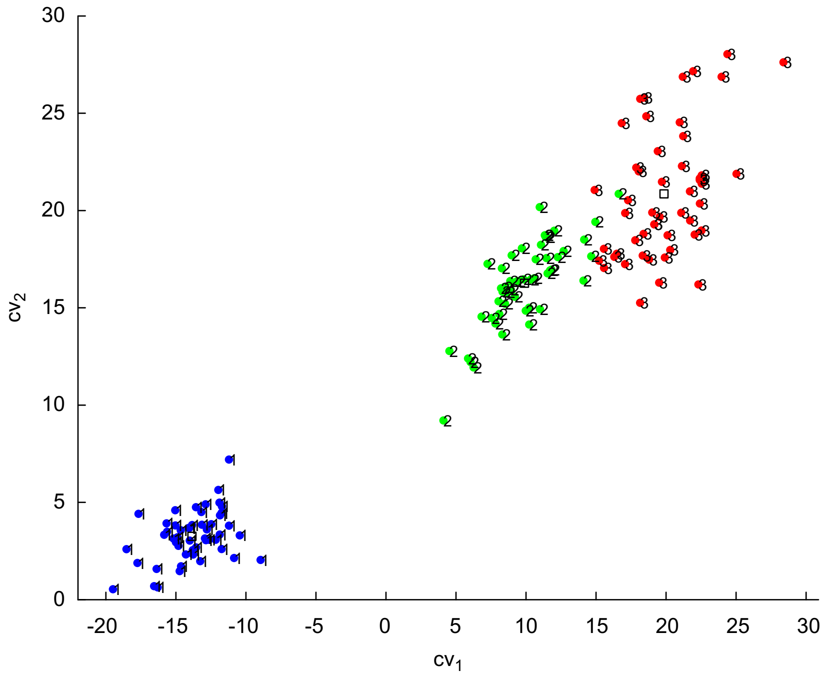

Fig. 3. *Iris* Data plotted against the first two orthogonal canonical variates.

Table 8
DALASS canonical variates for Fisher's Iris Data ($t = 1.2$)

Vars.	Raw coefficients		Standardized coefficients		Structure coefficients	
x_1	−.21	−.16	−1.09	−.83	.77	−.72
x_2	−.02	.34	−.06	1.15	−.51	.78
x_3	.31	.00	1.35	.00	.98	−.85
x_4	.13	−.00	.27	−.00	.97	−.78

Table 9
DALASS orthogonal canonical variates for Fisher's Iris Data ($t = 1.2$)

Vars.	Raw coefficients		Structure coefficients	
x_1	−.00	−.15	.81	.85
x_2	−.15	−.00	−.44	−.45
x_3	.06	.99	.97	1.00
x_4	.99	−.06	1.00	.96

5.2. *Vowel Data (Hastie et al., 2001)*

The vowel recognition data set, Vowel Data for short, is constructed from speaker independent recognition of the 11 steady state vowels of British English. The words (heed, hid, head, had, hard, hud, hod, hoard, hood, who'd, heard) were uttered by each of 15 speakers. Four male and four female speakers were used for training, and the other four male and three female speakers were used for testing the performance. Each vowel is eventually represented as a six-dimensional vector, resulting in training and test data sets of 528 (11 × 6 × 8) and 462 (11 × 6 × 7) records. For a more detailed

