Songklanakarin J. Sci. Technol.
41 (2), 319-331, Mar. – Apr. 2019

http://www.sjst.psu.ac.th

Original Article

Discriminant methods for high dimensional data

Poompong Kaewumpai* and Samruam Chongcharoen

Graduate School of Applied Statistics, National Institute of Development Administration (NIDA), Bang Kapi, Bangkok, 10240, Thailand

Received: 12 July 2017; Revised: 4 October 2017; Accepted: 16 November 2017

Abstract

The main purpose of discriminant analysis is to enable classification of new observations into one of *g* classes or populations. Discriminant methods suffer when applied to high dimensional data because the sample covariance matrix is singular. In this study, we propose two new discriminant methods for high dimensional data under the multivariate normal population with a block diagonal covariance matrix structure. As the first method, we approximate the sample covariance matrix as a singular matrix based on the idea of reducing the dimensionality of the observations to get a well-conditioned covariance matrix. As the second method, we use a block diagonal sample covariance matrix instead. The performances of these two methods are compared with some of the existing methods in a simulation study**.** The results show that both proposed methods outperform other comparative methods in various situations**.** In addition, the two new proposed methods for discriminant analysis are applied to a real dataset.

Keywords: discriminant analysis, high dimensional data, classification, inverse of covariance matrix, block diagonal matrix

1. Introduction

Discriminant analysis is a multivariate technique for sample classification. The objective of discriminant analysis is to construct appropriate rules for assigning new observations to one of *g* classes or populations. Assume that there are *g* different classes, $\Pi_1, \Pi_2, ..., \Pi_g$, each with a multivariate normal distribution with mean vectors $\underset{\sim}{\mu}_h$ and covariance matrices Σ_h, $h = 1, 2, ..., g$ in $p \times p$ dimensions**.** In this study, we are only interested in the population covariance matrices that have this block diagonal structure

$$\begin{pmatrix} \Sigma_{11h} & O_{12h} & \cdots & O_{1mh} \\ O_{21} & \Sigma_{22h} & \cdots & O_{2mh} \\ \vdots & \vdots & \ddots & \vdots \\ O_{m1h} & O_{m2h} & \cdots & \Sigma_{mmh} \end{pmatrix},$$

where Σ_{iih} is a $p_i \times p_i$ submatrix and O_{ijh} is a $p_i \times p_j$ zero matrix, $i, j = 1, 2, ..., m$.

Now, suppose that $\underset{\sim}{x}_{jh}^T = (x_{1jh}, x_{2jh}, ..., x_{pjh})$, $j = 1, 2, ..., n_h$, $n_h > p$, is a random sample of n_h observations that are collected from Π_h. The basic method for classifying new observations $\underset{\sim}{x}$ to one of the *g* classes is as follows:

Assign $\underset{\sim}{x}$ to Π_h if $D_h(\underset{\sim}{x}) < D_l(\underset{\sim}{x})$ for all $h \neq l = 1, 2, ..., g$, where $D_h(\underset{\sim}{x}) = (\underset{\sim}{x} - \underset{\sim}{\mu}_h)^T \Sigma_h^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}_h) + \ln|\Sigma_h| - 2\ln p_h$ is called

*Corresponding author
Email address: poompongk@gmail.com

the "discriminant score" and p_h is the prior probability of Π_h.

The $\underset{\sim}{\mu}_h$ and Σ_h are unknown, but the unbiased estimators of these parameters are $\underset{\sim}{\bar{x}}_h = \frac{1}{n_h}\sum_{j=1}^{n_h} \underset{\sim}{x}_{jh}$ and $S_h = \frac{1}{n_h - 1}\sum_{j=1}^{n_h}(\underset{\sim}{x}_{jh} - \underset{\sim}{\bar{x}}_h)(\underset{\sim}{x}_{jh} - \underset{\sim}{\bar{x}}_h)^T$. When the covariance matrices are equal, that is $\Sigma_h = \Sigma$ for all h, their common covariance matrix Σ is estimated by $S_{pooled} = \frac{1}{(N-g)}\sum_{h=1}^{g}(n_h - 1)S_h$, where S_{pooled} is the pooled sample covariance matrix and $N = \sum_{h=1}^{g} n_h$. In this study, we are interested in the case of common covariance matrix. Thus, the estimate of discriminant score is $\hat{D}_h(\underset{\sim}{x}) = (\underset{\sim}{x} - \underset{\sim}{\bar{x}}_h)^T S_{pooled}^{-1}(\underset{\sim}{x} - \underset{\sim}{\bar{x}}_h) - 2\ln p_h$.

Discriminant analysis is widely used in many scientific domains, such as medical research, financial analysis, computer vision, etc. These sources provide high dimensional data, which means that the number of observations is less than the dimensionality of the observations. In high dimensional data, discriminant analysis cannot be applied directly because the sample covariance matrix is singular, i.e. the inverse of the sample covariance matrix does not exist. Di Pillo (1976) stated that the performance of discriminant analysis in high dimensional data is far from optimal. The generalized inverse substitutes for the inverse of a singular sample covariance matrix. While simple, this method might have poor performance since the generalized inverse may be very unstable (Guo, Hastie, & Tibshirani, 2007).

