Statistics and Its Interface Volume 12 (2019) 149–166

Network-incorporated integrative sparse linear discriminant analysis

Xiaoyan Wang, Kuangnan Fang, Qingzhao Zhang, and Shuangge Ma[*]

Linear discriminant analysis (LDA) has been extensively applied in classification. For high-dimensional data, results generated from a single dataset may be unsatisfactory because of the small sample size. Under the regression framework, integrative analysis, which pools and analyses raw data from multiple datasets, has presented superior performance than single dataset analysis and meta-analysis. In this study, we conduct integrative analysis for LDA (iLDA). A network structure for variables is constructed to accommodate their interconnections, which have not been considered in many of the existing classification studies. We adopt the 1-norm group MCP method for simultaneous estimation and discriminative variable selection, and a Laplacian penalty to incorporate the network. The proposed method has intuitive formulations and can be computed using an effective coordinate descent algorithm. Simulation study shows that iLDA outperforms benchmarks with more accurate variable identification and classification. Analysis of three breast cancer datasets demonstrate that iLDA can improve prediction performance.

Keywords and phrases: Integrative analysis, Discriminant analysis, Network.

1. INTRODUCTION

Binary classification is commonly encountered in the field of biology and economics, such as the diagnosis of cancer, and recognition of credit default. Among the existing classification methods, linear discriminant analysis (LDA) is one of the most popular (Guo et al., 2006). In practice, high-dimensional data with the “large p, small n” characteristic appear more and more frequently. For example, a breast cancer study as analyzed in Section 4 may profile the expressions of $\sim 10^5$ genes on only tens of subjects. To analyze such data, a series of studies have proposed ways to modify LDA, including for example the nearest shrunken centroids classifier (NSC) (Tibshirani et al., 2002), l_1 penalized linear discriminant (l_1PLD) (Witten and Tibshirani, 2011), and direct sparse discriminant analysis (DSDA) (Mai and Zou, 2012). The goal of these methods is to identify variables that contribute the most to classification (referred to as discriminative variables) as well as to accommodate the high-dimensional characteristic. Despite many promising successes, it is still often observed that results generated from analyzing a single dataset are unsatisfactory (Guerra and Goldsterin, 2009). For instance, there may be a lack of reproducibility in variable identification. Many factors may contribute to the unsatisfactory performance, and in regression models, an important or perhaps the most important one is the small sample size of a single dataset. Existing studies suggest that pooling and analysing data from multiple datasets may effectively increase sample size and hence improve performance (Guerra and Goldsterin, 2009; Ma et al., 2011(b); Liu et al., 2014). As a family of multi-dataset analysis methods, integrative analysis pools and analyses raw data from multiple studies and outperforms single-dataset analysis and many other multi-dataset analysis approaches especially including the classic meta-analysis (Ma et al., 2011(a); Huang et al., 2012(a); Liu et al., 2013(a) and references therein). In classification, as our empirical study in Section 4 shows, there are indeed multiple independent studies sharing comparable designs. These studies publish their raw data at data warehouses such as GEO. This makes it feasible to pool multiple datasets and borrow information across them to conduct integrative analysis. With the importance of LDA, successes of integrative analysis, and convenience of data acquisition, it is naturally desirable to investigate integrative analysis for LDA (iLDA).

Variables can be interconnected. For example, genes belonging to the same pathways tend to have similar biological functions and correlated expressions. This can be partly observed from Figure 1 in our empirical study where genes present moderate to high correlations. Taking the interconnection into consideration when modeling can not only interpret the relationships among variables, but also lead to better variable selection and prediction performance (Huang et al., 2011; Liu et at., 2013(a)). There are multiple ways of describing the interconnections among variables, and in regression models, network has become a common choice (Huang et al., 2011; Liu et al., 2013(a); Shi et al., 2013). However, network-based analysis has not been well pursued in LDA.

This study develops the iLDA method which can incorporate the network structure to accommodate the interconnections among variables. It is related to but differs from

[*]Corresponding Author.

the previous ones in the following aspects. First, it belongs to the same integrative analysis paradigm as some previous studies (Liu et al., 2014; Zhao et al., 2015), with the aim of improving estimation and variable selection performance by pooling raw data from multiple studies. However, significantly different from the previous integrative analysis studies, the current study is the first to investigate integrative analysis under LDA, which is a very important and commonly applied classification method. Second, the aim of this study is not to develop a new LDA method. Rather, it is to significantly extend the traditional LDA to the integrative analysis paradigm. Third, this study is the first to incorporate a network structure for LDA under the integrative analysis paradigm to accommodate the interconnections among variables.

The rest of this study is organized as follows. In Section 2, the model framework and computation are described. In Section 3, simulation is conducted under various homogeneity and heterogeneity models. In Section 4, the analysis of three breast cancer datasets is conducted. In Section 5, brief discussions are provided.

2. INTEGRATIVE ANALYSIS FOR LDA

In this study, we conduct iLDA with M independent datasets with the goal of improving the performance of discriminative variable selection and estimation. It is noted that some similarity across these datasets is the basis of integrative analysis (in our study, similarity in the sets of discriminative variables), otherwise information borrowing across datasets may be unlikely. Hence, a proper selection of datasets is important. Some selection techniques such as analyzing meta-data have been discussed in the literature (Guerra and Goldstein, 2009; Tseng et al., 2015), and we do not reiterate here. It is also noted that, in our simulation study, under an extreme heterogeneity scenario where different datasets have completely different sets of discriminative variables, integrative analysis still presents satisfactory performance. For the simplicity of notation, we assume that the same set of variables is observed in all datasets. Missingness may occur in real data analysis, and different datasets may have overlapping but different variable sets. Following the studies in Tseng et al. (2015), our iLDA method can easily accommodate these scenarios.

2.1 Data and model settings

We use the superscript "m" to denote the mth dataset. In dataset $m\,(m = 1, \cdots, M)$, there are n^m iid observations, and the total sample size is $n = \sum_{m=1}^{M} n^m$. Let $Y^m = 1, 2$ be the class label, X^m be the length-p vector of variables, and n_1^m and n_2^m be the sample size of class 1 and 2, respectively. Under LDA, we denote the prior probabilities as $\mathrm{pr}(Y^m = 1) = \pi_1^m$, $\mathrm{pr}(Y^m = 2) = \pi_2^m$, and assume that a data point $x^m \in R^p$ has the conditional distribution $x^m | Y^m = g \sim N(\mu_g^m, \Sigma^m)$. Then, the Bayes rule for the mth dataset classifies x^m to class 2 if and only if

$$\begin{aligned}(1)\qquad &\left\{x^m - \left(\mu_1^m + \mu_2^m\right)/2\right\}^{\mathrm{T}} \left(\Sigma^m\right)^{-1} \left(\mu_2^m - \mu_1^m\right) \\ &\quad + \log\left(\pi_2^m / \pi_1^m\right) > 0.\end{aligned}$$

If the class labels $Y^m = 1, 2$ are coded as $-n^m/n_1^m$ and n^m/n_2^m, then the above classical LDA can be reconstructed via least squares (see Chapter 4 of Hastie et al., 2009 for detailed discussions). The following regression problem can be formulated:

(2)

$$\left(\hat{\boldsymbol{\beta}}^m_{ols}, \hat{\beta}^m_{0,ols}\right) = \arg\min_{\boldsymbol{\beta}^m, \beta_0^m} \sum_{i=1}^{n^m} \left(y_i^m - \beta_0^m - (x_i^m)^{\mathrm{T}} \boldsymbol{\beta}^m\right)^2,$$

where $x_i^m \in R^p$ is the ith ($i = 1, \cdots, n^m$) observation, $y_i^m \in \{-n^m/n_1^m, n^m/n_2^m\}$ is the ith relabelled response, $\boldsymbol{\beta}^m = (\beta_1^m, \cdots, \beta_p^m)^{\mathrm{T}}$ is the length-p vector of regression coefficients, and β_0^m is the intercept. Based on the least squares estimation, we can deduce that $\hat{\boldsymbol{\beta}}^m_{ols} = c(\hat{\sum}^m)^{-1}(\hat{\mu}_2^m - \hat{\mu}_1^m)$ for a positive constant c. In other words, the usual LDA direction $(\sum^m)^{-1}(\hat{\mu}_2^m - \hat{\mu}_1^m)$ of (1) can be exactly derived by the least squares estimation from (2).

Denote $\boldsymbol{\beta}_0 = \left(\beta_0^1, \cdots, \beta_0^M\right)^{\mathrm{T}}$ and $\boldsymbol{\beta} = \left(\boldsymbol{\beta}^{1\mathrm{T}}, \cdots, \boldsymbol{\beta}^{M\mathrm{T}}\right)^{\mathrm{T}}$. To conduct integrative analysis, we consider the overall loss function $L(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \sum_{m=1}^{M} \left(\sum_{i=1}^{n^m} \left(y_i^m - \beta_0^m - (x_i^m)^{\mathrm{T}} \boldsymbol{\beta}^m\right)^2\right)$. With this loss function, larger datasets have more contributions, which is intuitively reasonable.

2.2 Penalized variable selection

With high-dimensional data, the connection between LDA and linear regression may be lost, and the LDA direction is not well defined (Mai and Zou, 2012). In addition, among all of the candidate variables, some may be noises, and only a small subset is expected to contribute to classification. Thus, variable selection is needed. For this purpose, we consider penalization, which has been the choice of several related studies (Wu et al., 2009; Mai and Zou, 2012). Consider the penalized least squares problem

$$(3)\qquad \min_{\boldsymbol{\beta}_0, \boldsymbol{\beta}} \left\{L(\boldsymbol{\beta}_0, \boldsymbol{\beta})/2n + P\left(\boldsymbol{\beta}; \lambda_1\right)\right\}.$$

For penalty function $P(\cdot)$, we adopt the 1-norm group minimum concave penalty (gMCP, Huang et al., 2012(a)), which has been discussed in Liu et al., (2014) under an integrative regression analysis. Specifically,

$$(4)\qquad P\left(\boldsymbol{\beta}; \lambda_1\right) = \sum_{j=1}^{p} \rho\left(||\boldsymbol{\beta}_j||_1; \lambda_1, \gamma\right),$$

where $\rho(\cdot)$ is MCP (Zhang, 2010) with the form of $\rho\left(t; \lambda, \gamma\right) = \lambda \int_0^{|t|} \left(1 - \frac{x}{\lambda\gamma}\right)_+ dx$, λ_1 is the data-dependent tuning parameter, γ is the positive regularization parameter controlling the concavity of $\rho(\cdot)$, $\boldsymbol{\beta}_j$ is the effect vector of the jth variable across M datasets, and $||\boldsymbol{\beta}_j||_1 = \sum_{m=1}^{M} |\beta_j^{(m)}|$.

The sparsity structure of $\boldsymbol{\beta}$ can be described using the homogeneity and heterogeneity models. Under the homogeneity model, $I\left(\beta_j^1 = 0\right) = \cdots = I\left(\beta_j^M = 0\right)$ for any $j\,(= 1, \cdots, p)$. That is, all M datasets have the same sparsity structure. However, this model may be too restricted because heterogeneity inevitably exists across datasets (Knudsen, 2006). Under the heterogeneity model, a variable is allowed to be associated with the responses in some datasets, but not others. That is, for a variable X_j, $j = 1, \cdots, p$, there may exist two datasets (denoted as m_1 and m_2) in which $I\left(\beta_j^{m_1} = 0\right) \neq I\left(\beta_j^{m_2} = 0\right)$. Under this model, we first need to determine whether a variable contributes to classification in any dataset at all. In addition, for a selected variable, we need to identify the datasets in which it contributes to classification. That is, a bi-level of selection is needed. Motivated by these considerations, we adopt the above 1-norm gMCP which is a composite of the outer MCP and inner Lasso. The overall penalty is the sum of p individual penalties, with one for each variable. For a specific variable, the first level of selection is achieved using the outer MCP. In addition, for a selected variable, the second level of selection is accomplished with the inner Lasso penalty.

In (4), the Lasso penalty is adopted mainly because of its computational simplicity, and MCP is employed owing to its satisfactory performance (for example, more accurate selection) in single-dataset analysis. Note that the choice of penalty can vary. Some alternative composite and sparse group penalties, such as the composite MCP (Liu et al., 2014) and 1-norm group Bridge (Shi et al., 2013), are expected to present competitive results.

2.3 Accommodating the network structure

In practical data analysis, variables are complexly interconnected. This can be partly seen from our analyzed datasets where some genes present high correlations. However, the existing LDA studies have not sufficiently considered the interconnections. In the framework of regression, a series of studies have shown that taking the interconnections into consideration in integrative analysis can improve selection and estimation (Liu et al., 2013(a); Shi et al., 2013). Therefore, it is natural to incorporate the interconnections among variables in iLDA.

We adopt a network structure to describe the interconnections among variables. To construct a network, first we define a similarity measure between two variables. For continuous variables, we adopt the absolute value of the Pearson correlation coefficient. For ordinal variables, the Spearman's correlation may be adopted. Here denote $\hat{r}_{jk}$ as the Pearson's correlation coefficient between variables j and k computed using the M datasets. Second, the adjacency matrix $A = [a_{jk}]_{p\times p}$ can be computed. Specifically, we consider a dense and a sparse adjacency: (N.1) $a_{jk} = |\hat{r}_{jk}|^{\alpha}$, where $\alpha > 0$ can be determined by the scale-free topology criterion (Zhang and Horvath, 2005); (N.2) $a_{jk} = |\hat{r}_{jk}|\mathrm{I}\left(|\hat{r}_{jk}| > r\right)$, where r is the cut-off calculated from the Fisher transformation (Huang et al., 2011). N.2 may be sparse in that some entries are zero. It is noted that many other adjacencies exist, we refer to Liu et al. (2013(a)) and Huang et al. (2011) for more discussions.

