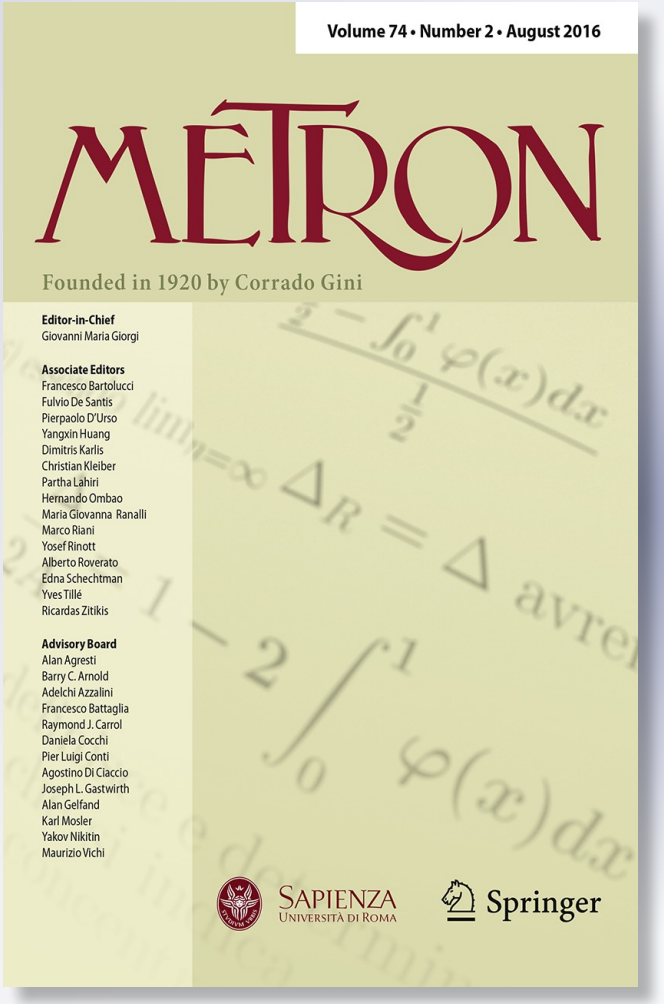
Recipes for sparse LDA of horizontal data

Nickolay T. Trendafilov & Tsegay Gebrehiwot Gebru

METRON

ISSN 0026-1424
Volume 74
Number 2

METRON (2016) 74:207-221
DOI 10.1007/s40300-016-0093-8



Springer

METRON (2016) 74:207–221
DOI 10.1007/s40300-016-0093-8

Recipes for sparse LDA of horizontal data

Nickolay T. Trendafilov[1] · **Tsegay Gebrehiwot Gebru**[1]

Received: 10 November 2015 / Accepted: 17 June 2016 / Published online: 1 July 2016

Abstract Many important modern applications require analyzing data with more variables than observations, called for short *horizontal*. In such situation the classical Fisher's linear discriminant analysis (LDA) does not possess solution because the within-group scatter matrix is singular. Moreover, the number of the variables is usually huge and the classical type of solutions (discriminant functions) are difficult to interpret as they involve *all* available variables. Nowadays, the aim is to develop fast and reliable algorithms for sparse LDA of horizontal data. The resulting discriminant functions depend on very few original variables, which facilitates their interpretation. The main theoretical and numerical challenge is how to cope with the singularity of the within-group scatter matrix. This work aims at classifying the existing approaches according to the way they tackle this singularity issue, and suggest new ones.

Keywords Diagonal within-group scatter · Function constrained LDA · Minimization of classification error · Common and proportional principal components

1 Introduction

Discriminant analysis (DA) is a descriptive multivariate technique for analyzing grouped data, i.e. data where the observations are divided into a number of groups that usually represent samples from different populations [14]. Recently DA has also been viewed as a promising dimensionality reduction technique. 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

✉ Nickolay T. Trendafilov
Nickolay.Trendafilov@open.ac.uk

Tsegay Gebrehiwot Gebru
Tsegay.Gebru@open.ac.uk

[1] Department of Mathematics and Statistics, The Open University, Milton Keynes, UK

groups in terms of canonical variates which are linear combinations of the original variables [14]. LDA requires solving a generalized eigenvalue problem [19, §8.7].

The interpretation of the canonical variates is based on the coefficients of the original variables in the linear combinations. The interpretation can be clear and obvious if the coefficients in the loadings vectors take one of a small number of values which includes exact zero. Unfortunately, in many applications this is not the case. The interpretation problem is exacerbated by the fact that there are three types of coefficient, raw, standardized and structure, which can be used to describe the canonical variates [37,44], where the disadvantages for their interpretation are also discussed. A modification of LDA, aiming for better discrimination and possibly interpretation, is considered in [11,27]. In this approach the vectors of coefficients in the canonical variates are constrained to be orthogonal.

These difficulties are similar to those encountered when interpreting principal component analysis (PCA) [25]. Last decade this problem was approached by developing PCA procedures that produce *sparse* component loadings, i.e. containing many zeros. Such techniques are commonly known as sparse PCA [42], and can be adapted for use in LDA. This was realized first by Trendafilov and Jolliffe [44] who obtained sparse discriminant functions. The non-zero entries correspond to the variables that dominate the discrimination. This method cannot be applied directly to horizontal data, but triggered active research in this direction, which we try to review here.

Horizontal data occur when the number of variables (p) is larger than the sample size (n). Such datasets are nowadays common in many applications. There are two main problems in using classical LDA on horizontal data. First, the within group covariance matrix $\mathbf{W}$ is singular or nearly singular and, hence, it cannot be inverted. This is because of the presence of many variables which are not useful for discrimination. Second, computations are very difficult if not impossible, hence deterring the applicability of classical LDA on horizontal data.

The paper is organized as follows. The classic LDA is briefly revised in Sect. 2. Section 3 briefly summarizes the idea for sparse solutions and the approaches to achieve sparseness. Section 4 is central and is divided into three parts. Section 4.1 reviews several approaches to do LDA of horizontal data by replacing the singular within-group scatter matrix $\mathbf{W}$ by its main diagonal. Another alternative to avoid the singular $\mathbf{W}$ is presented in Sect. 4.2, where sparse LDA is based on minimization of classification error. Section 4.3 lists techniques equivalent to LDA which do not need inverse $\mathbf{W}$ or $\mathbf{T}$, as optimal scaling and common principal components (CPC). Section 4.4 briefly reminds the application of multidimensional scaling is discrimination problems. Finally, sparse pattern with each original variable contributing to only one discriminant functions is discussed in Sect. 4.5. The last Sect. 5 briefly reports the performance three methods for sparse LDA on several data sets.

