

ADHERENTLY PENALIZED LINEAR DISCRIMINANT ANALYSIS

Hideitsu Hino* and Jun Fujiki†

ABSTRACT

A problem of supervised learning in which the data consist of p features and n observations is considered. Each observation is assumed to belong to either one of the two classes. Linear discriminant analysis (LDA) has been widely used for both classification and dimensionality reduction in this setting. However, when the dimensionality p is high and the observations are scarce, LDA does not offer a satisfactory result for classification. Witten & Tibshirani (2011) proposed the penalized LDA based on the Fisher's discriminant problem with sparsity penalization. In this paper, an elastic-net type penalization is considered for LDA, and the corresponding optimization problem is efficiently solved.

1. Introduction

Development and dissemination of measuring technologies has results in fine details and comprehensive measurements. A common property with this type of data is high dimension p and small sample size n . High dimensional data with small sample size are available in many places such as financial data and biomedical data. Analyzing such data requires developing new models as the traditional models in statistics usually perform poorly. For example, based on the theory of random matrices (Marcenko and Pastur 1967, Yin et al. 1988), linear discriminant analysis (LDA; Fisher (1936)), which is widely used for classification in many fields, does not classify accurately than a fair coin in this setup (Bickel and Levina 2004).

Many efforts have been devoted to make linear discriminant analysis in high dimensions reliable. Bickel and Levina (2004) suggest using only the diagonal elements of the empirical covariance matrix for LDA because the inconsistency of the covariance matrix is one of the main reasons for the failure of LDA in high dimensional settings. Using only the diagonal elements of the estimated covariance matrix is regarded as imposing extreme sparsity on the covariance structure, and the resultant classifier is referred to as the naïve Bayes LDA. Less sparse models are obtained by sparse inverse covariance selection methods such as **glasso** (Friedman et al. 2008), **quic** (Hsieh et al. 2011), and their supervised variant (Hino and Reyhani 2013). Witten and Tibshirani (2009) proposed to replace the sample covariance with a shrunk estimate, and formulated a general procedure for covariance regularized regression and classification. Dealing with the covariance matrix requires estimating large number of parameters, which is typically of order $O(p^2)$. For the cases where the objective

*Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan E-mail: hinohide@cs.tsukuba.ac.jp

†Department of Applied Mathematics, Fukuoka University, 8-19-1 Nanakuma, Jonan-ku, Fukuoka, 814-0180, Japan E-mail: fujiki@fukuoka-u.ac.jp

Key words: Linear Discriminant Analysis; Sparse Penalization; Continuous Optimization; Classification; Lasso

is not finding out the covariance structure of the variable and instead only to find a good linear classification axis, Cai and Liu (2011) proposed to directly estimate the classification axis for LDA with sparsity-promoting penalization, which requires one to deal with only $O(p)$ parameters. Mai et al. (2012) adopts the least squares model, which is also called the optimal scoring model (Breiman and Ihaka 1984, Hastie et al. 2009) for linear discriminant analysis. They cast the linear classification problem in the ℓ_1 -norm penalized regression problem (lasso) (Tibshirani 1996). Other sparse discriminant analysis methods have been proposed based on Fisher’s discriminant problem. Trendafilov and Jolliffe (2007) and Wu et al. (2009) considered the ℓ_1 -norm constrained Fisher’s discriminant problem, that is, the ℓ_1 -norm of the classification vector is constrained to be less than a certain pre-specified value. Witten and Tibshirani (2011) proposed the penalized LDA (PLDA), which minimizes the ℓ_1 -norm penalized between class variance under the constraint that the within class variance is less than one.

In this work, based on the penalized LDA framework proposed by Witten and Tibshirani (2011), we consider an elastic-net type penalization (Zou and Hastie 2005) that takes into account both sparsity and *adherency* to the normal discriminant model. The proposed penalization avoids too sparse solutions, which sometimes cause problems in penalized regression and classification, and it also shows adherence to the normal model solution. The proposed method is shown to work well via numerical experiments.

The rest of the paper is organized as follows. In Section 2, we define notations to describe the problem considered in this paper, and introduce two different formulations for linear discriminant analysis. Then, in Section 3, we explain the penalized linear discriminant analysis, which is the basis of the proposed method in this work. Section 4 introduces our approach for penalized linear discriminant analysis, and an efficient optimization algorithm for the problem is derived. In Sections 5, we show experimental results, and the last section is devoted to conclusion and discussion.

2. Notation and Preliminary

Let $X \in \mathbb{R}^{n \times p}$ be the design matrix composed of n observations of p -dimensional vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$, where n is the sample size and $\mathbf{x}_i \in \mathbb{R}^p$. Each observation is supposed to belong to one of two classes C_1 or C_2 . In the observed dataset $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, we assume there are n_1 observations in C_1 and $n_2 = n - n_1$ observations in C_2 . We also assume that the features are centered to have mean zero. Let $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ be the mean vectors of observations in class C_1 and class C_2 , respectively, and $\boldsymbol{\xi}$ be the difference of two mean vectors $\boldsymbol{\xi} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Σ_w and Σ_b are the within and the between class covariance matrices, respectively, and their sample estimates are defined by

$$\hat{\Sigma}_w = \frac{1}{n} \sum_{k=1}^2 \sum_{i \in C_k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^\top, \quad (1)$$

$$\hat{\Sigma}_b = \frac{1}{n} X^\top X - \hat{\Sigma}_w = \frac{1}{n} (n_1 \boldsymbol{\mu}_1 \boldsymbol{\mu}_1^\top + n_2 \boldsymbol{\mu}_2 \boldsymbol{\mu}_2^\top), \quad (2)$$

where $\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{i \in C_k} \mathbf{x}_i$, $k = 1, 2$ are sample estimates of the mean vectors of class C_k , $k = 1, 2$. We also define a positive-definite estimate of the within class covariance matrix Σ_w as $\tilde{\Sigma}_w$, which can be the diagonal version of the empirical estimate $\hat{\Sigma}_w$, or the shrunk estimate $\hat{\Sigma}_w + \epsilon I_p$ with $\epsilon > 0$. The empirical estimate of $\boldsymbol{\xi}$ is defined as $\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$.

