Advanced Review

A review of discriminant analysis in high dimensions

Qing Mai*

Linear discriminant analysis (LDA) is among the most classical classification techniques, while it continues to be a popular and important classifier in practice. However, the advancement of science and technology brings the new challenge of high-dimensional datasets, where the dimension can be in thousands. In such datasets, LDA is inapplicable. Recently, statisticians have devoted many efforts to creating high-dimensional LDA methods. These methods typically perform variable selection via regularization techniques. Various theoretical results, algorithms, and empirical results support the application of these methods. In this review, we provide a brief description of difficulties in extending LDA and present some successful proposals.

How to cite this article:
WIREs Comput Stat 2013, 5:190–197. doi: 10.1002/wics.1257

Keywords: classification; discriminant analysis; high dimension; ℓ_1 penalty; sparsity

INTRODUCTION

As one of the most classical classification techniques, linear discriminant analysis (LDA) is still widely used in contemporary statistical applications. Consider a pair of random variables $(Y, \mathbf{X})$, where $Y \in \{1, \ldots, K\}$ is the class label and $\mathbf{X}$ is a p-dimensional predictor. The LDA model assumes that

$$\mathbf{X} \mid Y = y \sim N(\mu_y, \Sigma), \quad (1)$$

with the prior probabilities $\pi_y = \Pr(Y = y)$. Under this model, the Bayes rule has the following linear form:

$$\delta_{\text{Bayes}}(\mathbf{X}) = \arg\max_k \{\log \pi_k + \mu_k^T \Sigma^{-1}(\mathbf{X} - \mu_k)\}. \quad (2)$$

In particular, when $K = 2$ as in the binary classification problem, the Bayes rule reduces to

$$\delta_{\text{Bayes}}(\mathbf{X}) = \mathbf{1}\left((\mathbf{X} - \frac{1}{2}(\mu_1 + \mu_2))^T \Sigma^{-1}(\mu_2 - \mu_1) + \log \frac{\pi_2}{\pi_2} > 0 \right) + 1, \quad (3)$$

where $\mathbf{1}(\cdot)$ is the indicator function. For simplicity, we focus on the binary classification problem unless stated otherwise.

Of course, the parameters $\Sigma, \mu_1, \mu_2, \pi_1$, and π_2 are unknown in practice. LDA substitutes them with the sample estimates ($k = 1, 2$):

$$\hat{\pi}_k = n_k/n, \quad \hat{\mu}_k = \sum_{Y_i = k} \mathbf{X}_i/n_k, \quad (4)$$

$$\hat{\Sigma} = \sum_k \sum_{Y_i = k} (\mathbf{X}_i - \hat{\mu}_k)(\mathbf{X}_i - \hat{\mu}_k)^T/(n-2), \quad (5)$$

where n is the sample size and n_k is the number of class-k observations. Then LDA is defined as

$$\delta_{\text{LDA}}(\mathbf{X}) = \mathbf{1}\left((\mathbf{X} - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2))^T \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) + \log \frac{\hat{\pi}_2}{\hat{\pi}_1} > 0 \right) + 1. \quad (6)$$

*Correspondence to: maixx034@umn.edu

School of Statistics, University of Minnesota, Minneapolis, MN, USA

Conflict of interest: The author has declared no conflicts of interest for this article.

Under its elegant assumptions, LDA is optimal in the traditional large-sample scenario, that is, if p is fixed and n tends to infinity. On the other hand, it has also been observed that LDA is very competitive in practice, where the assumptions may not hold. For example, Michie et al.[1] and Hand[2] reported that LDA is comparable to many sophisticated classifiers on many benchmark datasets. For a more comprehensive review of LDA, interested readers are referred to Ref 3.

However, modern scientists often encounter datasets with high dimensions, such as in genomics, economics, and machine learning.[4] It has been observed that simple linear classifiers are highly competitive on such datasets.[5,6] Therefore, it is desirable to extend LDA to high-dimensional data. The generalization of LDA, however, is far from trivial. It is obvious that LDA is inapplicable when $p > n$, because $\hat{\Sigma}$ will be singular. To solve this problem, Bickel and Levina[7] proposed the independence rule (IR) that replaces $\hat{\Sigma}$ with diag$(\hat{\Sigma})$, which is applicable for arbitrarily high dimensions. Yet, Fan and Fan[8] showed that IR will deteriorate to no better than random guessing when the dimension is too high. Instead, Fan and Fan[8] proposed the features annealed independence rule (FAIR) that performs feature selection while using a diagonal estimator of Σ. Another such proposal is nearest shrunken centroid classifier, which is commonly known as PAM.[9] Although PAM and FAIR are valuable tools, consequent researchers showed that ignoring the correlation structure in Σ could be misleading in both feature selection and classification.[10–12] In the meantime, many high-dimensional LDA methods were proposed that handle general correlation structure in Σ.[10–17]

