HHS Public Access

Author manuscript

Chemometr Intell Lab Syst. Author manuscript; available in PMC 2021 November 15.

Published in final edited form as:

Chemometr Intell Lab Syst. 2020 November 15; 206: . doi:10.1016/j.chemolab.2020.104142.

Sparse Linear Discriminant Analysis using the Prior-Knowledge-Guided Block Covariance Matrix

Jin Hyun Nam[a,b], **Donguk Kim**[c,*], **Dongjun Chung**[d,*]

[a]Division of Biostatistics and Bioinformatics, Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC 29412, United States of America;

[b]School of Pharmacy, Sungkyunkwan University, Suwon, Republic of Korea;

[c]Department of Statistics, Sungkyunkwan University, Seoul, Republic of Korea;

[d]Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio 43210, United States of America.

Abstract

There are two key challenges when using a linear discriminant analysis in the high-dimensional setting, including singularity of the covariance matrix and difficulty of interpreting the resulting classifier. Although several methods have been proposed to address these problems, they focused only on identifying a parsimonious set of variables maximizing classification accuracy. However, most methods did not consider dependency between variables and efficacy of selected variables appropriately. To address these limitations, here we propose a new approach that directly estimates the sparse discriminant vector without a need of estimating the whole inverse covariance matrix, by formulating a quadratic optimization problem. Furthermore, this approach also allows to integrate external information to guide the structure of covariance matrix. We evaluated the proposed model with simulation studies. We then applied it to the transcriptomic study that aims to identify genomic markers predictive of the response to cancer immunotherapy, where the

*Corresponding authors. **Author Contact:** Dongjun Chung, Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, chung.911@osu.edu, Phone: (614) 685-3183.

Author Statement:

Jin Hyun Nam: Methodology, Software, Formal analysis, Investigation, Writing - Original Draft, Visualization;
Donguk Kim: Methodology, Investigation, Writing - Original Draft, Supervision;
Dongjun Chung: Methodology, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Supervision, Project administration, Funding acquisition.

Author's contributions:

Jin Hyun Nam: Methodology, Software, Formal analysis, Investigation, Writing - Original Draft, Visualization; **Donguk Kim:** Methodology, Investigation, Writing - Original Draft, Supervision; **Dongjun Chung:** Methodology, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Supervision, Project administration, Funding acquisition

Supplementary Materials: The reader is referred to the online Supplementary Materials for additional Tables and Figures (Tables S1 – S5; Figures S1 – S3).

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

covariance matrix was constructed based on the prior knowledge available in the pathway database.

Keywords

Linear discriminant analysis; penalized approach; data integration; block covariance matrix; cancer immunotherapy

1. Introduction

This work was motivated by the study that aims to identify genomic markers predictive of the response to cancer immunotherapy. The cancer immunotherapy is treatment that uses certain parts of the immune system to fight against cancer. Specially, it either stimulates the patient's own immune system to work harder or smarter to attack cancer cells or introduces certain immune components to patients [1]. Recently it attracts a lot of attention because it turns out to be effective for certain types of cancer that was not easy to treat with other therapeutic approaches. Among those, immune checkpoint blockades (ICB), including Anti-PD-1/PDL1 and Anti-CTLA4, have revolutionized the cancer treatment. However, still significant degree of heterogeneity among patients has been reported and only less than one thirds of patients can be benefited from ICB [2]. Hence, it is of great interest to identify biomarkers predictive of how patients will respond to ICB [3].

Our motivating dataset was obtained from a large-scale phase II clinical trial study investigating major determinants of clinical outcome of metastatic urothelial cancer patients treated with atezolizumab, PDL1 blockade (IMvigor210) [4]. Here gene expressions were measured for tissue samples obtained from 298 patients, including 68 responders (complete or partial responses) or 230 non-responders (stable or progressive disease). Hence, identification of genomic markers to predict response to this cancer immunotherapy can be formulated as a high-dimensional variable selection problem (selection of genes) in the binary classification setting (responders vs. non-responders). There are various types of statistical approaches for this purpose, including linear discriminant analysis (LDA) [5, 6], quadratic discriminant analysis [6], support vector machine [7, 8], K-nearest neighbors [9], classification and regression trees [10], and random forest [11]. Among those, here we focus on LDA because of its stability, simplicity, and interpretability. First, because the final classifier of LDA is composed of a linear combination of predictors, it is easy to check the effectiveness of each predictor from the fitted LDA model. Second, because LDA considers a variance-covariance matrix in the model fitting, correlations among predictors can be taken into account. These properties are really powerful for a study to identify biomarkers based on a genomic study, which is known to have a complicated correlation structure. In this manuscript, we especially focus on penalized LDA approaches as they implement variable selection and classification simultaneously within a unified framework while maintaining the theoretical rigor and flexibility needed for the extension to the high-dimensional setting.

LDA can suffer from two problems when it is applied to high-dimensional data, which have a large number of variables compared to the number of observations. First of all, the standard LDA cannot be applied at all in this case because the estimate of the covariance

matrix tends to singular. Even when the estimate is nonsingular, the resulting classifier is unstable and often has poor classification performance due to the small sample size. Secondly, it is not straightforward to interpret the resulting classifier due to the large number of variables. The statistical approaches to address these problems can be categorized into three groups. First, we can reduce the dimension of data by using a feature selection method, e.g., pre-filtering [12–15]. Second, we can replace the full covariance matrix with a diagonal, regularized, or sparse covariance matrix [12, 16, 17]. Alternatively, we can replace the inverse covariance matrix with a sparse inverse covariance matrix [18]. Third, we can estimate the sparse discriminant vector by using a penalty function [19–21]. In these methods, the inverse covariance matrix is usually used to estimate the discriminant vector. Alternatively, the sparse discriminant vector can also be estimated without using an inverse covariance matrix [22, 23].

However, these approaches have still the multiple limitations. Specifically, the dimension reduction approach using feature selection methods do not consider correlations between variables. On the other hand, methods based on a modified covariance matrix do not implement the dimension reduction, which makes it hard to interpret the resulting classifier. Multiple penalized LDA approaches have been suggested to address these two problems. They include optimal scoring [19, 20], Fisher discriminant analysis [21], discriminant analysis based on the Bayes' rule [22], and direct approach to estimate the discriminant vector avoiding separate estimation of the covariance matrix [23, 24]. However, these methods still focus only on identifying a parsimonious set of variables maximizing classification accuracy and they do not consider the efficacy of selected variables appropriately.

To address this, here we propose a sparse linear discriminant analysis (SLDA) based on the Bayes' rule, which can be solved using a quadratic optimization. We further use a covariance matrix instead of an inverse covariance matrix in the quadratic optimization, and replace the full sample covariance matrix with a prior-knowledge-guided block covariance matrix, to improve the stability and reproducibility in variable selection. Also, we utilize Elastic Net as a penalty function to simultaneously promote sparseness of coefficients and address correlations among variables. This paper is structured as follows. In Section 2, we propose the sparse LDA framework with a penalty function. Section 3 evaluates the performance of SLDA with a prior-knowledge-guided block covariance matrix with simulation studies. In Section 4, we apply the proposed method to the real data (IMvigor210), along with the prior knowledge obtained from a public pathway database. Finally, In Section 5, we discuss strengths of the proposed approach and future directions.

2. Methodology

In this section, we describe our approach, a sparse linear discriminant analysis using a prior-knowledge-guided block covariance matrix with the Elastic Net penalty (BSLDA). We first define notations as follows. For sample i, $i = 1, \ldots, n$, we observe the p-dimensional vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ and the class label y_i, where x_{ij} represents the value of j-th variable from sample i and y_i is a categorical variable with K levels. The data with p number of variables

and n number of observations can be expressed as an $n \times p$ matrix $\boldsymbol{X} = \{(x_{ij}); i = 1, \ldots, n, j = 1, \ldots, p\}$.