In this situation, the cause of the problem with discriminant analysis is singularity of the sample covariance matrix. There are two often used ways to address this problem. The first is a subspace approach. For example, the well-known Fisherfaces method (Belhumeur, Hespanha, & Kriegman, 1997) and Chen, Liao, Ko, Lin, and Yu (2000) presented direct linear discriminant analysis (D-LDA). Lu, Plataniotis, and Venetsanopoulos (2003) proposed a new discriminant analysis method called regularized direct quadratic discriminant analysis (RD-QDA) by combining the D-LDA method with regularized discriminant analysis (RDA), previously proposed by Friedman (1989). The second is to apply linear algebra to solve the singularity problem. For example, Tian, Barbero, Gu, and Lee (1986) utilized the pseudoinverse as estimate of inverse of the sample covariance matrix. Additionally, Dudoit, Fridlyand, and Speed (2002) used a diagonal covariance matrix, and Srivastava and Kubokawa (2007) used an empirical Bayes estimator instead of the sample covariance matrix.

In this paper, two new discriminant methods are proposed to deal with high dimensional data. In the first proposed method, we reduce the dimensionality of the observations by taking linear combinations of $\underset{\sim}{x}_{jh}$ to create $\underset{\sim}{y}_{jh} = H^T \underset{\sim}{x}_{jh}$, where H is the matrix obtained from the RD-QDA method (Lu *et al.*, 2003), and find a well-conditioned estimator for the high dimensional covariance matrix by the expression given by Schäfer and Strimmer (2005). In the second proposed method, we use a block diagonal sample covariance matrix $S_{block} = diag(S_{11}, S_{22}, ..., S_{mm})$ where $S_{hh}, h = 1, ..., m$ are submatrices of the pooled sample covariance matrix in the discriminant score. We also review the method proposed by Dudoit *et al.* (2002), labeled DI, and the method proposed by Srivastava and Kubokawa (2007), labeled SK. Our two proposed methods are compared with the DI and SK methods in a simulation study.

Dudoit *et al.* (2002) in the DI method assumed independent variables and replaced every off-diagonal element in the sample covariance matrix with zero**.** Specifically, they replaced the pooled sample covariance

matrix in the discriminant score by $S_{DI} = diag(s_{11},...,s_{pp})$ for s_{ii}, $i = 1,2,...,p$, i.e., the diagonal part of the pooled sample covariance matrix S_{pooled}, and this gave the DI classification rule:

Assign $\underset{\sim}{x}$ to Π_h if $\hat{D}_h(\underset{\sim}{x}) < \hat{D}_l(\underset{\sim}{x})$ for all $l \neq h$, where $\hat{D}_h(\underset{\sim}{x}) = (\underset{\sim}{x} - \underset{\sim}{\bar{x}}_h)^T S_{DI}^{-1}(\underset{\sim}{x} - \underset{\sim}{\bar{x}}_h) - 2\ln p_h$.

Since S_{DI} above uses only the diagonal elements of S_{pooled}, the information represented by the off-diagonal elements was modified and lost.

Srivastava and Kubokawa (2007) derived the empi-rical Bayes estimator of Σ^{-1} as $S_{SK}^{-1} = \left(S_{pooled} + \frac{tr(S_{pooled})}{\min(n, p)} I \right)^{-1}$ and gave the SK classification rule:

Assign $\underset{\sim}{x}$ to Π_h if $\hat{D}_h(\underset{\sim}{x}) < \hat{D}_l(\underset{\sim}{x})$ for all $l \neq h$, where $\hat{D}_h(\underset{\sim}{x}) = (\underset{\sim}{x} - \underset{\sim}{\bar{x}}_h)^T S_{DI}^{-1}(\underset{\sim}{x} - \underset{\sim}{\bar{x}}_h) - 2\ln p_h$.

It may be noted that S_{SK}^{-1} exists irrespective of whether $n < p$ or $n > p$ and this method performed the best in their study.

The remainder of this paper is organized as follows. Two new discriminant methods for high dimensional data from a multivariate normal population with a block diagonal covariance matrix are proposed in Section 2. In Section 3, we assess the performance of the proposed methods and compare them with other methods in a simulation study and with real data. Section 4 is the conclusions and future work is described in Section 5.

2. The Proposed Methods

In this section, two new discriminant methods are proposed for dealing with high dimensional data. We consider only binary classification to two classes, on the condition that the population covariances of the classes are equal (two blocks on the diagonal).

2.1 The first proposed method

We reduce the dimensionality of the observations from p to q by taking linear combinations of $\underset{\sim}{x}_{jh}$ to create $\underset{\sim}{y}_{jh} = H^T \underset{\sim}{x}_{jh}$, where H is the matrix obtained from the RD-QDA method proposed by Lu *et al.* (2003). To obtain the matrix H, let $U = (\underset{\sim}{u}_1, \underset{\sim}{u}_2, ..., \underset{\sim}{u}_q)$ be the q eigenvectors of $S_b = \sum_{h=1}^{g} n_h (\underset{\sim}{\bar{x}}_h - \underset{\sim}{\bar{x}})(\underset{\sim}{\bar{x}}_h - \underset{\sim}{\bar{x}})^T$, where $\underset{\sim}{\bar{x}} = \sum_{h=1}^{g}\sum_{j=1}^{n_h} \underset{\sim}{x}_{jh} \Big/ N$, corresponding to the q nonzero eigenvalues $\omega_1, \omega_2, ..., \omega_q$.

The matrix H is given by $H = U\Lambda^{-1/2}$ where $\Lambda = diag(\omega_1, \omega_2, ..., \omega_q)$. The linear combination $\underset{\sim}{y}_{jh} = H^T \underset{\sim}{x}_{jh}$ has a multivariate normal distribution with mean vectors $H^T \underset{\sim}{\mu}_h$ and covariance matrices $\tilde{\Sigma} = H^T \Sigma H$ in $q \times q$ dimensions. Instead of finding the regularized sample covariance matrix (Friedman, 1989), which was used by Lu *et al.* (2003) in their classification rule, we use a well-conditioned estimator for high dimensional covariance matrices, namely the minimum mean squared error estimator defined by Ledoit and Wolf (2003). This it is always positive definite, even for high dimensional data (Schäfer & Strimmer, 2005).