2 Basic notations, definitions and assumptions of classical LDA

Consider the following linear combinations $\mathbf{XA}$ also called discriminant scores. This is a linear transformation of the original data $\mathbf{X}$ into another vector space. There is interest in finding 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 $\mathbf{XA}$ than with respect to any of the original variables. The number of transformed dimensions s is typically much smaller than the original p. Thus the transformation also achieves dimension reduction. Fisher's LDA works 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 $\mathbf{XA}$. Formally this is organized by maximizing:

$$\frac{\mathbf{a}^\top \mathbf{B}\mathbf{a}}{\mathbf{a}^\top \mathbf{W}\mathbf{a}}\,, \tag{1}$$

where

$$\begin{aligned}\mathbf{B} &= \mathbf{X}^\top \mathbf{G}^\top (\mathbf{G}\mathbf{G}^\top)^{-1}\mathbf{G}\mathbf{X} = ((\mathbf{G}\mathbf{G}^\top)^{-1/2}\mathbf{G}\mathbf{X})^\top ((\mathbf{G}\mathbf{G}^\top)^{-1/2}\mathbf{G}\mathbf{X})\,,\\ \mathbf{W} &= \mathbf{T} - \mathbf{B} = \mathbf{X}^\top (\mathbf{I}_n - \mathbf{G}^\top (\mathbf{G}\mathbf{G}^\top)^{-1}\mathbf{G})\mathbf{X} \qquad (2)\\ &= \mathbf{X}^\top (\mathbf{I}_n - \mathbf{G}^\top (\mathbf{G}\mathbf{G}^\top)^{-1}\mathbf{G})^\top (\mathbf{I}_n - \mathbf{G}^\top (\mathbf{G}\mathbf{G}^\top)^{-1}\mathbf{G})\mathbf{X}\,, \qquad (3)\end{aligned}$$

and $\mathbf{G}$ is the $g \times n$ group indicator matrix, i.e. $\mathbf{G}$ has $1/n_j$ at its (i, j) position if the ith observation (row of $\mathbf{X}$) belongs to the jth group, and 0 otherwise. Then, the matrix of the group means $\overline{\mathbf{X}} = \mathbf{G}\mathbf{X}$.

In other words LDA depends on the between-groups sums-of-squares matrix, $\mathbf{B}$, and the within-groups sums-of-squares matrix, $\mathbf{W}$, of the original data. $\mathbf{B}$ and $\mathbf{W}$ are also called between-and within-groups scatter matrices respectively.

The procedure of finding $\mathbf{A}$ from $\mathbf{B}$ and $\mathbf{W}$ is sequential. Suppose $\mathbf{a}$ is the first column of $\mathbf{A}$. One can show [14,27, §11.1] that $\mathbf{a}$ should be chosen so that (1) is maximized. The maximisation of objective function (1) is equivalent to the following generalized eigenvalue problem:

$$(\mathbf{B} - d\mathbf{W})\mathbf{a} = 0 \text{ or } (\mathbf{W}^{-1}\mathbf{B} - d\mathbf{I}_p)\mathbf{a} = 0\,. \tag{4}$$

Thus the maximum of objective function (1) is the largest eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$ and is achieved at the corresponding eigenvector $\mathbf{a}$. Successive columns of $\mathbf{A}$ are also eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$ and the corresponding values of objective function (1) are the corresponding eigenvalues.

The rank of $\mathbf{W}^{-1}\mathbf{B}$ is $r \leq \min(p, g-1)$, i.e. all the eigenvalues after the first r are 0s. The number r is called the dimension of the discriminant function representation. The number of useful dimensions for discriminating between groups, s, is smaller than r, and the transformation $\mathbf{A}$ is formed by the eigenvectors corresponding to the s largest eigenvalues ordered in decreasing order. Clearly the $(p \times s)$ transformation $\mathbf{A}$ determined by Fisher's LDA maximizes the discrimination among the groups and represents the transformed data in a lower s-dimensional space.

Note, that the Fisher's LDA problem (1) and (4) can be solved without forming $\mathbf{B}$ and $\mathbf{W}$ explicitly. For vertical data, the maximum of (1) can be found by the generalized SVD of $\mathbf{X_B} = (\mathbf{G}\mathbf{G}^\top)^{-1/2}\mathbf{G}\mathbf{X}$ and $\mathbf{X_W} = (\mathbf{I}_n - \mathbf{G}^\top (\mathbf{G}\mathbf{G}^\top)^{-1}\mathbf{G})\mathbf{X}$ [19, §8.7.3]. For large data this method is rather expensive as it requires $O((n+g)^2 p)$ operations. The method proposed by [12] seems a better alternative if one needs to avoid the calculation of large $\mathbf{B}$ and $\mathbf{W}$. Further related results can be found in [38].

The eigenvalue problems (4) can be rewritten in matrix terms as $\mathbf{BA} = \mathbf{WAD}$, where $\mathbf{D}$ is the $(s \times s)$ diagonal matrix of the s largest eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ ordered in decreasing order. This is not a symmetric eigenvalue problem and in general the columns of $\mathbf{A}$ are not orthogonal. However the matrix $\mathbf{A}^\top \mathbf{W}\mathbf{A}$ is diagonal i.e. the solution is orthogonal in the $\mathbf{W}$-space.

Note also that an important assumption for a valid LDA is that the population within-group covariance matrices are equal. This can be checked by using the likelihood-ratio test [27, p. 370] to compare each within-group covariance matrix to the common one. If the null hypothesis is rejected in some groups than the results from LDA are considered unreliable.

The common principal components (CPC) model has been introduced by Krzanowski [26] and Flury [15] to study discrimination problems with unequal group covariance matrices.

3 Interpretation and sparseness

It turns out that in the modern applications the typical data format is with more variables than observations. Such data are also commonly referred to as the small-sample or horizontal data. In other words horizontal data occur when the number of variables (p) is larger than the sample size (n). The following two data are examples of horizontal data.

1. *Ovarian cancer data* [9] are collected from women who have a high risk of ovarian cancer due to family or personal history of cancer. The objective is to distinguish ovarian cancer from non-cancer observations (women). The data contains 216 samples, 121 cancer samples and 95 normal samples. The number of variables is as many as 373,401. But only 4000 variables are considered in this study.
2. *Rice data* [29,36] have 100 variables and 62 observations. They have four groups (varieties) of rice with 7, 19, 9 and 27 observations in them.

The main problem with such data is that the within-groups scatter matrix is singular and the Fisher's LDA (1) is not defined. Moreover, the number of variables is usually huge (e.g. tens of thousands), and thus, it make sense to look for methods that produce sparse discriminant functions, i.e. involving only few of the original variables.

Broadly speaking a vector/matrix is called sparse when it has very few non-zero entries. The number of the non-zero entries is called cardinality of the vector/matrix. There are two main ways to impose sparseness on a vector/matrix solution: by specifying certain cardinality constrain on the solution, or by finding the solution subject to sparseness inducing penalties. The most popular sparseness inducing penalty is the Least Absolute Shrinkage and Selection Operator (LASSO), introduced by Tibshirani [39] for multiple regression problems. For a unit length vector $\mathbf{a}$ ($\|\mathbf{a}\|_2 = 1$), the LASSO has the form $\|\mathbf{a}\|_1 = \sum_i |a_i| < \tau$, where τ is called tuning parameter. By reducing τ, one forces the smaller entries of $\mathbf{a}$ to become exact zeros. Apparently, the sparsest $\mathbf{a}$ has only one non-zero entry equal to 1.

It is also possible to obtain sparse solution by prescribing in advance certain pattern of sparseness [40,47]. For example, one can be interested to find a sparse matrix $\mathbf{A}$ having a single non-zero entry in each raw, as considered in Sect. 4.5.