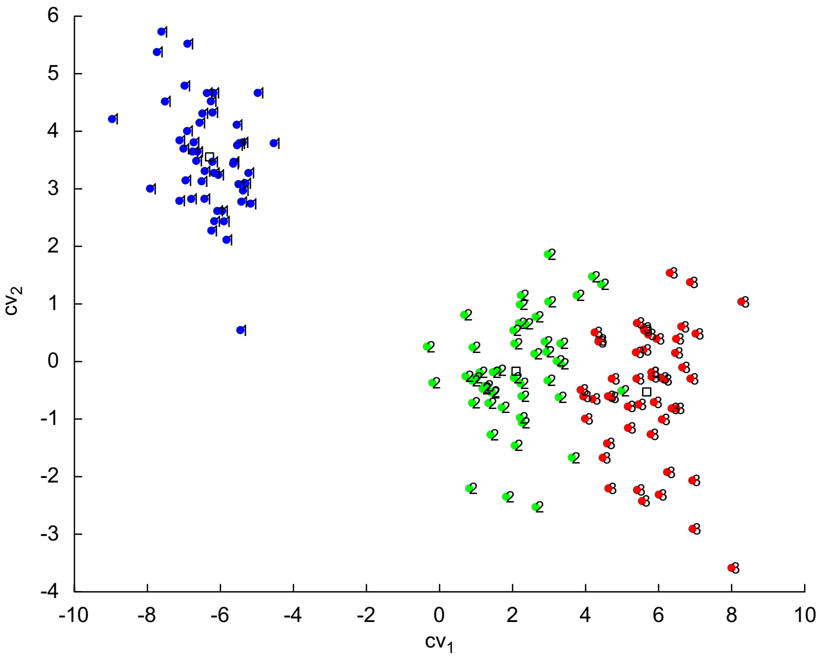

Fig. 4. *Iris* Data plotted against the first two DALASS canonical variates with tuning parameter 1.2.

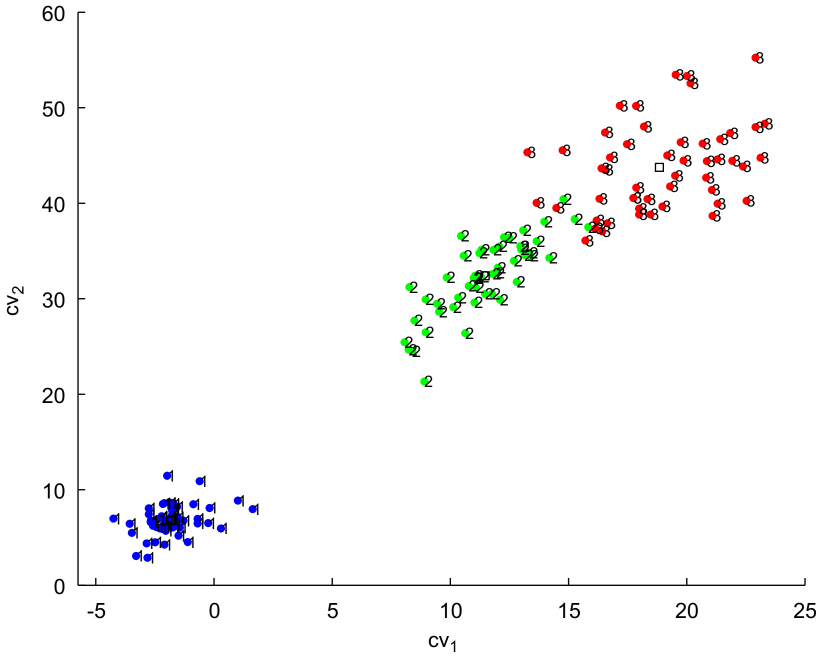

Fig. 5. *Iris* Data plotted against the first two DALASS orthogonal canonical variates with tuning parameter 1.2.

Table 10
Canonical variates for Vowel Data

Vars.	Raw coefficients		Standardized coefficients		Structure coefficients	
x_1	.90	1.08	.61	.73	.67	.45
x_2	−1.15	.35	−.79	.24	−.83	.37
x_3	−.54	−.48	−.38	−.34	−.26	−.11
x_4	−.02	−.61	−.02	−.44	.24	−.29
x_5	.01	−1.61	.00	−.87	.43	−.60
x_6	−.71	−1.48	−.40	−.83	.23	−.52
x_7	−.84	−.87	−.39	−.40	−.18	.04
x_8	−1.31	−.94	−.71	−.51	−.39	.07
x_9	−.97	−.51	−.57	−.30	−.33	.17
x_{10}	−.35	−.15	−.19	−.08	.12	.05

Table 11
Orthogonal canonical variates for Vowel Data

Vars.	Raw coefficients		Structure coefficients	
x_1	.36	−.03	.67	−.21
x_2	−.45	.39	−.83	.86
x_3	−.21	−.09	−.26	.12
x_4	−.01	−.30	.24	−.37
x_5	.00	−.73	.43	−.72
x_6	−.28	−.40	.23	−.53
x_7	−.33	−.22	−.18	.14
x_8	−.52	.01	−.39	.31
x_9	−.38	.04	−.33	.34
x_{10}	−.14	−.06	.12	−.07

explanation of the experiment, the problem, and the original references see (Hastie et al., 2001) and the related web site http://www-stat-class.stan-ford.edu/∼tibs/ElemStatLearn/.