2.1. Sparse LDA based on the Bayes' Rule

Assuming that the p-dimensional vector $\boldsymbol{x}_{k,i}$ follows the multivariate normal distribution with a mean vector $\boldsymbol{\mu}_k$ and a common variance-covariance matrix Σ, the discriminant function D_k ($k = 1, \ldots, K$) based on the Bayes' rule for K-category classification problem is defined as:

$$D_k(\boldsymbol{x}) = \left(\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_k\right)^T \boldsymbol{\beta}_k + \beta_{k0}, \tag{1}$$

where $\boldsymbol{\beta}_k = \Sigma^{-1}\boldsymbol{\mu}_k$ and $\beta_{k0} = \ln \boldsymbol{\pi}_k$. Then, the classification rule is $argmax_k D_k(x)$, i.e., the new observation $\boldsymbol{x}$ is allocated to k' if $argmax_k D_k(x) = k'$. Especially, when the data has two classes ($K = 2$), the discriminant function D will be:

$$D(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\beta} + \beta_0, \tag{2}$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$, $\boldsymbol{\beta} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, and $\beta_0 = \ln \boldsymbol{\pi}_1/\boldsymbol{\pi}_2$. We can assign the new observation $\boldsymbol{x}$ to the class 1 if $D(\boldsymbol{x}) > 0$, and assign it to the class 2 otherwise.

Here, our main task is to identify the discriminant vector, i.e., $\boldsymbol{\beta}_k = \Sigma^{-1}\boldsymbol{\mu}_k$ for the general case and $\boldsymbol{\beta} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ when $K = 2$. These discriminant vectors can be easily estimated given the inverse covariance matrix Σ. However, in the high dimensional setting, the existence of an inverse covariance matrix is often not guaranteed due to the low rank of covariance matrix. In addition, when estimating the discriminant vectors, it is desirable to eliminate redundant variables to facilitate interpretation. Note that since the discriminant function D_k (D for $K = 2$) is a linear function of discriminant vector $\boldsymbol{\beta}_k$ ($\boldsymbol{\beta}$ for $K = 2$), the contribution of j-th variable to the classification performance is negligible if the absolute value of its coefficient, $|\beta_{kj}|$, is small. To address these issues, we suggest a penalization approach to estimate the discriminant vector, which minimizes the sum of squared difference between $\Sigma\boldsymbol{\beta}_k$ and the mean vector $\boldsymbol{\mu}_k$ by considering the relationship that $\boldsymbol{\beta}_k = \Sigma^{-1}\boldsymbol{\mu}_k$. Similarly, when $K = 2$, we minimize the squared difference between $\Sigma\boldsymbol{\beta}$ and the mean difference vector $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ by considering the relationship that $\boldsymbol{\beta} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

For the estimation of the discriminant vector based on samples, we suggest the following optimization rule for the general K:

$$minimize_{\beta_k} \sum_{k=1}^{K} \left\{ \| \tilde{\Sigma}\boldsymbol{\beta}_k - \hat{\boldsymbol{\mu}}_k \|_2^2 + P(\boldsymbol{\beta}_k) \right\}, \tag{3}$$

where $\tilde{\Sigma} = \hat{\Sigma} + \Omega$, Ω is a positive definite matrix, $P(\boldsymbol{\beta}_k)$ is a penalty function of $\boldsymbol{\beta}_k$, and $\hat{\Sigma}$ and $\hat{\boldsymbol{\mu}}_k$ are estimates of the covariance matrix and the mean vector for class k, respectively.

Similarly, we suggest the following optimization rule for the estimation of a sparse discriminant vector when $K = 2$:

$$minimize_{\boldsymbol{\beta}} \left\| \widetilde{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) \right\|_2^2 + P(\boldsymbol{\beta}), \tag{4}$$

where $\widetilde{\Sigma} = \hat{\Sigma} + \Omega$, Ω is positive definite matrix, $P(\boldsymbol{\beta})$ is a penalty function of $\boldsymbol{\beta}$, and $\hat{\Sigma}$ and $\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$ are estimates of the covariance matrix and the mean difference vector between two classes, respectively.

In a high-dimensional setting, $\hat{\Sigma}$ does not have full rank and cannot be estimated reliably. The standard approach to make the covariance matrix estimation more stable is to regularize a covariance matrix by adding a positive definite matrix Ω. The positive definite matrix of usual choice is $\Omega = \gamma \mathrm{I}$, where γ is a regularization parameter and I is the identity matrix. Here the optimal value for the parameter γ depends on the problem of interest and has to be optimized by a researcher [16, 20, 22, 25, 26]. In this manuscript, we assume that the data is scaled, i.e., each variable of data $\boldsymbol{X}$ is centered to have mean zero and standard deviation one. In our simulation studies with various settings assuming the scaled data, we found that using the identity matrix for Ω ($\gamma = 1$) works well in practice in terms of both classification accuracy and variable selection performance. Thus, we recommend to use the identity matrix for Ω for the researcher's convenience assuming that the data is scaled.

2.2. Estimation of Sparse Discriminant Coefficients with the Elastic Net Penalty

In order to achieve sparseness in the discriminant vector, we utilize the Elastic Net penalty function [27], which combines the LASSO penalty [28] with the Ridge penalty [29]. The Elastic Net penalty function promotes sparsity in the discriminant vector with the LASSO penalty while addressing correlation among variables using the Ridge penalty. When the Elastic Net penalty function is incorporated, the objective functions (which correspond to Equations (3) and (4)) are obtained as:

$$L(\boldsymbol{\beta}, \alpha) = \begin{cases} \left\| \widetilde{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) \right\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2, K = 2 \\ \sum_{k=1}^{K} \left\{ \left\| \widetilde{\Sigma}\boldsymbol{\beta}_k - \hat{\boldsymbol{\mu}}_k \right\|_2^2 + \lambda_1 \|\boldsymbol{\beta}_k\|_1 + \lambda_2 \|\boldsymbol{\beta}_k\|_2^2 \right\}, K > 2 \end{cases}. \tag{5}$$

We utilize the LARS-EN algorithm [27] to solve this Elastic Net problem. The key idea of the LARS-EN algorithm is to convert an objective function with the Elastic Net penalty into a LASSO form. Let $\widetilde{\Sigma}^*_{(p+p)\times p} = (1+\lambda_2)^{-1/2} \begin{pmatrix} \widetilde{\Sigma} \\ \sqrt{\lambda_2}\boldsymbol{I} \end{pmatrix}$, $\hat{\boldsymbol{\mu}}^*_{(p+p)} = \begin{pmatrix} \hat{\boldsymbol{\mu}} \\ \mathbf{0} \end{pmatrix}$, $\gamma = \lambda_1/\sqrt{1+\lambda_2}$, and $\boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}$. Then the modified objective function can be written as

$$L(\boldsymbol{\beta}^*, \gamma) = \left\| \widetilde{\Sigma}^* \boldsymbol{\beta}^* - \hat{\boldsymbol{\mu}}^* \right\|_2^2 + \gamma \|\boldsymbol{\beta}^*\|_1 . \tag{6}$$