Another possible option is the employ the vector/matrix majorization [31], which intuitively is expressed by the following example for unit length vectors from $\mathbb{R}^3_+ = [0, \infty) \times [0, \infty) \times [0, \infty)$:

$$\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \prec \left(0, \frac{1}{2}, \frac{1}{2}\right) \prec (0, 0, 1)\ ,$$

i.e. the "smallest" vector has equal entries. One can use some procedure for generation of majorization [31, p. 128] in order to achieve sparseness. A benefit of such an approach is that sparseness can be achieved without tuning parameters. For example, the procedure to obtain sparse patterns by Trendafilov [41] is equivalent to what is known now as soft-thresholding. However, the threshold is found easily by the majorization construction, rather than by tuning different values. Such a pattern construction can be further related to the fit, the classification error, and/or other desired features of the solution.

4 LDA of horizontal data matrix ($p > n$)

4.1 Sparse LDA with diagonal W

The straightforward idea to replace the non-existing inverse of $\mathbf{W}$ by some kind of generalized inverse has many drawbacks, and thus is not satisfactory. For this reason, Witten and Tibshirani [49] adopted the idea proposed by Bickel and Levina [2] to circumvent this difficulty by replacing $\mathbf{W}$ with a diagonal matrix $\mathbf{W}_d$ containing its diagonal, i.e. $\mathbf{W}_d := \mathbf{I}_p \odot \mathbf{W}$. Note, that Dhillon et al. [10] were even more extreme and proposed doing LDA of high-dimensional data by simply taking $\mathbf{W} = \mathbf{I}_p$, i.e. PCA of $\mathbf{B}$. Such LDA version was adopted already by Trendafilov and Vines [45] to obtain sparse discriminant functions when $\mathbf{W}$ is singular.

4.1.1 Sparse LDA as a two-stage sparse PCA

Probably the simplest strategy can be based on the LDA approach proposed by Campbell and Reyment [7], where LDA is performed in two stages each consisting of eigenvalue decomposition (EVD) of a specific matrix. This approach was already applied by Krzanowski et al. [30] with quite reasonable success to LDA problems with singular $\mathbf{W}$. When $\mathbf{W}_d$ is adopted, the original two-stage procedure simplifies like this. At the first stage, the original data are transformed as $\mathbf{Y} = \mathbf{X}\mathbf{W}_d^{-1/2}$. Then, at the second stage, the between-groups scatter matrix $\mathbf{B}_\mathbf{Y}$ of the transformed data $\mathbf{Y}$ is formed:

$$\mathbf{B}_\mathbf{Y} = \sum_{i=1}^{g} n_i (\bar{\mathbf{y}}_i - \bar{\mathbf{y}})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})^\top , \tag{5}$$

and then, some kind of sparse PCA on $\mathbf{B}_\mathbf{Y}$ is applied. Let the resulting sparse components be collected in a $p \times \min\{p, g-1\}$ matrix $\mathbf{C}$. Then, the sparse canonical variates are given by $\mathbf{A} = \mathbf{W}_d^{-1/2}\mathbf{C}$. The sparseness achieved by $\mathbf{C}$ is inherited in $\mathbf{A}$ because $\mathbf{W}_d$ is diagonal. Note that the calculation of $\mathbf{B}_\mathbf{Y}$ in (5) is not really needed. Following (2), the sparse PCA can be performed directly on $(\mathbf{G}\mathbf{G}^\top)^{-1/2}\mathbf{G}\mathbf{Y}$.

Krzanowski [28] proposed a generalization of this two-stage procedure for the case of unequal within-group scatter matrices. He adopted the CPC model for each of the within-group scatter matrices in each group. For horizontal data, this generalized procedure results in a slightly different way of calculating $\mathbf{B}_\mathbf{Y}$ in (5). Now, $\bar{\mathbf{y}}_i = \mathbf{X}_i\mathbf{W}_{i,d}^{-1/2}$, where $\mathbf{X}_i$ is the data sub-matrix containing the observations of the ith group and $\mathbf{W}_{i,d} = \mathbf{I}_p \odot \mathbf{W}_i$, where $\mathbf{W}_i$ is the within-group scatter matrix of the ith group.

4.1.2 Function constrained LDA (FC-LDA)

By adopting the simplification $\mathbf{W} = \mathbf{W}_d$, the function-constraint reformulation of LDA is straightforward:

$$\begin{aligned} &\min \quad \|\mathbf{a}\|_1 + \tau(\mathbf{a}^\top\mathbf{B}\mathbf{a} - d)^2 , \\ &\mathbf{a}^\top\mathbf{W}_d\mathbf{a} = 1 \\ &\mathbf{a} \perp \mathbf{W}_{i-1} \end{aligned} \tag{6}$$

where $\mathbf{W}_0 = \mathbf{0}_{p\times 1}$ and $\mathbf{W}_{i-1} = \mathbf{W}_d[\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{i-1}]$, and d is found as a solution of the standard Fisher's LDA problem (1) with $\mathbf{W} = \mathbf{W}_d$.

Let $\mathbf{b} = \mathbf{W}_d^{1/2}\mathbf{a}$. Note that $\mathbf{b}$ are in fact the so-called raw coefficients [27, p. 298]. As $\mathbf{W}_d$ is diagonal, $\mathbf{a}$ and $\mathbf{b}$ have the same sparseness. Then, the modified Fisher's LDA problem (6) to produce sparse raw coefficients is defined as:

$$\min_{\substack{\mathbf{b}^\top\mathbf{b}=1\\ \mathbf{b}\perp\mathbf{b}_{i-1}}} \|\mathbf{b}\|_1 + \tau(\mathbf{b}^\top\mathbf{W}_d^{-1/2}\mathbf{B}\mathbf{W}_d^{-1/2}\mathbf{b} - d)^2 \,. \tag{7}$$

Thus, the problem (7) is in fact a function-constraint PCA problem [42]. For small data, as those considered in the following examples one can readily apply the dynamical system approach [43]. For this reason, one can employ some kind of smoothing of the ℓ_1 vector norm, e.g.:

$$\|\mathbf{b}\|_1 = \mathbf{b}^\top \text{sign}(\mathbf{b}) \approx \mathbf{b}^\top \tanh(\gamma\mathbf{b}) \,, \tag{8}$$

with some large $\gamma > 0$. Other smoothing options are considered elsewhere [22]. Let f denote the objective function from (7), i.e. $f(\mathbf{b}) = \|\mathbf{b}\|_1 + \tau(\mathbf{b}^\top\mathbf{W}_d^{-1/2}\mathbf{B}\mathbf{W}_d^{-1/2}\mathbf{b} - d_i)^2$. Then, the solution of (7) can be found as an initial value problem (IVP) for:

$$\frac{d\mathbf{b}_i}{dt} = \Pi_i \nabla_f(\mathbf{b}_i) \,,\ \mathbf{b}_i(0) = \mathbf{b}_i^0 \,, \tag{9}$$

where ∇_f denotes the gradient of f with respect to the standard (Frobenius) matrix inner product and

$$\Pi_i = \mathbf{I}_p - \mathbf{B}_i\mathbf{B}_i^T \quad \text{with} \quad \mathbf{B}_i = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_i] \,. \tag{10}$$

The current ordinary differential equations(ODE) solvers [32] are not suitable for solving large optimization problems. They track the whole trajectory defined by the ODE which is time-consuming and undesirable when the asymptotic state is of interest only [35]. Instead, one can employ numerical methods for optimization on matrix manifolds, and in particular on the Stiefel manifold [1], and employ some existing software [3,48].