To incorporate the network structure, we add a Laplacian penalty to (3). Consider the iLDA estimate

(5)
$$\left(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}\right) = \arg\min_{\boldsymbol{\beta}_0, \boldsymbol{\beta}} \left\{ L(\boldsymbol{\beta}_0, \boldsymbol{\beta})/2n + P\left(\boldsymbol{\beta}; \lambda_1\right) + P_N\left(\boldsymbol{\beta}; \lambda_2\right) \right\},$$

where

$$P_N\left(\boldsymbol{\beta}; \lambda_2\right) = \frac{1}{2}\lambda_2 \sum_{1\leq j<k\leq p} a_{jk}\left(||\boldsymbol{\beta}_j||_1 - ||\boldsymbol{\beta}_k||_1\right)^2, \tag{6}$$

$\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}^{1\mathrm{T}}, \cdots, \hat{\boldsymbol{\beta}}^{M\mathrm{T}}\right)^{\mathrm{T}}$, and $\hat{\boldsymbol{\beta}}_0 = \left(\hat{\beta}_0^1, \cdots, \hat{\beta}_0^M\right)^{\mathrm{T}}$. Denote $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^{\mathrm{T}} = \left(||\boldsymbol{\beta}_1||_1, \cdots, ||\boldsymbol{\beta}_p||_1\right)^{\mathrm{T}}$, we can express the non-negative quadratic form in $P_N\left(\boldsymbol{\beta}; \lambda_2\right)$ using a positive semi-definite matrix $\boldsymbol{L}$, which satisfies

$$\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{\theta} = \sum_{1\leq j<k\leq p} a_{jk}\left(\theta_j - \theta_k\right)^2, \forall \boldsymbol{\theta} \in R^p. \tag{7}$$

Let $\boldsymbol{G} = \mathrm{diag}\left(g_1, \cdots, g_p\right)$ with $g_j = \sum_{k=1}^{p} a_{jk}$. Then, $\boldsymbol{L} = \boldsymbol{G} - \boldsymbol{A}$. This matrix is associated with a labelled weighted graph $\mathcal{G} = (V, \mathcal{E}, \omega)$ with the vertex set $V = (1, \cdots, p)$ and edge set $\mathcal{E} = \{(j,k) : (j,k) \in V \times V\}$. Here a_{jk} is the weight of edge (j,k), g_j is the degree of vertex j, and $\boldsymbol{L}$ is called the Laplacian of $\mathcal{G}$ and $\boldsymbol{A}$ (Chung, 1996). For tightly connected nodes with a large a_{jk}, the Laplacian penalty encourages their groups of coefficients to be similar. Here it is noted that $\boldsymbol{L}$ is not normalized, meaning that g_i is not standardized to 1. In problems where variables should be treated without preference with respect to connectivity, we normalize the Laplacian such that $\boldsymbol{L}^* = \mathbf{I}_p - \boldsymbol{G}^{-1/2}\boldsymbol{A}\boldsymbol{G}^{-1/2}$, and then use the penalty $P_N\left(\boldsymbol{\beta}; \lambda_2\right) = \frac{1}{2}\lambda_2\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{L}^*\boldsymbol{\theta}$. In this study, we consider formulation (6). This is motivated by existing studies (Huang et at., 2011; Liu et al., 2013(a)) which suggest that it is prudent to provide more protection for variables with higher connectivity.

2.4 Computation

The classification rule for the mth dataset is to assign an observation x to class 2 if

$$x^{\mathrm{T}}\hat{\boldsymbol{\beta}}^m + \hat{\beta}_{0\mathrm{opt}}^m > 0. \tag{8}$$

Note that $\hat{\beta}_{0\mathrm{opt}}^m$ in (8) is different from $\hat{\beta}_0^m$ in (5). Consider the ordinary least squares (OLS) estimate and the usual LDA. Write $\hat{\boldsymbol{\beta}}_{ols}^m = c\hat{\boldsymbol{\beta}}_{LDA}^m$for some constant c, where $\hat{\boldsymbol{\beta}}_{LDA}^m = \left(\sum^m\right)^{-1}(\hat{\mu}_2^m - \hat{\mu}_1^m)$. Therefore we should use $\hat{\beta}_{0\mathrm{opt}}^m = c\hat{\beta}_{0LDA}^m$ in (8), where $\hat{\beta}_{0LDA}^m = \log\left(n_2^m/n_1^m\right) -$

$\{(\hat{\mu}_1^m + \hat{\mu}_2^m)/2\}^T \hat{\boldsymbol{\beta}}_{LDA}^m$, such that the OLS classification and LDA rule yield identical classification.

According to Proposition 2 in Mai et al. (2012), under the classification rule (8), if $(\mu_2^m - \mu_1^m)^{\mathrm{T}} \hat{\boldsymbol{\beta}}^m > 0$, the optimal estimate $\hat{\beta}_{0\mathrm{opt}}^m$ is

(9)
$$\hat{\beta}_{0\mathrm{opt}}^m = -\left(\hat{\mu}_1^m + \hat{\mu}_2^m\right)^{\mathrm{T}} \hat{\boldsymbol{\beta}}^m/2 + \hat{\boldsymbol{\beta}}^{m\mathrm{T}} \hat{\Sigma}^m \hat{\boldsymbol{\beta}}^m \left\{\left(\hat{\mu}_2^m - \hat{\mu}_1^m\right)^{\mathrm{T}} \hat{\boldsymbol{\beta}}^m\right\}^{-1} \log\left(n_2^m/n_1^m\right).$$

If $n_1^m = n_2^m$, then we can take $\hat{\beta}_{0\mathrm{opt}}^m = -(\hat{\mu}_1^m + \hat{\mu}_2^m)^{\mathrm{T}} \hat{\boldsymbol{\beta}}^m/2$. If the linear classifier actually yields $(\mu_2^m - \mu_1^m)^{\mathrm{T}} \hat{\boldsymbol{\beta}}^m < 0$, then we can always use $\hat{\boldsymbol{\beta}}_{\mathrm{new}}^m = -\hat{\boldsymbol{\beta}}^m$, which obeys $(\mu_2^m - \mu_1^m)^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\mathrm{new}}^m > 0$.

When solving the minimization problem (5), the 1-norm gMCP does not have a convenient form for updating individual parameters. Thus we adopt the local linear approximation (LLA) technique. By taking the first-order Taylor expansion at the current estimate $\hat{\boldsymbol{\beta}}_j$, 1-norm gMCP as a function of $|\beta_j^m|$ is approximately proportional to

$$\widetilde{\lambda}_j = \rho'\left(||\boldsymbol{\beta}_j||_1; \lambda_1, \gamma\right) = \lambda_1 \left(1 - \frac{||\boldsymbol{\beta}_j||_1}{\lambda_1 \gamma}\right)_+. \tag{10}$$

Denote $\boldsymbol{y} = \left(\boldsymbol{y}^{1\mathrm{T}}, \cdots, \boldsymbol{y}^{M\mathrm{T}}\right)^{\mathrm{T}}$, $\boldsymbol{y}^m = (y_1^m, \cdots, y_{n^m}^m)^{\mathrm{T}}$ $(m = 1, \cdots, M)$, $\boldsymbol{X} = diag(\boldsymbol{X}^1, \cdots, \boldsymbol{X}^M)$, $\boldsymbol{X}_j$ as the component of $\boldsymbol{X}$ that corresponds to $\boldsymbol{\beta}_j$, X_j^m as the mth column of $\boldsymbol{X}_j$, and $\hat{\boldsymbol{\beta}}_j^{-m}$ as the current estimate $\hat{\boldsymbol{\beta}}_j$ with its mth element $\hat{\beta}_j^m = 0$. For the estimation of β_j^m, only the terms involving it in the objective function (5) matter, and take the form