This review intends to discuss the issues in high-dimensional LDA and describe some representative new methods. The rest of this paper is organized as follows. First, in second to fourth sections, we introduce IR, FAIR, and PAM as early attempts to generalize LDA. With these methods in mind, in fifth section we explain in detail issues for high-dimensional LDA known in current literature. Then a class of proposals that resolve these issues is discussed in sixth to tenth sections. Finally, the eleventh section presents some empirical comparison between the high-dimensional LDA methods, while the twelfth section contains some concluding remarks. Throughout the paper, we assume $\hat{\pi}_k$, $\hat{\mu}_k$, and $\hat{\Sigma}$ are defined as in Eqs. 4 and 5, and n_k is the sample size within class k. Also, we define

$$\hat{\mathbf{D}} = \mathrm{diag}(\hat{\Sigma}), \tag{7}$$

which, in other words, states that $\hat{D}_{jj} = \hat{\Sigma}_{jj}$ for $j = 1, \ldots, p$ and $\hat{D}_{ij} = 0$ for $i \neq j$.

INDEPENDENCE RULE

A major concern with LDA is the singularity of $\hat{\Sigma}$ when p exceeds n. The IR, also known as the naive Bayes rule, has been proposed to generalize LDA to high dimensions by replacing $\hat{\Sigma}$ with a positive definite estimator.[7] IR estimates Σ with $\hat{\mathbf{D}}$ as in Eq. 7, which is always positive definite and hence invertible. In the context of the LDA model, this is equivalent to treating all the variables as independent within groups. Although this treatment can be a model misspecification, theoretical studies show the surprising result that IR can outperform a rule that intends to model all the correlation. In other words, if we keep the correlation structure and estimate Σ^{-1} with $\hat{\Sigma}^{-}$, the Moore–Penrose inverse, the resulting classifier will asymptotically have an error rate of 1/2. On the other hand, consider the following parameter space:

$$\begin{aligned}\Gamma(c, \mathbf{k}, B) = \Big\{&(\mu_1, \mu_2, \Sigma) : (\mu_2 - \mu_1)^T \\ &\times \Sigma^{-1}(\mu_2 - \mu_1) \geq c^2, \quad k_1 \leq \lambda_{\min}(\Sigma) \\ &\leq \lambda_{\max}(\Sigma) \leq k_2, \mu_i \in \mathbf{B}\Big\},\end{aligned} \tag{8}$$

where $\mathbf{B} = \mathbf{B}_{\mathbf{a},d} = \left\{\mu \in \ell_2 : \sum_{j=1}^{p} a_j \mu_j^2 \leq d^2\right\}, \mathbf{k} = (k_1, k_2)$ and $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ are the smallest and largest eigenvalues of Σ, respectively. If we estimate μ_{ij} with

$$\hat{\mu}_{ij}^{\mathrm{IR}} = (1 - r_{jn})_{+} \hat{\mu}_{ij} \tag{9}$$

for some properly chosen r_{jn} and set IR to be

$$\begin{aligned}\delta_{\mathrm{IR}}(\mathbf{X}) = \mathbf{1}\Bigg(&(\mathbf{X} - \frac{1}{2}(\hat{\mu}_2^{\mathrm{IR}} + \hat{\mu}_1^{\mathrm{IR}}))^T \hat{\mathbf{D}}^{-1}(\hat{\mu}_2^{\mathrm{IR}} - \hat{\mu}_1^{\mathrm{IR}}) \\ &+ \log(\frac{n_2}{n_1}) > 0\Bigg) + 1,\end{aligned} \tag{10}$$

then, for the risk bound of IR, $R_{\Gamma}(\delta_{\mathrm{IR}}) = \max_{\Gamma} \Pr(Y \neq \delta_{\mathrm{IR}})$, we have

$$\limsup_{n \to \infty} R_{\Gamma}(\delta_{\mathrm{IR}}) = 1 - \Phi(\frac{\sqrt{K_0}}{1 + K_0} c) \tag{11}$$

if $\log p/n \to 0$, where Φ is the cumulative distribution function for a standard normal random variable, and $K_0 = \max_{\Gamma}[\lambda_{\max}(\mathbf{D}^{-1/2}\Sigma\mathbf{D}^{-1/2})]/[\lambda_{\min}(\mathbf{D}^{-1/2}\Sigma\mathbf{D}^{-1/2})]$. Note that, as long as $K_0 < \infty$, IR will be better than random guessing and hence superior to LDA. In particular, if $K_0 = 1$, IR asymptotically achieves the Bayes error rate.

FEATURES ANNEALED INDEPENDENCE RULE

The FAIR[8] utilizes an independence screening method to perform variable selection in IR. Note that, since a diagonal matrix is used in IR, the only useful variables would be those with different means across classes. A most classical way to detect such features is of course the *t*-test. Define the *t*-statistic

$$t_j = \frac{\hat{\mu}_{2j} - \hat{\mu}_{1j}}{\sqrt{s_{1j}^2/n_1 + s_{2j}^2/n_2}}. \tag{12}$$

where s_{kj}^2 is the sample variance of X_j in class k. Then FAIR selects the variables in $\hat{S} = \{j : |t_j| \text{ is among the first } s_n\text{th largest}\}$. FAIR classifies an observation according to the following rule:

$$\delta_{\mathrm{FAIR}}(\mathbf{X}) = \mathbf{1}\left(\left(\mathbf{X}_{\hat{S}} - \frac{1}{2}(\hat{\mu}_{1\hat{S}} + \hat{\mu}_{2\hat{S}})\right)^T \times \hat{\mathbf{D}}_{\hat{S}\hat{S}}^{-1}(\hat{\mu}_{2\hat{S}} - \hat{\mu}_{1\hat{S}}) + \log\frac{n_2}{n_1} > 0\right) + 1. \tag{13}$$