In the following, we introduce two different formulations for linear discriminate analysis,

namely, *Fisher's Discriminant Model Formulation* and *Normal Model Formulation*. In their idealized situations, the classification axes obtained by these methods are identical. In our proposed method described in Section 4, an elastic-net type regularization is imposed on Fisher's discriminant model to keep the solution close to that of the normal model.

2.1. Fisher's Discriminant Model Formulation

Fisher's Discriminant Model Formulation of the linear discriminant problem aims at finding the classification vector (axis) maximizing the between class variance while minimizing the within class variance. This problem is often formulated as a minimization problem of the Fisher's criterion $J_F(\beta) = \beta^\top \Sigma_w \beta / \beta^\top \Sigma_b \beta$. This problem is equivalently formulated as the minimization problem

$$\begin{aligned} & \underset{\beta}{\text{minimize}} \quad \{-\beta^\top \hat{\Sigma}_b \beta\} \\ & \text{subject to} \quad \beta^\top \tilde{\Sigma}_w \beta = 1. \end{aligned} \quad (3)$$

Recalling the definition of the between class covariance matrix, the solution of the problem is $\beta^* = \hat{\Sigma}_w^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$ when $\hat{\Sigma}_w$ is full rank, and $\hat{\Sigma}_w$ should be replaced with a positive definite estimate $\tilde{\Sigma}_w$ when $\hat{\Sigma}_w$ is singular.

2.2. Normal Model Formulation

In the *Normal Model Formulation* of the linear discriminant problem, we suppose that the distributions of data in two classes C_1 and C_2 are Gaussians with means μ_1 and μ_2 and the same covariance matrix $\Sigma = \Sigma_w$. Then, the Fisher's discriminant analysis finds the best linear projection in the sense of Bayes error minimization.

We consider a classifier defined by

$$f(\mathbf{x}) = \text{sign} \left(\log \frac{P(C_1|\mathbf{x})}{P(C_2|\mathbf{x})} \right),$$

which achieves Bayes classification error by definition. Then we obtain

$$\begin{aligned} \log \frac{P(C_1|\mathbf{x})}{P(C_2|\mathbf{x})} &= \log \frac{p(\mathbf{x}|C_1)P(C_1)/p(\mathbf{x})}{p(\mathbf{x}|C_2)P(C_2)/p(\mathbf{x})} = \log \frac{P(C_1)}{P(C_2)} + \log \frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} \\ &= \log \frac{P(C_1)}{P(C_2)} \\ &\quad + \log \exp \left\{ -\frac{1}{2}(\mathbf{x} - \mu_1)^\top \Sigma_w^{-1}(\mathbf{x} - \mu_1) + \frac{1}{2}(\mathbf{x} - \mu_2)^\top \Sigma_w^{-1}(\mathbf{x} - \mu_2) \right\} \\ &= \{\Sigma_w^{-1}(\mu_1 - \mu_2)\}^\top \mathbf{x} - \frac{1}{2}\mu_1^\top \Sigma_w^{-1} \mu_1 + \frac{1}{2}\mu_2^\top \Sigma_w^{-1} \mu_2 + \log \frac{P(C_1)}{P(C_2)}. \end{aligned}$$

That is, the Bayes optimal classifier is given by a linear classifier and its projection vector is the same as that obtained by Fisher's criterion, and its empirical estimate is given by

$$\beta^* = \hat{\Sigma}_w^{-1}(\hat{\mu}_1 - \hat{\mu}_2) \quad (4)$$

when $\hat{\Sigma}_w$ is full rank, and $\hat{\Sigma}_w$ should be replaced with $\tilde{\Sigma}_w$ when $\hat{\Sigma}_w$ is singular.

3. Penalized Linear Discriminant Analysis

Witten and Tibshirani (2011) showed the equivalence of equality and inequality constraints in Fisher's discriminant model, and then extended the Fisher's discriminant model

formulation with an inequality constraint to have a sparsity induced penalty as

$$\begin{aligned} & \underset{\boldsymbol{\beta}}{\text{minimize}} && -\boldsymbol{\beta}^\top \hat{\Sigma}_b \boldsymbol{\beta} + P_1(\boldsymbol{\beta}) \\ & \text{subject to} && \boldsymbol{\beta}^\top \hat{\Sigma}_w \boldsymbol{\beta} \leq 1. \end{aligned} \quad (5)$$

The penalty term $P_1(\boldsymbol{\beta})$ is the weighted ℓ_1 -norm penalization defined by

$$P_1(\boldsymbol{\beta}) = \lambda_1 \sum_{i=1}^p |\hat{\sigma}_i \beta_i|, \quad \lambda_1 \geq 0, \quad (6)$$

where $\hat{\sigma}_i, i = 1, \dots, p$ is the sample standard deviation of $\mathbf{x}_i$. The problem (5) is named Penalized Linear Discriminant Analysis (PLDA).

Since the objective function $-\boldsymbol{\beta}^\top \hat{\Sigma}_b \boldsymbol{\beta} + P_1(\boldsymbol{\beta})$ to be minimized is not convex, the authors used an iterative optimization approach called the Majorization-Minimization (MM) algorithm (Hunter and Lange 2004). Consider the problem

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad f(\boldsymbol{\beta}). \quad (7)$$

If f is not a convex function, its minimization is in general difficult. MM algorithm is a general algorithm which first majorizes the objective function $f(\boldsymbol{\beta})$ by a function $g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$, which is said to majorize the function $f(\boldsymbol{\beta})$ at the point $\boldsymbol{\beta}^{(m)}$ as

$$f(\boldsymbol{\beta}) \leq g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)}), \quad f(\boldsymbol{\beta}^{(m)}) = g(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)}). \quad (8)$$

From the initial $\boldsymbol{\beta}^{(0)}$, the MM algorithm solves (7) by iteratively minimizing the majorized objective function as

$$\boldsymbol{\beta}^{(m+1)} = \underset{\boldsymbol{\beta}}{\text{argmin}} \quad g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)}). \quad (9)$$

