


Article

Classification in High Dimension Using the Ledoit–Wolf Shrinkage Method

Rasoul Lotfi ¹, Davood Shahsavani ¹ and Mohammad Arashi ^{2,3,*} 

¹ Department of Statistics, Faculty of Mathematical Sciences, Shahrood University of Technology, Shahrood 3619995161, Iran

² Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad 9177948974, Iran

³ Department of Statistics, Faculty of Natural and Agricultural Sciences, University of Pretoria, Pretoria 0002, South Africa

* Correspondence: arashi@um.ac.ir; Tel.: +98-915-102-3551

Abstract: Classification using linear discriminant analysis (LDA) is challenging when the number of variables is large relative to the number of observations. Algorithms such as LDA require the computation of the feature vector's precision matrices. In a high-dimension setting, due to the singularity of the covariance matrix, it is not possible to estimate the maximum likelihood estimator of the precision matrix. In this paper, we employ the Stein-type shrinkage estimation of Ledoit and Wolf for high-dimensional data classification. The proposed approach's efficiency is numerically compared to existing methods, including LDA, cross-validation, gLasso, and SVM. We use the misclassification error criterion for comparison.

Keywords: classification; linear discriminant analysis; high-dimensional data; Ledoit and Wolf shrinkage method; Stein-type shrinkage; misclassification error

MSC: 62H30; 68T09



Citation: Lotfi, R.; Shahsavani, D.; Arashi, M. Classification in High Dimension Using the Ledoit–Wolf Shrinkage Method. *Mathematics* **2022**, *10*, 4069. <https://doi.org/10.3390/math10214069>

Academic Editor: Christophe Chesneau

Received: 23 September 2022

Accepted: 25 October 2022

Published: 1 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

As one of the most widely used classification techniques, linear discriminant analysis (LDA) is still interesting because of its simplicity, stability, and prediction accuracy. Consider X as a p -dimensional predictor vector and the response $Y \in \{1, 2, \dots, K\}$ as the class labels. In LDA, it is assumed that $X|Y \sim N_p(\mu_Y, \Sigma)$, where $\mu_Y \in \mathbb{R}^p$ and $\Sigma > 0$; consequently, the Bayes decision rule involves Σ^{-1} . In the large sample setting, when $p < n$, the sample covariance matrix S is an unbiased estimator of Σ . However, in a high-dimensional setting, $p \approx n$ or $p > n$, S will be singular, and the likelihood estimator has many weaknesses such as inaccuracy (see [1,2]).

Many studies have been conducted using factor methods, sparse, graphical, and shrinkage methods for estimating Σ . Srivastava [3] examined multivariate theory in a high-dimensional state and used the Moore–Penrose inverse of the covariance matrix to solve the singularity problem of S . However, when some covariance matrix values are zero or close to zero, this idea does not work well.

The idea of estimating the precision matrix using a sparse method was first proposed by Dempster [4], and later, Meinshausen and Bühlmann [5] proposed the use of least absolute shrinkage and selection operator (Lasso) regression to identify the zeros of the inverse covariance matrix. Banerjee et al. [6] performed a penalized maximum likelihood estimation with the lasso penalty for sparse estimation of the inverse of the covariance matrix. Friedman et al. [7], proposed the graphical Lasso method, under the sparsity assumption of $\Theta = \Sigma^{-1}$ by using coordinate descent for the lasso penalty, through the objective function

$$\log \det \Theta - \text{tr} S \Theta - \lambda \|\Theta\|_1, \quad (1)$$

where $\Theta > 0$, $tr(A)$ denotes the trace of matrix A , and $\|\cdot\|_1$ is the norm one operator. Bickel and Levina [8] used the hard thresholding estimator for the sparse estimation of the covariance matrix Σ . Furthermore, Cai and Zhang [9] developed an optimality theory for LDA in the high-dimensional setting by considering a different approach to solve the problem of LDA. Instead of estimating $\delta = (\mu_2 - \mu_1)$, and $\Theta = \Sigma^{-1}$ separately, they proposed a data-driven and tuning free classification rule called AdaLDA by directly estimating the discriminant direction $\beta = \Theta\delta$ through solving an optimization problem. As the hard threshold estimator in regression provides inflexible estimators, Ratman et al. [10] refined a generalized threshold law by using the combination of the threshold method and the shrinkage method. Bin and Tibshirani [11] generalized the estimate of a sparse covariance matrix by simultaneously estimating the nonzero covariance and the graph structure (location of zeros). Refer to Fan et al. [12] for more related studies.

Apart from sparse covariance matrix estimation, a common approach to improve the estimation of Σ is the use of the class of shrinkage estimators, which was initially proposed by James and Stein [13] to define bias estimation in order to reduce the variance of S (see [14–16] for extensive reviews). Di Pillo [17] and Campbell [18] improved the estimate of Σ^{-1} using the ridge idea. Peck and Van Niss [2] proposed another type of shrinkage estimator of Σ^{-1} by reducing the Fisher’s classification error. Mkhadri [19] used the cross-validation (CV) method to estimate the shrinkage parameter for the estimation of Σ^{-1} in classification rule. However, Choi et al. [20] demonstrated that the use of the CV method may not lead to a positive definite estimate for the high-dimensional case $n \ll p$.

In the shrinkage method and graphical and factor models, additional information is needed in the estimation process (e.g., Beckel and Levina [21]; Khare and Rajaratnam [22]; Cai and Zhou [23]), whereas this surplus knowledge is not always available (Maurya [24]). Therefore, Ledoit and Wolf [25] proposed the rule of the optimal linear shrinkage estimator with optimal asymptotic properties by using the analysis of covariance matrix eigenvalues. For estimating the inverse covariance matrix (precision matrix) when $p \geq n$, other studies have been conducted such as Wang et al. [26], Hong and Kim [27], and Lee et al. [28].

This paper aims to classify high-dimensional observations using the LDA, where the inverse sample covariance matrix is singular and not invertible. We apply Ledoit and Wolf’s shrinkage method to estimate Θ and efficiently classify new observations in a high-dimensional regime. Thus, the plan for the rest of this paper is as follows. In Section 2, the proposed methodology, along with some theory, is given. Section 3 includes extensive numerical assessments for performance analysis and compares the proposed discriminant rule with other existing methods. We conclude with the significant results in Section 4; Appendix A is allocated for the proofs.