If Σ has a diagonal structure, FAIR can select all the useful variables with overwhelming probability even when $\log p = o(n)$. In practice, we first sort $|t_j|$'s in a decreasing order and set

$$s_n = \arg\max_s \frac{1}{\hat{\lambda}_{\max}^s} \frac{n\left[\sum_{j=1}^s t_j^2 + s(n_1 - n_2)/n\right]^2}{sn_1n_2 + n_1n_2\sum_{j=1}^s t_j^2}, \tag{14}$$

where $\hat{\lambda}_{\max}^s$ is the largest eigenvalue of the upper-left $s \times s$ block of $\hat{\mathbf{D}}^{-1/2}\hat{\Sigma}\hat{\mathbf{D}}^{-1/2}$, to minimize an upper bound of the classification error.

NEAREST SHRUNKEN CENTROIDS CLASSIFIER

The nearest shrunken centroids classifier[9] prediction analysis for microarrays (PAM) is another variant of LDA applicable to high-dimensional data. It was actually proposed prior to IR. But we discuss it after IR and FAIR for the sake of presentation. As FAIR, it estimates Σ with a diagonal matrix and performs variable selection, but is slightly different from FAIR in both directions. First, it estimates Σ with a diagonal estimator

$$\tilde{\Sigma} = \hat{\mathbf{D}} + s_0^2\mathbf{I} \tag{15}$$

for some small constant $s_0 > 0$, where $\mathbf{I}$ is the identity matrix and $\hat{\mathbf{D}}$ is defined as in Eq. 7. Second, it performs variable selection via soft-thresholding rather than hard-thresholding. Define

$$t_{kj}^* = \frac{\hat{\mu}_{kj} - \hat{\overline{\mu}}_j}{m_k(s_j + s_0)}, \quad k = 1, 2,\ j = 1, \ldots, p, \tag{16}$$

where $\overline{\mu}_j = \frac{1}{n}\sum_{i=1}^n X_j^i$, $m_k = \sqrt{(1/n_k) + (1/n)}$ and $s_j^2 = \hat{D}_{jj}$. PAM shrinks t_j toward zero:

$$t'_{kj} = \mathrm{sign}(t_{kj}^*)(|t_{kj}^*| - \Delta)_+. \tag{17}$$

Then

$$\hat{\mu}'_{kj} = \hat{\overline{\mu}}_j + m_k(s_j + s_0)t'_{kj}. \tag{18}$$

Finally, PAM substitutes Eqs. 15 and 18 into Eq. 3:

$$\delta_{\mathrm{PAM}}(\mathbf{X}) = \arg\max_k (\mathbf{X} - \hat{\mu}'_k)^T\tilde{\Sigma}^{-1}(\mathbf{X} - \hat{\mu}'_k) - 2\hat{\pi}_k. \tag{19}$$

The presence of s_0 in Eq. 16 protects X_j's from having large $|t_{kj}|$'s by chance. The formula 18 brings variable selection into PAM. When Δ is sufficiently large, one will typically have many X_j's with $\hat{\mu}'_{1j} = \hat{\mu}'_{2j} = \hat{\overline{\mu}}_j$. These variables will consequently have no effect on $\delta_{\mathrm{PAM}}(\mathbf{X})$. In practice, s_0 is set to be the median of s_j's, while the amount of soft thresholding Δ is determined by cross validation. An R package for PAM is `pamr`.

ISSUES IN HIGH-DIMENSIONAL LDA

Now we are at a point where we can fully understand the challenges for LDA in high dimensions. Again, the first issue is the singularity of $\hat{\Sigma}$, which is the motivation of IR. Friedman[18] is an even earlier proposal that replaces $\hat{\Sigma}^{-1}$ with better estimates, although it was not specifically developed for high-dimensional data. Recently, many more developments exist in covariance matrix estimation.[19–22] These covariance estimators could also be used to generalize LDA.