(11)
$$R\left(\beta_j^m\right) = \frac{1}{2n}||\boldsymbol{y} - \sum_{j=1}^p \boldsymbol{X}_j \boldsymbol{\beta}_j||_2^2 + \frac{K_{1j}}{2} \cdot \left(\beta_j^m\right)^2 + K_{2j}^m |\beta_j^m|,$$

where

$$K_{1j} = \lambda_2 \sum_{k \neq j} a_{jk},$$

$$K_{2j}^m = \widetilde{\lambda}_j + \lambda_2 \left\{\sum_{k \neq j} a_{jk} \left[||\hat{\boldsymbol{\beta}}_j^{-m}||_1 - ||\hat{\boldsymbol{\beta}}_k||_1\right]\right\}.$$

With the above LLA, we adopt the coordinate descent algorithm for parameter estimation. This algorithm is iterative and optimizes over one parameter at a time. For the update with each β_j^m, we have an univariate Lasso penalized estimation

$$\hat{\beta}_j^m = \frac{1}{\frac{n^m}{n} + K_{1j}} S\left(\frac{1}{n} X_j^{m\mathrm{T}} \boldsymbol{r} + \frac{n^m}{n} \hat{\beta}_j^m, K_{2j}^m\right), \tag{12}$$

where $\boldsymbol{r} = \boldsymbol{y} - \sum_{j=1}^p \boldsymbol{X}_j \hat{\boldsymbol{\beta}}_j$, and $S(z, c) = \mathrm{sign}(z)(|z| - c)_+$ is the soft-thresholding operator.

With fixed γ, $\lambda_1$and λ_2, the coordinate descent algorithm proceeds as follows:

```
1. Let the superscript $s$ be the number of iterations. Set
   $s = 0$, $\hat{\boldsymbol{\beta}}^{(0)} = \left(\hat{\boldsymbol{\beta}}_1^{(0)\mathrm{T}}, \cdots, \hat{\boldsymbol{\beta}}_p^{(0)\mathrm{T}}\right)^{\mathrm{T}}$, and $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)}$.
2. For $j = 1, \cdots, p$, $m = 1, \cdots, M$,
   (a) Calculate $K_{1j}$ and $K_{2j}^m$;
   (b) Update $\hat{\beta}_j^{m,(s+1)}$ using expression (12);
   (c) Update $\boldsymbol{r} \leftarrow \boldsymbol{r} - X_j^{m\mathrm{T}}\left(\hat{\beta}_j^{m,(s+1)} - \hat{\beta}_j^{m,(s)}\right)$.
3. Update $s \leftarrow s + 1$.
4. Repeat Step 2 and 3 until convergence.
```

Convergence of this algorithm follows from Tseng (2001). It is achieved in all simulations and real data analysis. The objective function can be expressed as $f(\boldsymbol{\beta}) = f_0(\boldsymbol{\beta}) + P(\boldsymbol{\beta}; \lambda_1)$, where $f_0(\boldsymbol{\beta}) = L(\boldsymbol{\beta}_0, \boldsymbol{\beta})/2n + P_N(\boldsymbol{\beta}; \lambda_2)$ is regular and continuously differentiable in the sense of Tseng (2001) and $P(\boldsymbol{\beta}; \lambda_1) = \sum_{j=1}^p \rho\left(||\boldsymbol{\beta}_j||_1; \lambda_1, \gamma\right)$ is separable. As such, results in Tseng (2001) are directly applicable.

iLDA involves three tuning parameters: γ, λ_1, and λ_2. For γ in MCP, the existing studies (Zhang, 2010; Liu et at., 2013(a); Liu et al., 2014) suggest examining a few different values or fixing its value. In this study, we set $\gamma = 6$ as suggested in published studies. For λ_1 and λ_2, we apply a two-dimensional search and select the optimal value using cross validation. R code is available at https://github.com/shuanggema/. The algorithm is computationally affordable. For example, in a simulation study with $p = 300$ and $M = 4$, one simulation replicate takes no more than 100 seconds using a regular laptop.

3. SIMULATION

Simulation is conducted to gauge performance of the proposed approach and compare with alternatives. Although the proposed approach is designed for the heterogeneity model, we also consider the homogeneity model (a special case of heterogeneity model). Under the homogeneity model (Simulation 1), we consider two scenarios: Homo.A is a general scenario, where each dataset has different coefficients; Homo.B is a simplified case, where three datasets have the same coefficients. In Simulation 2, we consider the heterogeneity model where the datasets have overlapping but different discriminative variables. Here three scenarios are considered. Under scenario Hetero.A, four datasets share seven common discriminative variables. In addition, datasets 3 and 4 have three more common discriminative variables. Under scenario Hetero.B, three datasets share four common discriminative variables, and have four, six, and eight dataset-specific ones, respectively. Under scenario Hetero.C, each dataset has ten unique discriminative variables. These three simulated scenarios cover the whole spectrum of high, low, to no overlapping discriminative variables. Here the coefficients in all three scenarios are dominantly set as positive. In Simulation 3, we consider a heterogeneity model with both negative and positive coefficients randomly blended for all three datasets. Across these datasets, a moderate overlapping among discriminative variables is presented. Under the above models, we mainly focus on the differences/similarities among datasets in terms of sparsity structures. Under the sensitivity model (Simulation 4), we consider a scenario where there exists a high degree of difference among the coefficients of a specific variable. In Simulation 5, we consider five scenarios where the normality assumption on covariates is violated. Under scenario Nonnorm.A, there are two variables following non-normal distributions and both of them are with non-zero coefficients. Under scenario Nonnorm.B, among all the important variables, 50% are generated from non-normal distributions. Under scenario Nonnorm.C, 80% of the important variables are from non-normal distributions. Under scenarios Nonnorm.D and Nonnorm.E, all non-normal variables are with zero coefficients, and account for 50% and 80% of all variables, respectively. In Simulation 6, a scenario with sample sizes similar to those in the real data is analyzed.

For each subject, the class label is randomly generated from a Bernoulli distribution with probability 0.5. In Simulation 1–4 and 6, in each dataset, covariates are generated to have a multivariate normal distribution with covariance matrix $\Sigma = (\rho_{jk})_{p\times p}$. Here two correlation structures are considered. The first is the auto-regressive (AR) correlation, where $\rho_{jk} = \rho^{|j-k|}$ with ρ = 0.1, 0.5, and 0.8, corresponding to weak, moderate, and strong correlations, respectively. The second is the banded correlation (BC). Here two cases are considered: (BC(i)) $\rho_{jk} = 0.33$, if $|j-k| = 1$ and 0 otherwise; (BC(ii)) $\rho_{jk} = 0.6$, if $|j-k| = 1$, 0.33 if $|j-k| = 2$, and 0 otherwise. Further, based on the covariance matrix, we set the mean vectors $\mu_1^m = 0$ and $\mu_2^m = \Sigma\beta^m$ for the mth dataset. In Simulation 5, some variables follow non-normal distributions, including asymmetric ($\chi^2(5,a)$, $t(5,a)$, $t(30,a)$, and Lognormal$(a,1)$), symmetric (Uniform$(a-1,a+1)$), and multimodal (normal mixture $0.5N(0,1)+0.5N(a,1)$) distributions. Here, a is the non-centrality parameter for skewed χ^2- and $t-$distributions. To make these variables discriminative for classification, we set a differently in subgroups $y=1$ and $y=2$. For example, under scenario Nonnorm.A, X_5 is an important variable and set to follow a non-normal distribution. Its observations in the mth dataset can be generated from standard $t(5)$ and $t(5,\beta_5^m)$ for $y=1$ and $y=2$, respectively. The settings of sample size, variable dimension, and β for each model are as follows.