Schäfer and Strimmer (2005) suggested using the linear shrinkage approach to obtain a well-conditioned covariance matrix Σ^* using a weighted average of the pooled sample covariance matrix and the shrinkage target matrix $T = [t_{ij}]_{q \times q}$:

$$\Sigma^* = \lambda T + (1 - \lambda) S_{y,pooled},$$

where $S_{y,pooled} = \sum_{h=1}^{g}\sum_{j=1}^{n_h}(\underset{\sim}{y}_{jh} - \underset{\sim}{\bar{y}}_h)(\underset{\sim}{y}_{jh} - \underset{\sim}{\bar{y}}_h)^T \Big/ N-g$, λ is the shrinkage intensity, and $\underset{\sim}{\bar{y}}_h = \sum_{j=1}^{n_h}\underset{\sim}{y}_{jh} \Big/ n_h$. After that, they chose λ by optimising

$$\min_{\lambda} E\left(\left\|\Sigma^* - \tilde{\Sigma}\right\|^2\right) = \min_{\lambda} E\left\|\lambda T + (1-\lambda)S_{y,pooled} - \tilde{\Sigma}\right\|^2$$
$$= \min_{\lambda} E\left[\sum_{i=1}^{q}\sum_{j=1}^{q}(\lambda t_{ij} + (1-\lambda)s_{y,ij} - \tilde{\sigma}_{ij})^2\right],$$

where $\tilde{\sigma}_{ij}$ represents the element at the i^{th} row and j^{th} column of $\tilde{\Sigma}$. Schäfer and Strimmer (2005) showed that the optimal shrinkage intensity λ^* is given by

$$\lambda^* = \frac{\sum_{i=1}^{q}\sum_{j=1}^{q}\left[Var(s_{y,ij}) - Cov(s_{y,ij}, t_{ij})\right]}{\sum_{i=1}^{q}\sum_{j=1}^{q}E\left[(t_{ij} - s_{y,ij})^2\right]}.$$

We need to estimate λ^*, and Schäfer and Strimmer (2005) emphasized the need to compute the optimal shrinkage intensity estimator $\hat{\lambda}^*$ by replacing all expectations, variances, and covariances in λ^* with their unbiased estimates.

In this study, we use the three shrinkage target matrices for $S_{y,pooled}$ that were compiled by Schäfer and Strimmer (2005) and the respective $\hat{\lambda}^*$ are as follows:

1) T = A: "Diagonal unit variance". In this case, we do not need to estimate the parameter because T is the identity matrix. Thus,

$$t_{ij} = \begin{cases} 1 & if\ i = j \\ 0 & if\ i \neq j \end{cases}$$

and $\hat{\lambda}^* = \frac{\sum_{i\neq j}Var(s_{y,ij}) + \sum_{i=1}^{q}Var(s_{y,ii})}{\sum_{i\neq j}s_{y,ij}^2 + \sum_{i=1}^{q}(s_{y,ii}-1)^2}$,

2) T = B: "Diagonal common variance". Now we need to estimate the diagonal element of T (i.e., the common variance α).

$$t_{ij} = \begin{cases} \alpha = \left(\sum_{i=1}^{q}s_{y,ii}\right)\Big/ q & if\ i = j \\ 0 & if\ i \neq j \end{cases} \text{ and } \quad \hat{\lambda}^* = \frac{\sum_{i\neq j}Var(s_{y,ij}) + \sum_{i=1}^{q}Var(s_{y,ii})}{\sum_{i\neq j}s_{y,ij}^2 + \sum_{i=1}^{q}(s_{y,ii}-\alpha)^2}, \text{ and}$$

3) T = C: "Diagonal common variance and common covariance". The shrinkage target matrix is provided by two parameters, namely the common variance α and the common covariance β, and we need to estimate both parameters. Thus,

$$t_{ij} = \begin{cases} \alpha = \left(\sum_{i=1}^{q} s_{y,ii}\right) \Big/ q & if\ i = j \\ \beta = \left(\sum_{i \neq j} s_{y,ij}\right) \Big/ q(q-1) & if\ i \neq j \end{cases} \quad \text{and} \quad \hat{\lambda}^* = \frac{\sum_{i \neq j} Var(s_{y,ij}) + \sum_{i=1}^{q} Var(s_{y,ii})}{\sum_{i \neq j} (s_{y,ij} - \beta)^2 + \sum_{i=1}^{q} (s_{y,ii} - \alpha)^2}.$$

When $\hat{\lambda}^*$ is computed, a well-conditioned estimator of the covariance matrix is given by

$$S^* = \hat{\lambda}^* T + (1 - \hat{\lambda}^*) S_{y,pooled},$$

where S^* is always positive definite, even for high-dimensional data, and has minimum mean squared error (Schäfer & Strimmer, 2005).

The first proposed classification rule is:

Assign $\underset{\sim}{x}$ to Π_h if $\hat{D}_h(\underset{\sim}{x}) < \hat{D}_l(\underset{\sim}{x})$ for all $l \neq h$, where

$$\hat{D}_h(\underset{\sim}{x}) = (H^T \underset{\sim}{x} - H^T \underset{\sim}{\bar{x}}_h)^T S^{*-1} (H^T \underset{\sim}{x} - H^T \underset{\sim}{\bar{x}}_h) - 2 \ln p_h.$$

From here on, we use symbols TA, TB, and TC for the classification rules that use the shrinkage targets A, B, and C, respectively.