However, an accurate estimate of Σ does not guarantee better classification rule. Fan and Fan[8] provided proofs that show, even when the independence structure is true, the signal can be swamped by the noises from estimating the means. Consider the following parameter space

$$\Gamma^* = \{(\mu_1, \mu_2, \Sigma) : (\mu_2 - \mu_1)^T\Sigma^{-1}(\mu_2 - \mu_1) \geq c_p, \lambda_{\max}(\mathbf{D}^{-1/2}\Sigma\mathbf{D}^{-1/2}) \leq b_0, \min D_{jj} > 0\}. \tag{20}$$

Note that Γ^* is larger than Γ in Eq. 8. Because now there is no restriction on the ℓ_2-norms of $\hat{\mu}_1, \hat{\mu}_2$, consider the IR with $\hat{\mu}_j$'s as estimators of μ_j's. We refer to this rule as IR*. Fan and Fan[8] showed that $R_{\Gamma^*}(\delta_{\mathrm{IR}^*}) \to 1 - \Phi(D_0/2\sqrt{b_0})$, where $D_0 = \lim_{n\to\infty}(\sqrt{n_1 n_2/np})c_p$. Then $R_{\Gamma^*}(\mathrm{IR}^*) \to 1/2$ if $D_0 \to 0$ and IR* will also be no better than random guessing. For example, in the extreme case that only the first s variables are useful, it can be easily shown that IR* is asymptotically no better than random guessing. Therefore, variable selection is still critical even if we can estimate Σ accurately, which is why FAIR explicitly regularizes the means by hard thresholding. Other proposals that regularize both Σ and μ_j include Refs 14 and 23.

The last and most severe issue is that individually regularizing Σ and μ_j can actually be misleading in terms of both variable selection and classification. Note that, with the assumption that Σ is diagonal, the target of FAIR and PAM is essentially $S = \{j : \mu_{1j} \neq \mu_{2j}\}$. However, from Eq. 3 we see that $\mathbf{X}$ affects classification only through its projection on the discriminant direction, $\beta^{\mathrm{Bayes}} = \Sigma^{-1}(\mu_2 - \mu_1)$. Therefore, the important variables are those in the set $D = \{j : \beta_j^{\mathrm{Bayes}} \neq 0\}$. Mai et al.[12] proved that D and S can be very different. In particular, $D \subset S$ if and only if $\Sigma_{S^c,S}\Sigma_{S,S}^{-1}(\mu_{2,S} - \mu_{1,S}) = 0$, while $S \subset D$ if and only if $\mu_{2,D^c} = \mu_{1,D^c}$ or $\Sigma_{D^c,D}\Sigma_{D,D}^{-1}(\mu_{2,D} - \mu_{1,D}) = 0$. Hence, examples can be easily constructed such that variable selection based on the IR could select a wrong set of variables. On the other hand, Cai and Liu[10] illustrated through a bivariate example that IR with true parameters will have an error rate close to 1/2, while the Bayes error rate is almost 0.

A class of high-dimensional LDA methods have been proposed that assume the sparsity of β^{Bayes}, that is, $|D| \ll p$. Then these methods borrow the well-developed idea of penalization in regression to regularize the estimated discriminant direction. See Refs 24–33 for an incomplete list of works in penalization. In the following sections, we are going to present the linear programming discriminant (LPD) rule[10] regularized optimal affine discriminant analysis[11] (ROAD) and direct sparse discriminant analysis[12] (DSDA) for binary classification, and the ℓ_1-Fisher's discriminant analysis[16] (FSDA) and sparse optimal scoring[13] (SOS) for multiclass classification, as some examples of high-dimensional LDA methods that do not require Σ to be diagonal.

LINEAR PROGRAMMING DISCRIMINANT

The LPD rule[10] aims to find a sparse approximation of β^{Bayes}. It is derived from the observation that

$$\Sigma\beta^{\mathrm{Bayes}} = \mu_2 - \mu_1. \tag{21}$$

Hence, LPD estimates β^{Bayes} by

$$\hat{\beta}^{\mathrm{LPD}} = \arg\min_{\beta} \|\beta\|_1, \text{ s.t. } \|\hat{\Sigma}\beta - (\hat{\mu}_2 - \hat{\mu}_1)\|_\infty \leq \lambda. \tag{22}$$

The resulting $\hat{\beta}^{\mathrm{LPD}}$ is typically sparse. Note that Eq. 22 is similar to Dantzig selector[34] in the regression context. Indeed, $\hat{\beta}^{\mathrm{LPD}}$ can be found by the primal-dual interior-point method[35] as the Dantzig selector.

By assuming $\pi_1 = \pi_2$, LPD classifies an observation to Class 2 if

$$(\mathbf{X} - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2))^T\hat{\beta}^{\mathrm{LPD}} > 0.$$

For theoretical consideration, LPD is consistent in the sense that the error rate of LPD will tend to the Bayes error rate under proper regularity conditions.

REGULARIZED OPTIMAL AFFINE DISCRIMINANT

ROAD analysis generalizes LDA from a different angle from LPD. It intends to minimize the classification error rate. As in LPD, ROAD assumes that $\pi_1 = \pi_2 = 1/2$. Then for any $\tilde{\beta}$, the expected error rate is

$$R(\tilde{\beta}) = 1 - \Phi\left(\frac{\tilde{\beta}^T(\mu_2 - \mu_1)}{2(\tilde{\beta}^T\Sigma\tilde{\beta})^{1/2}}\right). \tag{23}$$

Therefore, by minimizaing $\tilde{\beta}^T\Sigma\tilde{\beta}$ subject to $\tilde{\beta}^T(\mu_2 - \mu_1)/2 = 1$, we can minimize the expected error rate.