Replacing (7) by

$$\min_{\substack{\mathbf{b}^\top\mathbf{b}=1\\ \mathbf{b}\perp\mathbf{b}_{i-1}}} \|\mathbf{b}\|_1 - \tau\mathbf{b}^\top\mathbf{W}_d^{-1/2}\mathbf{B}\mathbf{W}_d^{-1/2}\mathbf{b} \,, \tag{11}$$

increases considerably the speed but usually increases the classification error.

In experiments with simulated and real data, solving (11) outperforms [49] in any case (probably to blame the MM optimization method), and is comparable to [8].

Example 1 The data in the following examples are centered, and normalized to variables with unit length. Iris data [14] have four variables and three groups with 50 observations each. First we solve the original Fisher's LDA (1). The effective number of discriminant functions for this problem is $\min(4, 3-1) = 2$. The first two eigenvalues are 32.1919 and .2854 (32.4773 in total), and the raw coefficients are depicted in the first two columns of Table 1. The projection of the data onto the space spanned by the first two discriminant functions is given in the (1,1) panel of Fig. 1. It is well known that there are three misclassified points (52, 103 and 104) for this solution, i.e. 2 % misclassification. Then, we solve the original Fisher's LDA with $\mathbf{W} = \mathbf{W}_d$. The first two eigenvalues are 31.0969 and .3125 (31.4094 in total), and the raw coefficients are depicted in the second two columns of Table 1. There are six misclassified points (9, 31, 50, 52, 103 and 119) for this solution, i.e. 4 % misclassification. The discriminant plot of the data is given in the (1,2) panel of Fig. 1. Next, we solve (7) with $\tau = 1.2$. The minimum of the objective function in (7) is 1.0680. The first two eigenvalues 31.0969 and .3125 are approximated by 30.7763 and .4407 respectively. The sparse raw coefficients are depicted in the third two columns of Table 1. There are five misclassified points (9, 31, 50, 52, 103) for this solution, i.e. 3.3 % misclassification. The discriminant plot of the data is given in the (2,1) panel of Fig. 1. Finally, we solve (7) with $\tau = .5$. The

Table 1 Different raw coefficients for Fisher's Iris data

Vars	$\mathbf{W}$		$\mathbf{W}_d$		$\text{Sparse}_{1.2}$		$\text{Sparse}_{.5}$		SDP		SDP	
x_1	−.22	−.31	−.23	−.17	−.17	0	−.15	0	0	−.50	.80	0
x_2	.28	−.82	.12	−.89	.04	−1.0	0	−1.0	1.0	0	−.60	0
x_3	−.81	.07	−.72	.23	−.74	−.05	−.74	0	0	−.62	0	−.71
x_4	−.46	−.47	−.65	−.35	−.65	0	−.66	0	0	−.61	0	−.71

minimum of of the objective function in (7) is 1.0579. The first two eigenvalues 31.0969 and .3125 are approximated by 30.502 and .616, respectively. The sparse raw coefficients are depicted in the last two columns of Table 1. The same five points are misclassified in this solution. The discriminant plot of the data is given in the (2,2) panel of Fig. 1. It seems, that the LDA (1) with $\mathbf{W} = \mathbf{W}_d$ gives the worst solution, while the sparse LDA with $\tau = .5$ is most satisfying both in terms of fit and interpretability.

Example 2 Rice data [29,36] have 100 variables (wavelengths) and four groups of rice with 7, 19, 9 and 27 observations in them. The effective number of discriminant functions for this problem is $\min(100, 4-1) = 3$. The first three eigenvalues are 25.3009, 1.6737 and .0077, which indicates that the discrimination power of the second and the third discriminant functions are not high. There are 37 misclassified points for this solution, i.e. 59.68 % misclassification. This solution is worse than the results obtained by [29] and employing PCA as a preprocessing (reduction the number of variables). The projection of the data onto the space spanned by the first two discriminant functions is given in the (1,1) panel of Fig. 2. The panel (1,2) contains the raw coefficients of these discriminant functions. Next, we solve (7) with $\tau = .5$. The minimum of the objective function in (7) is 1.1896. The first three eigenvalues are approximated by 23.6843, 0.0874 and 0.0803, respectively. The discriminant plot of the data is given in the (2,1) panel of Fig. 2. There are 40 misclassified points for this solution, i.e. 64.52 % misclassification. The panel (2,2) contains the raw coefficients of these discriminant functions, and the first ones are not sparse at all. Finally, we solve (7) with $\tau = .01$. The minimum of the objective function in (7) is 1.0000. The first three eigenvalues are approximated by .4260, .1437 and .2418, respectively. The discriminant plot of the data is given in the (3,1) panel of Fig. 2. There are again 37 misclassified points for this solution, i.e. 59.68 % misclassification. The panel (3,2) contains the sparse raw coefficients of these discriminant functions. It is really surprising to achieve such discrimination by two variables only! They are probably too sparse and one can look for a better τ.

4.2 Sparse LDA based on minimization of the classification error

Fan et al. [13] argued that ignoring the covariances (the off-diagonal entries in $\mathbf{W}$) as suggested by Bickel and Levina [2] may not be a good idea. In order to avoid redefining Fisher's LDA for singular $\mathbf{W}$, Fan et al. [13] proposed working with (minimizing) the classification error instead of the Fisher's LDA ratio (1). The method is called for short ROAD (from Regularized Optimal Affine Discriminant) and is developed for two groups. Let $\mathbf{m}_1$ and $\mathbf{m}_2$ denote the group means and form $\mathbf{d} = \frac{\mathbf{m}_1 - \mathbf{m}_2}{2}$, and, as before, $\mathbf{T} = \mathbf{B} + \mathbf{W}$. To avoid Fisher's LDA with singular $\mathbf{W}$, Fan et al. [13] consider general linear classifier $\delta_{\mathbf{a}}(\mathbf{x}) = \mathbb{I}\{\mathbf{a}^\top(\mathbf{x} - \mathbf{m}) > 0\}$, where $\mathbf{m} = \frac{\mathbf{m}_1 + \mathbf{m}_2}{2}$ is the groups mean and $\mathbb{I}$ is indicator function. The misclassification error