The solution $\hat{\boldsymbol{\beta}}^*$ is called the naïve Elastic Net coefficient and can be obtained using the least angle regression (LARS) algorithm [30]. The LARS algorithm efficiently solves the entire LASSO solution path using the same order of computations as the single ordinary least squares fit. Finally, Zou and Hastie (2005) [27] suggested to use the rescaled solution that corrects the naive Elastic Net coefficients. The final solution is defined as

$$\hat{\boldsymbol{\beta}} = \sqrt{1+\lambda_2}\hat{\boldsymbol{\beta}}^*.$$

2.3. Parameter Tuning

While the LARS-EN algorithm provides an efficient solution to identify the sparse discriminant vector, two problems still remain. First, it is not guaranteed that the LASSO solutions obtained using the approach described above provides the best discriminant coefficients in the sense of the classification performance. Second, it is not straightforward to find the optimal combination of the two tuning parameters λ_1 and λ_2 and this can potentially affect stability of variable selection. To address these problems, we use the following reparameterization tricks. First, we replace the tuning of the pair (λ_1, λ_2) with the pair (s, λ_2), where a fraction s is the ratio of the L_1 norm of the coefficient vector relative to the norm of the full least square solution. Note that here the range of s is restricted to (0, 1). Second, we also restrict the parameter space of λ_2 to (0, 1). Although it is possible that this restriction might give slightly sub-optimal penalized regression coefficient estimates, this makes the parameter tuning significantly more convenient and also improves stability in model fitting by reducing the parameter space to be searched. Third, in order to further stabilize the variable selection results, we set the fraction $\lambda_2 = s$. Hence, it suffices to only tune the parameter s to control both sparseness and shrinkage of coefficients, where the smaller s value gives more weight on the LASSO penalty while the larger s value gives more weight on the Ridge penalty. Again, although it is possible that restricting the λ_2 parameter space in this way might result in slightly sub-optimal discriminant coefficient estimates in the sense of classification performance, this approach makes the model selection procedure much more intuitive and also improves stability and robustness of variable selection. Note that essentially here we aim to incorporate the L_2 norm penalty proportional to the degree of sparseness for the purpose of considering the correlation between variables within each block in the covariance matrix. In this sense, this proposed tuning approach can be considered as an Elastic Net formula with weak L_2 penalization, or equivalently, a modified LASSO problem. Combining all of these tricks together, we now need to tune only a single parameter s within the range between 0 and 1. We choose the value of s that maximizes the cross-validation classification accuracy.

2.4. Prior-Knowledge-Guided Block Covariance Matrix and Software Implementation

When solving Equations (3) and (4) with respect to the discriminant vector $\boldsymbol{\beta}$, we can use the sample mean vector $\hat{\boldsymbol{\mu}}_k = \sum_{i=1}^{n_k} \boldsymbol{x}_{k,i}/n_k$ and the covariance matrix $\hat{\Sigma} = \sum_{k=1}^{K}(n_k-1)\hat{\Sigma}_k/(n-k)$, where $\hat{\Sigma}_k = \sum_{i=1}^{n_k}(\boldsymbol{x}_{k,i}-\hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_{k,i}-\hat{\boldsymbol{\mu}}_k)^T/(n_k-1)$. However, this standard estimation of the sample covariance matrix can be unstable in the high dimensional setting while it can also include a significant number of redundant covariance terms. As a result, the resulting classifier suffers from the poor performance.

Multiple types of covariance structure have been proposed to improve this covariance matrix estimation, especially by eliminating redundant covariance terms. The simplest approach is to consider only the diagonal covariance matrix by assuming all variables are independent

[12, 31–34]. A more relaxed approach is to use a block structure of covariance matrix, which assumes that variables are independent between different blocks while correlations between variables are allowed within each block [34–38]. Other proposed covariance matrix structures include banded covariance matrices [39–42], sparse and low-rank covariance matrices [43–45]. In order to make the covariance matrix estimation more stable, we use a block diagonal covariance matrix, where the covariance matrix with $(L + 1)$ blocks is defined as

$$\tilde{\Sigma} = \begin{bmatrix} \tilde{\Sigma}_1 & 0 & 0 & \cdots & 0 \\ 0 & \tilde{\Sigma}_2 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \cdots & 0 \\ \vdots & \vdots & \vdots & \tilde{\Sigma}_L & 0 \\ 0 & 0 & 0 & 0 & \tilde{\Sigma}_{L+1} \end{bmatrix}.$$

The block diagonal covariance matrix assumes that the data $\boldsymbol{x}$ is composed of several mutually independent subsets. In other words, this approach assumes the variables belonging to different blocks to be mutually independent while allowing the variables to be correlated with each other within each block. This might sound like a strong assumption but it has been reported that such ignorance of weak associations between blocks actually rather improves the classification performance [16, 46].

To determine the number of blocks and the block size, we utilize exterior prior knowledge. For example, in the context of our motivating example that aims to identify genomic markers predictive of the response to cancer immunotherapy, the Kyoto Encyclopedia of Genes and Genomes (KEGG; https://www.genome.jp/kegg/) pathway annotations can be utilized to construct blocks in the covariance matrix. Alternatively, these blocks can be obtained based on gene modules or subnetworks identified using text mining of biomedical literature [47, 48]. The proposed sparse linear discriminant analysis using prior-knowledge-guided covariance matrix was implemented as an R package ‘bslda’, which is currently publicly available at https://elflini.github.io/bslda/.

3. Simulation Study

We first performed simulation studies to evaluate performance of the proposed sparse LDA using prior-knowledge-guided block covariance matrix (BSLDA). We also considered a sparse LDA without prior-knowledge-guided block covariance matrix (SLDA), a sparse discriminant analysis (SDA, [20]), a penalized LDA (PLDA, [21]), and a diagonal LDA (DLDA, [12]) as competing approaches. Both SDA and PLDA are penalized LDA. SDA utilizes optimal scoring criterion with Elastic Net penalty, where the discriminant vector is identified by converting a classification problem into a regression form. On the other hand, PLDA is based on Fisher’s discriminant problem with LASSO penalty. Both methods implemented regularization of the covariance matrix by adding a positive definite matrix to stabilize the covariance matrix. Finally, DLDA does not have variable selection procedure (i.e., uses all the variables) while only diagonal terms of covariance matrix are considered instead of the regularization.

Since BSLDA is one type of classification methods, we used test data prediction accuracy as the criterion to compare classification performances between BSLDA and other penalized discriminant analysis approaches. Specifically, prediction accuracy on the testing dataset is defined as:

$$\text{Accuracy} = [\ \#\text{ of correctly classified observations }]/[\ \#\text{ of total observations }].$$

We also evaluated true positive rate (TPR) and false positive rate (FPR) to compare variable selection performances between BSLDA and other penalized discriminant analysis approaches. TPR and FRR are defined as:

$$\text{TPR} = [\ \#\text{ of selected signal variables }]/[\ \#\text{ of signal variables }],$$

$$\text{FPR} = [\ \#\text{ of selected non-signal variables }]/[\ \#\text{ of non-signal variables }].$$

Here, we considered the two-class problem (K=2), where there are 50 observations corresponding to each class (total 100 observations) with 400 variables. For each setting, we generated 100 sets of training and test data and used the 5-fold cross-validation to tune the parameters.

3.1. Performance Comparison

Here we assumed that the data for the 1st class $\boldsymbol{x}_{1,i} \sim N(\mathbf{0}, \Sigma)$ and the data for the 2nd class $\boldsymbol{x}_{2,i} \sim N(\boldsymbol{\mu}, \Sigma)$. For the setting #1, we assumed that $\boldsymbol{\mu} = (\mathbf{1}_{40}, \mathbf{0}_{360})$, i.e., the first 40 variables have mean of 1 and while the remaining 360 variables have mean of 0. In addition, we assumed the correlation structure that $\Sigma_{jj'} = 0.5^{|j-j'|}$, $1 \le j, j' \le 400$. For the setting #2, we assumed that $\boldsymbol{\mu} = (\mathbf{1}_{30}, \mathbf{0.5}_{40}, \mathbf{0}_{330})$, i.e., the first 30 variables have mean of 1 and the second 40 variables have mean of 0.5, while the remaining 330 variables have mean of 0. We again considered the same covariance matrix as the simulation setting #1. For the settings #3 and #4, we used the same mean vectors used for the settings #1 and #2, respectively, while we assumed the covariance matrix Σ estimated using the real data (Section 4) [4] to mimic the real data situation. For the BSLDA which needs prior knowledge to guide the covariance matrix structure, we assumed that there exist two variable groups (each with 40 and 360 variables) for the settings #1 and #3, and three variable groups (each with 30, 40, and 330 variables) for the settings #2 and #4. Figure 1 visualizes the covariance structures we assumed for each simulation setting and Figure S1 (in the Supplementary Materials) visualizes the sample covariance estimated using block structures guided by the prior knowledge.