In order to encourage sparsity in high-dimensional datasets, ROAD uses the following formula to estimate the direction.

$$\begin{aligned}\hat{\beta}^{\mathrm{ROAD}} &= \arg\min_{\beta} \beta^T\hat{\Sigma}\beta, \text{ s.t. } (\hat{\mu}_2 - \hat{\mu}_1)^T\beta/2 \\ &= 1 \text{ and } \sum_{j=1}^{p} |\beta_j| \leq \tau.\end{aligned} \tag{24}$$

The constraint $\sum_{j=1}^{p} |\beta_j| \leq \tau$ usually leads to sparse $\hat{\beta}^{\mathrm{ROAD}}$. ROAD is shown to be able to asymptotically achieve the Bayes error rate as n tends to infinity, without extra conditions on Σ. In the case that we have prior knowledge of Σ, we can directly substitute an appropriate estimator in 24. With $\hat{\beta}^{\text{ROAD}}$, the prediction is

$$\hat{Y} = \mathbf{1}((\mathbf{X} - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2))^T \hat{\beta}^{\text{ROAD}} > 0) + 1.$$

For implementation, ROAD approximates Eq. 24 by

$$\hat{\beta}_{\gamma}^{\text{ROAD}} = \arg\min \frac{1}{2}\beta^T \hat{\Sigma}\beta + \lambda\|\beta\|_1 + \frac{1}{2}\gamma(\beta^T(\hat{\mu}_2 - \hat{\mu}_1)/2 - 1)^2. \quad (25)$$

Ideally, if $\gamma \to \infty$, $\hat{\beta}_{\gamma}^{\text{ROAD}} \to \hat{\beta}^{\text{ROAD}}$. In practice, $\hat{\beta}_{\gamma}^{\text{ROAD}}$ is insensitive to the choice of γ as long as it is reasonably large. For a fixed γ, ROAD uses coordinate descent to compute $\hat{\beta}_{\gamma}^{\text{ROAD}}$.[36,37] The matlab code for ROAD can be found at http://www.mathworks.com/matlabcentral/fileexchange/40047. Also, ROAD is theoretically justified. The original paper provides the rates of convergence for both $\hat{\beta}^{\text{ROAD}}$ and the error rate.

A slightly different formula was proposed by Wu et al.[17]:

$$\tilde{\beta}^{\text{ROAD}} = \arg\min_{\beta} \beta^T \hat{\Sigma}\beta, \text{s.t. } (\hat{\mu}_2 - \hat{\mu}_1)^T\beta = 1 \text{ and } \|\beta\|_1 \le \tau \quad (26)$$

for some tuning parameter $\tau > 0$. For the same purpose as in ROAD, Wu et al.[17] added an ℓ_1 constraint to Eq. 30. Their implementation is available at http://www.bios.unc.edu/mwu/software/sLDA/SLDA Pathway.R.

DIRECT SPARSE DISCRIMINANT ANALYSIS

DSDA[12] is another high-dimensional LDA method. As LPD and ROAD, it is theoretically justified, while it is more convenient in terms of computation, because it recasts LDA as a linear regression problem. Define

$$(\hat{\beta}_0^{\text{ols}}, \hat{\beta}^{\text{ols}}) = \arg\min_{(\beta_0,\beta)} \sum_{i=1}^{n} (Y_i - \beta_0 - \mathbf{X}_i^T\beta)^2, \quad (27)$$

where Y_i is still the class label but treated as continuous. Then, by Hastie et al.,[3] β^{ols} and $\hat{\beta}^{\text{Bayes}}$ share the same direction, that is, $\hat{\beta}^{\text{ols}} \propto \hat{\beta}^{\text{Bayes}}$. Therefore, DSDA obtains a sparse estimator by

$$\hat{\beta}^{\text{DSDA}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \beta_0 - \mathbf{X}_i^T\beta)^2 + P_{\lambda}(\beta), \quad (28)$$

where P_{λ} is a penalty function. In general, any penalty function can be applied in DSDA to achieve specific goals for variable selection. Because Eq. 28 has the same form as the least squares estimator, efficient algorithms for penalized linear regression can be directly applied. For example, when P_{λ} is taken to be the Lasso penalty, $P_{\lambda}(\beta) = \lambda\|\beta\|_1$, one could make use of the extremely efficient algorithms in `lars`[38] or `glmnet`[39] to compute the solution path for DSDA. Also, with overwhelming probability, DSDA with the Lasso penalty will consistently select all the useful variables and estimate the direction of β^{Bayes}. So will DSDA with the smooth clipped absolute deviation penalty.[25]

Another feature of DSDA that distinguishes it from LPD and ROAD is that it can deal with problems with unequal prior probabilities. With $\hat{\beta}^{\text{DSDA}}$, DSDA further computes

$$\hat{\beta}_0^{\text{DSDA}} = -\frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)^T \hat{\beta}^{\text{DSDA}} + (\log\frac{n_2}{n_1})\frac{(\hat{\beta}^{\text{DSDA}})^T \hat{\Sigma}\hat{\beta}^{\text{DSDA}}}{(\hat{\mu}_2 - \hat{\mu}_1)^T \hat{\beta}^{\text{DSDA}}}. \quad (29)$$