2. Materials and Methods

In discriminant analysis, a set of observations are classified into predetermined categories using a function called the decision function or discriminant function. In other words, discriminant analysis seeks to identify linear or nonlinear combinations of independent variables that are best able to separate groups of observations using the discriminant rule.

Consider distinct populations Π_1, \dots, Π_K with density function $f_j(x); j = 1, 2, \dots, K$ and prior probabilities $\pi_j = Pr(Y \in \Pi_j)$. An observation x is classified into Π_i if

$$x \in \Pi_i \iff i = \arg \max_j \pi_j f_j(x). \tag{2}$$

In the simplest case, $K = 2$, it is assumed that $\Pi_j \sim N_p(\mu_j, \Sigma_j); j = 1, 2$, so that Π_1 is independent of Π_2 , $\mu_j \in \mathbb{R}^p$, and $\Sigma_j > 0$. In the LDA, it is also assumed $\Sigma_1 = \Sigma_2 = \Sigma$. According to Equation (2), $x \in \Pi_1$ if $f_1(x) > f_2(x)$; so, the discriminant function is obtained as follows

$$D_{12}(x) = (\mu_1 - \mu_2)^T \Sigma^{-1} \left(x - \frac{\mu_1 + \mu_2}{2} \right), \tag{3}$$

where μ_j and Σ are unknown, and we estimate them using the training sample by the mean vector \bar{x}_j and the pooled sample covariance matrix S , respectively, where $\bar{x}_j = \frac{1}{n_j} \sum_{i \in \Pi_j} x_i$, and

$$NS = \sum_{i \in \Pi_1} (x_i - \bar{x}_1)(x_i - \bar{x}_1)^T + \sum_{i \in \Pi_2} (x_i - \bar{x}_2)(x_i - \bar{x}_2)^T; N = n_1 + n_2 - 2.$$

Based on Equation (3), the classification function can be considered as a linear function W

$$W(x) = (\bar{x}_1 - \bar{x}_2)^T S^{-1} \left(x - \frac{\bar{x}_1 + \bar{x}_2}{2} \right). \tag{4}$$

Hence, an observation x is classified into Π_2 if $W(x) > 0$, and it is classified into the population Π_1 otherwise. Therefore, the probability of misclassification (PMC) depends on the sample values \bar{x}_1, \bar{x}_2 , and S . If the S estimate is weak, the PMC will not be minimized and in high-dimension ($p \geq n$); S^{-1} either cannot be calculated or it is not efficient. In this case, we use the approach of shrinkage methods.

2.1. Ledoit and Wolf Shrinkage Estimators

The James and Stein shrinkage estimator [13] is a convex combination of a sample covariance matrix and a target matrix T as follows

$$S^* = (1 - \lambda)S + \lambda T, \tag{5}$$

where $\lambda \in (0, 1)$ is the shrinkage parameter and T is positive definite. The target matrix T should be chosen to have several properties. The target matrix must be structured, positive definite, and well-conditioned, representing our application’s true covariance matrix. The T matrix may be biased; however, with its well-defined structure, it has a low variance. Given that the matrix T is predetermined, determining the shrinkage parameter λ is important and should be chosen in such a way that the variance of the shrinkage estimator is less than the variance of S . If $n > p$, the variance S^* must be less than the variance of the target matrix, i.e., $\lambda \rightarrow 0$, and if $p > n$, the target matrix must have less variance than the variance S^* , i.e., $\lambda \rightarrow 1$. Therefore, the λ values have a significant effect on the degree of the misclassification error.

In the category of Stein-type shrinkage estimators; Ledoit and Wolf [25] proposed the estimation of the shrinkage parameter λ using the following result, when $p \geq n$.

Theorem 1. Suppose x_1, x_2, \dots, x_n is a random sample from $N_p(\mu_j, \Sigma)$, $\mu_j \in \mathbb{R}^p$, $\Sigma > 0$, $j = 1, 2, \dots, K$ and $\delta^2 = E[\|S - I\|^2]$; $\alpha^2 = E[\|\Sigma - I\|^2]$ and $\beta^2 = E[\|S - \Sigma\|^2]$; then,

1. $\delta^2 = \alpha^2 + \beta^2$;
2. assuming $\Sigma^* = (1 - \lambda)S + \lambda I$, the optimal shrinkage parameter that minimizes the risk value of Σ^* is equal to $\hat{\lambda}^{LW} = \frac{\hat{\beta}^2}{\delta^2}$, where

$$\hat{\beta}^2 = \frac{1}{n} \hat{a}_2 + \frac{p}{n} \hat{a}_1^2 \quad ; \quad \delta^2 = \frac{n+1}{n} \hat{a}_2 + \frac{p}{n} \hat{a}_1^2 - 2\hat{a}_1 + 1$$

in which, based on Srivastava [29],

$$\hat{a}_1 = \frac{1}{p} tr(S) \quad ; \quad \hat{a}_2 = \frac{n^2}{p(n-1)(n+2)} (tr(S^2) - \frac{1}{n} (tr(S))^2).$$

See Ledoit and Wolf [25] for details.

2.2. Improved Linear Discriminant Rules

Since the shrinkage parameter, λ in the Ledoit and Wolf’s approach is obtained using the optimization method, in contrast to the CV method used by Mkhaderi [19], it will always lead to a positive definiteness of the sample covariance matrix. It also does not require additional information about explanatory variables and their independence. As a result, it has an advantage over other shrinkage estimation methods (Ledoit and Wolf [30]); therefore, the proposed method for reducing the misclassification error and the discriminant analysis in the high-dimensional case $p \geq n$ leads to the following classification rule

$$\tilde{W}(x, \lambda^{LW}) = (\bar{x}_1 - \bar{x}_2)^T \tilde{\mathbf{S}}^{-1}(\lambda^{LW}) \left(x - \frac{\bar{x}_1 + \bar{x}_2}{2} \right), \tag{6}$$

where $\tilde{\mathbf{S}}(\lambda^{LW}) = (1 - \lambda^{LW})\mathbf{S} + \lambda^{LW}\mathbf{T}$ can be obtained from the Equation (5), and

$$\lambda^{LW} = \min(\hat{\lambda}^{LW}, 1). \tag{7}$$

2.3. Properties of the Improved Discriminant Rule

Given that in the discriminant problem, $W(x)$ has a normal distribution with the mean

$$E(W(x|II_1)) = (\bar{x}_1 - \bar{x}_2)^T \mathbf{S}^{-1} \mu_1 - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)^T \mathbf{S}^{-1}(\bar{x}_1 + \bar{x}_2) \tag{8}$$

and variance of

$$\text{var}(W(x|II_1)) = (\bar{x}_1 - \bar{x}_2)^T \mathbf{S}^{-1} \boldsymbol{\Sigma} \mathbf{S}^{-1} (\bar{x}_1 - \bar{x}_2), \tag{9}$$

we have the following results.