- Simulation 1:
  - Homo.A, $n^1 = n^2 = n^3 = 300$ and $p = 800$. There are 21 nonzero coefficients with the true values of $(\beta_1^1,\cdots,\beta_7^1) = (0.9,0.8,0.7,0.6,0.5,0.4,0.3)$, $(\beta_1^2,\cdots,\beta_7^2) = (0.2,0.35,0.5,0.65,0.8,0.95,1.1)$, $(\beta_1^3,\cdots,\beta_7^3) = (0.5,0.55,0.6,0.65,0.7,0.75,0.8)$.
  - Homo.B, $n^1 = \cdots = n^4 = 200$ and $p = 600$. All datasets have the same coefficients, where the nonzero elements for the mth dataset are $\beta_{1\sim 10}^m = (0.5,0.55,0.6,0.65,0.75,0.8,0.75,0.7,0.65,0.6)$.
- Simulation 2:
  - Hetero.A, $n^1 = n^3 = 100$, $n^2 = n^4 = 200$ and $p = 300$. There are a total of 34 nonzero coefficients across the four datasets:
$$\begin{aligned}(\beta_1^1,\cdots,\beta_7^1) &= (0.6,0.65,0.5,0.55,0.6,0.4,0.3),\\ (\beta_1^2,\cdots,\beta_7^2) &= (0.65,0.7,0.75,0.85,0.5,0.75,0.7),\\ (\beta_1^3,\cdots,\beta_{10}^3) =&(0.3,0.35,0.4,0.45,0.75,\\ &0.8,0.75,0.7,0.65,0.6),\\ (\beta_1^4,\cdots,\beta_{10}^4) =&(1,0.85,0.8,0.75,0.7,\\ &0.65,0.6,0.5,0.7,0.75).\end{aligned}$$
  - Hetero.B, $n^1 = n^2 = n^3 = 100$ and $p = 600$. There are a total of 30 true positives across all three datasets. The nonzero coefficients are randomly generated from Uniform$(0.4,1.0)$.
  - Hetero.C, $n^1 = n^2 = n^3 = 200$ and $p = 600$. There are 30 nonzero coefficients across all three datasets, specifically
$$\begin{aligned}(\beta_1^1,\cdots,\beta_{10}^1) =&(0.3,0.333,0.367,0.4,0.433,0.467,\\ &0.5,0.533,0.567,0.6),\\ (\beta_{11}^2,\cdots,\beta_{20}^2) =&(0.35,0.4,0.45,0.5,0.55,0.6,\\ &0.65,0.7,0.75,0.8),\\ (\beta_{21}^3,\cdots,\beta_{30}^3) =&(0.3,0.35,0.4,0.45,0.5,0.55,0.6,\\ &0.65,0.7,0.75).\end{aligned}$$
- Simulation 3: $n^1 = n^2 = n^3 = 200$ and $p = 600$. The nonzero coefficients are generated as follows: the absolute values of $\beta_{1\sim 8}^1$ are randomly generated from Uniform$(0.4,0.8)$, and their signs (positive or negative) are generated using a Bernoulli distribution with probability 0.5. Similar data generation mechanism is adopted for $\beta_{3\sim 10}^2$ and $\beta_{5\sim 12}^3$.
- Simulation 4: all three datasets have sample size 100 and $p = 400$. Connected variables have opposite contributions to classification in different datasets. The following nonzero coefficients are set according to a single dataset study [19].
$$\begin{aligned}(\beta_1^1,\cdots,\beta_{10}^1) =&0.556\times\\ &(3,1.5,0,0,2,1.8,1.78,1.74,1.72,1.68),\\ (\beta_1^2,\cdots,\beta_8^2) =&0.582\times\\ &(3,2.5,-2.8,1.6,1.5,1.4,1.3,-2),\end{aligned}$$
$(\beta_1^3,\cdots,\beta_4^3) = 0.556\times(0,0,2,2.2)$, and $(\beta_5^3,\cdots,\beta_{10}^3)$ are generated from Uniform$(0.556\times 1.5, 0.556\times 2.5)$. Here both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ have large differences among their values in different datasets.

Table 1. Simulation 1: the first row is the number of true positives (standard deviation), the second row is the number of the false positives (standard deviation), and the third row is the mean prediction error (standard deviation).

Correlation	Bayes	iLDA		NSC	l_1PLD	DSDA
		N.1	N.2			
			Homo.A			
AR(0.1)		21(0)	21(0.31)	20(0.89)	20(0.71)	18(1.09)
		3(3.85)	0(1.05)	57(53.67)	93.5(49.12)	5(7.85)
	17.44(0.01)	18.33(0.01)	17.72(0.02)	19.11(0.02)	17.89(0.02)	18.50(0.01)
AR(0.5)		21(0.57)	21(1.60)	20(1.09)	21(0.18)	18(1.08)
		1(1.68)	0(0.18)	22.5(131.72)	6(149.39)	0(2.04)
	9.33(0.01)	9.39(0.01)	9.44(0.01)	9.37(0.01)	8.90(0.01)	9.83(0.01)
AR(0.8)		20(1.07)	20(1.03)	19(1.57)	21(0)	17(1.56)
		2(1.20)	2(1.02)	4(7.85)	20.5(135.01)	0(1.58)
	3.67(0.01)	3.78(0.01)	3.72(0.01)	3.89(0.01)	4.11(0.01)	3.78(0.01)
BC(i)		21(0.18)	21(0.25)	20(0.76)	21(0.35)	18(1.11)
		0(1.00)	0(1.05)	57.5(73.25)	10(31.10)	0(3.08)
	14.33(0.01)	13.89(0.01)	13.72(0.01)	14.61(0.01)	13.94(0.01)	14.78(0.01)
BC(ii)		21(3.13)	21(0.48)	19.5(1.36)	21(0.25)	17(1.09)
		0(0.43)	0(0)	11(129.69)	4(114.07)	0(1.07)
	9.33(0.01)	8.56(0.01)	9.11(0.01)	9.28(0.01)	8.83(0.01)	9.50(0.01)
			Homo.B			
AR(0.1)		40(0.73)	40(0.73)	40(0.68)	40(0)	38(1.55)
		1(5.27)	0(0.63)	110.5(64.57)	356.5(138.75)	21.5(13.85)
	13.94(0.01)	14.13(0.01)	13.88(0.01)	15.94(0.01)	14.69(0.02)	15.63(0.02)
AR(0.5)		40(1.22)	40(1.38)	40(0.78)	40(0)	38(1.42)
		0(1.22)	0(4.74)	113(169.43)	84(101.64)	3(4.03)
	6.50(0.01)	6.31(0.01)	6.63(0.01)	6.38(0.01)	6.13(0.01)	6.63(0.01)
AR(0.8)		40(1.51)	40(1.70)	40(0.77)	40(0)	39(1.07)
		0(1.02)	0(1.22)	110(281.10)	15(70.68)	0(5.48)
	2.88(0.01)	2.88(0.01)	2.50(0.01)	2.81(0.01)	2.50(0.01)	2.75(0.01)
BC(i)		40(1.22)	40(1.63)	40(0.94)	40(0)	37(1.31)
		0(0.76)	0(0.25)	82(58.97)	82.5(75.44)	8(12.59)
	9.94(0.01)	10.06(0.01)	9.81(0.01)	10.56(0.01)	9.31(0.01)	10.81(0.01)
BC(ii)		40(1.99)	40(2.49)	39(1.41)	40(0.25)	36(1.69)
		0(4.36)	0(2.15)	86.5(206.80)	72(104.07)	2.5(4.35)
	5.25(0.01)	5.00(0.01)	4.69(0.01)	5.00(0.01)	4.69(0.01)	5.13(0.01)