2.2 The second proposed method

Recall that the population covariance matrices are assumed to be block diagonal, and the proposed method is based on constructing sample covariance matrices of similar pattern. The pooled sample covariance matrix S_{pooled} is partitioned as

$$S_{pooled} = \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1m} \\ S_{21} & S_{22} & \cdots & S_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ S_{m1} & S_{m2} & \cdots & S_{mm} \end{pmatrix} = [S_{ij}]_{p \times p},$$

where S_{ij} are submatrices of S_{pooled}, for $i, j = 1, 2, ..., m$, and the dimensions of S_{ij} are $p_i \times p_j$ and $\sum_{i=1}^{m} p_i = p$. S_{pooled} is partitioned in the same manner as Σ so that the block sizes of S_{ij} and Σ_{ij} are equal. Now, define the block diagonal matrix S_{block} as

$$S_{block} = diag(S_{11}, S_{22}, ..., S_{mm}) = \begin{pmatrix} S_{11} & O_{12} & \cdots & O_{1m} \\ O_{21} & S_{22} & \cdots & O_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ O_{m1} & O_{m2} & \cdots & S_{mm} \end{pmatrix}_{p \times p}.$$

For classifying two classes, the pooled sample covariance matrix is nonsingular when $p < \nu$ where ν is the degrees of freedom. The S_{pooled} has $n_1 + n_2 - 2$ degrees of freedom, and $S_{ii}, i = 1,2,...,m$ are submatrices of p_i dimensions with ν degrees of freedom. If we specify that $p_i < \nu$, then $S_{ii}, i = 1,2,...,m$ are all invertible (Dempster, 1958). As a result, S_{block} has the inverse

$$S_{block}^{-1} = diag(S_{11}^{-1}, S_{22}^{-1}, ..., S_{mm}^{-1}) = S_{block} = \begin{pmatrix} S_{11}^{-1} & O_{12} & \cdots & O_{1m} \\ O_{21} & S_{22}^{-1} & \cdots & O_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ O_{m1} & O_{m2} & \cdots & S_{mm}^{-1} \end{pmatrix}_{p \times p}.$$

Now, S_{block}^{-1} is used instead of S_{pooled}^{-1} because the latter does not exist for high dimensional data.

The second proposed classification rule is:

Assign $\underset{\sim}{x}$ to Π_h if $\hat{D}_h(\underset{\sim}{x}) < \hat{D}_l(\underset{\sim}{x})$ for all $l \neq h$, where

$$\hat{D}_h(\underset{\sim}{x}) = (\underset{\sim}{x} - \bar{\underset{\sim}{x}}_h)^T S_{block}^{-1} (\underset{\sim}{x} - \bar{\underset{\sim}{x}}_h) - 2\ln p_h .$$

From here on, the symbol BD is used for a classification rule that uses a block diagonal sample covariance matrix.

3. Simulation Study

In this section, the performances of the two proposed methods are compared with the DI (Dudoit *et al.*, 2002) and the SK (Srivastava & Kubokawa, 2007) methods in a simulation study, by considering their misclassification rates with 1000 iterations. The misclassification rate (M) is defined as

$$\mathrm{M} = 1 - \frac{\text{number of correct classification}}{\text{number of all observations}} .$$

Its value ranges from 0 to 1. For M=0 all the observations were assigned correct classes, while for M=1 all cases were called incorrectly. Therefore, the higher the misclassification rate the poorer the method.

3.1 The generated datasets

The datasets were generated as follows: $\underset{\sim}{x}_{j1} \sim i.i.d.N_p(\underset{\sim}{\mu}_1, \Sigma)$ and $\underset{\sim}{x}_{j2} \sim i.i.d.N_p(\underset{\sim}{\mu}_2, \Sigma)$, $j = 1,2,...,n$, where, $\underset{\sim}{\mu}_1 = (0,0,...,0)^T$, $\underset{\sim}{\mu}_2 = (\underset{\sim}{m},0,...,0)^T$, $\underset{\sim}{m}$ is a r-dimensional vector generated from Uniform(-1.5,1.5), $r = 0.05p$. We study only the binary classification to two classes with equal prior probabilities Π_1 and Π_2, i.e. each observation has equal probabilities to represent Π_1 or Π_2.

Two different forms of the population covariance matrix are used as follows:

1) The first form of covariance matrix is $\Sigma_1 = diag(\Sigma_{11}, \Sigma_{22}, ..., \Sigma_{mm})$ and $\Sigma_{ii} = (1-\theta)I_{p_i} + \theta J_{p_i}$, $\theta = 0.1, 0.5, 0.9,\ i = 1,2,...,m,$ in which J is a matrix where all elements are 1's and the dimensions of Σ_{ii} are $p_i \times p_i$ and $\sum_{i=1}^{m} p_i = p.$

2) The second form of covariance matrix is $\sum_2 = diag(\sum_{11}, \sum_{22}, ..., \sum_{mm})$ and $\sum_{ii} = [\rho_{kl}]$, $\rho_{kl} = (\theta)^{|k-l|}$, $\theta = 0.9$, $i = 1, 2, ..., m$, and $k, l = 1, 2, ..., p_i$, in which the dimensions of $\sum_{ii}$ are $p_i \times p_i$ and $\sum_{i=1}^{m} p_i = p$.