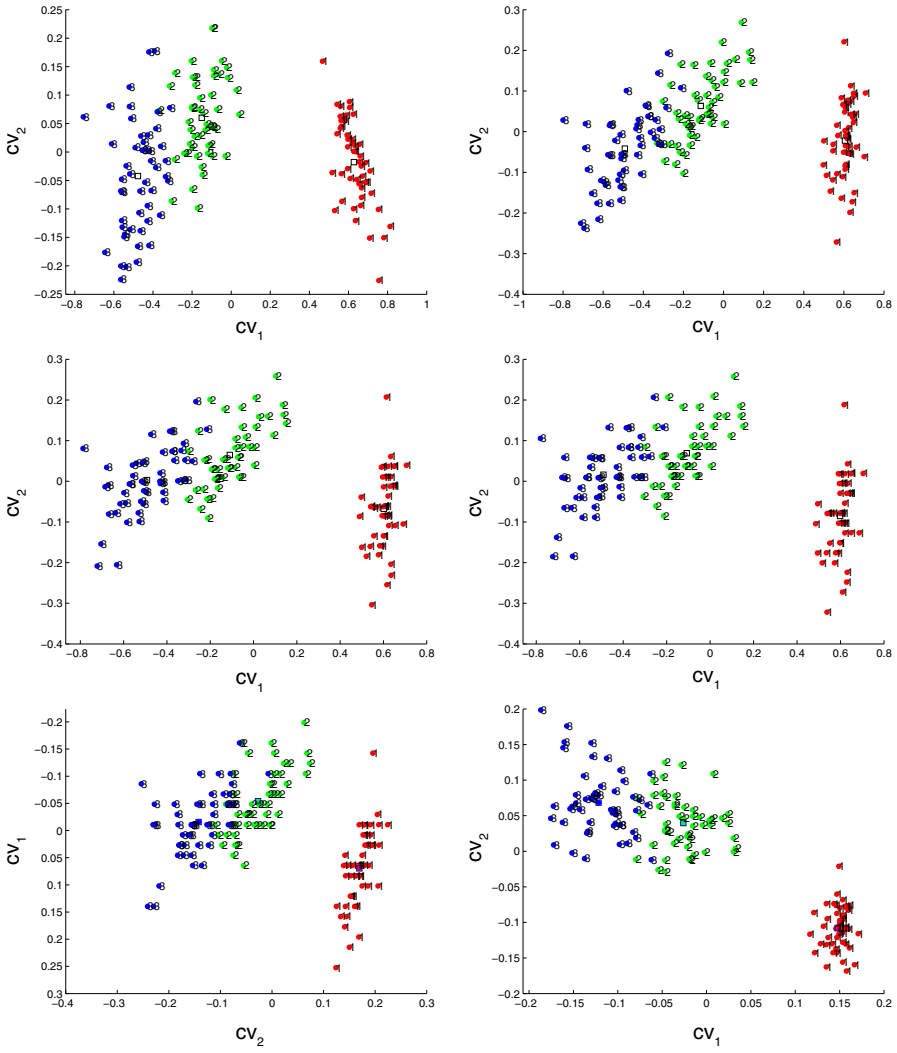


Fig. 1 *Iris* data plotted against two CVs. 1 = *Iris setosa*, 2 = *Iris versicolor*, 3 = *Iris virginica*. *Squares* denote group means. The (1, 1) *panel* uses the original CVs (with $\mathbf{W}$). The (1, 2) *panel* uses the CVs with $\mathbf{W}_d$. The *panels* (2, 1) and (2, 2) use sparse CVs with $\tau = 1.2$ and $\tau = .5$ respectively. For *panels* (3, 1) and (3, 2) see Example 3 in Sect. 4.5

of $\delta_{\mathbf{a}}(\mathbf{x})$ is $1 - \Phi\left\{\frac{\mathbf{d}^\top \mathbf{a}}{\sqrt{\mathbf{a}^\top \mathbf{T}\mathbf{a}}}\right\}$. To minimize the classification error of $\delta_{\mathbf{a}}(\mathbf{x})$, one can maximize $\frac{\mathbf{d}^\top \mathbf{a}}{\sqrt{\mathbf{a}^\top \mathbf{T}\mathbf{a}}}$, or minimize $\mathbf{a}^\top \mathbf{T}\mathbf{a}$ subject to $\mathbf{d}^\top \mathbf{a} = 1$.

Then, the ROAD problem is to find such a minimizer $\mathbf{a}$, which is moreover *sparse*. Thus, the ROAD minimizer $\mathbf{a}$ is sought subject to a LASSO-type constraint introduced as a penalty term, i.e.:

$$\min_{\mathbf{d}^\top \mathbf{a}=1} \mathbf{a}^\top \mathbf{T}\mathbf{a} + \tau \|\mathbf{a}\|_1 \,. \tag{12}$$