The MM approach is applied to solve the PLDA problem (5). Let $f(\boldsymbol{\beta}) = -\boldsymbol{\beta}^\top \hat{\Sigma}_b \boldsymbol{\beta}$. For a fixed $\boldsymbol{\beta}^{(m)}$, $f(\boldsymbol{\beta})$ is majorized as

$$\begin{aligned} f(\boldsymbol{\beta}) &\leq f(\boldsymbol{\beta}^{(m)}) + (\boldsymbol{\beta} - \boldsymbol{\beta}^{(m)})^\top \nabla f(\boldsymbol{\beta}^{(m)}) \\ &= \boldsymbol{\beta}^{(m)\top} \hat{\Sigma}_b \boldsymbol{\beta}^{(m)} - 2\boldsymbol{\beta}^\top \hat{\Sigma}_b \boldsymbol{\beta}^{(m)}. \end{aligned}$$

Therefore, the PLDA problem is solved by iteratively minimizing the subproblem

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \mathbf{d}^{(m)\top} \boldsymbol{\beta} + P_1(\boldsymbol{\beta}) \quad (10)$$

$$\text{subject to} \quad \boldsymbol{\beta}^\top \hat{\Sigma}_w \boldsymbol{\beta} \leq 1, \quad (11)$$

where $\mathbf{d}^{(m)} = -2\hat{\Sigma}_b \boldsymbol{\beta}^{(m)}$. The solution to the subproblem is given by the soft-thresholding operator when $\hat{\Sigma}_w$ is the diagonal estimate of Σ_w , and by the coordinate descent method (Friedman et al. 2007) when $\hat{\Sigma}_w$ is a non-diagonal estimate of Σ_w .

4. Adherently Penalized Linear Discriminant Analysis

4.1. Adherent Penalization

Following the extension by Witten and Tibshirani (2011), we consider the penalized Fisher's problem of the form

$$\begin{aligned} & \underset{\boldsymbol{\beta}}{\text{minimize}} && -\boldsymbol{\beta}^\top \hat{\Sigma}_b \boldsymbol{\beta} + P(\boldsymbol{\beta}) \\ & \text{subject to} && \boldsymbol{\beta}^\top \hat{\Sigma}_w \boldsymbol{\beta} \leq 1. \end{aligned} \quad (12)$$

Then, we consider the penalization $P(\beta) = P_1(\beta) + P_2(\beta)$, where $P_1(\beta)$ is the weighted ℓ_1 -norm term (6) which is the same as in PLDA, while

$$P_2(\beta) = \lambda_2 \|\tilde{\Sigma}_w^{-1/2} \beta - \tilde{\Sigma}_w^{-1/2} \hat{\xi}\|_2^2, \quad \lambda_2 \geq 0, \quad (13)$$

which we call the *adherent penalization*. Note that $P_2(\beta)$ is a quadratic form in β and the penalization $P_1(\beta) + P_2(\beta)$ is similar to the elastic-net (Zou and Hastie 2005). The rationale behind this penalization is in the fact that the optimal LDA classification axis in the normal model is $\beta^* = \Sigma_w^{-1}(\mu_1 - \mu_2)$, and it is reasonable to keep $\|\tilde{\Sigma}_w \beta - \hat{\xi}\|$ as small as possible. Here $\hat{\xi} = \hat{\mu}_1 - \hat{\mu}_2$ and $\|\cdot\|$ is a certain vector norm. We note that $P_2(\beta) = \lambda_2(\beta^\top \tilde{\Sigma}_w \beta - 2\hat{\xi}^\top \beta + \hat{\xi}^\top \tilde{\Sigma}_w^{-1} \hat{\xi}) \leq \lambda_2(1 - 2\hat{\xi}^\top \beta + \hat{\xi}^\top \tilde{\Sigma}_w^{-1} \hat{\xi})$ when the constraint $\beta^\top \tilde{\Sigma}_w \beta \leq 1$ in eq. (12) holds, hence the objective function of eq. (12) is naturally bounded from above by a linear form with respect to β . Considering that $\tilde{\Sigma}_w^{-1/2}(\tilde{\Sigma}_w^{1/2} \beta - \tilde{\Sigma}_w^{-1/2} \hat{\xi}) = \beta - \tilde{\Sigma}_w^{-1} \hat{\xi}$, small $P_2(\beta)$ implies that the desired relation $\tilde{\Sigma}_w \beta = \hat{\xi}$ approximately holds. Cai and Liu (2011) proposed to use the max-norm to measure the deviation from the optimal solution of the normal model, and formulated the optimization problem with respect to β as a Dantzig selector-type problem (Candes and Tao 2007) and solved it by linear programming. In this work, considering that $P_1(\beta)$ is a weighted ℓ_1 -norm penalization, and the elastic net-type penalization is shown to find less sparse solutions that can be overlooked by lasso, we adopt the ℓ_2 -norm and defined $P_2(\beta) = \lambda_2 \|\tilde{\Sigma}_w^{1/2} \beta - \tilde{\Sigma}_w^{-1/2} \hat{\xi}\|_2^2$, $\lambda_2 \geq 0$. We call the problem (12) with $P(\beta) = P_1(\beta) + P_2(\beta)$ the Adherently Penalized Linear Discriminant Analysis (APLDA) henceforth.