- Simulation 5: we set the sample sizes and coefficients as Homo.A and Hetero.B. The variables with normal distribution are generated to have an AR(0.5) correlation structure.
- Simulation 6: $n^1 = 86$, $n^2 = 30$, $n^3 = 29$, and $p = 600$. There are 12, 8, and 10 important variables for the three datasets, respectively. The nonzero coefficients are randomly generated from Uniform (0.3,1).

In our simulation, we mainly focus on the accuracy of variable identification, which can be measured using the number of true positives and number of false positives. In addition, prediction performance is also of interest. For this purpose, for each simulation replicate, M independent testing datasets under the same settings as the training datasets are randomly generated. By adopting cross validation, we select tuning parameters using training data and conduct estimation, and then make prediction for subjects in the testing data and compute the mean prediction error. For comparison, we adopt three single-dataset sparse classification methods, including NSC [23], l_1PLD [27] and DSDA [19], and then combine analysis results across datasets (i.e., a meta-analysis strategy). Here existing R packages are used, including *pamr* for NSC, *penalizedLDA* for l_1PLD, and *glmnet* for DSDA. We also compute classification errors with the Bayes rule using the true discriminative variable set. This is the oracle benchmark. Summary statistics based on 100 replicates are shown in Table 1–6 and Table 9–12 (in Appendix).

The result in Table 1 shows that for the homogeneity models, iLDA can identify all of the true positives except for scenario Homo.A with AR(0.8). DSDA may have inferior performance in terms of true positives, especially it has trouble accommodating high correlations. When looking at false positives, compared to NSA and l_1PLD, iLDA presents overwhelming superiority with almost zero false positives. In

Table 2. Simulation 2: the first row is the number of true positives (standard deviation), the second row is the number of the false positives (standard deviation), and the third row is the mean prediction error (standard deviation).

Correlation	Bayes	iLDA		NSC	l_1PLD	DSDA
		N.1	N.2			
			Hetero.A			
AR(0.1)		34(1.26)	34(1.14)	31.5(2.38)	33.5(0.67)	29(2.24)
		31.5(9.78)	11(6.38)	89.5(79.84)	85.5(84.00)	20.5(11.70)
	17.88(0.02)	17.75(0.02)	18.56(0.02)	21.06(0.02)	18.75(0.02)	20.31(0.02)
AR(0.5)		34(0.75)	34(1.17)	33(1.95)	34(0.77)	31(1.22)
		16(6.66)	6(8.68)	79(70.31)	62(56.24)	8(11.18)
	10.75(0.01)	10.63(0.01)	10.63(0.01)	11.00(0.01)	10.13(0.01)	12.06(0.01)
AR(0.8)		34(0.48)	32.5(1.93)	32(2.03)	34(0.75)	32(1.03)
		6(7.14)	5.5(4.72)	41.5(119.63)	38(47.79)	3.5(3.30)
	5.88(0.01)	6.06(0.01)	5.69(0.01)	6.00(0.01)	5.56(0.01)	6.44(0.01)
BC(i)		32(1.75)	32(2.27)	33(1.47)	34(0.00)	27(1.98)
		4(4.20)	3.5(2.14)	64.5(48.67)	78.5(73.09)	7.5(10.42)
	14.19(0.01)	14.25(0.01)	13.69(0.02)	14.88(0.01)	13.69(0.01)	16.81(0.02)
BC(ii)		32(3.13)	32(2.38)	32(2.29)	34(0.25)	25(1.72)
		4(3.23)	4(4.23)	16(56.05)	40(101.15)	1(4.79)
	8.31(0.01)	8.12(0.01)	8.12(0.01)	8.44(0.01)	7.94(0.01)	9.25(0.01)
			Hetero.B			
AR(0.1)		26(1.56)	28(1.79)	29(1.05)	30(0.00)	26.5(1.50)
		6(3.45)	9.5(4.61)	64(74.89)	170.5(246.72)	9(10.08)
	15.67(0.01)	17.44(0.02)	17.06(0.01)	18.50(0.02)	16.78(0.02)	18.61(0.02)
AR(0.5)		27.5(1.68)	28(2.56)	29(0.92)	29(0.45)	25(1.80)
		7(2.76)	9(11.16)	110.5(139.83)	23.5(157.13)	2(2.67)
	9.17(0.01)	9.5(0.01)	8.67(0.01)	9.44(0.01)	9.00(0.01)	9.88(0.01)
AR(0.8)		27.5(1.30)	27(2.67)	24(3.31)	29(0.57)	19(1.85)
		8(1.98)	8.5(2.50)	29(181.84)	212(196.99)	1(18.73)
	7.11(0.01)	7.06(0.01)	6.83(0.01)	8.61(0.01)	7.94(0.01)	7.89(0.01)
BC(i)		25(3.12)	23(2.09)	28(1.68)	30(0.31)	21.5(2.40)
		4.3(3.27)	1.5(3.50)	55.5(101.96)	150.5(56.63)	6.5(12.79)
	10.78(0.02)	12.18(0.02)	12.28(0.01)	12.22(0.02)	11.22(0.02)	13.89(0.02)
BC(ii)		27((2.38)	27(3.47)	29(1.49)	30(0.00)	21(1.59)
		4.5(2.86)	3(2.48)	72(130.92)	19.5(219.89)	3(6.64)
	4.56(0.01)	4.02(0.01)	3.98(0.01)	4.22(0.01)	3.61(0.01)	4.83(0.01)
			Hetero.C			
AR(0.1)		26(2.03)	27(2.26)	27(1.80)	30(0.41)	22(2.71)
		44.5(19.46)	26.5(30.39)	68(130.09)	268.5(277.53)	11.5(18.11)
	19.44(0.01)	20.61(0.01)	19.94(0.02)	23.05(0.02)	22.17(0.03)	22.67(0.02)
AR(0.5)		29(1.17)	28(1.05)	29(1.21)	30(0.00)	23.5(1.70)
		2(1.05)	1(0.75)	33.5(103.44)	19(191.10)	0.5(2.88)
	10.22(0.01)	9.89(0.01)	9.67(0.01)	9.89(0.01)	9.83(0.01)	10.72(0.01)
AR(0.8)		30(0.45)	27(1.36)	26(3.64)	30(0.00)	20(1.82)
		7.5(2.06)	3.5(2.19)	8(207.87)	93.5(193.59)	0(1.86)
	3.33(0.01)	2.83(0.01)	3.11(0.01)	3.28(0.01)	3.28(0.01)	3.17(0.01)
BC(i)		29(1.14)	29(1.59)	29(1.38)	30(0.25)	24(1.92)
		13(15.06)	3(4.41)	81(129.13)	132.5(126.20)	5(8.89)
	15.78(0.01)	15.17(0.01)	15.33(0.01)	17.44(0.01)	15.83(0.02)	17.67(0.02)
BC(ii)		29(1.12)	29(0.79)	29(0.85)	30(0.00)	22(2.19)
		6.5(5.19)	5.5(3.73)	125.5(129.87)	16(201.28)	0(5.74)
	10.11(0.01)	9.22(0.01)	9.28(0.01)	9.78(0.01)	9.67(0.01)	10.94(0.01)