Lemma 1. Under the assumptions of Section 2.2, $\tilde{W}(x, \lambda^{LW})$ has a p -dimensional normal distribution with mean

$$E(\tilde{W}(x, \lambda^{LW})|II_1) = (\bar{x}_1 - \bar{x}_2)^T \tilde{\mathbf{S}}^{-1}(\lambda^{LW}) \mu_1 - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)^T \tilde{\mathbf{S}}^{-1}(\lambda^{LW})(\bar{x}_1 + \bar{x}_2)$$

and variance

$$\text{var}(\tilde{W}(x, \lambda^{LW})|II_1) = (\bar{x}_1 - \bar{x}_2)^T \tilde{\mathbf{S}}^{-1}(\lambda^{LW}) \boldsymbol{\Sigma} \tilde{\mathbf{S}}^{-1}(\lambda^{LW})(\bar{x}_1 - \bar{x}_2)$$

Proof. Refer to Appendix A. □

Theorem 2. Under the assumption of Section 2.2, using Lemma 1, we have

$$E(\tilde{W}(x, \lambda^{LW})|II_1) = E(W(x)|II_1) + B$$

$$\text{var}(\tilde{W}(x, \lambda^{LW})|II_1) \leq \text{var}(W(x)|II_1),$$

where $B = \frac{\lambda^{LW}(\lambda^{LW} - 2)}{n} \Delta^2$ and $\Delta = [(\mu_1 - \mu_2)^T \boldsymbol{\Sigma}^{-1}(\mu_1 - \mu_2)]^{\frac{1}{2}}$.

Proof. Refer to Appendix A. □

3. Numerical Studies

To assess the performance of the estimator (5) in classification, we conducted a simulation study and analyzed some real data.

3.1. Simulation Study

Data were generated from two populations $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ and $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $p = 12$ in which $\mathbf{0}$ is a p -dimensional zero vector, $\boldsymbol{\mu}$ is a p -dimensional desired vector, and $\boldsymbol{\Sigma}$ is a p -dimensional square matrix. When the covariance matrix is almost singular, discriminant analysis is likely to be sensitive to different choices of the mean vector; so in this

paper, two different forms for the mean vector μ are selected, namely, $(m^*, 0, \dots, 0)^T$ and $(m, m, \dots, m)^T$ (named Mod1 and Mod2, respectively). The values of m^* and m were chosen so that the Mahalanobis distance, $D = (\mu^T \Sigma^{-1} \mu)^{\frac{1}{2}}$, was the same for each case, and the μ were chosen using different values of D (0.5, 1.5 and 2.5). The covariance matrix was also considered as $\Sigma = (1 - \rho)\mathbf{I} + \rho\mathbf{J}$, where $\frac{-1}{p-1} \leq \rho \leq 1$, \mathbf{I} is the identity matrix, and \mathbf{J} is the unit matrix of dimension $p \times p$. In order to determine the correlation role of the explanatory variables in the estimator, two values 0.2 and 0.4 for ρ were considered. For each population and different combinations of μ and Σ , we generated 10 p -dimensional training and 50 p -dimensional test data vectors. We chose the target matrix as $\mathbf{T} = \mathbf{I}$, so the sample covariance matrix shrank to the identity matrix. This target imposed no variance to the shrinkage estimator. This simulation was repeated 1000 times, and the performance of proposed methodology LW was compared with linear discriminant analysis (LDA), the Mkhaderi method [19] (CV), the graphic lasso method (gLasso), and support vector machine (SVM).

The results for each case μ , i.e., Mod1 and Mod2, are summarized in Tables 1 and 2. The column ‘Test’ shows the average value of the misclassification test errors for each parameter value of ρ . Further, in these tables, the mean value of the shrinkage parameter λ for each discriminant rule is shown. The quantities in parentheses are the standard deviations of the respective means. Bold values are the smallest among all, showing the best method.

Table 1. Misclassification error and shrinkage parameter values for Mod1.

		$\rho = 0.2$		$\rho = 0.4$	
		Test	$\bar{\lambda}$	Test	$\bar{\lambda}$
$D = 0.5$	LDA	0.487(0.037)	---	0.475(0.072)	---
	CV	0.462(0.042)	0.949(0.043)	0.472(0.052)	0.957(0.031)
	LW	0.443(0.039)	0.796(0.146)	0.462(0.070)	0.409(0.161)
	gLasso	0.478(0.040)	0.007(0.001)	0.479(0.072)	0.009(0.002)
	SVM	0.504(0.035)	---	0.506(0.034)	---
$D = 1.5$	LDA	0.406(0.064)	---	0.369(0.055)	---
	CV	0.374(0.072)	0.974(0.023)	0.340(0.043)	0.952(0.029)
	LW	0.351(0.073)	0.846(0.160)	0.318(0.034)	0.439(0.121)
	gLasso	0.390(0.066)	0.007(0.002)	0.345(0.046)	0.009(0.002)
	SVM	0.461(0.052)	---	0.441(0.057)	---
$D = 2.5$	LDA	0.268(0.058)	---	0.251(0.060)	---
	CV	0.195(0.042)	0.975(0.033)	0.235(0.068)	0.969(0.032)
	LW	0.183(0.050)	0.833(0.177)	0.189(0.045)	0.591(0.153)
	gLasso	0.233(0.041)	0.007(0.002)	0.233(0.057)	0.008(0.002)
	SVM	0.315(0.064)	---	0.332(0.087)	---

The improvement of the shrinkage algorithm strongly depends on the Mahalanobis distance between two populations. When the Euclidean distance between the means is small, the mean estimation error caused by the poor estimate of Σ is very damaging to the classification. Therefore, as the Euclidean distance increases, the means move further apart, and it does not have much relative effect on the classification.

According to Table 1, by increasing D , the misclassification error decreased. Apparently, for each D , the shrinkage method LW had a lower classification error compared to the LDA, CV, gLasso, and SVM methods. This means that the Ledoit and Wolf method had better performance in determining and assigning new observations to populations.