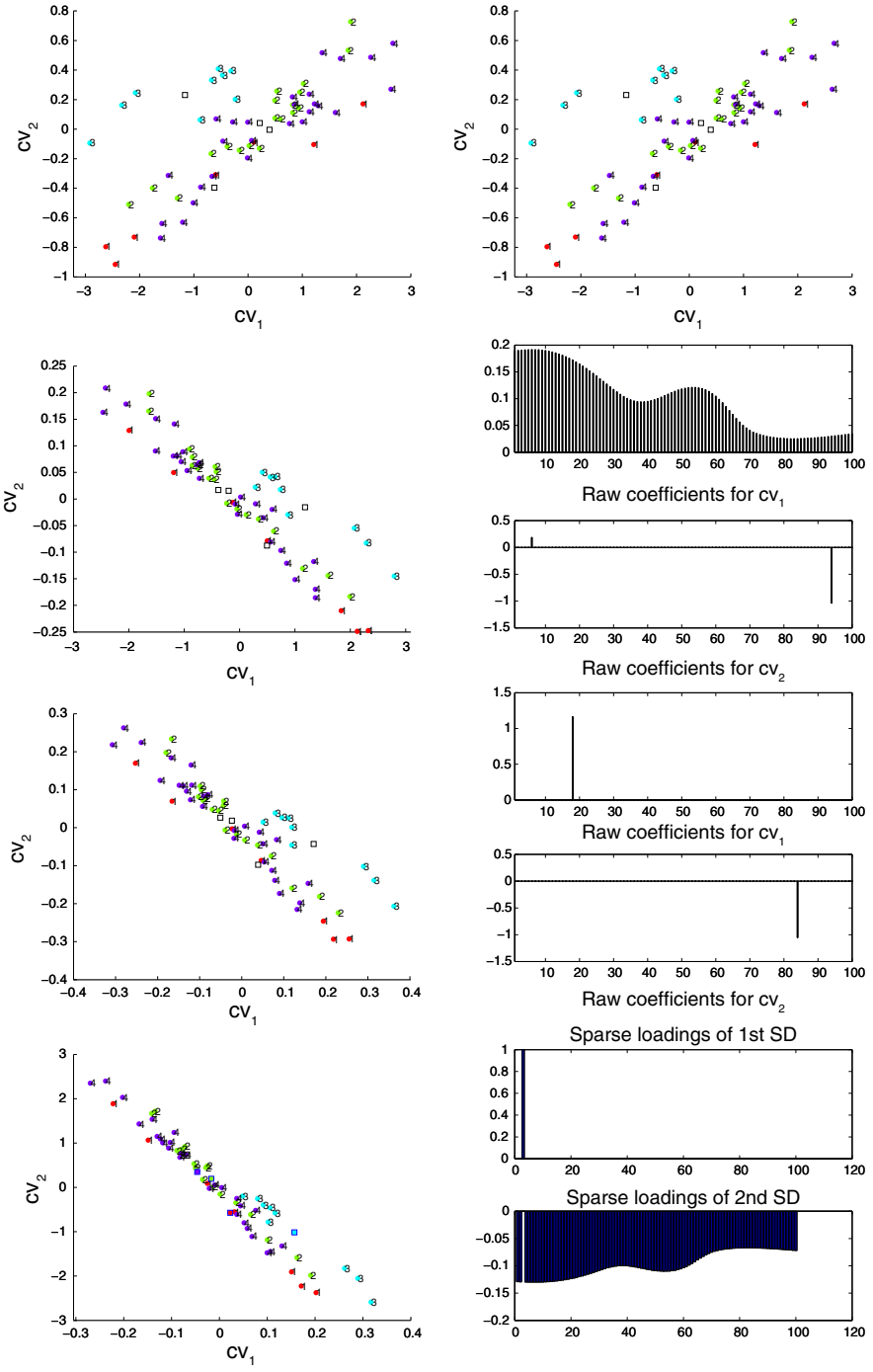


Fig. 2 *Rice* data plotted against two CVs. The groups are 1 = *France*, 2 = *Italy*, 3 = *India*, 4 = *USA*. *Squares* denote group means. The (1, 1) *panel* uses the CVs with $\mathbf{W}_d$. The *panels* (2, 1) and (2, 2), and (3, 1) and (3, 2) use sparse CVs with $\tau = .5$ and $\tau = .01$, respectively. For *panels* (4, 1) and (4, 2) see Example 3 in Sect. 4.5

Further on, Fan et al. [13] replace the affine constrained problem (12) by a quadratic penalty term, which results in the following unconstraint problem:

$$\min \mathbf{a}^\top \mathbf{T}\mathbf{a} + \tau\|\mathbf{a}\|_1 + \upsilon(\mathbf{d}^\top \mathbf{a} - 1)^2 \,. \tag{13}$$

Let us forget for a while for the sparseness of $\mathbf{a}$, and solve (13) for the Iris data with $\tau = 0$. Then, the first group *Iris setosa* is perfectly separated by the cloud composed by the rest two groups of *Iris versicolor* and *Iris virginica*. The difference between the means of these two groups is $\mathbf{d} = (-.1243, .1045, -.1598, -.1537)$. The discriminant is $\mathbf{a} = (.6623, 1.4585, -5.1101, -.7362)$. One can check that $\mathbf{a}^\top \mathbf{d} = 1$. However, this solution of (13) is not convenient for interpretation because one cannot assess the relative sizes of the elements of $\mathbf{a}$. Another related problem is that the LASSO constraint may not work well with vectors $\mathbf{a}$ with arbitrary length.

Thus, it seems reasonable to consider a constrained version of problem (13) subject to $\mathbf{a}^\top \mathbf{a} = 1$. Solution of the following related "dense" problem (with $\tau = 0$)

$$\begin{array}{l} \min \quad \mathbf{a}^\top \mathbf{T}\mathbf{a} \\ \mathbf{d}^\top \mathbf{a} = 1, \\ \mathbf{a}^\top \mathbf{a} = 1 \end{array} \tag{14}$$

is available by Gander et al. [17]. One can consider "sparsifying" their solution to produce unit length ROAD discriminants.

Other works in this direction exploit the fact that the classified error depends on $\mathbf{T}^{-1}$ and $\mathbf{d}$ only through their product $\mathbf{T}^{-1}\mathbf{d}$ [5,23]. As the ROAD approach, they are also designed for discrimination into two groups. This is helpful for obtaining asymptotic results, however not quite helpful for complicated applications involving several groups.

Finally, the function-constraint reformulation of ROAD is:

$$\min_{\mathbf{a}^\top \mathbf{a}=1} \|\mathbf{a}\|_1 + \tau(\mathbf{a}^\top \mathbf{T}\mathbf{a} - d) + \upsilon(\mathbf{d}^\top \mathbf{a} - 1)^2 \,, \tag{15}$$

where d is a standard eigenvalue of $\mathbf{T}$.

4.3 Indirect methods for discriminant analysis

The main purpose of this class of approaches is to avoid the explicit use of the singular matrices $\mathbf{T}^{-1}$ and/or $\mathbf{W}^{-1}$.

4.3.1 Equivalent definitions of discriminant analysis

Clemmensen et al. [8] make use of the LDA re-formulation as optimal scoring, discussed in detail by Hastie et al. [24]. The optimal scoring problem does not require $\mathbf{W}^{-1}$, and thus, is applicable for $p > n$. Let $\mathbf{Y}$ denote an $n \times g$ group indicator matrix. The sparse DA solution is obtained by solving:

$$\min_{\mathbf{a},\mathbf{b}} \{\|\mathbf{Y}\mathbf{b} - \mathbf{X}\mathbf{a}\|^2 + \omega(\mathbf{a}^T \mathbf{P}\mathbf{a}) + \tau\|\mathbf{a}\|_1\}$$

subject to $\frac{1}{n}\mathbf{b}^T \mathbf{Y}^T \mathbf{Y}\mathbf{b} = 1$ and $\mathbf{b}^T \mathbf{Y}^T \mathbf{Y}\mathbf{b}_i = 0$, where τ and ω are nonnegative tuning parameters, and $\mathbf{P}$ is a positive-definite penalty matrix.

4.3.2 Discriminant analysis with CPC

Common principal components (CPC) are developed by Flury [15] and can be used to discriminate several groups of observations with *different* covariance matrices in each group. Zou [50] already considered briefly such an option. In a simulated study, Flury et al. [16] demonstrated that even a simpler CPC model with proportional covariance matrices [15, Ch 5] can provide quite competitive discrimination compared to other more complicated methods.

4.4 Sparse LDA based on metric scaling

Gower [20] showed that metric scaling of the matrix of Mahalanobis distances between all pairs of groups will recover the canonical variate configuration of group means. However, the Mahalanobis distances use the pooled within-group scatter matrix, and thus, this approach is not applicable for horizontal data. It was mentioned before, that Dhillon et al. [10] avoided this problem by simply doing PCA of the between-group scatter matrix **B** to obtain LDA results. Trendafilov and Vines [45] considered sparse version of this LDA procedure.

The above approach can still be applied if the equality of the population covariance matrices of the groups cannot be assumed. A particularly elegant solution, employing Hellinger distances, can be obtained if the CPC hypothesis is appropriate for the different covariance matrices [28].

Another unexplored option would be to consider linear discrimination employing within- and between-group distance matrices [21], which have sizes $n \times n$.

4.5 Sparse LDA without sparse inducing penalty

In this section we consider a new procedure for sparse LDA. The sparseness of the discriminant functions **A** will be achieved without employing sparse inducing penalties. Instead, we will look for a solution **A** with specific pattern of sparseness, with only one nonzero entry in each row of **A**. The methods is inspired by the recent works of Timmerman et al. [40] and Vichi and Saporta [47].