First, in order to assess benefits of guiding the covariance matrix using prior knowledge, we compared the prediction accuracy, TPR, and FPR between BSLDA and SLDA, i.e., SLDA models with and without the prior-knowledge-guided block covariance matrix, respectively. Table 1 shows TPR, FPR, and prediction accuracy of BSLDA and SLDA in each simulation setting. In the case of settings #1 and #2, TPR and FPR were comparable between BSLDA and SLDA and the prediction accuracy was also comparably high for both approaches (about

0.95). In contrast, in the settings #3 and #4, BSLDA provides the better FPR compared to SLDA with some sacrifice of prediction accuracy, which might imply that BSLDA can be more effective in eliminating noise variables.

Second, by considering that BSLDA is one type of penalized linear discriminant models, we compared the performance of BSLDA with those of SDA and PLDA (Table 1, Table S1, and Table S2 in the Supplementary Materials). In all settings, BSLDA, SDA and PLDA perform comparably or outperform DLDA in the sense of prediction accuracy. On the other hand, the TPR and FPR of SDA vary significantly across the simulation settings, which implies that the performance of SDA is highly data-dependent. In the case of PLDA, in spite of high TPR, it also showed high FPR consistently across all the settings. Moreover, its standard deviations of FPR were highest among all the approaches we considered, which implies that PLDA is not stable in the sense of FPR control. In contrast, BSLDA showed relatively high TPR and the lowest FPR consistently across all the settings. These two results indicate strengths of BSLDA in the sense of stability and robustness with respect to data characteristics.

Third, by considering that the stability of BSLDA can be affected by the ratio of signal to non-signal variables, we performed the following additional simulation setting that considers various ratios of non-signal variables. Here we modified the simulation settings #1 and #2 by considering different numbers of non-signal variables. Specifically, we generated 160 and 260 non-signal variables (instead of 360 non-signal variables) for the setting #1, and 130 and 230 non-signal variables (instead of 330 non-signal variables) for the setting #2, so that the total numbers of variables are 200 and 300, respectively. Table S3 in the Supplementary Materials shows TPR, FPR, and prediction accuracy of BSLDA, SLDA, SDA, PLDA and DLDA. Overall, SLDA outperformed SDA and PLDA in the sense of FPR whereas SLDA also showed high TPR. However, the performance of SLDA was still affected by noise variables. In contrast, BSLDA showed high TPR, low FPR, and high prediction accuracy consistently regardless of ratios of signal to non-signal variables. These results indicate that BSLDA provides great stability in the sense of variable selection while it is also robust to existence of noise variables.

3.2. Studies of Stability and Reproducibility

Here we generated signal variables $\boldsymbol{x}_I \sim N(\mathbf{0}, \Sigma)$ with $\Sigma_{jj'} = 0.5^{|j-j'|}$, $1 \le j, j' \le 40$ and set the discriminant function $D(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x}$, where $\boldsymbol{\beta} = (\mathbf{1.2}_{10}, \mathbf{1}_{10}, \mathbf{0.8}_{10}, \mathbf{0.5}_{10})$. Then, we obtained 50 observations with $D(\boldsymbol{x}) > 0$ and 50 observations with $D(\boldsymbol{x}) < 0$, and set them to classes 1 and 2, respectively. Finally, we added independent 360 variables following the standard normal distribution. Thus, we simulated 100 observations with 400 variables and the class label vector **y** consists 50 observations of class 1 and 50 observations of class 2. For BSLDA, we further assumed that there are two variable groups with 40 and 360 variables, respectively (Figure S2 in the Supplementary Materials). Table S4 in the Supplementary Materials shows TPR, FPR, and prediction accuracy of BSLDA, SLDA, SDA, and PLDA for this setting.

First, by considering that the performance of BSLDA can be affected by the sample collection, we subsampled the 100 observations generated above with different subsampling rates. Specifically, within each class, observations were subsampled without replacement

with different sampling rates, and subsampling rate was considered from 0.2 to 0.9. For example, the subsampling rate of 0.9 means that we chose 45 observations (=50 × 0.9) in each of classes 1 and 2. Table 2 (and Table S5 in the Supplementary Materials) shows TPR, FPR and prediction accuracy of BSLDA, SLDA, SDA, and PLDA (We did not include DLDA in this comparison because it does not provide variable selection) for different resampling rates. First, in the case of SDA, FPR was well controlled at the low level but its TPR was also the lowest among the methods we considered. On the other hand, PLDA showed relatively high TPR compared to other methods but its FPR and corresponding standard deviations were also the highest among the methods we considered. In contrast, BSLDA maintained very low FPR in general and its performance was not degraded in spite of decreasing sample sizes. SLDA showed relatively lower TPR and higher FPR compared to BSLDA, which might imply benefits of using prior knowledge to guide the covariance matrix.

Second, by considering that the performance of BSLDA can be affected by misspecification of the covariance matrix structure, we performed additional simulation study that changed proportion of signal variables within each variable group. Specifically, we first assumed 40 signal variables and 360 non-signal variables. Then, we chose subsets of signal variables with different subsampling rates and considered these subsets as one variable group. The remaining non-selected signal variables and non-signal variables are considered as another variable group. For example, for the subsampling rate of 0.1, randomly selected 4 signal variables (= 40 × 0.1) were considered as one variable group while the remaining 36 signal variables (= 40 − 4) and 360 non-signal variables were considered as another variable group. This process was implemented 50 times for each subsampling rate and we considered the subsampling rate between 0.1 and 0.9. Table 3 shows TPR, FPR, and prediction accuracy of BSLDA. We can see that in spite of the misspecified block structure for the covariance matrix, BSLDA did not suffer from loss of prediction accuracy and FPR. Although TPR still decreases as the size of correctly specified signal variables decreases, its rate of decrease was still moderate.

4. Real Data Analysis

We applied the proposed BSLDA to the transcriptomic study ‘IMvigor210’ that aims to identify genomic markers predictive of the response to cancer immunotherapy [4, 49, 50]. We obtained the data using the R package ‘IMvigor210CoreBiologies’, which is available at http://research-pub.gene.com/IMvigor210CoreBiologies. There are two groups of patients, including responders (complete or partial responses) and non-responders (stable or progressive diseases). 68 patients correspond to responders and the remaining 230 samples correspond to non-responders, resulting in the total sample size of 298. We utilized the Kyoto Encyclopedia of Genes and Genomes (KEGG; https://www.genome.jp/kegg/) pathway annotations as a prior-knowledge to guide the covariance matrix, which are available from the Molecular Signatures database (MSigDB; http://software.broadinstitute.org/gsea/msigdb/). In this analysis, we considered 4792 genes, which overlapped the KEGG categories. KEGG pathway database (https://www.genome.jp/kegg/pathway.html) was constructed as a manually drawn pathway maps and it consists of the 7 main categories, including metabolism, genetic information processing, environmental

information processing, cellular processes, organismal systems, human diseases, and drug development. Among those, we considered 6 categories except for drug development. Then, in order to address overlaps of genes between pathway categories, we extracted genes that appear in more than one of these 6 pathway categories and made up a new category consisting of these overlapping genes. Finally, we utilized these 7 categories as prior knowledge to construct the block covariance matrix. Figure 2 shows a heatmap of the data and Figure S3 in the Supplementary Materials shows a heatmap of the prior-knowledge-guided block sample covariance matrix constructed using the 7 KEGG pathway categories. We used the 5-fold cross-validation for the parameter tuning.