DALASS is applied to the training data set. First the standard LDA is applied. The maximum number of canonical variates that could be used in this analysis is $r = 10$. The first two eigenvalues give more than 91% of the total sum of the non-zero eigenvalues, so we restrict attention to these $s = 2$ dimensions. The data can be plotted against the first two canonical variates. In Tables 10 and 11 are given the standard and the orthogonal two-dimensional canonical variates solutions. Fisher's LDA objective function (7) for the first CV is 209.49 and for the second CV is 131.23 (total 340.71), i.e. both CVs are important for this analysis. The objective values at the orthogonal CVs are 209.49 and 160.87 with a total of 370.36.

These two solutions are depicted on Figs. 6 and 7, respectively. The eleven groups on the plot Fig. 7 using two orthogonal variates are made more compact and thus seem collapsed and more difficult to distinguish. Unfortunately, the discriminant function coefficients for both of the solutions have too many original variables included non-trivially for there to be a simple interpretation.

Now standard and orthogonal CVs are obtained applying DALASS with $t = 1.8$. In Tables 12 and 13 are given the standard and the orthogonal two-dimensional DALASS canonical variates solutions. The objective function at the two standard CVs is 171.07 and 108.10, with a total of 279.16. The objective function at the two orthogonal CVs is 164.34 and 151.57. The total is 315.57, and thus both DALASS representations have nearly the same discrimination quality—about 85% of the original one.

These two solutions are depicted on Figs. 8 and 9, respectively. Despite the drop of 15% in the value of the total objective function these DALASS solutions provide clearer representation of the groups. The DALASS canonical variates provide more easily interpreted discriminant functions and standardized coefficients. The corresponding graphical representations are also clearer. They both suggest that the first CV is composed mainly of x_2 and x_5, and the second

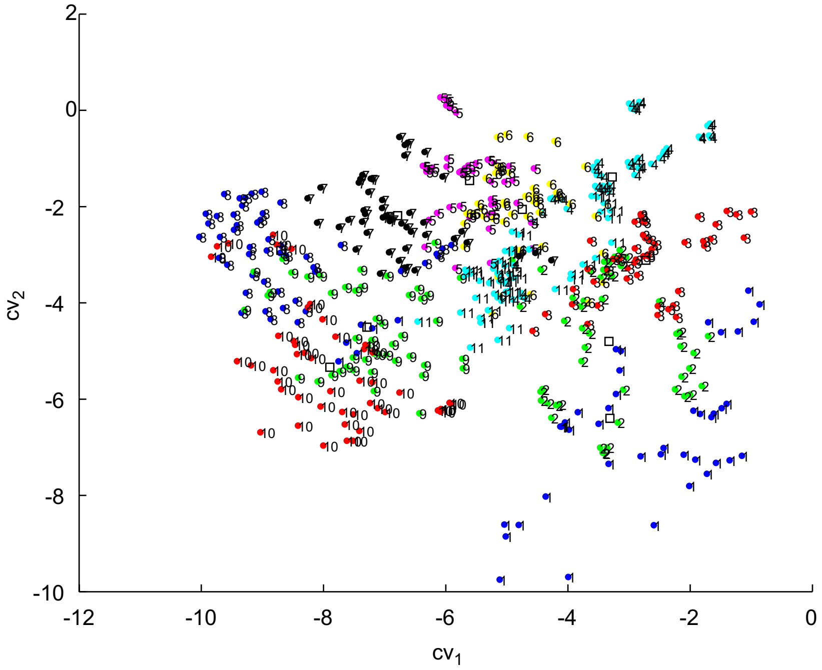

Fig. 6. *Vowel* training data plotted against the first two canonical variates.