The sparsity penalty $P_1(\beta)$ and adherency penalty in the ℓ_2 -norm form $P_2(\beta)$ complement each other. The sparsity penalty prefers a sparse structure in the general sense to control its complexity while the adherency penalty prefers a discriminant vector β as close as possible to the solution of the normal model. By combining these two penalties, we obtain the following objective function to be minimized:

$$\begin{aligned} J(\beta) &= -\beta^\top \hat{\Sigma}_b \beta + P_1(\beta) + P_2(\beta) \\ &= -\beta^\top \hat{\Sigma}_b \beta + \lambda_1 \sum_{i=1}^p |\hat{\sigma}_i \beta_i| + \lambda_2 (\beta^\top \tilde{\Sigma}_w \beta - 2\hat{\xi}^\top \beta) + \text{const.} \\ &\leq -\beta^\top \hat{\Sigma}_b \beta + \lambda_1 \sum_{i=1}^p |\hat{\sigma}_i \beta_i| + \lambda_2 - 2\lambda_2 \hat{\xi}^\top \beta + \text{const.} \end{aligned} \quad (14)$$

$$\leq \beta^{(m)\top} \hat{\Sigma}_b \beta^{(m)} - 2\beta^{(m)\top} \hat{\Sigma}_b \beta - 2\lambda_2 \hat{\xi}^\top \beta + \lambda_1 \sum_{i=1}^p |\hat{\sigma}_i \beta_i| + \text{const.} \quad (15)$$

$$\begin{aligned} &= -2(\beta^{(m)\top} \hat{\Sigma}_b + \lambda_2 \hat{\xi}^\top) \beta + \lambda_1 \sum_{i=1}^p |\hat{\sigma}_i \beta_i| + \text{const.} \\ &= c_{\lambda_2}^{(m)\top} \beta + \lambda_1 \sum_{i=1}^p |\hat{\sigma}_i \beta_i| + \text{const.} \end{aligned} \quad (16)$$

In the above derivation, we used the inequality constraint $\beta^\top \tilde{\Sigma}_w \beta \leq 1$ for the first inequality, and the second inequality is the result of majorization. In the final line of the above equations, $c_{\lambda_2}^{(m)} = -2(\hat{\Sigma}_b \beta^{(m)} + \lambda_2 \hat{\xi})$. At the m -th iteration step, the majorized objective function

$$J^{(m)}(\boldsymbol{\beta}) = \mathbf{c}_{\lambda_2}^{(m)\top} \boldsymbol{\beta} + \lambda_1 \sum_{i=1}^p |\hat{\sigma}_i \beta_i| \quad (17)$$

is minimized by solving

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad J^{(m)}(\boldsymbol{\beta}) \quad \text{subject to} \quad \boldsymbol{\beta}^\top \tilde{\Sigma}_w \boldsymbol{\beta} \leq 1. \quad (18)$$

This subproblem is iteratively solved until convergence. As is shown in the next subsection, the solution $\boldsymbol{\beta}$ obtained by our problem satisfies $\boldsymbol{\beta}^\top \tilde{\Sigma}_w \boldsymbol{\beta} = 1$ except the case $\boldsymbol{\beta} = \mathbf{0}$, and the majorized objective function $J^{(m)}(\boldsymbol{\beta})$ satisfies $J^{(m)}(\boldsymbol{\beta}^{(m)}) = J(\boldsymbol{\beta}^{(m)})$. That is, the sequence $\{\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}, \dots\}$ of solutions of the subproblem produces a non-increasing sequence $\{J(\boldsymbol{\beta}^{(0)}), J(\boldsymbol{\beta}^{(1)}), \dots\}$ with $J(\boldsymbol{\beta}^{(0)}) \geq J(\boldsymbol{\beta}^{(1)}) \geq \dots$.

4.2. Soft Thresholding Operator for Solving the Subproblem

To solve the subproblem eq. (17), we consider the Karush-Kuhn-Tucker (KKT) conditions of the problem, which are given by

$$\mathbf{c}_{\lambda_2}^{(m)} + \lambda_1 \boldsymbol{\gamma} + 2\eta \tilde{\Sigma}_w \boldsymbol{\beta} = \mathbf{0}, \quad (19)$$

$$\eta(\boldsymbol{\beta}^\top \tilde{\Sigma}_w \boldsymbol{\beta} - 1) = 0, \quad (20)$$

$$\boldsymbol{\beta}^\top \tilde{\Sigma}_w \boldsymbol{\beta} - 1 \leq 0, \quad \eta \geq 0, \quad (21)$$

where $\boldsymbol{\gamma} \in \mathbb{R}^p$ is a vector composed of the sub-gradients of $\sum_{j=1}^p |\hat{\sigma}_j \beta_j|$ with respect to β_j , $j = 1, \dots, p$ defined by

$$\gamma_j = \begin{cases} \hat{\sigma}_j, & \beta_j > 0, \\ -\hat{\sigma}_j, & \beta_j < 0, \end{cases} \quad (22)$$

and $\gamma_j \in [-\hat{\sigma}_j, \hat{\sigma}_j]$ when $\beta_j = 0$. It is seen that $|\mathbf{c}_{\lambda_2}^{(m)}|_j| > \lambda_1 \hat{\sigma}_j$ for some j implies $\eta \tilde{\Sigma}_w \boldsymbol{\beta} \neq \mathbf{0}$. Then, $\eta > 0$ and $\boldsymbol{\beta}^\top \tilde{\Sigma}_w \boldsymbol{\beta} = 1$ by the complementary condition. Now the KKT conditions for the problem eq. (17) reduce to

$$\mathbf{c}_{\lambda_2}^{(m)} + \lambda_1 \boldsymbol{\gamma} + 2\eta \tilde{\Sigma}_w \boldsymbol{\beta} = \mathbf{0}, \quad (23)$$

$$\boldsymbol{\beta}^\top \tilde{\Sigma}_w \boldsymbol{\beta} = 1, \quad \eta > 0. \quad (24)$$

Then, the solution for the problem eq. (17) is obtained by firstly solving the unconstrained problem

$$\underset{\boldsymbol{\alpha} \in \mathbb{R}^p}{\text{minimize}} \quad \left(\boldsymbol{\alpha}^\top \tilde{\Sigma}_w \boldsymbol{\alpha} + \mathbf{c}_{\lambda_2}^{(m)\top} \boldsymbol{\alpha} + \lambda_1 \sum_{i=1}^p |\hat{\sigma}_i \alpha_i| \right), \quad (25)$$

then normalizing as

$$\boldsymbol{\beta} = \begin{cases} \frac{\boldsymbol{\alpha}}{\sqrt{\boldsymbol{\alpha}^\top \tilde{\Sigma}_w \boldsymbol{\alpha}}}, & \boldsymbol{\alpha} \neq \mathbf{0}, \\ \mathbf{0}, & \boldsymbol{\alpha} = \mathbf{0}. \end{cases} \quad (26)$$