The simulations were conducted for $p \in \{100, 200, 300, 400\}$ with $n \in \{35, 70\}$. Each experiment consisted of a training dataset with 25, 50 observations, and the testing datasets had 10, 20 observations from each class. The classification rules had their parameters estimated using the training dataset, after which the classification rules were tested on the testing dataset. For each (p, n) combination, both equal and mixed block sizes were considered. In the equal block size case, all the $\sum_{ii}$ are of size $p_i = 5, 10, 25$ with p/p_i blocks, and in the mixed block size case there are two different block sizes in the matrix. The two block sizes of submatrix $\sum_{ii}$ are chosen from $p_i, p_j = 5, 10, 25$, in which size p_i has $p/2p_i$ blocks and size p_j has $p/2p_j$ blocks. The number of blocks rather than block size is considered in order to assess trends.

In each simulation, 1,000 iterations were done, and the performance of each method was evaluated based on misclassification rate.

3.2 Simulation study results

The misclassification rates (M) are reported in Tables 1-8. When $\sum = \sum_1$, the values of M are shown in Tables 1-6. For any θ, when p and n increase, the values of M for TA, TB, TC, and BD methods decrease. For fixed p and n, when θ, increases, the TA, TB, and TC methods get higher (poorer) values of M than the BD method.

When $\theta = 0.1$ and p, n are fixed with a decrease in the number of blocks, M for the BD method increased while those for TA, TB, and TC increased slightly and are stable in the mixed block sizes case. On comparing the two proposed methods with the DI and SK methods, all of these obtain similar values of M except for BD, which has slightly higher M than the others when the number of blocks decreases.

Table 1. The misclassification rate (M) of TA, TB, TC, BD, SK, and DI when $\sum = \sum_1$ and $\theta = 0.1$ with equal block sizes.

n	p	p_i	TA, TB, TC	BD	DI	SK
35	100	5	0.3026	0.3082	0.3072	0.3089
		10	0.3033	0.3173	0.3059	0.3091
		25	0.3136	0.3529	0.3150	0.3118
	200	5	0.2294	0.2358	0.2328	0.2333
		10	0.2322	0.2472	0.2378	0.2359
		25	0.2369	0.2958	0.2396	0.2323
	300	5	0.1893	0.1966	0.1931	0.1908
		10	0.1875	0.2070	0.1933	0.1859
		25	0.1946	0.2607	0.2003	0.1921
	400	5	0.1509	0.1552	0.1552	0.1519
		10	0.1495	0.1643	0.1534	0.1495
		25	0.1618	0.2293	0.1679	0.1582
70	100	5	0.2649	0.2639	0.2659	0.2751
		10	0.2629	0.2649	0.2662	0.2716
		25	0.2713	0.2789	0.2724	0.2688
	200	5	0.1856	0.1855	0.1880	0.1941
		10	0.1816	0.1845	0.1839	0.1904
		25	0.1943	0.2046	0.1968	0.1885
	300	5	0.1329	0.1340	0.1359	0.1395
		10	0.1361	0.1365	0.1387	0.1393
		25	0.1416	0.1512	0.1436	0.1362
	400	5	0.0984	0.0987	0.1002	0.1011
		10	0.0991	0.1009	0.1016	0.1016
		25	0.1075	0.1171	0.1105	0.1036

Table 2. The misclassification rate (M) of TA, TB, TC, BD, SK, and DI when $\sum = \sum_1$ and $\theta = 0.1$ with mixed block sizes.

n	p	p_i	p_j	TA, TB, TC	BD	DI	SK
35	100	5	10	0.3069	0.3201	0.3097	0.3075
		5	25	0.3082	0.3288	0.3130	0.3084
		10	25	0.3066	0.3750	0.3080	0.3064
	200	5	10	0.2312	0.2421	0.2352	0.2354
		5	25	0.2370	0.3160	0.2415	0.2365
		10	25	0.2380	0.3358	0.2426	0.2370
	300	5	10	0.1873	0.2010	0.1931	0.1902
		5	25	0.1878	0.2843	0.1926	0.1859
		10	25	0.1850	0.2945	0.1912	0.1865
	400	5	10	0.1504	0.1700	0.1555	0.1527
		5	25	0.1589	0.2298	0.1656	0.1605
		10	25	0.1594	0.2631	0.1658	0.1584
70	100	5	10	0.2698	0.2716	0.2707	0.2761
		5	25	0.2654	0.2829	0.2680	0.2741
		10	25	0.2683	0.2864	0.2706	0.2723
	200	5	10	0.1829	0.1790	0.1860	0.1903
		5	25	0.1887	0.2011	0.1918	0.1942
		10	25	0.1857	0.2113	0.1877	0.1858
	300	5	10	0.1376	0.1363	0.1411	0.1432
		5	25	0.1392	0.1483	0.1396	0.1395
		10	25	0.1402	0.1520	0.1415	0.1392
	400	5	10	0.0995	0.1020	0.1021	0.1040
		5	25	0.1054	0.1150	0.1079	0.1039
		10	25	0.1053	0.1141	0.1087	0.1055

When $\theta = 0.5$ and p, n are fixed with a decrease in the number of blocks, M for the TA, TB, and TC methods increased, while for BD method it only increased when n was small. For n large, the BD method achieves similar M for any number of blocks. When the proposed methods are compared with the DI and SK methods, the BD method performs the best in this situation with the lowest M, while the SK method performs better than the DI, TA, TB, and TC methods, since these had the highest M and poorest performance.