On the other hand, according to Table 2, by changing the strategy and considering Mod2, the results obtained in Mod1 were still valid. Thus, changing all the values of the mean vector μ was established (better efficiency and performance of the proposed method of this research than the studied methods). Figure 1 simply shows the results stated in Tables 1 and 2. Bold values are the smallest among all, showing the best method.

Table 2. Misclassification error and shrinkage parameter values for Mod2.

		$\rho = 0.2$		$\rho = 0.4$	
		Test	$\bar{\lambda}$	Test	$\bar{\lambda}$
$D = 0.5$	LDA	0.491(0.057)	---	0.502(0.050)	---
	CV	0.420(0.091)	0.938(0.039)	0.456(0.065)	0.915(0.040)
	LW	0.399(0.099)	0.791(0.204)	0.445(0.053)	0.389(0.120)
	gLasso	0.479(0.042)	0.007(0.001)	0.479(0.055)	0.010(0.002)
	SVM	0.438(0.0679)	---	0.467(0.086)	---
$D = 1.5$	LDA	0.214(0.077)	0.00	0.324(0.066)	0.00
	CV	0.104(0.030)	0.965(0.035)	0.218(0.059)	0.950(0.036)
	LW	0.100(0.035)	0.716(0.226)	0.214(0.036)	0.679(0.273)
	gLasso	0.186(0.044)	0.007(0.001)	0.303(0.082)	0.008(0.002)
	SVM	0.140(0.078)	---	0.215(0.030)	---
$D = 2.5$	LDA	0.097(0.076)	---	0.183(0.054)	---
	CV	0.024(0.013)	0.989(0.014)	0.085(0.027)	0.960(0.030)
	LW	0.023(0.012)	0.501(0.142)	0.081(0.031)	0.580(0.180)
	gLasso	0.063(0.025)	0.007(0.001)	0.147(0.041)	0.008(0.001)
	SVM	0.031(0.020)	---	0.083(0.029)	---

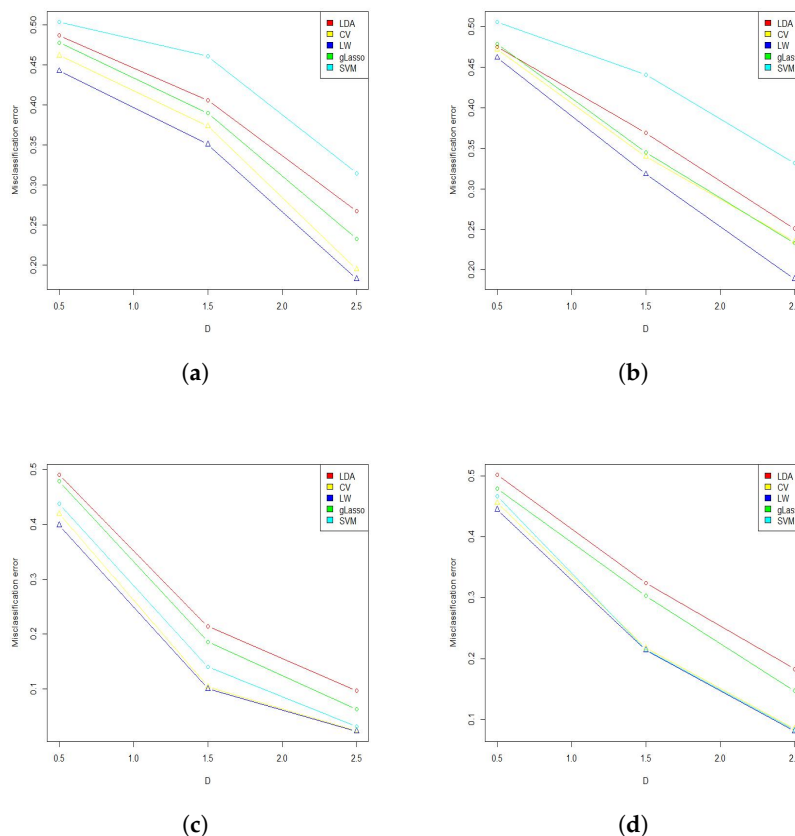
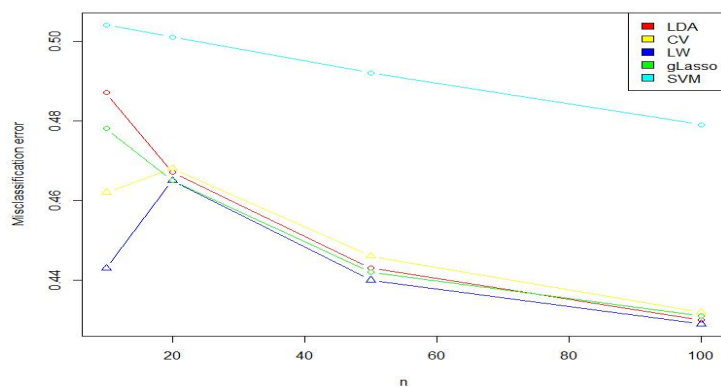
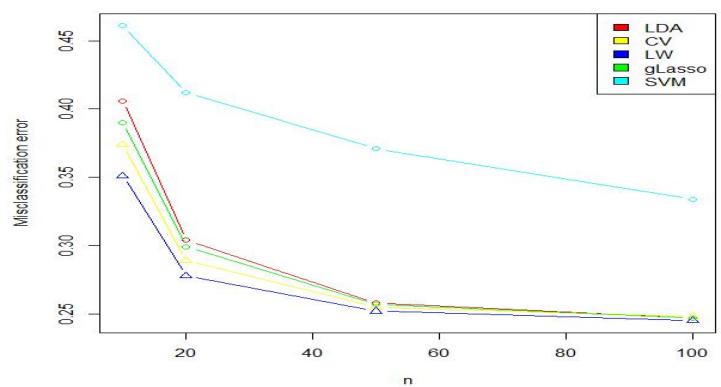


Figure 1. Misclassification error by increasing the Mahalanobis distance. (a) Mod1, $\rho = 0.2, n = 10$; (b) Mod1, $\rho = 0.4, n = 10$; (c) Mod2, $\rho = 0.2, n = 10$; (d) Mod2, $\rho = 0.4, n = 10$.