Since we are using the diagonal estimate of the within-class covariance matrix $\tilde{\Sigma}_w$, the minimizer of the problem (25) is obtained by applying the soft thresholding operator (Donoho and Johnstone 1995) component-wise:

$$\hat{\alpha}_i^{t+1} = \begin{cases} \frac{1}{2\hat{\sigma}_i^2} \left([2\hat{\Sigma}_b \boldsymbol{\beta}^{(m)} + 2\lambda_2 \hat{\boldsymbol{\xi}}]_i - \lambda_1 \hat{\sigma}_i \right), & [2\hat{\Sigma}_b \boldsymbol{\beta}^{(m)} + 2\lambda_2 \hat{\boldsymbol{\xi}}]_i > \lambda_1 \hat{\sigma}_i, \\ 0, & -\lambda_1 \hat{\sigma}_i \leq [2\hat{\Sigma}_b \boldsymbol{\beta}^{(m)} + 2\lambda_2 \hat{\boldsymbol{\xi}}]_i \leq \lambda_1 \hat{\sigma}_i, \\ \frac{1}{2\hat{\sigma}_i^2} \left([2\hat{\Sigma}_b \boldsymbol{\beta}^{(m)} + 2\lambda_2 \hat{\boldsymbol{\xi}}]_i + \lambda_1 \hat{\sigma}_i \right), & [2\hat{\Sigma}_b \boldsymbol{\beta}^{(m)} + 2\lambda_2 \hat{\boldsymbol{\xi}}]_i < -\lambda_1 \hat{\sigma}_i, \end{cases} \quad (27)$$

where $[\cdot]_i$ indicates the i -th element of a vector.

By investigating the eq. (27), we can see the effect of the additional penalization term P_2 . To concentrate on the effect of P_2 , let us consider the case $\lambda_1 = 0$ in eq. (27) for now, and consider the vector updating formula

$$\boldsymbol{\beta}^{(m+1)} \leftarrow \tilde{\Sigma}_w^{-1} \left(\hat{\Sigma}_b \boldsymbol{\beta}^{(m)} + \lambda_2 \hat{\boldsymbol{\xi}} \right). \quad (28)$$

Recalling the fact that we are treating a two-class problem and the features are centered, we have $\hat{\boldsymbol{\mu}}_2 = -\frac{n_1}{n_2} \hat{\boldsymbol{\mu}}_1$, $\hat{\Sigma}_b = \frac{n_1}{n_2} \hat{\boldsymbol{\mu}}_1 \hat{\boldsymbol{\mu}}_1^\top$, and $\hat{\boldsymbol{\xi}} = \frac{n}{n_2} \hat{\boldsymbol{\mu}}_1$. Substituting these estimates into eq. (28), we obtain

$$\boldsymbol{\beta}^{(m+1)} \leftarrow \tilde{\Sigma}_w^{-1} \left(\frac{n_1}{n_2} \hat{\boldsymbol{\mu}}_1 \hat{\boldsymbol{\mu}}_1^\top \boldsymbol{\beta}^{(m)} + \lambda_2 \frac{n}{n_2} \hat{\boldsymbol{\mu}}_1 \right). \quad (29)$$

When $\lambda_1 = 0$, the proposed penalization has essentially no effect on the solution of LDA, since the above formula gives a vector proportional to $\tilde{\Sigma}_w^{-1} \hat{\boldsymbol{\mu}}_1$ and also to $\tilde{\Sigma}_w^{-1} \hat{\boldsymbol{\xi}}$ whenever $\lambda_2 \neq -\frac{n_1}{n} \hat{\boldsymbol{\mu}}_1^\top \boldsymbol{\beta}^{(m)}$. Combined with the penalty term P_1 with positive λ_1 , the proposed penalization affects the sparsity-inducing mechanism. The thresholding operation is applied to $[2\hat{\Sigma}_b \boldsymbol{\beta}^{(m)} + \lambda_2 \hat{\boldsymbol{\xi}}]_j$, which is more aligned to the direction $\tilde{\Sigma}_w^{-1} \hat{\boldsymbol{\xi}}$ than that of $[2\hat{\Sigma}_b \boldsymbol{\beta}^{(m)}]_j$.

5. Experimental Results

In this section, we perform several numerical experiments on dichotomy problems to show the effect of the proposed penalty term, and that the proposed method is comparable or superior to conventional sparse discriminant analysis methods.

5.1. Results with artificial datasets

In this subsection, with simple artificial datasets, we see the effect of the proposed penalization. Since the effect of the proposed penalization term is expected to make the solution $\hat{\boldsymbol{\beta}}$ close to the Bayes optimal projection axis $\boldsymbol{\beta}^* = \Sigma_w^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ with the ground truth values of Σ_w , $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, we define *alignment* of the estimate $\hat{\boldsymbol{\mu}}$ with the Bayes optimal value $\boldsymbol{\beta}^*$ by

$$a(\boldsymbol{\beta}^*, \hat{\boldsymbol{\beta}}) = \left| \frac{\hat{\boldsymbol{\beta}}^\top \boldsymbol{\beta}^*}{\|\hat{\boldsymbol{\beta}}\|_2 \|\boldsymbol{\beta}^*\|_2} \right| \in [0, 1]. \quad (30)$$

We generated n_1 observations from a p -dimensional Gaussian distribution with unit covariance matrix and mean vector $\boldsymbol{\mu}_1 = (0.5 + 1/p, \dots, 0.5 + 1/p, 0, \dots, 0)$. The first $p \times (1 - s)$ elements of the vector are filled with $0.5 + 1/p$, and rest $p \times s$ elements are filled with 0. Also, we generated n_2 observations from the Gaussian distribution with unit covariance matrix and mean vector $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$. We varied dimension $p = 10, 100, 1000$, the sample size $n_1 = n_2 = 20$ and $n_1 = n_2 = 200$, and the degree of sparseness $s = 0, 0.1, \dots, 0.8$.