BSLDA selected 2386 genes with 1224 genes with positive coefficient estimates and 1162 genes with negative coefficient estimates. Figure 3 shows a heatmap of the data with selected genes. We first implemented a gene set enrichment analysis of the 1162 genes with negative coefficient estimates (i.e., markers associated with non-response to the treatment) using the ToppGene Suite (https://toppgene.cchmc.org/). Some of the key pathways associated with these genes include MAPK signaling pathway, cytokine-cytokine receptor interaction, and WNT signalling pathway (Bonferroni-adjusted p-value = 2.07e–14, 1.27e–13, and 3.3e–08, respectively). Among those, Mariathasan et al. [4] reported the cytokine-cytokine receptor interaction as the key pathway associated with the non-response for the cancer immunotherapy. They reported some key marker genes in this pathway including IFNGR1, TGFB1, ACVR1, and TGFBR2 and all of them were also selected by our approach (coefficient estimates = −0.1195, −0.0739, −0.0750, and −0.0515, respectively). These genes were reported to be associated with non-response and also with reduced overall survival [4].

We next checked enrichment of 1224 genes with positive coefficient estimates (i.e., markers associated with response to the treatment) with respect to the gene sets profiled by Mariathasan et al. (Supplementary Table 8 of [4]) using a hypergeometric test. Significantly associated gene sets include $CD8^+$ T-effector signature and cell cycle (Bonferroni-adjusted p-value = 2.95e–05 and 7.67e–04, respectively). Among those, Mariathasan et al. [4] reported that $CD8^+$ T-effector ($CD8^+$ T_{eff}) signatures were highly correlated with PD-L1 expression on immune cells and also associated with increased overall survival.

The mouse study implemented by Mariathasan et al. [4] further showed that while a blockade of each of PD-L1 or TGFβ alone had little or no effect, mice treated with antibodies against both PD-L1 and TGFβ resulted in a significant reduction in tumour burden. Moreover, this combined antibody blockade significantly increased tumour-infiltrating T cells, especially $CD8^+$ T_{eff} cells, and $CD8^+$ T_{eff} signature was also increased in mouse tumours treated with this combined antibody blockade. Thus, we could validate some of the genomic markers we identified using BSLDA based on the literature. Currently we are also working on investigation of novel genomic markers identified using BSLDA.

5. Conclusions

In this paper, we proposed the sparse linear discriminant analysis using prior-knowledge-guided block covariance matrix with Elastic Net penalty (BSLDA). First, we address the

issue of inability of constructing a discriminant vector and its estimation instability in high-dimensional setting by bypassing the estimation of an inverse covariance matrix and using a covariance matrix based on a Bayes' rule. Second, we employ the Elastic Net penalty to eliminate the redundant covariance terms. Third, we utilize prior knowledge in construction of a block covariance matrix. Our simulation studies showed that this approach can provide more stable and reproducible variable selection and prediction results. Finally, we applied BSLDA to the study for identifying genomic markers associated with response to the cancer immunotherapy. The genomic markers we identified include those that were previously reported, along with a large set of potential genomic markers that can be further investigated.

We are currently working on multiple directions to further improve BSLDA. First, in this manuscript, we utilized a static and well-established pathway annotation (KEGG) to guide the covariance matrix. However, there are other annotation datasets with richer information and those related to different biological aspects. Hence, it will be of interest to investigate utilization of other databases and address relevant issues. Second, in this manuscript, we only investigated gene expression data to identify genomic markers associated with response to the cancer immunotherapy. However, other relevant datasets are also available, such as genetic alterations, and it will be of interest to investigate integration of these datasets with the gene expression data. Finally, our real data analysis identifies multiple potential candidates associated with response to the cancer immunotherapy and we plan to further investigate and validate these novel candidates. In summary, we believe that BSLDA can be a powerful tool for identifying genomic markers associated with response to the cancer immunotherapy, and for variable selection and classification in the high-dimensional setting in general.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding:

This work was supported by the National Institutes of Health [grant numbers R01-GM122078, R21-CA209848, U01-DA045300]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abbreviations:

ICB	immune checkpoint blockades
LDA	linear discriminant analysis
BSLDA	sparse LDA using prior-knowledge-guided block covariance matrix
SLDA	sparse LDA without prior-knowledge-guided block covariance matrix
SDA	sparse discriminant analysis
PDA	penalized LDA
DLDA	diagonal LDA

TPR	true positive rate
FPR	false positive rate

References

[1]. American Cancer Society, What Is Cancer Immunotherapy?. https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/immunotherapy/what-is-immunotherapy.html, 2020 (accessed 2 Feb 2020). .

[2]. Sharma P, Hu-Lieskovan S, Wargo JA, Ribas A, Primary, adaptive, and acquired resistance to cancer immunotherapy, Cell, 168 (2017) 707–723. [PubMed: 28187290]

[3]. Nishino M, Ramaiya NH, Hatabu H, Hodi FS, Monitoring immune-checkpoint blockade: response evaluation and biomarker development, Nature reviews Clinical oncology, 14 (2017) 655.

[4]. Mariathasan S, Turley SJ, Nickles D, Castiglioni A, Yuen K, Wang Y, Kadel III EE, Koeppen H, Astarita JL, Cubas R, TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells, Nature, 554 (2018) 544. [PubMed: 29443960]

[5]. Mardia KV, Multivariate analysis, 1979.

[6]. Hastie T, Tibshirani R, Friedman J, The elements of statistical learning: data mining, inference, and prediction, Springer Science & Business Media 2009.

[7]. Cortes C, Vapnik V, Support vector machine, Machine learning, 20 (1995) 273–297.

[8]. Vapnik V, The nature of statistical learning theory. 2000, There is no corresponding record for this reference, (1997).

[9]. Altman NS, An introduction to kernel and nearest-neighbor nonparametric regression, The American Statistician, 46 (1992) 175–185.

[10]. Breiman L, Friedman J, Stone CJ, Olshen RA, Classification and regression trees, CRC press 1984.

[11]. Breiman L, Random forests, Machine learning, 45 (2001) 5–32.

[12]. Dudoit S, Fridlyand J, Speed TP, Comparison of discrimination methods for the classification of tumors using gene expression data, Journal of the American statistical association, 97 (2002) 77–87.

[13]. Tibshirani R, Hastie T, Narasimhan B, Chu G, Diagnosis of multiple cancer types by shrunken centroids of gene expression, Proceedings of the National Academy of Sciences, 99 (2002) 6567–6572.

[14]. Chai H, Domeniconi C, An evaluation of gene selection methods for multi-class microarray data classification, Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics, 2004, pp. 3–10.

[15]. Guyon I, Weston J, Barnhill S, Vapnik V, Gene selection for cancer classification using support vector machines, Machine learning, 46 (2002) 389–422.

[16]. Guo Y, Hastie T, Tibshirani R, Regularized linear discriminant analysis and its application in microarrays, Biostatistics, 8 (2007) 86–100. [PubMed: 16603682]

[17]. Bien J, Tibshirani RJ, Sparse estimation of a covariance matrix, Biometrika, 98 (2011) 807–820. [PubMed: 23049130]

[18]. Cai T, Liu W, Luo X, A constrained $\ell 1$ minimization approach to sparse precision matrix estimation, Journal of the American Statistical Association, 106 (2011) 594–607.

[19]. Hastie T, Buja A, Tibshirani R, Penalized discriminant analysis, The Annals of Statistics, (1995) 73–102.

[20]. Clemmensen L, Hastie T, Witten D, Ersbøll B, Sparse discriminant analysis, Technometrics, 53 (2011) 406–413.

[21]. Witten DM, Tibshirani R, Penalized classification using Fisher's linear discriminant, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73 (2011) 753–772. [PubMed: 22323898]

[22]. Cai T, Liu W, A direct estimation approach to sparse linear discriminant analysis, Journal of the American statistical association, 106 (2011) 1566–1577.

[23]. Mai Q, Zou H, Yuan M, A direct approach to sparse discriminant analysis in ultra-high dimensions, Biometrika, 99 (2012) 29–42.