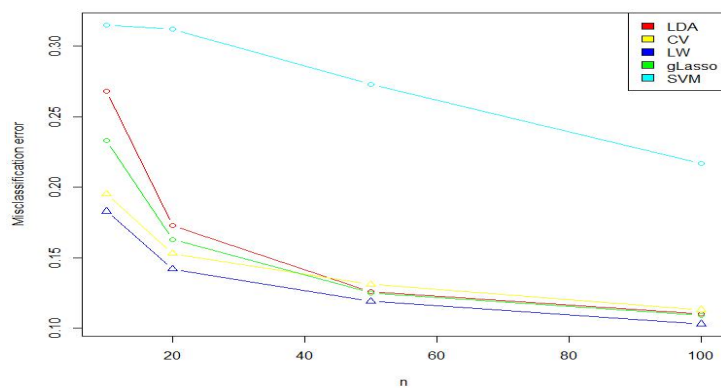
Figure 2 depicts the misclassification error for varying sample sizes. Not surprisingly, as n increased, the misclassification error decreased for all the methods in this study, but surprisingly, none of the methods had a smaller error compared to the LW. As shown in Figure 2, with the increasing sample size, the misclassification error in all methods was higher than the proposed method. Moreover, as the Mahalanobis distance increased, it became easier to identify which population the new observation x^* belonged to, which improved the classification in the proposed method.



(a)



(b)



(c)

Figure 2. Misclassification error by increasing sample size and $p = 12$. (a) Mod1, $\rho = 0.2, D = 0.5$; (b) Mod1, $\rho = 0.2, D = 1.5$; (c) Mod1, $\rho = 0.2, D = 2.5$.

To evaluate the efficiency of the LW method, simulations were performed using $p = 16, 30, 50, 100, 500$ values. These results are summarized in Table 3. As the dimension increased, the classical method of linear discriminant analysis as well as the proposed cross-validation method of Mkhaderi [19] could not be used due to the singularity of the covariance matrix. With an increase in the value of D , the misclassification error

decreased, and for each value D , the shrinkage method of LW was significantly better than the other methods.

Table 3. Misclassification error and shrinkage parameter values for $\rho = 0.4$ in Mod1 state.

		LDA	CV	LW	gLasso	SVM
$D = 0.5$	$p = 12$	0.475(0.072)	0.472(0.052)	0.462(0.070)	0.479(0.072)	0.506(0.034)
	$p = 16$	0.494(0.060)	0.494(0.042)	0.450(0.050)	0.477(0.054)	502(0.050)
	$p = 30$	--- ^a	---	0.478(0.066)	0.495(0.076)	0.511(0.030)
	$p = 50$	---	---	0.470(0.063)	0.493(0.061)	0.504(0.037)
	$p = 100$	---	---	0.496(0.045)	0.504(0.039)	0.515(0.034)
	$p = 500$	---	---	0.502(0.044)	0.519(0.056)	0.505(0.034)
$D = 1.5$	$p = 12$	0.369(0.055)	0.340(0.043)	0.318(0.034)	0.345(0.046)	0.441(0.057)
	$p = 16$	0.413(0.077)	0.537(0.098)	0.334(0.049)	0.373(0.074)	0.458(0.045)
	$p = 30$	---	---	0.372(0.041)	0.410(0.070)	0.501(0.053)
	$p = 50$	---	---	0.409(0.073)	0.417(0.061)	0.490(0.037)
	$p = 100$	---	---	0.415(0.055)	0.422(0.040)	0.502(0.027)
	$p = 500$	---	---	0.451(0.059)	0.462(0.038)	0.518(0.027)
$D = 2.5$	$p = 12$	0.251(0.060)	0.235(0.068)	0.189(0.045)	0.233(0.057)	0.332(0.087)
	$p = 16$	0.338(0.072)	0.592(0.115)	0.213(0.054)	0.260(0.057)	0.391(0.078)
	$p = 30$	---	---	0.235(0.062)	0.276(0.070)	0.410(0.079)
	$p = 50$	---	---	0.266(0.055)	0.286(0.061)	0.466(0.051)
	$p = 100$	---	---	0.326(0.067)	0.330(0.053)	0.482(0.044)
	$p = 500$	---	---	0.428(0.069)	0.430(0.076)	0.488(0.061)

^a The covariance matrix is singular.

3.2. Real Data Analyses

In this section, we assess the performance of the five methods in classification for the datasets in Table 4.

Table 4. Datasets (Accessed on 30 November 2022).

	DataName	Specification	Link
Data1	BreastCancer	$n = 116$ $p = 10$	www.UCIMachineLearning.com
Data2	Insurance	$n = 36,634$ $p = 17$	www.Kaggle.com
Data3	LSVT	$n = 126$ $p = 309$	www.UCIMachineLearning.com
Data4	mRNA	$n = 219$ $p = 1650$	www.UCIMachineLearning.com

As shown in Table 5, the LW shrinkage method for classification was superior comparatively (marked as bold). In Data 4, the execution time in the system with specifications CPU: i7 – 4720HQ and Ram: 8 GB for the gLasso method took more than 10 h, and for the LW method it was less than 5 min, which was a sign of the rapidity of the shrinkage method in the classification of high-dimensional data.

Table 5. Misclassification error for the real datasets in Table 4.

	LDA	CV	SVM	gLasso	LW
<i>Data1</i>	0.31707	0.26829	0.34146	0.29268	0.26829
<i>Data2</i>	0.00063	0.00063	0.00009	0.00018	0.00009
<i>Data3</i>	NaN ^a	NaN	0.21053	0.18421	0.10526
<i>Data4</i>	NaN	NaN	0.00000	0.00000	0.00000

^a The covariance matrix is singular.

Figure 3 depicts the average of misclassification errors for the five methods, discussed in the paper.