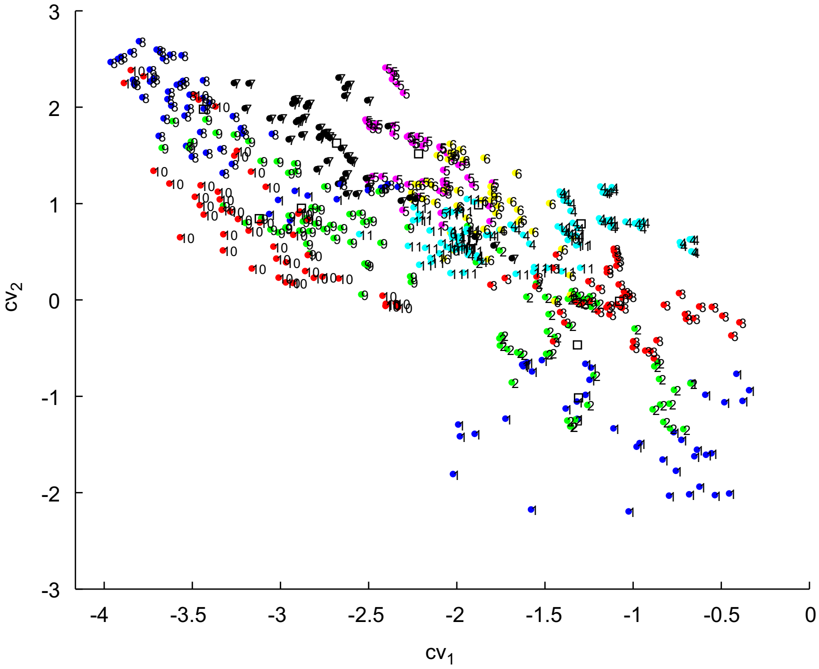

Fig. 7. *Vowel* training data plotted against the first two orthogonal canonical variates.

Table 12
DALASS canonical variates for Vowel Data ($t = 1.8$)

Vars.	Raw coefficients		Standardized coefficients		Structure coefficients	
x_1	.37	1.17	.25	.79	.39	.77
x_2	−1.07	−.32	−.73	−.22	−.92	−.17
x_3	.05	−1.03	.03	−.73	−.07	−.33
x_4	.46	−.27	.33	−.20	.40	−.10
x_5	1.02	−.65	.56	−.36	.64	−.24
x_6	−.05	-1.85	−.03	−1.04	.34	−.31
x_7	−.16	−.16	−.07	−.07	−.25	.06
x_8	−.36	−.77	−.20	−.42	−.37	−.13
x_9	−.41	−.00	−.24	−.00	−.32	.10
x_{10}	−.00	−.00	−.00	−.00	.17	.09

Table 13
DALASS orthogonal canonical variates for Vowel Data ($t = 1.8$)

Vars.	Raw coefficients		Structure coefficients	
x_1	.61	.00	.74	−.23
x_2	−.25	.49	−.76	.86
x_3	.00	.00	−.28	.09
x_4	.00	−.17	.14	−.33
x_5	.00	−.83	.33	−.75
x_6	−.00	−.15	.22	−.45
x_7	−.00	−.00	−.04	.20
x_8	−.72	−.16	−.44	.26
x_9	−.22	−.00	−.24	.34
x_{10}	−.00	−.00	.13	−.07

CV of x_1, x_3 and x_6. The DALASS orthogonal canonical variates are again easier to interpret: the first CV is composed mainly of x_1 and x_8, and the second CV of x_2 and x_5.

The standard CVs wrongly classify 35.04% of the original observations, while the orthogonal CVs misclassify 42.05%. The DALASS CVs wrongly classify 39.77% of the original observations, while the DALASS orthogonal CVs misclassify 40.15%.

Now one can check the performance of the constructed discriminant rules (based on the training data set) on the test data set of 462 records. The standard CVs wrongly classify 49.13% of the test observations, while the orthogonal CVs—51.73%. The DALASS CVs wrongly classify 42.21% of the test observations, while the DALASS orthogonal CVs—54.11%. The classification of the test data based on the DALASS CVs (and depicted in Fig. 10) is surprisingly successful with 57.79% correctly classified observations, compared to 54.76% which is the highest reported in the original study (see Hastie et al., 2001), obtained using square node network classifier.