Table 3. The misclassification rate (M) of TA, TB, TC, BD, SK, and DI when $\sum = \sum_1$ and $\theta = 0.5$ with equal block sizes.

n	p	p_i	TA, TB, TC	BD	DI	SK
35	100	5	0.3404	0.2374	0.3415	0.3056
		10	0.3709	0.2371	0.3722	0.2942
		25	0.3905	0.2806	0.3921	0.2603
	200	5	0.2753	0.1538	0.2787	0.2538
		10	0.3121	0.1593	0.3144	0.2580
		25	0.3542	0.2048	0.3554	0.2308
	300	5	0.2359	0.1082	0.2411	0.2200
		10	0.2741	0.1071	0.2763	0.2287
		25	0.3357	0.1545	0.3363	0.2280
	400	5	0.2039	0.0732	0.2079	0.1911
		10	0.2499	0.0735	0.2516	0.2103
		25	0.3077	0.1190	0.3075	0.2153
70	100	5	0.2994	0.1958	0.2997	0.2407
		10	0.3225	0.1992	0.3241	0.2120
		25	0.3571	0.1945	0.3582	0.1890
	200	5	0.2214	0.1059	0.2231	0.1781
		10	0.2566	0.1089	0.2589	0.1584
		25	0.3110	0.1104	0.3116	0.1267
	300	5	0.1794	0.0639	0.1814	0.1445
		10	0.2191	0.0590	0.2197	0.1370
		25	0.2716	0.0660	0.2726	0.1049
	400	5	0.1458	0.0390	0.1479	0.1212
		10	0.1783	0.0328	0.1794	0.1126
		25	0.2521	0.0404	0.2528	0.0967

Table 4. The misclassification rate (M) of TA, TB, TC, BD, SK, and DI when $\sum = \sum_1$ and $\theta = 0.5$ with mixed block sizes.

n	p	p_i	p_j	TA, TB, TC	BD	DI	SK
35	100	5	10	0.3490	0.2509	0.3499	0.2968
		5	25	0.3672	0.2974	0.3688	0.2919
		10	25	0.3754	0.3282	0.3743	0.2737
	200	5	10	0.3011	0.1460	0.3047	0.2573
		5	25	0.3286	0.2381	0.3295	0.2485
		10	25	0.3442	0.2489	0.3468	0.2541
	300	5	10	0.2590	0.1059	0.2616	0.2255
		5	25	0.2944	0.1597	0.2956	0.2183
		10	25	0.3153	0.1714	0.3194	0.2399
	400	5	10	0.2253	0.0717	0.2319	0.1982
		5	25	0.2726	0.1543	0.2742	0.2093
		10	25	0.2811	0.1860	0.2836	0.2116
70	100	5	10	0.3110	0.1803	0.3118	0.2209
		5	25	0.3306	0.1890	0.3321	0.2268
		10	25	0.3404	0.1840	0.3412	0.1976
	200	5	10	0.2450	0.1141	0.2473	0.1800
		5	25	0.2761	0.1163	0.2762	0.1425
		10	25	0.2846	0.1125	0.2859	0.1451
	300	5	10	0.2049	0.0704	0.2056	0.1502
		5	25	0.2396	0.0662	0.2410	0.1163
		10	25	0.2481	0.0671	0.2502	0.1268
	400	5	10	0.1665	0.0422	0.1675	0.1212
		5	25	0.2041	0.0517	0.2064	0.1120
		10	25	0.2209	0.0429	0.2207	0.1073

When $\theta = 0.9$ and p, n are fixed with a decrease in the number of blocks, M for TA, TB, and TC methods increased while the BD method obtained the least values among the methods. In particular, the BD method is able to classify the test set nearly 100% correctly when p, n are large. The SK method performs better than the DI, TA, TB, and TC methods, and the TA, TB, and TC methods obtain the highest values of M (similar to the DI method).

Table 5. The misclassification rate (M) of TA, TB, TC, BD, SK, and DI when $\sum = \sum_1$ and $\theta = 0.9$ with equal block sizes.

n	p	p_i	TA, TB, TC	BD	DI	SK
35	100	5	0.3720	0.0426	0.3802	0.2719
		10	0.4060	0.0294	0.4067	0.2196
		25	0.4343	0.0581	0.4349	0.1436
	200	5	0.3356	0.0067	0.3384	0.2803
		10	0.3714	0.0054	0.3763	0.2472
		25	0.4156	0.0127	0.4160	0.1757
	300	5	0.2931	0.0012	0.3011	0.2548
		10	0.3501	0.0007	0.3545	0.2612
		25	0.3954	0.0030	0.3957	0.1989
	400	5	0.2738	0.0001	0.2760	0.2358
		10	0.3304	0.0001	0.3315	0.2567
		25	0.3828	0.0009	0.3834	0.2208
70	100	5	0.3396	0.0303	0.3395	0.1459
		10	0.3785	0.0182	0.3792	0.0811
		25	0.4071	0.0207	0.4083	0.0344
	200	5	0.2807	0.0037	0.2826	0.1445
		10	0.3272	0.0015	0.3293	0.0829
		25	0.3821	0.0016	0.3833	0.0303
	300	5	0.2438	0.0003	0.2456	0.1383
		10	0.2983	0.0002	0.2994	0.0971
		25	0.3627	0.0001	0.3629	0.0342
	400	5	0.2110	0.0001	0.2133	0.1312
		10	0.2710	0.0000	0.2723	0.1046
		25	0.3415	0.0000	0.3420	0.0440

Table 6. The misclassification rate (M) of TA, TB, TC, BD, SK, and DI when $\sum = \sum_1$ and $\theta = 0.9$ with mixed block sizes.