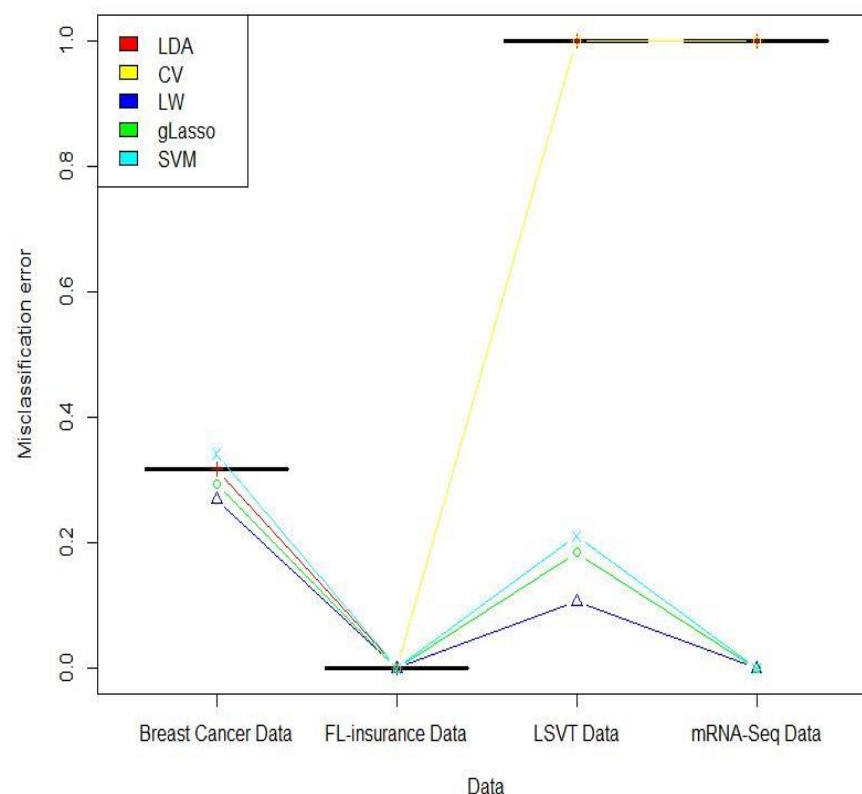


Figure 3. Misclassification error for the four real datasets.

4. Discussion and Conclusions

In the present paper, we aimed to improve the efficiency of the LDA classification method in high-dimensional situations, where the singularity of the inverse covariance matrix produced problems. We proposed to estimate the precision matrix with the shrinkage approach of Ledoit and Wolf (LW) [25], where the sample covariance matrix and a target matrix were linearly combined through a penalty factor in LDA classification.

The implementation of the suggested method on simulation data showed that for different scenarios, the proposed approach was superior to two powerful competitors, gLasso and SVM, in the sense of misclassification error. By increasing the Mahalanobis distance and using $p = 100$ and $p = 500$, we became more confident that the LW method could be considered as an alternative to some well-known methods for better classification.

In the real data analyses, three results were considerable. First, the analysis of Data1 and Data2 showed that LW was as good as the other competitors or much better when $n \gg p$. Second, the misclassification of Data3 ($n = 126, p = 309$) exhibited that our method was more reliable than the others in high-dimensional regimes ($n < p$). Third, the LW method had the same misclassification error for Data4 when $n \ll p$, compared to

others. However, the execution time of the LW method was much shorter. In addition to high accuracy, the LW method was strongly recommended from the computation burden point of view, especially in a high-dimensional setting, where computing precision matrix estimation is challenging.

Lastly, if the covariance matrices of the populations are not the same ($\Sigma_i \neq \Sigma_j; i, j = 1, 2, \dots, K$), instead of LDA, it is possible to use the quadratic discriminant analysis (QDA). Friedman [31] proposed the method of regularized discriminant analysis (RDA) by using the trace estimator, $p^{-1}tr(\Sigma_i)\mathbf{I}_p$, and applying twice the shrinking sample covariance matrix. Since the trace estimator pools the diagonal elements of sample covariance matrices and ignores off-diagonal elements, Wu et al. [32] introduced the ppQDA estimator, which pools all elements in the covariance matrix, and it does not need to impose sparse assumptions. Although the ppQDA method has good asymptotic properties, it may not have a good performance for data classification. Therefore, considering (5) and Friedman’s [31] method as follows

$$\mathbf{S}_i(\lambda, \gamma) = (1 - \gamma)\mathbf{S}_i(\lambda) + \gamma\left(p^{-1}tr(\mathbf{S}_i(\lambda))\mathbf{I}_p\right),$$

where λ, γ are tuning parameters, and $\mathbf{S}_i(\lambda) = \frac{(1-\lambda)(n_i-1)\mathbf{S}_i + \lambda(n_1+n_2-2)\mathbf{S}}{(1-\lambda)(n_i-1) + \lambda(n_1+n_2-2)}$, it seems that by using the shrinkage method presented in this article, estimation of Σ_i can be improved and it is possible to reduce the misclassification error of high-dimensional data in the quadratic discriminant analysis mode.

Author Contributions: Conceptualization, R.L., D.S. and M.A.; Funding acquisition, M.A.; Methodology, R.L., D.S. and M.A.; Software, R.L.; Supervision, D.S. and M.A.; Visualization, R.L., D.S. and M.A.; Formal analysis, R.L., D.S. and M.A.; Writing—original draft preparation, R.L.; Writing—review and editing, R.L., D.S. and M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was based upon research supported in part by the National Research Foundation (NRF) of South Africa, SARCHI Research Chair UID: 71199, the South African DST-NRF-MRC SARCHI Research Chair in Biostatistics (Grant No. 114613), and STATOMET at the Department of Statistics at the University of Pretoria, South Africa. The third author’s research (M. Arashi) is supported by a grant from Ferdowsi University of Mashhad (N.2/58266). The opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are publicly available.

Acknowledgments: The authors would like to sincerely thank the anonymous reviewers for their constructive comments, which helped to improve the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Here, we give the sketch of the proofs of Lemma 1 and Theorem 2.

Proof of Lemma 1. According to (3), replacing \mathbf{S}^{-1} with $\tilde{\mathbf{S}}^{-1}(\lambda^{LW})$, the decision function $\tilde{W}(x, \lambda^{LW})$ is given by

$$\begin{aligned} \tilde{W}(x, \lambda^{LW}) &= (\bar{x}_1 - \bar{x}_2)^T \tilde{\mathbf{S}}^{-1}(\lambda^{LW}) \left(x - \frac{\bar{x}_1 + \bar{x}_2}{2}\right) \\ &= (\bar{x}_1 - \bar{x}_2)^T \tilde{\mathbf{S}}^{-1}(\lambda^{LW}) x - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)^T \tilde{\mathbf{S}}^{-1}(\lambda^{LW})(\bar{x}_1 + \bar{x}_2). \end{aligned}$$