6. Choice of tuning parameter t

The choice of the tuning parameter t is very important for obtaining an interpretable solution that provides a reasonable fit to the data. The following heuristic procedure is employed and is also based on the DALASS numerical algorithm which is applied sequentially. As has been explained in Trendafilov and Jolliffe (2006), if one needs to find a solution for some tuning parameter $t_0 \in [1, \sqrt{p}]$ then a list of steps $t_{init} = \sqrt{p} > t' > t'' > \cdots > t_0$ is created first, and then DALASS solves a sequence of problems corresponding to this list of tuning values. Of course, the solution of DALASS with $t = \sqrt{p}$ is in fact solution of the original LDA problem. In order to decide which tuning parameter might be most appropriate for a particular problem one can plot the values of the LDA objective function for the first and second canonical variates, as well as their totals, and the number of the "zero" loadings for each of the solutions/tuning

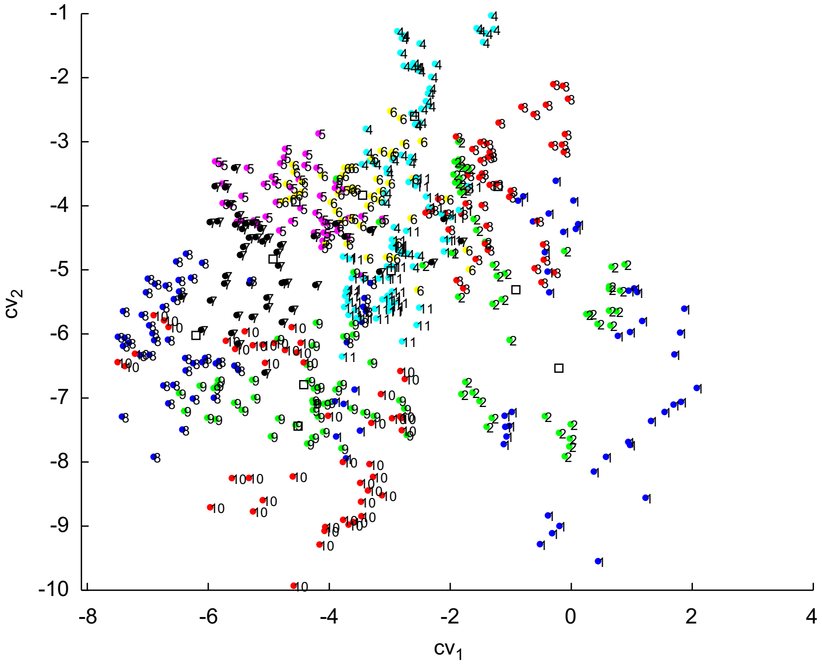

Fig. 8. *Vowel* training data plotted against the first two DALASS canonical variates with tuning parameter 1.8.

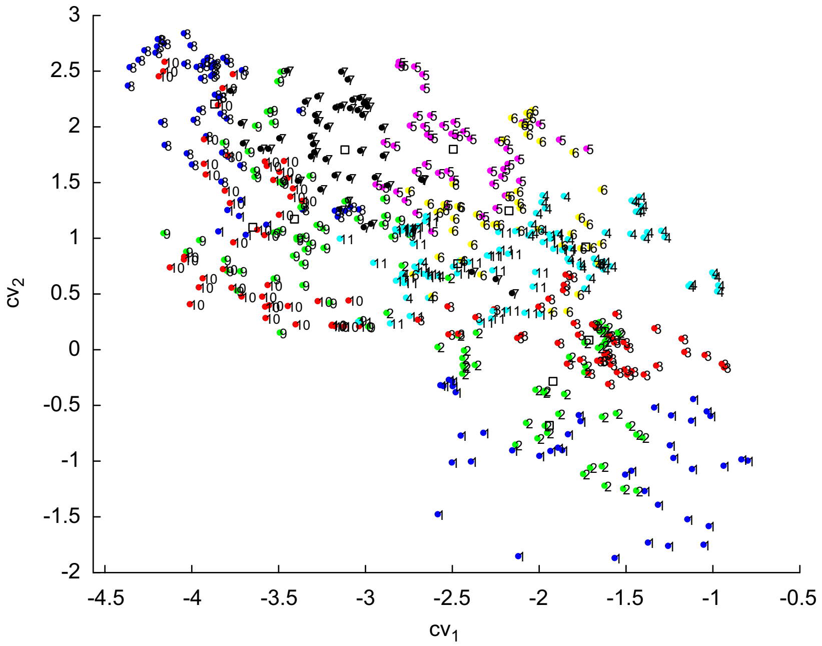

Fig. 9. *Vowel* training data plotted against the first two DALASS orthogonal canonical variates with tuning parameter 1.8.