In Figure 1, we show boxplots of alignments of the estimates $\hat{\boldsymbol{\beta}}$ obtained by PLDA and the proposed APLDA in different settings. Parameters λ_1 for PLDA and $\lambda_i, i = 1, 2$ for APLDA are optimized by 10-fold cross-validation using the given observations. From this figure, we see that when the degree of sparseness is low, the proposed penalization offers

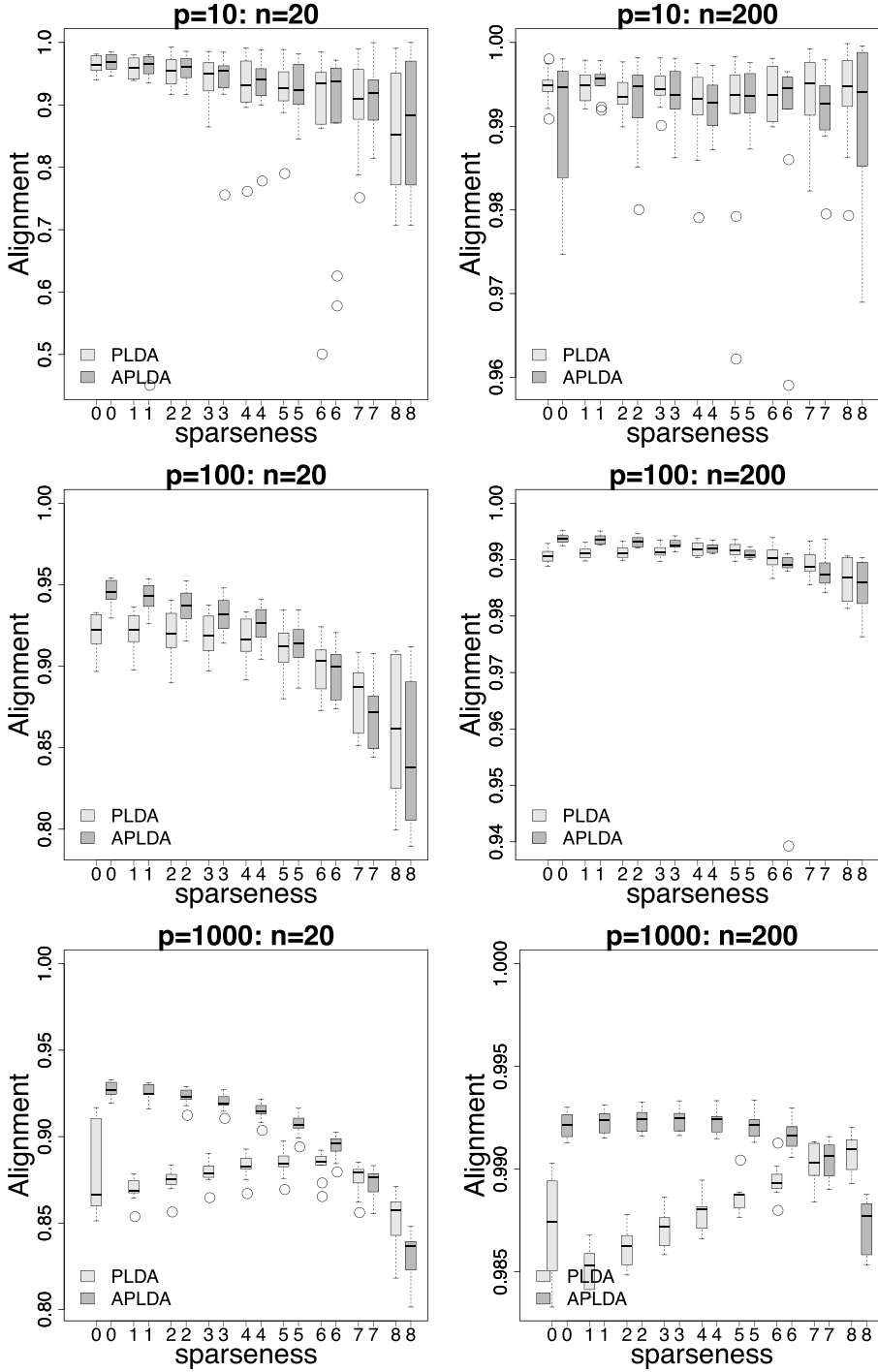


Fig. 1: Effects of the proposed penalization term for datasets with different dimensionalities, sample sizes, and degree of sparseness.

a more aligned solution compared to those of PLDA. On the other hand, when the degree of sparseness is high, namely the Bayes optimal projection vector β^* is very sparse, the proposed method is inferior to the conventional PLDA in terms of alignment as expected. This tendency is particular when the dimension p is high. We can also see, from the range of the alignment shown in the top and bottom panel of Figure 1, that the above discussed tendency is remarkable when the sample size is small relative to the dimensionality of the data.

Finally, we note that the proposed method is derived by bounding the original objective twice, namely, by using the inequality constraints in eq. (14) and eq. (15), which cause the convergence of the optimization to be slower compared to PLDA. We count the number of iterations of MM steps required for convergence in both PLDA and APLDA in the experiments in this subsection. The averages and standard deviations of the iteration count were 4.09(0.851) for PLDA and 4.37(2.298) for APLDA, respectively. The number of iterations was increased on average, while the order remained the same.

5.2. Results with real-world datasets

In this subsection, we show experimental results with six real-world high-dimensional datasets with small number of observations. We consider the two-class discriminant problem and measure the performance of discriminant methods by their area under the curves (AUCs).