Then DSDA classifies an observation to Class 2 if

$$\mathbf{X}^T \hat{\beta}^{\text{DSDA}} + \hat{\beta}_0^{\text{DSDA}} > 0.$$

THE ℓ_1-FISHER'S DISCRIMINANT ANALYSIS

All the methods discussed above are designed for binary classification except for PAM. Now we introduce two methods that deal with K-class problems, where $K \ge 2$. We start with the FSDA.[16] Define $\mathbf{Y}^{\text{dm}}$ as an $n \times K$ matrix of dummy variables with $\mathbf{Y}_{ik}^{\text{dm}} = 1(Y_i = k)$. FSDA regularizes a variant of Fisher's discriminant analysis:

$$\hat{\beta}_k = \arg\max_{\beta_k} \beta_k^T \hat{\Sigma}_b^k \beta_k \text{ s.t. } \beta_k^T \tilde{\Sigma}\beta_k \le 1, \quad (30)$$

where

$$\hat{\Sigma}_b^k = \mathbf{X}^T\mathbf{Y}^{\text{dm}}((\mathbf{Y}^{\text{dm}})^T\mathbf{Y}^{\text{dm}})^{-1/2} \times \mathbf{\Omega}_k((\mathbf{Y}^{\text{dm}})^T\mathbf{Y}^{\text{dm}})^{-1/2}(\mathbf{Y}^{\text{dm}})^T\mathbf{X} \quad (31)$$

and $\mathbf{\Omega}_k$ is the identity matrix if $k = 1$ and otherwise an orthogonal projection matrix with column space orthogonal to $((\mathbf{Y}^{\text{dm}})^T\mathbf{Y})^{-1/2}\mathbf{Y}^T\mathbf{X}\hat{\beta}_l$ for all $l < k$.

To generalize Eq. (30) to high dimensions, FSDA estimates the discriminant directions by

$$\hat{\beta}_k = \arg\max_{\beta_k} \beta_k^T \hat{\Sigma}_b^k \beta_k \text{s.t. } \beta_k^T \tilde{\Sigma} \beta_k \leq 1, \tag{32}$$

where

$$\begin{aligned} \hat{\Sigma}_b^k &= \mathbf{X}^T \mathbf{Y}^{\mathrm{dm}}((\mathbf{Y}^{\mathrm{dm}})^T \mathbf{Y}^{\mathrm{dm}})^{-1/2} \\ &\quad \times \mathbf{\Omega}_k((\mathbf{Y}^{\mathrm{dm}})^T \mathbf{Y}^{\mathrm{dm}})^{-1/2}(\mathbf{Y}^{\mathrm{dm}})^T \mathbf{X}, \qquad (33) \\ \hat{\beta}_k^{\mathrm{FSDA}} &= \arg\max_{\beta_k} \beta_k^T \hat{\Sigma}_b \beta_k \\ &\quad - P_{\lambda_k}(\beta_k), \text{ s.t. } \beta_k^T \tilde{\Sigma} \beta_k \leq 1, \qquad (34) \end{aligned}$$

with $\tilde{\Sigma}$ being a positive definite estimation of Σ, such as $\hat{\mathrm{D}}$ in IR, and $P_{\lambda_k}(\beta_k)$ a penalty function such as Lasso and fused Lasso.

The estimator $\hat{\beta}^{\mathrm{FSDA}}$ can be found by iteration. First, we initialize Eq. 34 with $\hat{\beta}^{(0)}$ equal to the first eigenvector of $\tilde{\Sigma}^{-1}\hat{\Sigma}_b$. Then one can iteratively solve for $\beta^{(m)}$ as follows:

$$\beta_k^{(m)} = \max_{\beta_k}\{2\beta_k \hat{\Sigma}_b^k \beta_k^{(m-1)} - P_{\lambda_k}(\beta_k)\}, \text{s.t. } \beta^T \tilde{\Sigma} \beta_k \leq 1. \tag{35}$$

This is indeed the algorithm implemented in the R package `penalizedLDA`.

SPARSE OPTIMAL SCORING

SOS[13] also tackles multiclass problems. It adds a penalty to the optimal scoring formulation. Again, $\mathbf{Y}^{\mathrm{dm}}$ is a matrix of dummy variables. Further define a K-dimensional vector θ_k of scores. The optimal scoring problem is formulated as

$$\begin{aligned} (\hat{\theta}_k, \hat{\beta}_k) &= \arg\min_{\theta_k, \beta_k} \sum_{i=1}^{n} (\mathbf{Y}^{\mathrm{dm}}\theta_k - \mathbf{X}\beta_k)^2 \\ &\text{s.t. } \frac{1}{n}\theta_k^T (\mathbf{Y}^{\mathrm{dm}})^T \mathbf{Y}^{\mathrm{dm}} \theta_k = 1, \\ &\theta_k^T (\mathbf{Y}^{\mathrm{dm}})^T \mathbf{Y}^{\mathrm{dm}} \theta_l = 0, l < k. \qquad (36) \end{aligned}$$