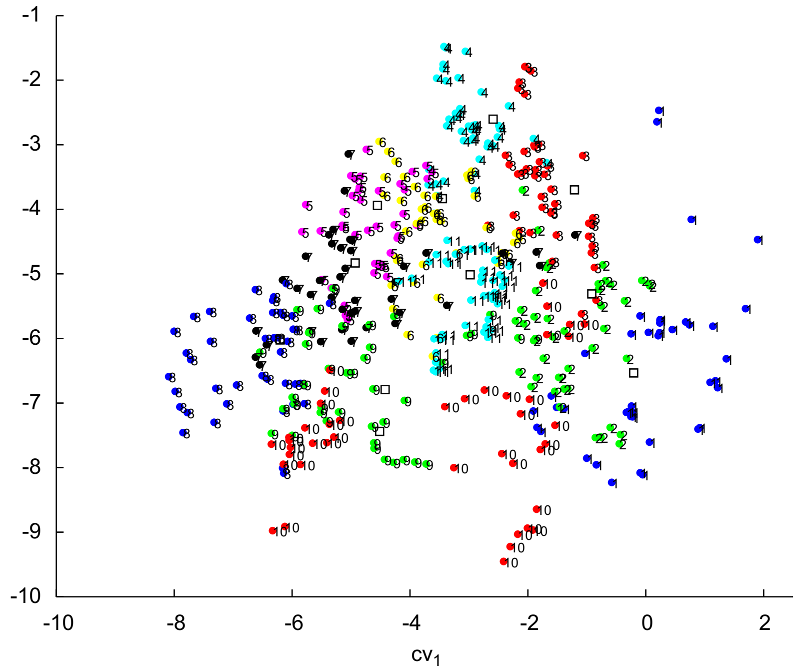

Fig. 10. *Vowel* test data plotted against the first two DALASS canonical variates with tuning parameter 1.8.

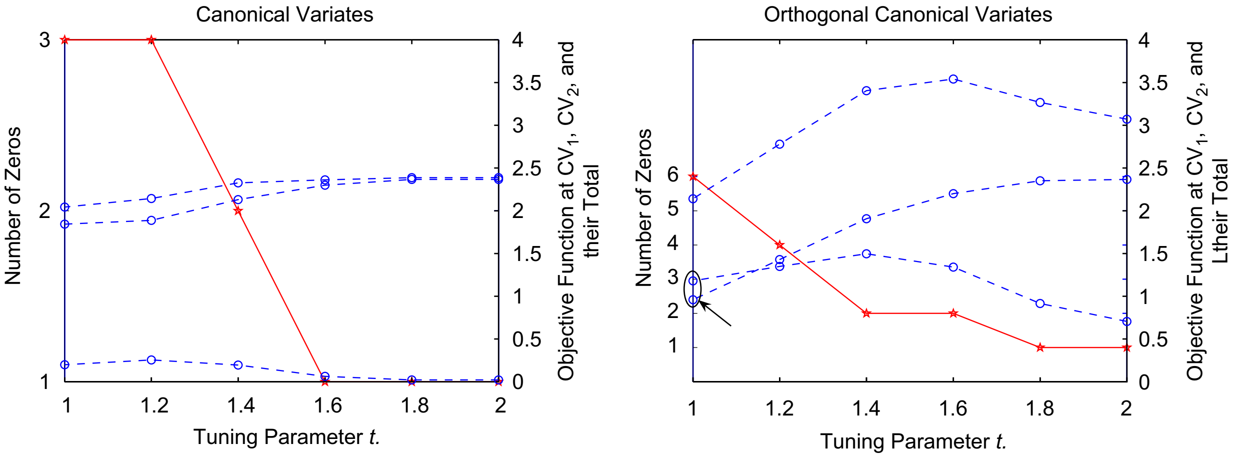

Fig. 11. The values of the LDA objective function at the first and the second CVs, and their totals are plotted by blue "○"; the number of the loadings smaller than .06 in magnitude is denoted by red "⋆".

parameters. Then one looks for the tuning parameter t which produces largest amount of "zero" loadings for reasonable drop of the objective function.

For illustration, consider the choice of the tuning parameter t for the Iris Data. The "zero" loadings are assumed to have magnitude less then .06. The following Fig. 11 is helpful to understand the choice of $t = 1.2$ for the Iris Data for both the standard and orthogonal canonical variates.