Isolet a dataset for a spoken letter recognition problem (Fanty and Cole 1991). From spoken letter data, $p = 617$ dimensional features are extracted as explanatory variable. The task here is to discriminate between “A” and “B”.

p53 a dataset for the problem of predicting the mutant p53 transcriptional activity (active vs inactive) based on data extracted from biophysical simulations (Danziger et al. 2009). The dimensionality of observations is $p = 5,398$.

prostate a dataset of the prostate gene expression, consists of prostate tumor and normal samples. The dimensionality of observations is $p = 6,033$.

secom a dataset from a semi-conductor manufacturing process. Each observation is composed of a selection of sensor outputs placed at the production process, where each observation represents a single production entity with associated measured features. The notion of class here represents {pass, fail}, yielded by certain in house line testing for the products. The dimensionality of observations is $p = 591$.

USPS a dataset for handwritten digit recognition problem (Hull 1994), originally collected by United States Postal Service. The dimensionality of observations is $p = 256$. We consider to discriminate two characters “1” and “2”.

Gisette a dataset also for handwritten digit recognition problem, where the problem is to separate the highly confusable digits “4” and “9”. The dimensionality of observations is $p = 5,000$.

Isolet, **secom**, and **p53** datasets are available from <https://archive.ics.uci.edu/ml/datasets/>. For all datasets, we randomly sampled $n_1 = n_2 = 50$ data for training, and used the rest for estimating the AUC. We repeat this procedure 10 times, and the results are shown in Figure 2 as boxplots.

The following five methods are included for comparison:

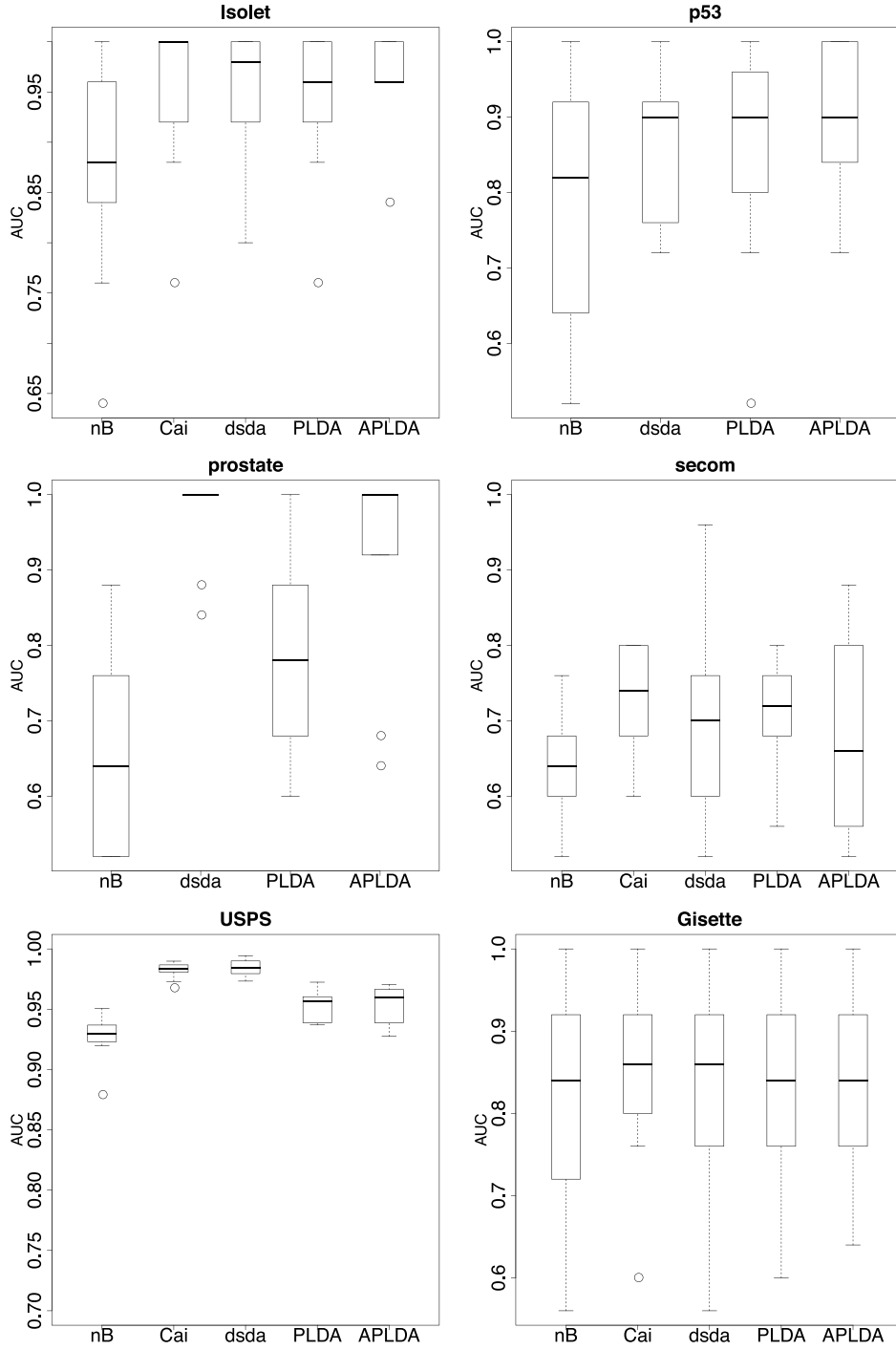


Fig. 2: AUCs of different methods for six different real-world datasets.

nB naïve Bayes approach suggested by Bickel and Levina (2004),

Mai direct approach for sparse discriminant analysis proposed by Mai et al. (2012),

Cai Dantzig selector-based direct method for discriminant analysis proposed by Cai and Liu (2011),

PLDA penalized LDA proposed by Witten and Tibshirani (2011), and

APLDA our proposed adherently penalized LDA.

We note that conventional LDA is also applied to all the problems, but its AUC was far inferior to those obtained by the above five methods. For the sake of legibility, we do not report the results by conventional LDA. For the **p53** and **prostate** datasets, the method proposed by Cai did not return a solution because of the high dimensionality, hence the method is excluded from comparison. For all methods, tuning parameters such as λ_1 and λ_2 are determined by cross-validation solely using the training dataset.