On the other hand, since $x \sim N_p(\mu, \Sigma)$ is independent of \bar{x}_1, \bar{x}_2 , and \mathbf{S} , the conditional distribution $\tilde{W}(x, \lambda^{LW})$ given I_{I_1} has the following mean and variance

$$E(\tilde{W}(x, \lambda^{LW})|I_{I_1}) = (\bar{x}_1 - \bar{x}_2)^T \tilde{\mathbf{S}}^{-1}(\lambda^{LW})E(x|I_{I_1}) - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)^T \tilde{\mathbf{S}}^{-1}(\lambda^{LW})(\bar{x}_1 + \bar{x}_2).$$

$$\text{var}(\tilde{W}(x, \lambda^{LW})|I_{I_1}) = (\bar{x}_1 - \bar{x}_2)^T \tilde{\mathbf{S}}^{-1}(\lambda^{LW})\text{var}(x|I_{I_1})\tilde{\mathbf{S}}^{-1}(\lambda^{LW})(\bar{x}_1 - \bar{x}_2)$$

The proof is complete. \square

Proof of Theorem 2. Considering $d = \bar{x}_1 - \bar{x}_2$ and $d^* = \frac{\bar{x}_1 + \bar{x}_2}{2}$ in Equations (8) and (9), we have

$$E(W(x|I_{I_1})) = d^T \mathbf{S}^{-1} \mu_1 - d^T \mathbf{S}^{-1} d^*$$

$$\text{var}(W(x|I_{I_1})) = d^T \mathbf{S}^{-1} \Sigma \mathbf{S}^{-1} d.$$

Let $x^* = Ax + b$, where A is a nonsingular matrix, A and b are considered in such a way that $\Sigma \rightarrow \mathbf{I}; \mu_1 - \mu_2 \rightarrow \delta, \delta = (\Delta, \mathbf{0}, \dots, \mathbf{0}); \Delta = ((\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2))^{\frac{1}{2}}$; and $\mu_1 \rightarrow \mathbf{0}$; we also define the variables Y, V and Z as follows

$$\mathbf{S} = \mathbf{I} + \frac{1}{\sqrt{n}}V \quad ; \quad \bar{x}_1 = \frac{1}{\sqrt{n}}Z \quad ; \quad d = \delta - \frac{1}{\sqrt{n}}Y.$$

As a result, Equations (8) and (9) are equal to

$$E(W(x|I_{I_1})) = -d^T \mathbf{S}^{-1} d^*$$

and

$$\text{var}(W(x|I_{I_1})) = d^T \mathbf{S}^{-2} d.$$

Using the Taylor’s series expansion, we have

$$\begin{aligned} E(W(x|I_{I_1})) &= -d^T \mathbf{S}^{-1} d^* \\ &= -d^T \left(\mathbf{I} - \frac{1}{n^{\frac{1}{2}}}V + \frac{1}{n}V^2 - \frac{1}{n^{\frac{3}{2}}}V^3 + \dots \right) d^* \\ &= -d^T d^* + \frac{1}{n^{\frac{1}{2}}}d^T V d^* - \frac{1}{n}d^T V^2 d^* + O(n^{-\frac{3}{2}}) \end{aligned}$$

and

$$\begin{aligned} \text{var}(W(x|I_{I_1})) &= d^T \mathbf{S}^{-2} d \\ &= d^T \left(\mathbf{I} - \frac{2}{n^{\frac{1}{2}}}V + \frac{3}{n}V^2 - \frac{4}{n^{\frac{3}{2}}}V^3 + \dots \right) d \\ &= d^T d - \frac{2}{n^{\frac{1}{2}}}d^T V d - \frac{3}{n}d^T V^2 d + o(n^{-\frac{3}{2}}). \end{aligned}$$

Using Ledoit and Wolf [25] and $\tilde{\mathbf{S}}(\lambda^{LW}) = (1 - \lambda^{LW})\mathbf{S} + \lambda^{LW}\mathbf{T}$, we obtain

$$\begin{aligned}
 E(\tilde{W}(x, \lambda^{LW})|I_1) &= -\mathbf{d}^T \tilde{\mathbf{S}}^{-1}(\lambda^{LW}) \mathbf{d}^* \\
 &= -\mathbf{d}^T \mathbf{d}^* + \frac{1 - \lambda^{LW}}{n^{\frac{1}{2}}} \mathbf{d}^T \mathbf{V} \mathbf{d}^* - \frac{(1 - \lambda^{LW})^2}{n} \mathbf{d}^T \mathbf{V}^2 \mathbf{d}^* \\
 &\quad + O((1 - \lambda^{LW})^3 n^{-\frac{3}{2}}) \\
 &= -\mathbf{d}^T \mathbf{d}^* + \frac{1}{n^{\frac{1}{2}}} \mathbf{d}^T \mathbf{V} \mathbf{d}^* - \frac{1}{n} \mathbf{d}^T \mathbf{V}^2 \mathbf{d}^* - \frac{\lambda^{LW}}{n^{\frac{1}{2}}} \mathbf{d}^T \mathbf{V} \mathbf{d}^* \\
 &\quad - \frac{\lambda^{LW}(\lambda^{LW} - 2)}{n} \mathbf{d}^T \mathbf{V}^2 \mathbf{d}^* + O((1 - \lambda^{LW})^3 n^{-\frac{3}{2}}) \\
 &= E(W(x|I_1)) + \mathbf{B} + O(\lambda^{LW} n^{-\frac{1}{2}}),
 \end{aligned}$$

where $\mathbf{B} = \frac{\lambda^{LW}(\lambda^{LW}-2)}{n} \Delta^2$.

Moreover,

$$\begin{aligned}
 \text{var}(\tilde{W}(x, \lambda^{LW})|I_1) &= \mathbf{d}^T \tilde{\mathbf{S}}^{-2}(\lambda^{LW}) \mathbf{d} \\
 &= \mathbf{d}^T \mathbf{d} - \frac{2(1 - \lambda^{LW})}{n^{\frac{1}{2}}} \mathbf{d}^T \mathbf{V} \mathbf{d} - \frac{3(1 - \lambda^{LW})^2}{n} \mathbf{d}^T \mathbf{V}^2 \mathbf{d} + o((1 - \lambda^{LW})^3 n^{-\frac{3}{2}}) \\
 &= \mathbf{d}^T \mathbf{d} - \frac{2}{n^{\frac{1}{2}}} \mathbf{d}^T \mathbf{V} \mathbf{d} + \frac{3}{n} \mathbf{d}^T \mathbf{V}^2 \mathbf{d} + \frac{3\lambda^{LW}(\lambda^{LW} - 2)}{n} \mathbf{d}^T \mathbf{V}^2 \mathbf{d} \\
 &\quad + o(\lambda^{LW} n^{-\frac{1}{2}}) \\
 &= \text{var}(W(x|I_1)) + \boldsymbol{\psi}(\lambda^{LW}) \Delta^2 + o(\lambda^{LW} n^{-\frac{1}{2}}),
 \end{aligned}$$

in which $\boldsymbol{\psi}(\lambda^{LW}) = \frac{3\lambda^{LW}(\lambda^{LW}-2)}{n}$. Since $0 < \lambda^{LW} < 1$, we obtain $\boldsymbol{\psi}(\lambda^{LW}) < 0$, and the proof is complete. \square