To perform variable selection, SOS combines the optimal scoring with ℓ_1 penalty:

$$\begin{aligned} (\hat{\theta}_k, \hat{\beta}_k^{\mathrm{SOS}}) &= \arg\min_{\theta_k, \beta_k} \sum_{i=1}^{n} (\mathbf{Y}^{\mathrm{dm}}\theta_k - \mathbf{X}\beta_k)^2 + \lambda \|\beta_k\|_1 \\ &\text{s.t. } \frac{1}{n}\theta_k^T (\mathbf{Y}^{\mathrm{dm}})^T \mathbf{Y}^{\mathrm{dm}} \theta_k = 1, \\ &\theta_k^T (\mathbf{Y}^{\mathrm{dm}})^T \mathbf{Y}^{\mathrm{dm}} \theta_l = 0, l < k. \qquad (37) \end{aligned}$$

Because SOS assumes that $\mathbf{X}$ is centered, Eq. (37) does not involve the intercept term. Then SOS applies LDA to $(\mathbf{X}^T \hat{\beta}_1^{\mathrm{SOS}}, \ldots, \mathbf{X}^T \hat{\beta}_{K-1}^{\mathrm{SOS}})$.

SOS can also be solved by iterative algorithms, because for fixed β_k, θ_k can be easily found, and vice versa. Such an algorithm is implemented in the R package `sparseLDA`.

EMPIRICAL COMPARISON

We compare the performance of some high-dimensional LDA methods on two benchmark datasets, the colon cancer dataset[40] and the prostate cancer dataset.[41] The goal on both of these datasets is to predict whether a person has the specific sort of cancer. To test the classifiers, both datasets are split with a 2:1 ratio to form training and testing sets. The numerical comparison includes are five methods: DSDA, FSDA, ROAD, PAM, and FAIR. Note that, ROAD in Table 1 is actually the implementation of Ref 17. LPD was not included because there is no publicly available implementation, while SOS was left out because it is essentially DSDA in the binary classification problem.[42]

A comparison of accuracies in classification and variable selection is listed in Table 1. These numerical results are directly quoted from Mai et al.[12] Overall, all the five methods have similar performances on the colon cancer dataset, while DSDA is the best on the

TABLE 1 | Comparison of high-dimensional LDA methods. Errors and fitted model sizes are medians of 100 replicates. Numbers in parentheses are standard errors.

		DSDA	FSDA	ROAD	PAM	FAIR
Colon	Error (%)	86.4	86.4	84.1	86.4	86.4
		(1.54)	(0.49)	(2.17)	(1.20)	(0.61)
	Fitted model size	5	10	1	89	11
		(0.63)	(1.39)	(0)	(29.95)	(1.19)
Prostate	Error (%)	94.1	91.2	91.2	91.2	76.5
		(0.55)	(0.24)	(0.70)	(0.96)	(0.54)
	Fitted model size	10	18	1	10	4
		(0.77)	(4.45)	(0)	(0.84)	(0.40)

prostate cancer dataset. Also, it is worth pointing out that, on the colon cancer dataset, the accuracy of 86.4% is the highest in the literature, while, before the developments of the high-dimensional method, BagBoost[5] was the most accurate classifier on the prostate cancer dataset, with an accuracy of 92.5%. However, DSDA has an accuracy of 94.1% on this dataset and outperforms BagBoost. These empirical results further suggest that LDA is a competitive candidate even for high-dimensional data.

DISCUSSION

Generalization of LDA to high dimensions has attracted much attention of statisticians. Such proposals are rejuvenating the study of this classical classifier. In general, there are three issues with LDA in high dimensions: the singularity of $\hat{\Sigma}$, the estimation accuracy of $\hat{\mu}_j$ and the simultaneous modeling of Σ and μ_j. To tackle some or all of these issues, a common theme in high-dimensional LDA proposals is variable selection via regularization, either on the covariance, the means or the discriminant direction. Therefore, high-dimensional LDA is closely related to many hot topics in modern statistics, including penalization methods and covariance estimation. Moreover, as pointed out by a referee, high-dimensional LDA methods are also important topics outside statistics. These methods typically modify Fisher's view of discriminant analysis in two ways. They use kernel methods to relax the normality assumption and employ proper regularization to overcome the impact of high dimensionality. Some such papers are Refs 43–46. Therefore, although many successful high-dimensional LDA methods are present, we expect further advancement in this area in the future.

ACKNOWLEDGMENTS

The author thanks the editor and three referees for their helpful comments and suggestions. This work is supported in part by a grant from National Science Foundation of USA.