p	p_i	p_j	TA, TB, TC	BD	DI	SK
100	5	10	0.3886	0.0453	0.3902	0.2582
	5	25	0.4116	0.0455	0.4113	0.2109
	10	25	0.4211	0.0532	0.4215	0.1892
200	5	10	0.3626	0.0053	0.3649	0.2692
	5	25	0.3871	0.0052	0.3887	0.2464
	10	25	0.4048	0.0080	0.4065	0.2192
300	5	10	0.3250	0.0011	0.3300	0.2549
	5	25	0.3663	0.0035	0.3675	0.2373
	10	25	0.3752	0.0031	0.3796	0.2411
400	5	10	0.3102	0.0001	0.3115	0.2514
	5	25	0.3471	0.0009	0.3529	0.2351
	10	25	0.3582	0.0004	0.3616	0.2369
100	5	10	0.3615	0.0314	0.3632	0.1247
	5	25	0.3773	0.0339	0.3761	0.1069
	10	25	0.3928	0.0183	0.3948	0.0503
200	5	10	0.3087	0.0022	0.3100	0.1116
	5	25	0.3455	0.0013	0.3461	0.0665
	10	25	0.3578	0.0020	0.3583	0.0598
300	5	10	0.2739	0.0003	0.2735	0.1123
	5	25	0.3146	0.0002	0.3150	0.0779
	10	25	0.3318	0.0002	0.3318	0.0665
400	5	10	0.2430	0.0000	0.2457	0.1162
	5	25	0.2998	0.0000	0.3008	0.0965
	10	25	0.3092	0.0000	0.3110	0.0811

The results from simulation study when $\sum = \sum_2$ are given in Tables 7-8. As p and n increase, M for the proposed methods decreased. M for TA, TB, and TC methods increased when the number of blocks decreased with any p. For the BD method, M increased when the number of blocks decreased and p and n are small. On comparing the proposed methods with the previously reported ones, the results are almost the same as from $\sum = \sum_1$ with $\theta = 0.9$, for which the BD method performs the best with this form of the population covariance matrix.

Table 7. The misclassification rate (M) of TA, TB, TC, BD, SK, and DI when $\sum = \sum_2$ with equal block sizes.

n	p	p_i	TA, TB, TC	BD	DI	SK
35	100	5	0.3753	0.0529	0.3753	0.2863
		10	0.3914	0.0530	0.3918	0.2677
		25	0.4004	0.0824	0.4008	0.2460
	200	5	0.3252	0.0094	0.3261	0.2690
		10	0.3496	0.0087	0.3507	0.2571
		25	0.3609	0.0219	0.3628	0.2457
	300	5	0.2836	0.0022	0.2873	0.2479
		10	0.3247	0.0014	0.3281	0.2572
		25	0.3451	0.0063	0.3467	0.2505
	400	5	0.2603	0.0003	0.2629	0.2261
		10	0.2952	0.0003	0.2973	0.2384
		25	0.3202	0.0020	0.3239	0.2422
70	100	5	0.3326	0.0406	0.3343	0.1659
		10	0.3543	0.0324	0.3551	0.1403
		25	0.3646	0.0429	0.3664	0.1327
	200	5	0.2723	0.0048	0.2730	0.1528
		10	0.2999	0.0041	0.3001	0.1289
		25	0.3158	0.0043	0.3158	0.1095
	300	5	0.2274	0.0007	0.2301	0.1413
		10	0.2612	0.0005	0.2646	0.1235
		25	0.2821	0.0005	0.2844	0.1013
	400	5	0.2004	0.0002	0.2018	0.1336
		10	0.2266	0.0001	0.2278	0.1115
		25	0.2653	0.0001	0.2676	0.1077

Table 8. The misclassification rate (M) of TA, TB, TC, BD, SK, and DI when $\sum = \sum_2$ with mixed block sizes.

n	p	p_i	p_j	TA, TB, TC	BD	DI	SK
35	100	5	10	0.3860	0.0486	0.3878	0.2743
		5	25	0.3856	0.0739	0.3879	0.2734
		10	25	0.3947	0.0748	0.3965	0.2538
	200	5	10	0.3354	0.0120	0.3364	0.2692
		5	25	0.3490	0.0312	0.3537	0.2629
		10	25	0.3596	0.0349	0.3615	0.2611
	300	5	10	0.3055	0.0021	0.3044	0.2502
		5	25	0.3201	0.0036	0.3231	0.2456
		10	25	0.3319	0.0050	0.3350	0.2521
	400	5	10	0.2859	0.0002	0.2889	0.2418
		5	25	0.2986	0.0049	0.3014	0.2410
		10	25	0.3077	0.0018	0.3121	0.2403
70	100	5	10	0.3471	0.0397	0.3472	0.1592
		5	25	0.3517	0.0477	0.3507	0.1537
		10	25	0.3618	0.0404	0.3620	0.1349
	200	5	10	0.2904	0.0047	0.2914	0.1431
		5	25	0.2988	0.0064	0.2991	0.1338
		10	25	0.3144	0.0045	0.3162	0.1224
	300	5	10	0.2491	0.0006	0.2507	0.1349
		5	25	0.2607	0.0010	0.2614	0.1284
		10	25	0.2779	0.0009	0.2801	0.1177
	400	5	10	0.2156	0.0001	0.2157	0.1265
		5	25	0.2319	0.0003	0.2330	0.1227
		10	25	0.2488	0.0002	0.2499	0.1141

Note that M was similar for TA, TB and TC methods with the same combination of dimensionality p and number of observations n, i.e. the choice of shrinkage target matrix in the first proposed method did not affect performance in this simulation study.

From this simulation study, we observe that the TA, TB, and TC methods perform well, similar to DI and SK methods when θ is small. The BD method performs the best when θ is greater than 0.5.

3.3 A real data example

In this section, we use a real dataset to assess the performances of the four methods: 1) SK; 2) DI; 3) TA, TB, and TC; and 4) BD. The Notterman Carcinoma dataset used for this study was taken from a gene expression project at Princeton University, New Jersey, by Notterman, Alon, Sierk, and Levine (2001). These data consist of p = 7,457 genes' expression in 18 paired colon tissue samples (18 tumor tissues n_1 , and 18 normal tissues n_1).