References

1. Clemmensen, L.; Hastie, T.; Witten, D.; Ersbøll, B. Sparse discriminant analysis. *Technometrics* **2011**, *53*, 406–413. [[CrossRef](#)]
2. Peck, R.; Van Ness, J. The use of shrinkage estimators in linear discriminant analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1982**, *5*, 530–537. [[CrossRef](#)] [[PubMed](#)]
3. Srivastava, M.S. Multivariate theory for analyzing high dimensional data. *J. Jpn. Stat. Soc.* **2007**, *37*, 53–86. [[CrossRef](#)]
4. Dempster, A.P. Covariance selection. *Biometrics* **1972**, *28*, 157–175. [[CrossRef](#)]
5. Meinshausen, N.; Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *Ann. Stat.* **2006**, *34*, 1436–1462. [[CrossRef](#)]
6. Banerjee, O.; El Ghaoui, L.; d’Aspremont, A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **2008**, *9*, 485–516.
7. Friedman, J.; Hastie, T.; Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **2008**, *9*, 432–441. [[CrossRef](#)]
8. Bickel, P.J.; Levina, E. Covariance regularization by thresholding. *Ann. Stat.* **2008**, *36*, 2577–2604. [[CrossRef](#)]
9. Cai, T.T.; Zhang, L. High dimensional linear discriminant analysis: Optimality, adaptive algorithm and missing data. *J. R. Stat. Soc. Ser. (Stat. Methodol.)* **2019**, *89*, 675–705.
10. Rothman, A.J.; Levina, E.; Zhu, J. Generalized thresholding of large covariance matrices. *J. Am. Stat. Assoc.* **2009**, *104*, 177–186. [[CrossRef](#)]
11. Bien, J.; Tibshirani, R. Sparse estimation of a covariance matrix. *Biometrika* **2011**, *98*, 807–820. [[CrossRef](#)] [[PubMed](#)]
12. Fan, J.; Liao, Y.; Liu, H. An overview of the estimation of large covariance and precision matrices. *Econom. J.* **2016**, *19*, C1–C32. [[CrossRef](#)]
13. Stein, C.; James, W. Estimation with quadratic loss. In Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 20–30 June 1961; Volume 1, pp. 361–379.
14. Efron, B. Biased versus unbiased estimation. *Adv. Math.* **1975**, *16*, 259–277. [[CrossRef](#)]
15. Efron, B.; Morris, C. Data analysis using Stein’s estimator and its generalizations. *J. Am. Stat. Assoc.* **1975**, *70*, 311–319. [[CrossRef](#)]
16. Efron, B.; Morris, C. Multivariate empirical Bayes and estimation of covariance matrices. *Ann. Stat.* **1976**, *4*, 22–32. [[CrossRef](#)]
17. Di Pillo, P.J. The application of bias to discriminant analysis. *Commun. Stat. Theory Methods* **1976**, *5*, 843–854. [[CrossRef](#)]

18. Campbell, N.A. Shrunken estimators in discriminant and canonical variate analysis. *J. R. Stat. Soc. Ser. (Appl. Stat.)* **1980**, *29*, 5–14. [[CrossRef](#)]
19. Mkhadri, A. Shrinkage parameter for the modified linear discriminant analysis. *Pattern Recognit. Lett.* **1995**, *16*, 267–275. [[CrossRef](#)]
20. Choi, Y.-G.; Lim, J.; Roy, A.; Park, J. Fixed support positive-definite modification of covariance matrix estimators via linear shrinkage. *J. Multivar. Anal.* **2019**, *171*, 234–249. [[CrossRef](#)]
21. Bickel, P.J.; Levina, E. Regularized estimation of large covariance matrices. *Ann. Stat.* **2008**, *36*, 199–227. [[CrossRef](#)]
22. Khare, K.; Rajaratnam, B. Wishart distributions for decomposable covariance graph models. *Ann. Stat.* **2011**, *39*, 514–555. [[CrossRef](#)]
23. Cai, T.; Zhou, H. Minimax estimation of large covariance matrices under ℓ_1 -norm. *Stat. Sin.* **2012**, *22*, 1319–1349.
24. Maurya, A. A well-conditioned and sparse estimation of covariance and inverse covariance matrices using a joint penalty. *J. Mach. Learn. Res.* **2016**, *17*, 4457–4484.
25. Ledoit, O.; Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.* **2004**, *88*, 365–411. [[CrossRef](#)]
26. Wang, C.; Pan, G.; Tong, T.; Zhu, L. Shrinkage estimation of large dimensional precision matrix using random matrix theory. *Stat. Sin.* **2015**, *25*, 993–1008. [[CrossRef](#)]
27. Hong, Y.; Kim, C. Recent developments in high dimensional covariance estimation and its related issues, a review. *J. Korean Stat. Soc.* **2018**, *47*, 239–247. [[CrossRef](#)]
28. Le, K.T.; Chaux, C.; Richard, F.; Guedj, E. An adapted linear discriminant analysis with variable selection for the classification in high-dimension, and an application to medical data. *Comput. Stat. Data Anal.* **2020**, *152*, 107031. [[CrossRef](#)]
29. Srivastava, M.S. Some tests concerning the covariance matrix in high dimensional data. *J. Jpn. Stat. Soc.* **2005**, *35*, 251–272. [[CrossRef](#)]
30. Ledoit, O.; Wolf, M. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Stat.* **2012**, *40*, 1024–1060. [[CrossRef](#)]
31. Friedman, J.H. Regularized discriminant analysis. *J. Am. Stat. Assoc.* **1989**, *88*, 165–175. [[CrossRef](#)]
32. Wu, Y.; Qin, Y.; Zhu, M. Quadratic discriminant analysis for high-dimensional data. *Stat. Sin.* **2019**, *29*, 939–960. [[CrossRef](#)]