The following model represents the original data **X** by only the group means projected onto the reduced space, formed by the orthonormal discriminant functions **A**. The model can be formally written as:

$$\mathbf{X} = \mathbf{U}(\overline{\mathbf{X}}\mathbf{A})\mathbf{A}^\top = \mathbf{U}\overline{\mathbf{X}}(\mathbf{A}\mathbf{A}^\top) \,, \tag{16}$$

where $\overline{\mathbf{X}}$ is the $g \times p$ matrix of group means and **U** is the $n \times g$ indicator matrix of the groups, such that $\mathbf{G} = (\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top$. In this notations, one has $\overline{\mathbf{X}} = \mathbf{G}\mathbf{X} = (\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{X}$, and the model (16) can be rewritten as:

$$\mathbf{X} = \mathbf{P}\mathbf{X}(\mathbf{A}\mathbf{A}^\top) \,, \tag{17}$$

where one notes that $\mathbf{P} = \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top$ is a projector. The $p \times r$ orthonormal matrix **A** contains the orthonormal "raw coefficients" of the problem, and r is the number of required discriminant functions.

We want to find sparse raw coefficients **A** but without relying on sparseness inducing constrains as in the previous sections. In general, this is unsolvable problem, but it can be easily tackled if we restrain ourselves to a particular pattern of sparseness: each row of **A** should posses a *single* nonzero entry. Thus, the total number of nonzero entries in **A** will be p. To construct **A** with such a pattern, we introduce a $p \times r$ binary (of 0's and 1's)

membership matrix $\mathbf{V}$, indicating which variables have nonzero loadings on each particular discriminant function i.e. in each column of $\mathbf{A}$. Then, $\mathbf{A}$ will be sought in the form of a product $\mathbf{A} = \text{Diag}(\mathbf{b})\mathbf{V}$, where $\text{Diag}(\mathbf{b})$ is a diagonal matrix formed by the vector $\mathbf{b}$. The ith element of $\mathbf{b}$ gives the nonzero value at the ith row of $\mathbf{A}$. In other words, $\mathbf{V}$ is responsible for the locations of the nonzero entries in $\mathbf{A}$, while $\mathbf{b}$ will give their values. Apparently, the choice of $\mathbf{V}$ and $\mathbf{b}$ will affect the fit of the model (17). Thus, we need to solve the following least-squares problem:

$$\min_{\mathbf{V},\mathbf{b}} \|\mathbf{X} - \mathbf{PX}[\text{Diag}(\mathbf{b})\mathbf{V}\mathbf{V}^\top \text{Diag}(\mathbf{b})]\| , \tag{18}$$

which will be called for short SDP (Sparse Discriminative Projection).

Example 3 We apply SDP to the Iris data. Two solutions (matrices $\mathbf{A}$) are depicted in the last four columns of Table 1, two for each solution. The first pair of columns is the solution $\mathbf{A}$ for which the SDP objective function (17) is minimal (1.0563) among several random starts. However, this solution produces 11 misclassified observations, which is 7.33 % misclassification rate. This solution looks less satisfying compared to previous ones, reported in Example 1. The last two columns of Table 1 give another SDP solution for which the objective function (17) is 1.1121, but the misclassification is 4 % only, with six misclassified observations (9, 31, 50, 52, 103 and 104). The quality of this solution resembles the (dense) LDA solution with $\mathbf{W} = \mathbf{W}_d$ from Example 1. It's clear that SDP performance is not satisfying for this data set.

Now, we apply SDP to the Rice data. The best solution (among several random starts) produces 26 misclassified observations, which is only 41.94 % misclassification rate. This result looks much better than what was achieved by other approaches, and probably needs further checking. The discriminant plot of the data is given in the (4,1) panel of Fig. 2. The panel (4,2) contains the raw coefficients of these discriminant functions: the first ones has a single nonzero entry, while the second is not sparse at all. The separation achieved by these discriminant functions is quite satisfying, but not the sparseness.

One can develop a better SDP method if the classification error is minimized instead of fitting the data matrix $\mathbf{X}$ or its projection onto the subspace spanned by the discriminant functions. Nevertheless, the main weakness of SDP is that for large p the SDP solutions are not sparse enough, and thus, not attractive for application.

5 Comparison of existing methods

We consider three sparse discriminant analysis methods for comparison using five datasets. The three methods are:

- Function constrained linear discriminant analysis (FC-LDA) which is introduced in Sect. 4.1.2.
- Sparse discriminant analysis (SDA) which is proposed by Clemmensen et al. [8]. It is considered in Sect. 4.3.1.
- Penalized classification using Fisher's linear discriminant (PLDA). This method is proposed by Witten and Tibshirani [49] for penalizing the discriminant vectors in Fisher's discriminant problem.

In Table 2 we summarize the results from numerical experiments with the three methods referred above.

Table 2 Results from three methods for sparse LDA applied to several data sets

Data	vars	obs	groups	FC-LDA		SDA		PLDA	
	n	p	g	Error (%)	Time	Error (%)	Time	Error (%)	Time
Iris	150	4	3	3.33	0.0018	3.33	0.0870	4.00	0.0020
Rice	62	100	4	33.50	0.0065	30.65	0.2230	34.50	0.0060
Leukemia	85	10,000	6	22.09	35.3201	27.65	19.9700	27.33	35.2000
Ovarian cancer	216	4000	2	19.03	59.1958	19.31	58.3452	20.65	60.1024
Ramaswamy	198	16,063	14	13.13	115.1903	16.16	116.5012	–	–

The solutions produced by FC-LDA and SDA have about 5 % non-zero entries. From this table we see that FC-LDA works as well as SDA. The reason that FC-LDA does not show superiority as compared with SDA may be due to the fact that FC-LDA uses diagonal within group covariance matrix. The results in [49] have higher percentage of non-zero entries, so not quite comparable with the other two.

6 Connection with Gini's transvariation

This paper mainly revises sparse LDA methods on horizontal data. The main objective of LDA is to find the linear combination of p variables which maximizes group separation. In other words, it is obtained by maximizing the ratio of between to within covariance matrices [14].

Alternatively, the linear combinations can also be obtained in terms of Gini's transvariation [33]. Gini [18] defined that two groups are said to be transvariate on a variable X if the sign of the difference of any two values of X from different groups is opposite to the sign of their corresponding mean difference. Any difference satisfying this condition is called a transvariation [46]. Montanari [33] has shown that transvariation measures can be used to discriminate between groups. Moreover, Caló [6] has used the transvariaon method to measure group separablity. Other authors such as [34] and Bragoli et al. [4] have also applied transvariation for group separation and classification.

These references show that LDA is related to the Gini's transvariation due to the fact that linear discriminant function can be derived as the linear combination which minimize transvariation probability or area. Therefore, we can see that our method, FC-LDA is also related to Gini's transvariation. It is also possible to impose sparsity penalty on the transvariation method so as to find only few important variables in the case of horizontal data. The nature of the trasvariation formulation most likely will require non-parametric methods. Nowadays, Bayesian methods with sparseness inducing priors are widely used for sparse PCA and factor analysis. This could be a new contribution of Gini's transvariation to LDA and in general, to discrimination and classification problems.

Acknowledgments The work of the first author is supported by a Grant RPG-2013-211 from The Leverhulme Trust, UK.