REFERENCES

1. Michie D, Spiegelhalter D, Taylor C. *Machine Learning, Neural and Statistical Classification*. 1st edn. Ellis Horwood; New York, 1994.
2. Hand DJ. Classifier technology and the illusion of progress. *Stat Sci* 2006, 21:1–14.
3. Hastie T, Tibshirani R, Friedman JH. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edn. Springer Verlag; New York, 2008.
4. Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. *Stat Sinica* 2010, 20:101–148.
5. Dettling M. Bagboosting for tumor classification with gene expression data. *Bioinformatics* 2004, 20:3583–3593.
6. Hall P, Marron JS, Neeman A. Geometric representation of high dimension, low sample size data. *J R Stat Soc B* 2005, 67:427–444.
7. Bickel PJ, Levina E. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* 2004, 10:989–1010.
8. Fan J, Fan Y. High dimensional classification using features annealed independence rules. *Ann Stat* 2008, 36:2605–2637.
9. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci* 2002, 99:6567–6572.
10. Cai T, Liu W. A direct estimation approach to sparse linear discriminant analysis. *J Am Stat Assoc* 2011, 106:1566–1577.
11. Fan J, Feng Y, Tong X. A ROAD to classification in high dimensional space. *J R Stat Soc Ser B* 2012, 74:745–771.
12. Mai Q, Zou H, Yuan M. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* 2012, 99:29–42.
13. Clemmensen L, Hastie T, Ersbøll B. Sparse discriminant analysis. *Technometrics* 2011, 53:406–413.
14. Shao J, Wang Y, Deng X, Wang S. Sparse linear discriminant analysis with high dimensional data. *Ann Stat* 2011, 39: 1241–1265.
15. Trendafilov NT, Jolliffe IT. DALASS: Variable selection in discriminant analysis via the lasso. *Comput Stat Data Anal* 2007, 51:3718–3736.
16. Witten D, Tibshirani R. Penalized classification using Fisher's linear discriminant. *J R Stat Soc Ser B* 2011, 73:753–772.
17. Wu M, Zhang L, Wang Z, Christiani D, Lin X. Sparse linear discriminant analysis for simultaneous testing for

the significance of a gene set/pathway and gene selection. *Bioinformatics* 2008, 25:1145–1151.

18. Friedman JH. Regularized discriminant analysis. *J Am Stat Assoc* 1989, 84:165–175.
19. Bickel PJ, Levina E. Regularized estimation of large covariance matrices. *Ann Stat* 2008, 36:199–227.
20. Cai T, Zhang C, Zhou H. Optimal rates of convergence for covariance matrix estimation. *Ann Stat* 2010, 38:2118–2144.
21. Rothman A, Bickel P, Levina E, Zhu J. Sparse permutation invariant covariance estimation. *Electron J Stat* 2008, 2:494–515.
22. Cai T, Zhou H. Minimax estimation of large covariance matrices under ℓ_1 norm (with discussion). *Stat Sinica* 2012, 22:1319–1378.
23. Guo Y, Hastie T, Tibshirani R. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* 2007, 8:86–100.
24. Frank I, Friedman JH. A statistical view of some chemometrics regression tools (with discussion). *Technometrics* 1993, 35:109–148.
25. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001, 96:1348–1360.
26. Lv J, Fan Y. A unified approach to model selection and sparse recovery using regularized least squares. *Ann Stat* 2009, 37:3498–3528.
27. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B* 2005, 67:301–320.
28. Zou H. The adaptive Lasso and its oracle properties. *J Am Stat Assoc* 2006, 101:1418–1429.
29. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B* 1996, 58:267–288.
30. Tibshirani R, Saunders M, Rosset S, Zhu J, Keith K. Sparsity and smoothness via the fused lasso. *J R Stat Soc Ser B* 2005, 67:91–108.
31. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc Ser B* 2006, 68:49–67.
32. Zhang C. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat* 2010, 38:894–942.
33. Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann Stat* 2008, 36:1509–1533.
34. Candes E, Tao T. The Dantzig selector: statistical estimation when p is much larger than n. *Ann Stat* 2007, 35:2313–2351.
35. Boyd S, Vandenberghe L. *Convex Optimization*. Cambridge University Press; New York, 2004.
36. Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann Appl Stat* 2011, 232–253.
37. Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. *J Optim Theory Appl* 2001, 109:47–494.
38. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat* 2004, 32:407–499.
39. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2008, 33:1–22.
40. Alon U, Barkai N, Notterman D, Gish K, Mack S, Levine J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci* 1999, 96:6745–6750.
41. Singh D, Febbo P, Ross K, Jackson D, Manola J, Ladd C, Tamayo P, Renshaw A, D'Amico A, Richie J, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 2002, 1:203–209.
42. Mai Q, Zou H. On the connection and equivalence of three sparse linear discriminant analysis methods. *Technometrics* In press. 2012.
43. Lu J, Plataniotis K, Venetsanopoulos A, Wang J. An efficient kernel discriminant analysis method. *Pattern Recogn* 2005, 38:1788–1790.
44. Yu H, Yang J. A direct LDA algorithm for high-dimensional data—with application to face recognition. *Pattern Recogn* 2001, 34:2067–2070.
45. Ye J, Li T, Xiong T, Janardan R. Using uncorrelated discriminant analysis for tissue classification with gene expression data. *IEEE/ACM Trans Comput Biol Bioinform* 2004, 1:181–190.
46. Ji S, Ye J. Generalized linear discriminant analysis: a unified framework and efficient model selection. *IEEE Trans Neural Netw* 2008, 19:1768–1782.