[24]. Gaynanova I, Booth JG, Wells MT, Simultaneous sparse estimation of canonical vectors in the p≫ N setting, Journal of the American Statistical Association, 111 (2016) 696–706.

[25]. Friedman JH, Regularized discriminant analysis, Journal of the American statistical association, 84 (1989) 165–175.

[26]. Tax DM, Regularizing the covariance matrix using spatial information, Proceedings of the Sixteenth Belgium-Netherlands Conference on Artificial Intelligence, Citeseer, 2004, pp. 179–186.

[27]. Zou H, Hastie T, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2005) 301–320.

[28]. Tibshirani R, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological), (1996) 267–288.

[29]. Hoerl AE, Kennard RW, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics, 12 (1970) 55–67.

[30]. Efron B, Hastie T, Johnstone I, Tibshirani R, Least angle regression, The Annals of statistics, 32 (2004) 407–499.

[31]. Speed T, Statistical analysis of gene expression microarray data, CRC Press 2003.

[32]. Ye J, Li T, Xiong T, Janardan R, Using uncorrelated discriminant analysis for tissue classification with gene expression data, IEEE/ACM Transactions on computational biology and bioinformatics, 1 (2004) 181–190. [PubMed: 17051700]

[33]. Lee JW, Lee JB, Park M, Song SH, An extensive comparison of recent classification tools applied to microarray data, Computational Statistics & Data Analysis, 48 (2005) 869–885.

[34]. Pang H, Tong T, Zhao H, Shrinkage based diagonal discriminant analysis and its applications in high dimensional data, Biometrics, 65 (2009) 1021–1029. [PubMed: 19302409]

[35]. Storey JD, Tibshirani R, Estimating the positive false discovery rate under dependence, with applications to DNA microarrays, Department of Statistics, Stanford University Stanford, CA 2001.

[36]. Langaas M, Lindqvist BH, Ferkingstad E, Estimating the proportion of true null hypotheses, with application to DNA microarray data, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2005) 555–572.

[37]. Hu P, Bull S, Jiang H, Gene network modules-based liner discriminant analysis of microarray gene expression data, International Symposium on Bioinformatics Research and Applications, Springer, 2011, pp. 286–296.

[38]. Pang H, Tong T, Ng M, Block-diagonal discriminant analysis and its bias-corrected rules, Statistical applications in genetics and molecular biology, 12 (2013) 347–359. [PubMed: 23735433]

[39]. Wu WB, Pourahmadi M, Nonparametric estimation of large covariance matrices of longitudinal data, Biometrika, 90 (2003) 831–844.

[40]. Bickel PJ, Levina E, Regularized estimation of large covariance matrices, The Annals of Statistics, 36 (2008) 199–227.

[41]. Cai T, Liu W, Adaptive thresholding for sparse covariance matrix estimation, Journal of the American Statistical Association, 106 (2011) 672–684.

[42]. Xue L, Ma S, Zou H, Positive-definite ℓ_1-penalized estimation of large covariance matrices, Journal of the American Statistical Association, 107 (2012) 1480–1491.

[43]. Richard E, Savalle P-A, Vayatis N, Estimation of simultaneously sparse and low rank matrices, arXiv preprint arXiv:1206.6474, (2012).

[44]. Zhou S, Xiu N, Luo Z, Kong L, Sparse and Low-Rank Covariance Matrices Estimation, arXiv preprint arXiv:1407.4596, (2014).

[45]. Niu YS, Hao N, Dong B, A new reduced-rank linear discriminant analysis method and its applications, Statistica Sinica, (2018) 189–202.

[46]. Nam JH, Kim D, Modified linear discriminant analysis using block covariance matrix in high-dimensional data, Communications in Statistics-Simulation and Computation, 46 (2017) 1796–1807.

[47]. Chung D, Lawson A, Zheng WJ, A statistical framework for biomedical literature mining, Statistics in medicine, 36 (2017) 3461–3474. [PubMed: 28675924]

[48]. Couch D, Yu Z, Nam JH, Allen C, Ramos PS, da Silveira WA, Hunt KJ, Hazard ES, Hardiman G, Lawson A, GAIL: An interactive webserver for inference and dynamic visualization of gene-gene associations based on gene ontology guided mining of biomedical literature, PloS one, 14 (2019) e0219195. [PubMed: 31260503]

[49]. Rosenberg JE, Hoffman-Censits J, Powles T, Van Der Heijden MS, Balar AV, Necchi A, Dawson N, O'Donnell PH, Balmanoukian A, Loriot Y, Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: a single-arm, multicentre, phase 2 trial, The Lancet, 387 (2016) 1909–1920.

[50]. Balar AV, Galsky MD, Rosenberg JE, Powles T, Petrylak DP, Bellmunt J, Loriot Y, Necchi A, Hoffman-Censits J, Perez-Gracia JL, Atezolizumab as first-line treatment in cisplatin-ineligible patients with locally advanced and metastatic urothelial carcinoma: a single-arm, multicentre, phase 2 trial, The Lancet, 389 (2017) 67–76.

Highlights:

- While the linear discriminant analysis (LDA) has been popularly used for many classification problems, its application in the high-dimensional setting remains challenging due to singularity of the covariance matrix and difficulty of interpreting the resulting classifier.
- In this paper, we propose BSLDA, a novel penalized LDA approach that addresses this problem by directly estimating the sparse discriminant vector without a need of estimating the whole inverse covariance matrix and integrating external information to guide the structure of covariance matrix.
- We found that BSLDA provides more stable and reproducible variable selection compared to competing LDA approaches.
- In our application of BSLDA to a large-scale phase II clinical trial study that aims to identify genomic markers for metastatic urothelial cancer patients treated with atezolizumab, PDL1 blockade (IMvigor210), we could not only reproduce the findings reported in the literature, but also identify novel genomic markers that can be considered for future investigation and validation.