From this figure, we see that the proposed method is superior to PLDA the in **prostate** dataset, and inferior to PLDA in the **secom** dataset. APLDA offers almost the same result to PLDA in other datasets. Overall, the proposed method is comparable to or better than other methods in some situations, and it can be one of the potential candidates for linear discriminant analysis.

6. Conclusion

In the framework of penalized linear discriminant analysis (Witten and Tibshirani 2011), we have proposed a penalization term, which is a combination of ℓ_1 -norm penalization and ℓ_2 -norm penalization similar to the elastic-net (Zou and Hastie 2005). The rationale behind the proposed penalization is that in addition to the sparsity-inducing ℓ_1 -norm penalization, it is reasonable to keep the solution close to the optimal solution for the normal model of the linear discriminant analysis, even when the sparsity inducing penalization is imposed. The penalized objective function is reduced to a simple subproblem by using the MM algorithm approach. Then the subproblem is solved by simple soft thresholding operations.

The aim of this paper is in introducing a variant of the elastic-net type regularization to the linear discriminant problem, which we called the adherent regularization. Theoretical investigation of sparse linear discriminant analysis is difficult because LDA is not based on a familiar loss function such as quadratic loss, logistic loss, or hinge loss. In this work, we experimentally see that the proposed method works well in a high dimensionality setting and is comparable to or sometimes improves PLDA. The reason for relatively better or inferior performance of each method, and dependency of the performance on the dimensionality of the data will be explored in our future work. It is also our important future work to develop asymptotic theory of the proposed adherently regularized linear discriminant analysis.

Acknowledgement

The authors would like to express their special thanks to the editor and anonymous reviewers whose comments led to valuable improvements of the manuscript. Part of this work was supported by JSPS Grant-in-Aid for Scientific Research on Innovative Areas No. 26120504.

REFERENCES

- Peter J. Bickel and Elizaveta Levina. Some theory for Fisher’s linear discriminant function, ‘naive Bayes’, and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 12 2004. doi: 10.3150/bj/1106314847. URL <http://dx.doi.org/10.3150/bj/1106314847>.
- L. Breiman and R. Ihaka. Nonlinear discriminant analysis via ACE and scaling. Technical Report 40, Dept of Statistics, University of California, Berkeley, 1984.
- Tony Cai and Weidong Liu. A direct estimation approach to sparse linear discriminant analysis. *JASA. Journal of the American Statistical Association*, 106(496):1566–1577, 2011. ISSN 0162–1459. doi: 10.1198/jasa.2011.tm11199.
- Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n . *The Annals of Statistics*, 35(6):pp. 2313–2351, 2007. ISSN 00905364. URL <http://www.jstor.org/stable/25464587>.
- Samuel A. Danziger, Roberta Baronio, Lydia Ho, Linda Hall, Kirsty Salmon, G. Wesley Hatfield, Peter Kaiser, and Richard H. Lathrop. Predicting positive p53 cancer rescue regions using most informative positive (mip) active learning. 2009.
- David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. 90(432):1200–1224, December 1995. ISSN 0162-1459 (print), 1537-274X (electronic). URL <http://www.jstor.org/stable/2291512>.
- Mark A. Fanty and Ronald A. Cole. Spoken letter recognition. In R.P. Lippmann, J.E. Moody, and D.S. Touretzky, editors, *Advances in Neural Information Processing Systems 3*, pages 220–226. Morgan-Kaufmann, 1991. URL <http://papers.nips.cc/paper/295-spoken-letter-recognition.pdf>.
- Ronald Aylmer Fisher. The use of multiple measurements in taxonomic problems. *Annals Eugen.*, 7:179–188, 1936.
- Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, July 2008. URL <http://dx.doi.org/10.1093/biostatistics/kxm045>.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning — Data Mining, Inference, and Prediction*. Springer, second edition, 2009.
- Hideitsu Hino and Nima Reyhani. Supervised covariance selection for linear discriminant analysis. In *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*, pages 1–6, Sept 2013. doi: 10.1109/MLSP.2013.6661942.
- Cho-Jui Hsieh, Mátyás A. Sustik, Inderjit S. Dhillon, and Pradeep D. Ravikumar. Sparse inverse covariance matrix estimation using quadratic approximation. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *NIPS*, pages 2330–2338, 2011.
- J. J. Hull. A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(5):550–554, May 1994. ISSN 0162-8828. doi: 10.1109/34.291440. URL <http://dx.doi.org/10.1109/34.291440>.
- D.R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58:30–37, 2004. URL <http://EconPapers.repec.org/RePEc:bes:amstat:v:58:y:2004:m:february:p:30-37>.
- Qing Mai, Hui Zou, and Ming Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42, 2012. URL

- <http://EconPapers.repec.org/RePEc:oup:biomet:v:99:y:2012:i:1:p:29-42>.
- Vladimir Marcenko and Leonid Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967. URL <http://stacks.iop.org/0025-5734/1/i=4/a=A01>.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.
- Nickolay T. Trendafilov and Ian T. Jolliffe. DALASS: Variable selection in discriminant analysis via the LASSO. *Computational Statistics & Data Analysis*, 51(8):3718–3736, May 2007. URL <http://ideas.repec.org/a/eee/csdata/v51y2007i8p3718-3736.html>.
- Daniela M. Witten and Robert Tibshirani. Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):615–636, 2009. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2009.00699.x. URL <http://dx.doi.org/10.1111/j.1467-9868.2009.00699.x>.
- Daniela M. Witten and Robert Tibshirani. Penalized classification using fisher’s linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772, 2011. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2011.00783.x. URL <http://dx.doi.org/10.1111/j.1467-9868.2011.00783.x>.
- Michael C.Wu, Lingsong Zhang, ZhaoxiWang, David Christiani, and Xihong Lin. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, 25(9):1145–1151, 2009.
- Y.Q. Yin, Z.D. Bai, and P.R. Krishnaiah. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability Theory and Related Fields*, 78:509–521, 1988. ISSN 0178-8051. doi: 10.1007/BF00353874. URL <http://dx.doi.org/10.1007/BF00353874>.
- Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

(Received: December 18, 2014, Accepted: September 9, 2015)