References

1. Absil, P.-A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton (2008)
2. Bickel, P., Levina, E.: Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli **10**, 989–1010 (2004)
3. Boumal, N., Mishra, B., Absil, P.-A., Sepulchre, R.: MANOPT: a Matlab toolbox for optimization on manifolds. J. Mach. Learn. Res. **15**, 1455–1459 (2014)
4. Bragoli, D., Ganugi, P., Ianulardo, G.: Ginis transvariation analysis: an application on financial crises in developing countries. Empirica **40**(1), 153–174 (2013)
5. Cai, T., Liu, W.: A direct estimation approach to sparse linear discriminant analysis. J. Am. Stat. Assoc. **106**, 1566–1577 (2011)
6. Calò, D.G.: On a transvariation based measure of group separability. J. Classification **23**(1), 143–167 (2006)
7. Campbell, N.A., Reyment, R.A.: Discriminant analysis of a Cretaceous foraminifer using shrunken estimator. Math. Geol. **10**, 347–359 (1981)
8. Clemmensen, L., Hastie, T., Witten, D., Ersbøll, B.: Sparse discriminant analysis. Technometrics **53**, 406–413 (2011)
9. Conrads, T.P., Zhou, M., III, E.F.P., Liotta, L., Veenstra, T.D.: Cancer diagnosis using proteomic patterns. Expert Rev. Mol. Diagn. **3**(4), 411–420 (2003)
10. Dhillon, I.S., Modha, D.S., Spangler, W.S.: Class visualization of high-dimensional data with applications. Comput. Stat. Data Anal. **41**, 59–90 (2002)
11. Duchene, J., Leclercq, S.: An optimal transformation for discriminant and principal component analysis. IEEE Trans. Pattern Anal. Mach. Intel. **10**, 978–983 (1988)
12. Duintjer Tebbens, J., Schlesinger, P.: Improving implementation of linear discriminant analysis for the high dimension/small sample size problem. Comput. Stat. Data Anal. **52**, 423–437 (2007)
13. Fan, J., Feng, Y., Tong, X.: A road to classification in high dimensional space: the regularized optimal affine discriminant. J. R. Stat. Soc. B **74**, 745–771 (2012)
14. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugenics **7**, 179–184 (1936)
15. Flury, B.: Common Principal Components and Related Multivariate Models. Wiley, New York (1988)
16. Flury, B., Schmid, M.J., Narayanan, A.: Error rates in quadratic discrimination with constraints on the covariance matrices. J. Classification **11**, 101–120 (1994)
17. Gander, W., Golub, G., von Matt, U.: A constrained eigenvalue problem. Linear Algebra Appl. **114**, 815–839 (1989)
18. Gini, C.: Il Concetto di" transvariazione" e le sue prime applicazioni. Athenaeum (1916)
19. Golub, G.H., Van Loan, C.F.: Matrix Computations. Johns Hopkins University Press, Baltimore (1996)
20. Gower, J.C.: Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika **53**, 325–338 (1966)
21. Gower, J.C., Krzanowski, W.J.: Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. J. R. Stat. Soc. Ser. C (Appl. Stat.) **48**, 505–519 (1999)
22. Hage, C., Kleinsteuber, M.: Robust PCA and subspace tracking from incomplete observations using ℓ_0-surrogates. Comput. Stat. **29**, 467–487 (2014)
23. Hao, N., Dong, B., Fan, J.: Sparsifying the Fisher linear discriminant by rotation. J. R. Stat. Soc. B **77**, 827–851 (2015)
24. Hastie, T., Buja, A., Tibshirani, R.: Penalized discriminant analysis. Ann. Stat. **23**, 73–102 (1995)
25. Jolliffe, I.T., Trendafilov, N.T., Uddin, M.: A modified principal component technique based on the LASSO. J. Comput. Graph. Stat. **12**, 531–547 (2003)
26. Krzanowski, W.J.: Principal component analysis in the presence of group structure. Appl. Stat. **33**, 164–168 (1984)
27. Krzanowski, W.J.: Principles of Multivariate Analysis: A User's Perspective. Oxford University Press, Oxford (1988)
28. Krzanowski, W.J.: Between-group analysis with heterogeneous covariance matrices: the common principal components model. J. Classification **7**, 81–98 (1990)
29. Krzanowski, W.J.: Antedependence models in the analysis of multi-group high-dimensional data. J. Appl. Stat. **26**, 59–67 (1999)

30. Krzanowski, W.J., Jonathan, P., McCarthy, W.V., Thomas, M.R.: Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. J. R. Stat. Soc. C **44**, 101–115 (1995)
31. Marshall, A., Olkin, I.: Inequalities: Theory of Majorization and its Applications. Academic Press, London (1979)
32. MATLAB: MATLAB R2014b. The MathWorks Inc, New York (2014)
33. Montanari, A.: Linear discriminant analysis and transvariation. J. Classification **21**(1), 71–88 (2004)
34. Mussard, S., Savard, L.: The Gini multi-decomposition and the role of Gini's transvariation: application to partial trade liberalization in the Philippines. Appl. Econ. **44**(10), 1235–1249 (2012)
35. Ng, M., Li-Zhi, L., Zhang, L.: On sparse linear discriminant analysis algorithm for high-dimensional data classification. Numer. Linear Algebra Appl. **18**, 223–235 (2011)
36. Osborne, B.G., Mertens, B., Thomson, M., Fearn, T.: The authentication of basmati rice using near infrared spectroscopy. J. Near Infrared Spectrosc. **1**, 77–83 (1993)
37. Rencher, A.: Interpretation of canonical discriminant functions, canonical variates, and principal components. Am. Stat. **46**, 217–225 (1992)
38. Shin, H., Eubank, R.: Unit canonical correlations and high-dimensional discriminant analysis. J. Stat. Comput. Simul. **81**, 167–178 (2011)
39. Tibshirani, R.: Regression shrinkage and selection via the LASSO. J. R. Stat. Soc. **58**, 267–288 (1996)
40. Timmerman, M.E., Ceulemans, E., Kiers, H.A.L., Vichi, M.: Factorial and reduced k-means reconsidered. Comput. Stat. Data Anal. **54**, 1858–1871 (2010)
41. Trendafilov, N.T.: A simple method for Procrustean rotation in factor analysis using majorization theory. Multivar. Behav. Res. **29**, 385–408 (1994)
42. Trendafilov, N.T.: From simple structure to sparse components: a review. Comput. Stat. **29**, 431–454 (2014)
43. Trendafilov, N.T., Jolliffe, I.T.: Projected gradient approach to the numerical solution of the SCoTLASS. Comput. Stat. Data Anal. **50**, 242–253 (2006)
44. Trendafilov, N.T., Jolliffe, I.T.: DALASS: variable selection in discriminant analysis via the LASSO. Comput. Stat. Data Anal. **51**, 3718–3736 (2007)
45. Trendafilov, N.T., Vines, K.: Simple and interpretable discrimination. Comput. Stat. Data Anal. **53**, 979–989 (2009)
46. Vichi, M.: Between data science and applied data analysis (2002)
47. Vichi, M., Saporta, G.: Clustering and disjoint principal component analysis. Comput. Stat. Data Anal. **53**, 3194–3208 (2009)
48. Wen, Z., Yin, W.: A feasible method for optimization with orthogonality constraints. Math. Progr. **142**, 397–434 (2013)
49. Witten, D.M., Tibshirani, R.: Penalized classification using Fisher's linear discriminant. J. R. Stat. Soc. B **73**, 753–772 (2011)
50. Zou, M.: Discriminant analysis with common principal components. Biometrika **93**, 1018–1024 (2006)