terms of classification error, iLDA is close to the Bayes rule. For the heterogeneity models, both Table 2 and 3 suggest that, iLDA can effectively identify the majority of true positives and has a small number of false positives. Under some scenarios (for example, Hetero.A under case BC(i)), NSC and l_1PLD, which can conduct the classification of high-dimensional single datasets, may identify a few more true positives, however, at the price of a large number of false

Table 3. Simulation 3: the first row is the number of true positives (standard deviation), the second row is the false positives (standard deviation), and the third row is prediction mean classification error (standard deviation).

Correlation	Bayes	iLDA		NSC	l_1PLD	DSDA
		N.1	N.2			
AR(0.1)		20.5(1.78)	21(1.65)	21(1.92)	24(0.38)	19(1.93)
		6.5(4.65)	7(2.64)	63(98.94)	347(138.61)	14(15.95)
	19.50(0.01)	20.22(0.02)	19.94(0.02)	27.39(0.03)	26.44(0.02)	23.22(0.02)
AR(0.5)		21(2.45)	22(1.51)	17(1.79)	20(1.87)	14(1.45)
		7.5(3.80)	9(3.91)	39.5(69.74)	172(104.76)	3(7.66)
	20.11(0.01)	20.83(0.01)	20.56(0.01)	26.16(0.02)	25.75(0.02)	26.22(0.02)
AR(0.8)		21.5(3.23)	20(1.99)	15(2.81)	22(1.79)	9(1.22)
		16(9.23)	12.5(4.73)	10(49.76)	166.5(119.12)	3.5(5.92)
	21.83(0.01)	21.88(0.01)	22.17(0.01)	27.89(0.02)	28.00(0.02)	26.12(0.02)
BC(i)		20(2.86)	22(2.52)	19(1.51)	22(1.67)	15(2.14)
		6.5(6.62)	10(3.32)	83(91.55)	223(124.22)	5(14.28)
	20.72(0.01)	22.39(0.01)	21.33(0.01)	24.89(0.02)	24.11(0.02)	24.33(0.01)
BC(ii)		20(2.89)	22(2.32)	13.5(2.45)	20(2.05)	9(2.09)
		8(9.36)	11.5(4.21)	62(91.56)	231(154.34)	5.5(10.98)
	20.57(0.01)	20.89(0.01)	20.61(0.01)	30.22(0.02)	29.44(0.02)	30.11(0.02)

Table 4. Simulation 4: the first row is the number of true positives (standard deviation), the second row is the number of the false positives (standard deviation), and the third row is the mean prediction error (standard deviation).

Correlation	Bayes	iLDA		NSC	l_1PLD	DSDA
		N.1	N.2			
AR(0.1)		24(1.93)	24(1.40)	24(0.48)	24(0)	21(0.71)
		11(3.29)	6(3.12)	45(36.84)	39.5(55.60)	28(18.64)
	5.50(0.02)	5.67(0.02)	5.33(0.01)	5.83(0.02)	5.17(0.02)	5.67(0.02)
AR(0.5)		18.5(3.04)	19.5(3.59)	19.5(2.76)	23(0.67)	18(0.95)
		8(8.31)	6(3.13)	1(47.30)	84(98.37)	1(7.34)
	3.00(0.01)	3.67(0.01)	3.00(0.01)	4.00(0.01)	3.50(0.00)	3.67(0.01)
AR(0.8)		13.5(3.30)	15.5(3.06)	13(2.05)	24(0)	15(1.63)
		3.5(2.06)	3.5(3.28)	5(40.01)	94(93.45)	2(7.47)
	1.67(0.01)	1.67(0.01)	1.33(0.01)	2.33(0.01)	2.00(0.01)	2.00(0.01)
BC(i)		22(1.66)	23(2.02)	23(1.66)	24(0)	23(0.94)
		8.5(4.72)	6(2.91)	10(58.04)	32.5(91.05)	20(9.97)
	4.33(0.01)	3.17(0.01)	4.00(0.01)	5.00(0.01)	4.17(0.01)	4.33(0.01)
BC(ii)		19.5(1.80)	18.5(3.22)	20(1.68)	24(0.31)	22(0.91)
		7.5(3.16)	4(2.61)	35.5(151.22)	209.5(160.25)	11(10.01)
	3.17(0.01)	3.83(0.01)	3.00(0.01)	4.33(0.01)	3.83(0.01)	3.67(0.01)

positives. Under the sensitivity model, as shown in Table 4, iLDA misses some true positives, but is comparable to the benchmarks under some cases (for example, AR(0.1) and BC(i)). We also observe that the differences between iLDA with N.1 and N.2 are not big. Under the non-normal model, Table 5 shows that under scenario Nonnorm.A, iLDA still outperforms DSDA and is superior to NSC and l_1PLD in terms of both classification error and FP. Similar results can be observed under the other four scenarios from Table 9–12 in Appendix. Moreover, as Table 5, 9, and 10 show, the proportion of non-normal important variables may have little impact on iLDA in terms of TP except for the case where some variables follow a mixture distribution. An increased proportion of non-normal variables may result in an increase in FP. When the sample sizes are set similar as those in the real data, as shown in Table 6, under most cases iLDA still presents superiority in terms of FP and classification error, although an increased FP is observed compared to other simulations. This is reasonable with smaller sample sizes.

We have also compared the computational efficiency of different methods. It is observed that iLDA has a significant advantage over NSC, but is a little less efficient than l_1PLD and DSDA. For example, in Simulation 3, given the tunings λ_1 and λ_2, iLDA takes 3.2 and 3.7 seconds to complete overall estimation under network N.1 and N.2, respectively, compared to 10.9 (NSC), 0.9 (l_1PLD), and 1.9 (DSDA). In terms of CPU time, iLDA costs 0.3 and 0.4 seconds un-

Table 5. Simulation 5 under Nonnorm.A: the first row is the number of true positives (standard deviation), the second row is the false positives (standard deviation), and the third row is prediction mean classification error (standard deviation).