From the dataset, 100 genes for 10 tumor and normal tissues were selected for the training set, and 5 tumor and normal tissues for the testing set. These data were assumed to be multivariate normal. Recall that in the simulation study the BD method performed well when the correlation coefficient between variables in the same block was higher than 0.5. Thus, the variables of this dataset are arranged in block order so that the correlation coefficient between any two adjacent variables in the same block is greater than or equal to 0.5. The block sizes are mixed with a maximum of 17 and a minimum of 1, and the number of blocks is 14. There are 6 blocks of dimension one.

Before performing discriminant analysis on this dataset, the assumptions that the block diagonal covariance matrix structure and the equal covariances of both classes need to be checked. The test statistic proposed by Hyodo, Shutoh, Nishiyama, and Pavlenko (2015) was used to test the assumption of block diagonal covariance matrix for each class (tumor and normal tissues). The test statistics were: for tumor tissues 0.8479 (p-value $\approx$ 0.1983) and 0.5490 (p-value $\approx$0.2915) for normal tissues.

It can be concluded that the covariance matrices are block diagonal. The equality of covariances for the classes can be checked by the test statistic proposed by Chaipitak and Chongcharoen (2013). The test statistic was -0.9383 (p-value $\approx$ 0.3481), which indicates that the covariances of the classes are equal.

After checking for block diagonal covariance matrices and equality of covariances by class, the TA, TB, TC, BD, DI, and SK methods were applied to this dataset. The results show that the TA, TB, TC, DI, and SK methods gave zero M, i.e. 100% correct classification calls, while the BD method achieves M = 0.3, indicating that the TA, TB, TC, DI, and SK methods perform better than the BD method with this dataset.

The above experimental results show that the BD method successfully produces the lowest misclassification rate when the population covariance matrix is block diagonal and the values of off-diagonal elements in the blocks are large, preferably larger than 0.5. When the off-diagonal elements in the blocks are small, then TA, TB, and TC methods are more suitable than BD. In the real-life dataset tested, the covariance matrix had many blocks of dimension one, and the misclassification rates for TA, TB, and TC methods were lower (better) than for the BD method.

4. Conclusions

In this paper, we proposed two new discriminant methods for the classification of high dimensional data. In the first method **(**TA, TB, and TC**)**, we reduce the dimensionality of the observations and use a well-conditioned covariance matrix approximation guaranteeing minimum mean squared error**.** In the second method **(**BD**)**, we use a block diagonal part of the sample covariance matrix in place of the sample covariance matrix**.** We compared our methods with the DI and SK methods proposed by Dudoit *et al.* (2002) and Srivastava and Kubokawa (2007) in a simulation study and with a real data set, based on misclassification rates. The simulation study showed that the TA, TB, and TC methods perform well when the correlations among variables in a block are weak, but is inappropriate for classification when these correlations are strong. The BD method was superior when the correlation among variables in a block were strong. Finally, the two proposed discriminant methods (TA, TB, TC, and BD) were applied to a real dataset, the Notterman Carcinoma dataset.

5. Future Work

In this study, two new discriminant methods were proposed for a binary classification problem with equal covariances, so in future work we could examine classification with 3 or more classes and/or an unequal covariances. Regarding the real-life dataset, when the number of variables is large, then it is quite hard to find the block size and the number of blocks in the population covariance matrix. Cluster analysis could be considered to arrange the variables into blocks, in order to find the block size and the number of blocks in the population covariance matrix. Finally, the proposed methods can be tested on further real datasets.

Acknowledgements

The author would like to express sincere gratitude to the advisor Prof. Dr. Samruam Chongcharoen for continuous support of research, for patience, motivation, and immense knowledge.

References

Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(7), 711–720.

Chaipitak, S., & Chongcharoen, S. (2013). A test for testing the equality of two covariance matrices for high-dimensional data. *Journal of Applied Sciences*, *13*(2), 270–277.

Chen, L.-F., Liao, H.-Y. M., Ko, M.-T., Lin, J.-C., & Yu, G.-J. (2000). A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, *33*(10), 1713–1726.

Dempster, A. P. (1958). A high dimensional two sample significance test. *The Annals of Mathematical Statistics*, 995–1010.

Di Pillo, P. J. (1976). The application of bias to discriminant analysis. *Communications in Statistics-Theory and Methods*, *5*(9), 843–854.

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, *97*(457), 77–87.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, *84*(405), 165–175.

Guo, Y., Hastie, T., & Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, *8*(1), 86–100.

Hyodo, M., Shutoh, N., Nishiyama, T., & Pavlenko, T. (2015). Testing block-diagonal covariance structure for high-dimensional data. *Statistica Neerlandica*, *69*(4), 460–482.

Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, *10*(5), 603–621.

Lu, J., Plataniotis, K. N., & Venetsanopoulos, A. N. (2003). Regularized discriminant analysis for the small sample size problem in face recognition. *Pattern Recognition Letters*, *24*(16), 3079–3087.

Notterman, D. A., Alon, U., Sierk, A. J., & Levine, A. J. (2001). Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Research*, *61*(7), 3124–3130.

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1), 32.

Srivastava, M. S., & Kubokawa, T. (2007). Comparison of Discrimination Methods for High Dimensional Data. *Journal of the Japan Statistical Society*, *37*(1), 123–134.

Tian, Q., Barbero, M., Gu, Z.-H., & Lee, S. H. (1986). Image classification by the Foley-Sammon transform. *Optical Engineering*, *25*(7), 257834.