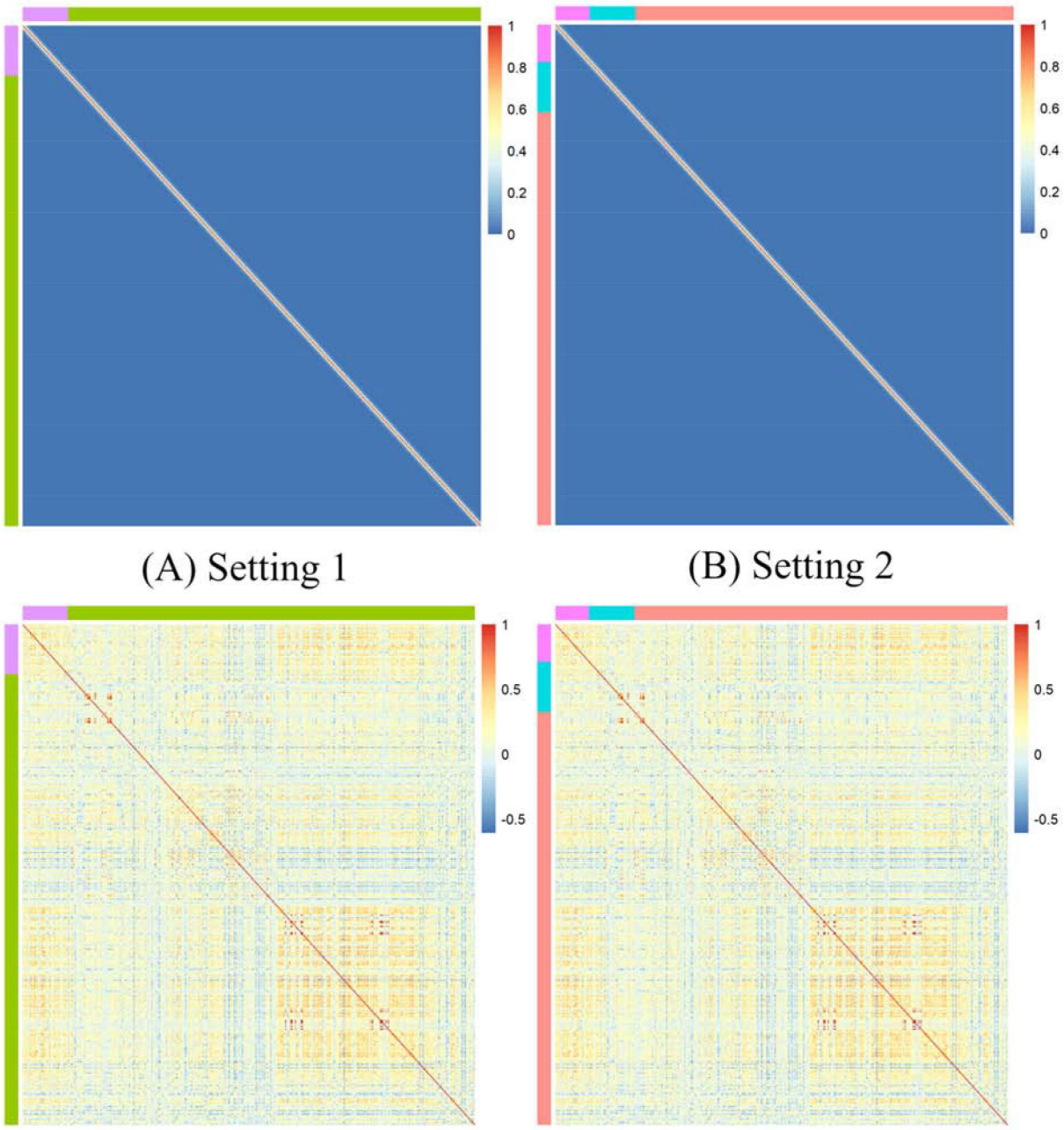

Figure 1.
Heatmaps of the covariance matrix assumed for each simulation setting.

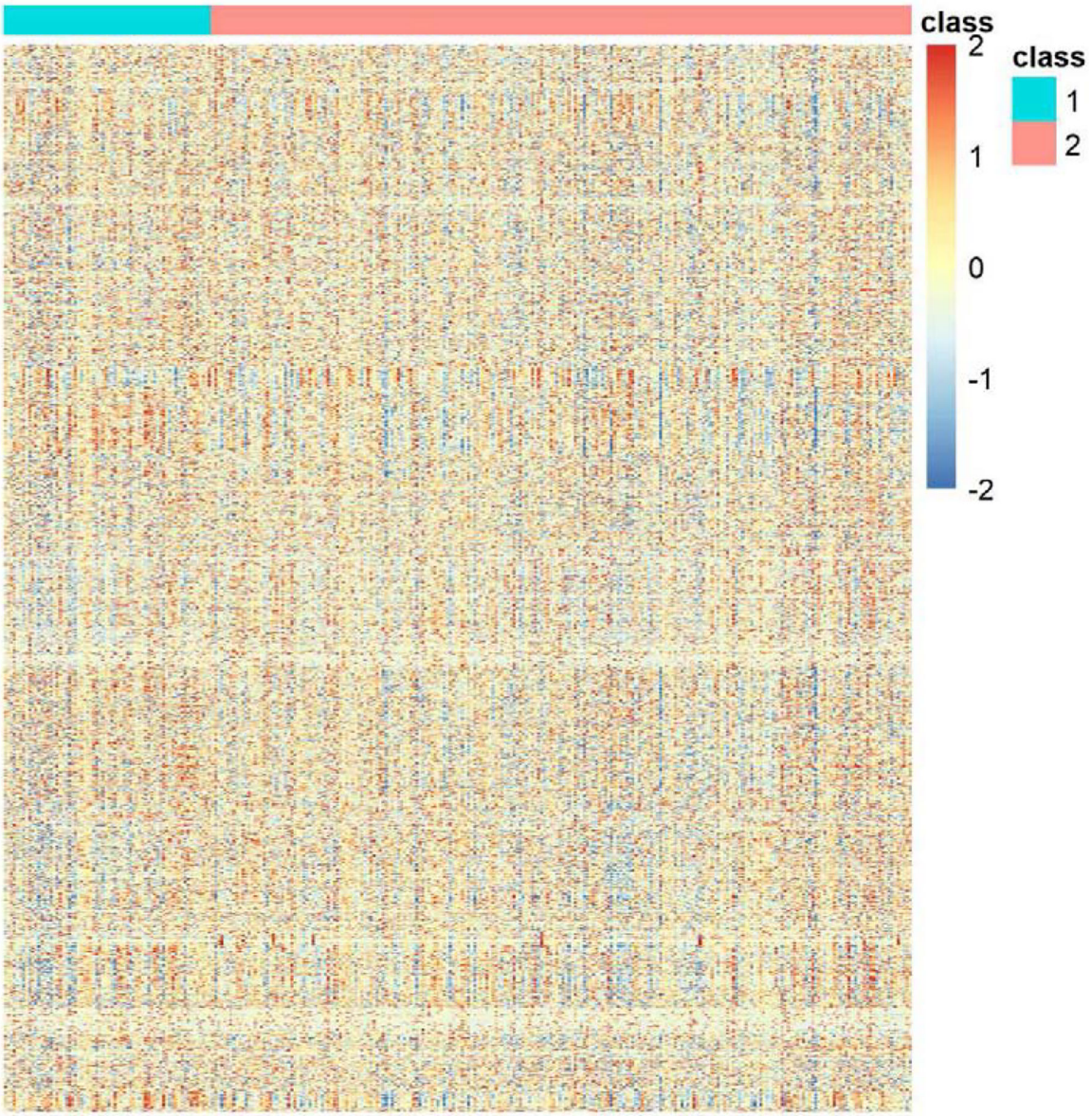

Figure 2.

A heatmap of the real data for 298 samples (columns) and 4792 genes (rows), which aims to identify genomic markers predictive of response to the cancer immunotherapy. Green and red colors indicate responders and non-responders, respectively.