Note from the right-hand side plot for the orthogonal CVs that at $t = 1$ the DALASS produces solution with 6 "zero" loadings. This extreme solution is given in Table 14.

Table 14
DALASS orthogonal canonical variates for Fisher's Iris Data ($t = 1$)

Vars.	Raw coefficients		Structure coefficients	
x_1	−.00	−.00	.82	.87
x_2	−.00	−.00	−.37	−.43
x_3	.00	1.00	.96	1.00
x_4	1.00	−.00	1.00	.96

Obviously, one might prefer this 0–1 solution for its simplicity, sacrificing an additional portion of the fit. Such types of solution are especially interesting for LDA of large data. They can be obtained by cheaper integer programming techniques and will be considered elsewhere. Note that for the solution with $t = 1$ the first orthogonal CV attains an objective function value of .96, which is smaller than that (1.18) of the second orthogonal CV. Trying different starting values does not help to invert their order. Finally, note that the structure coefficients are almost the same as those obtained with $t = 1.2$. The standard CVs do not gain any further simplification with $t = 1$.

Application of this procedure to the Vowel Data suggests the choice of $t = 1.8$ for both standard and orthogonal CVs. A simpler solution can be found for the orthogonal CVs if this choice is changed to $t = 1.55$ with a reasonable drop of the goodness-of-fit from 85% to 78%. Of course, the extreme case of $t = 1$ produces 0–1 orthogonal CVs.

The procedure for choosing t is quite subjective and in a sense similar to the choice of the number of principal components to be included in particular analysis. However, the tuning parameter problem is even more complicated because one can try different values t_i for each canonical variate.

7. Concluding remarks

In this paper we introduced and implemented linear discriminant analysis subject to LASSO constraints. We named this new technique DALASS. The LASSO inequality constraint is tackled by introducing an exterior penalty function. The transformed objective functions are then maximized subject to equality constraints, making use of a continuous-time algorithm approach which follows precisely the geometry of the constraints.

A heuristic procedure is proposed for choosing the tuning parameter t such that the DALASS solution exhibits reasonable interpretability, while still retaining a considerable proportion of the "discriminating variance".

In modern DA applications—chemometrics (Krzanowski, 1995), gene expressions (Hastie et al., 2001), etc.—the number of variables p can be much greater than the number of observations n. The proposed modified discriminant analysis DALASS can be readily applied to such problems following the data pre-processing suggested in Krzanowski (1995).

Acknowledgments

We are grateful to the Associate Editor and referees for helpful comments on the draft versions of this paper.

References

Dhillon, I.S., Modha, D.S., 2001. Concept decompositions for large sparse text data using clustering. Mach. Learning 42, 143–175.
Everitt, B.S., Dunn, G.M., 2001. Applied Multivariate Data Analysis. second ed. Arnold, London.
Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7, 179–184.
Golub, G.H., Van Loan, Ch.F., 1991. Matrix Computations. second ed. The John Hopkins University Press, Baltimore, London.
Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer, New York.
Jolliffe, I.T., 2002. Principal Component Analysis. second ed. Springer, New York.
Krzanowski, W.J., 1995. Orthogonal canonical variates for discrimination and classification. J. Chemometrics 9, 509–520.
Krzanowski, W.J., 2003. Principles of Multivariate Analysis. revised ed. Oxford University Press, Oxford.
MATLAB, 2002. Using MATLAB. Version 6. The MathWorks Inc.
McLachlan, G.J., 1992. Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York.
MINITAB, 2000. MINITAB 13.1. The Minitab Inc.

Parlett, B.N., 1980. The Symmetric Eigenvalue Problem. Prentice-Hall, Englewood Cliffs, NJ.

Pedhazur, E., 1982. Multiple Regression in Behavioral Research. second ed. Holt, Rinehart and Winston, Inc., Fort Worth, TX.

Rencher, A., 2002. Methods of Multivariate Analysis. Wiley, New York.

SPSS, 2001. SPSS 11.0. SPSS Inc.

Trendafilov, N.T., Jolliffe, I.T., 2006. Projected gradient approach to the numerical solution of the SCoTLASS. Comput. Statist. Data Anal. 50, 242–253.