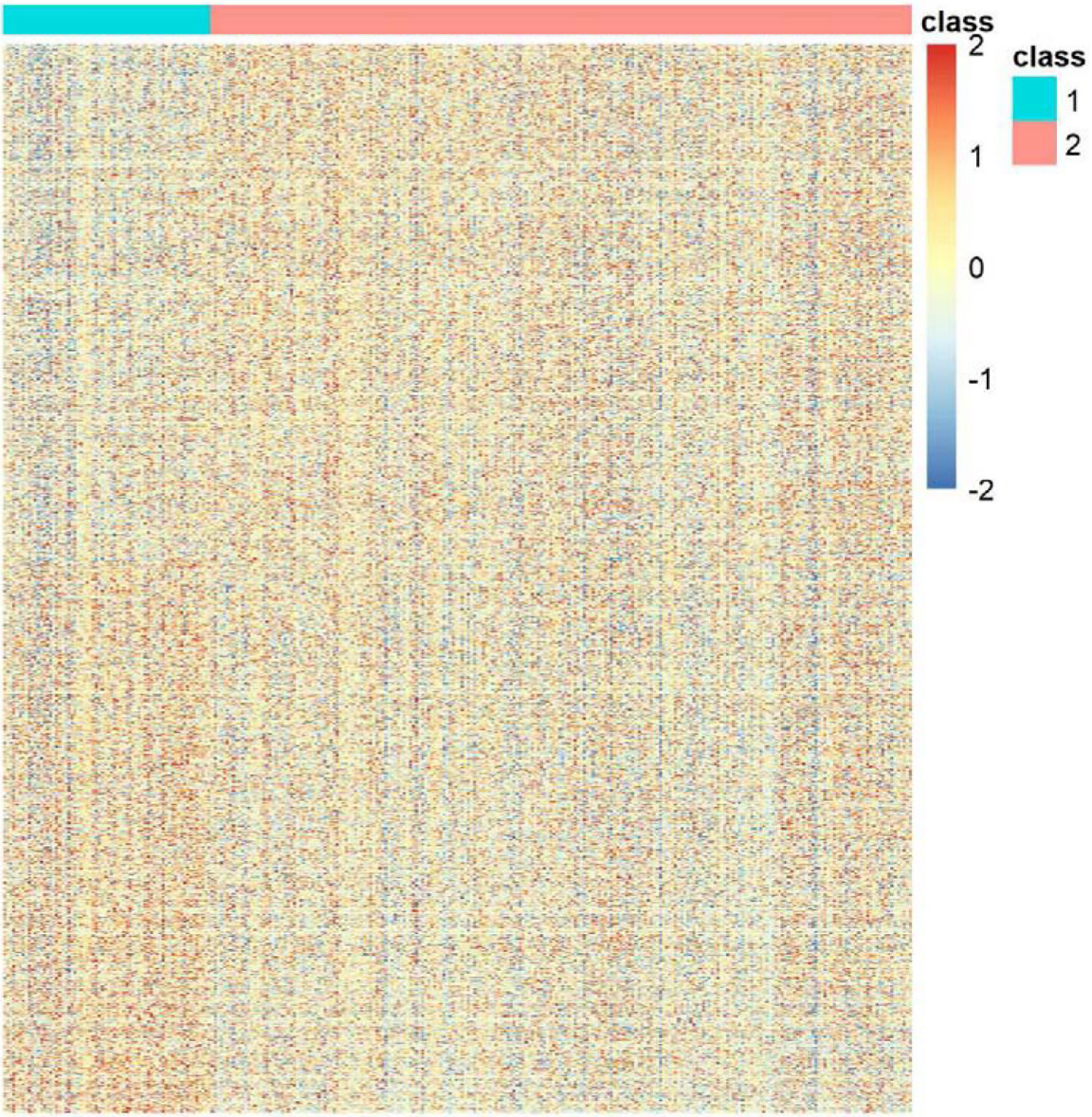

Figure 3.

A heatmap indicating the variable selection results for the real data, which shows 2386 selected genes (rows). Green and pink colors in the color bar above the heatmap indicate responders and non-responders, respectively.

Table 1.

Variable selection and classification performance comparison of BSLDA (SLDA using prior-knowledge-guided covariance matrix) and competing penalized LDA methods, including SLDA (the version without prior-knowledge-guided covariance matrix), sparse discriminant analysis (SDA), penalized linear discriminant analysis (PLDA), and diagonal linear discriminant analysis (DLDA). True positive rate (TPR) and false positive rate (FPR) are reported for the evaluation of variable selection performance, and the test data prediction accuracy is reported for the evaluation of classification performance. For each criterion, average and SD (within parenthesis) calculated over 100 simulated datasets are reported.

Method	Setting #1			Setting #2		
	TPR	FPR	Accuracy	TPR	FPR	Accuracy
BSLDA	0.91 (0.14)	0.11 (0.12)	0.96 (0.02)	0.78 (0.15)	0.13 (0.14)	0.95 (0.02)
SLDA	0.90 (0.10)	0.18 (0.19)	0.95 (0.02)	0.74 (0.15)	0.17 (0.18)	0.94 (0.02)
SDA	0.94 (0.10)	0.29 (0.22)	0.96 (0.02)	0.51 (0.08)	0.38 (0.24)	0.95 (0.02)
PLDA	1.00 (0.00)	0.32 (0.37)	0.96 (0.02)	0.87 (0.12)	0.39 (0.40)	0.95 (0.02)
DLDA	-	-	0.96 (0.02)	-	-	0.95 (0.02)
Method	**Setting #3**			**Setting #4**		
	TPR	**FPR**	**Accuracy**	**TPR**	**FPR**	**Accuracy**
BSLDA	0.81 (0.08)	0.15 (0.10)	0.87 (0.04)	0.82 (0.08)	0.23 (0.16)	0.84 (0.05)
SLDA	0.98 (0.04)	0.37 (0.14)	0.99 (0.02)	0.87 (0.09)	0.45 (0.15)	0.99 (0.01)
SDA	0.79 (0.15)	0.15 (0.06)	0.99 (0.01)	0.98 (0.07)	0.57 (0.07)	0.85 (0.06)
PLDA	1.00 (0.00)	0.32 (0.41)	0.86 (0.05)	0.85 (0.11)	0.27 (0.40)	0.84 (0.05)
DLDA	-	-	0.83 (0.08)	-	-	0.80 (0.07)

Table 2.

Variable selection and classification performance comparison of BSLDA (SLDA using prior-knowledge-guided covariance matrix) and competing penalized LDA methods, including SLDA (without prior-knowledge-guided covariance matrix), sparse discriminant analysis (SDA), and penalized linear discriminant analysis (PLDA), for different subsampling rates of observations. True positive rate (TPR) and false positive rate (FPR) are reported for the evaluation of variable selection performance, and the test data prediction accuracy is reported for the evaluation of classification performance. For each criterion, average and SD (within parenthesis) calculated over 100 simulated datasets are reported.

Method	Subsampling rate = 90%			Subsampling rate = 80%		
	TPR	FPR	Accuracy	TPR	FPR	Accuracy
BSLDA	0.85 (0.08)	0.18 (0.05)	0.98 (0.02)	0.80 (0.12)	0.19 (0.08)	0.98 (0.02)
SLDA	0.81 (0.11)	0.31 (0.19)	0.99 (0.02)	0.74 (0.14)	0.29 (0.19)	0.99 (0.02)
SDA	0.63 (0.13)	0.17 (0.08)	1.00 (0.00)	0.63 (0.16)	0.18 (0.13)	1.00 (0.00)
PLDA	0.74 (0.34)	0.55 (0.45)	0.93 (0.10)	0.80 (0.32)	0.67 (0.41)	0.95 (0.11)
Method	**Subsampling rate = 70%**			**Subsampling rate = 60%**		
	TPR	**FPR**	**Accuracy**	**TPR**	**FPR**	**Accuracy**
BSLDA	0.74 (0.14)	0.19 (0.08)	0.98 (0.02)	0.71 (0.14)	0.19 (0.10)	0.98 (0.02)
SLDA	0.68 (0.16)	0.28 (0.20)	0.99 (0.03)	0.70 (0.19)	0.37 (0.24)	0.99 (0.04)
SDA	0.58 (0.17)	0.19 (0.15)	1.00 (0.00)	0.58 (0.20)	0.23 (0.18)	1.00 (0.00)
PLDA	0.76 (0.34)	0.64 (0.44)	0.95 (0.10)	0.81 (0.33)	0.73 (0.39)	0.95 (0.13)

Table 3.

Variable selection and classification performances of BSLDA (SLDA using prior-knowledge-guided block covariance) with a misspecified block covariance matrix, where the rates indicate proportions of correctly specified signal variables. True positive rate (TPR) and false positive rate (FPR) are reported for the evaluation of variable selection performance, and the test data prediction accuracy is reported for the evaluation of classification performance. For each criterion, average and SD (within parenthesis) calculated over 50 simulated datasets are reported.

Rate (%)	TPR	FPR	Accuracy	Rate (%)	TPR	FPR	Accuracy
10	0.80 (0.09)	0.20 (0.08)	0.98 (0.03)	20	0.81 (0.11)	0.24 (0.15)	0.98 (0.03)
30	0.80 (0.09)	0.18 (0.08)	0.97 (0.03)	40	0.80 (0.09)	0.17 (0.09)	0.96 (0.03)
50	0.82 (0.08)	0.18 (0.07)	0.97 (0.02)	60	0.85 (0.09)	0.20 (0.08)	0.97 (0.03)
70	0.83 (0.09)	0.17 (0.08)	0.97 (0.03)	80	0.84 (0.09)	0.16 (0.08)	0.97 (0.02)
90	0.86 (0.07)	0.17 (0.05)	0.98 (0.02)	100	0.90	0.20	0.99