Scenario	Bayes	iLDA		NSC	l_1PLD	DSDA
		N.1	N.2			
			χ^2 (5)			
Homo.A		18(1.49)	18(1.50)	15(1.83)	18(2.55)	13(0.85)
		10(0.16)	7(5.23)	39.5(94.78)	1(14.90)	1(4.47)
	10.88(0.01)	11.56(0.01)	11.28(0.01)	11.78(0.01)	14.83(0.03)	11.33(0.01)
Hetero.B		24(2.67)	27(2.67)	25(1.50)	29.5(2.47)	21(1.62)
		8(15.05)	11(2.45)	36.5(146.72)	5(99.66)	1.5(3.55)
	4.67(0.01)	6.33(0.01)	6.05(0.01)	4.89(0.01)	16.39(0.03)	6.61(0.01)
			Log-normal			
Homo.A		21(0.94)	21(1.00)	19(1.53)	16(0.58)	13(1.07)
		8.5(3.47)	7(2.38)	8(139.38)	0(0.35)	1(4.95)
	9.72(0.01)	11.56(0.01)	11.22(0.01)	10.50(0.01)	16.67(0.02)	11.67(0.01)
Hetero.B		24(2.99)	26(2.59)	28(1.99)	16(0.78)	20.5(1.48)
		6(3.70)	8(4.54)	39(173.70)	0(0.53)	1(12.59)
	5.25(0.01)	5.67(0.01)	5.44(0.01)	6.44(0.01)	16.44(0.01)	6.22(0.01)
			Two-component normal mixture			
Homo.A		17(1.32)	17(1.40)	14(0.92)	20(0.79)	12(1.08)
		8.5(3.73)	7(3.20)	19.5(23.11)	4.5(4.42)	1(9.29)
	11.39(0.01)	11.73(0.01)	11.44(0.01)	12.00(0.01)	20.50(0.02)	11.40(0.01)
Hetero.B		22(3.35)	24(2.49)	24(1.33)	30(1.46)	21(1.22)
		4(3.01)	3(4.32)	27.5(83.32)	19(41.83)	3(1.92)
	5.33(0.01)	5.44(0.07)	5.56(0.01)	5.50(0.01)	15.33(0.02)	6.00(0.01)
			Uniform			
Homo.A		21(0.57)	20(0.56)	20(1.10)	21(0.41)	17(1.56)
		1(3.11)	2(1.01)	56.5(88.12)	36(54.62)	0(3.53)
	6.72(0.01)	11.67(0.01)	10.44(0.01)	9.28(0.01)	9.83(0.01)	14.11(0.01)
Hetero.B		25(1.68)	27.5(0.92)	29(1.25)	30(0.48)	25.5(1.56)
		21.5(7.32)	15(6.98)	28.5(153.57)	67.5(134.48)	2(7.02)
	3.5(0.01)	7.17(0.01)	6.00(0.01)	5.06(0.01)	5.05(0.01)	9.22(0.01)
			t (5)			
Homo.A		21(0.75)	21(0.46)	20(1.87)	21(0)	15(1.69)
		7(6.21)	5(4.23)	57(113.51)	23.5(60.02)	1.5(5.32)
	8.83(0.01)	10.33(0.01)	10.00(0.01)	10.28(0.01)	9.11(0.01)	11.06(0.01)
Hetero.B		25(2.66)	29(1.73)	28(1.96)	30(0.48)	24(1.83)
		7(13.50)	10(15.43)	25(129.22)	51(78.25)	2(4.26)
	4.5(0.01)	5.39(0.01)	5.06(0.01)	4.68(0.01)	4.94(0.01)	6.00(0.01)
			t (30)			
Homo.A		21(0.18)	21(0.35)	19(1.59)	21(0.55)	16(1.66)
		6(2.71)	5(3.74)	36(50.64)	35(87.76)	1.5(6.05)
	9.56(0.01)	9.72(0.01)	9.33(0.01)	11.17(0.01)	10.39(0.01)	10.44(0.01)
Hetero.B		27(2.73)	29(1.74)	27(2.45)	30(1.06)	26(1.89)
		9(13.38)	8(9.12)	61.5(192.57)	89(193.72)	3(4.69)
	4.72(0.01)	5.98(0.01)	5.57(0.01)	5.28(0.01)	5.78(0.01)	6.22(0.01)

der N.1 and N.2, respectively, compared to 3.3 (NSC), 0.1 (l_1PLD), and 0.1 (DSDA).

4. BREAST CANCER DATA ANALYSIS

Breast cancer is the second leading cause of cancer deaths. Many gene profiling studies have been conducted. Here three gene expression datasets (referred to D1–D3) are integrated, which are collected from three different studies. Although the same platform was used in all three datasets, researchers are not able to directly merge them because of differences in regional, environmental, clinical, and other factors. The first study reported the breast cancer gene expression profiles of 86 Malaysian women (Pau et al., 2010), among whom 43 are case samples and 43 are control samples. The second study was organized by Boston University School of Medicine. There are 30 laser capture microdissected breast tissue samples, among which 15 are case

Table 6. Simulation 6: the first row is the number of true positives (standard deviation), the second row is the false positives (standard deviation), and the third row is prediction mean classification error (standard deviation).

Correlation	Bayes	iLDA		NSC	l_1PLD	DSDA
		N.1	N.2			
		16(4.11)	16(5.41)	23(3.25)	29(4.24)	8(2.51)
AR(0.1)		24(40.87)	29(15.26)	172(201.45)	245(172.42)	3(2.84)
	17.24(0.04)	20.69(0.03)	20.34(0.03)	27.39(0.04)	32.78(0.03)	27.14(0.03)
		22(2.57)	24(2.64)	24(3.80)	30(0)	15(2.92)
AR(0.5)		14(18.45)	6(12.35)	66(100.56)	163(103.23)	2(5.63)
	3.52(0.02)	4.82(0.01)	3.79(0.01)	6.55(0.02)	6.98(0.03)	8.64(0.03)
		24(3.95)	26(2.71)	20(4.84)	30(0)	15(1.81)
AR(0.8)		44(36.85)	38(25.29)	56(59.39)	56(17.93)	2(2.67)
	1.39(0.01)	2.04(0.01)	2.18(0.01)	2.56(0.02)	2.56(0.01)	2.79(0.01)
		23(3.31)	26(2.63)	26(2.93)	30(2.48)	10(3.27)
BC(i)		45(46.19)	50(22.17)	103(154.49)	291(141.20)	4(8.39)
	13.79(0.02)	15.51(0.03)	14.13(0.03)	14.13(0.04)	20.00(0.04)	17.24(0.03)
		22(3.62)	25(3.60)	28(1.95)	30(0)	14(2.38)
BC(ii)		30(38.41)	26(14.25)	35(105.24)	225(124.30)	4(4.22)
	6.55(0.02)	6.55(0.03)	6.89(0.03)	7.24(0.03)	8.28(0.03)	8.28(0.02)

samples and the remaining are control samples (Graham et al., 2011). In the third study, 14 of the 29 samples were from the epitheliums adjacent to breast tumours, and 15 samples were obtained from patients undergoing reduction mammoplasty without apparent breast cancer (Tripathi et al., 2008). A total of 22,283 probe sets are profiled in all three datasets. It is expected that a large number of the genes are noises, and they may create problems such as false selection and high computational cost. Hence, we conduct an unsupervised screening and rank the genes using their variations and select the top 800 for analysis.

In previous studies, it has been shown that there exist strong correlations among genes (Liu et al., 2013(a)). The frequency of the absolute values of Pearson correlations among all genes across all the datasets is presented in the left panel of Figure 1, and that of one randomly selected gene (HLA-DRB1) is presented in the right panel. Moderate to high correlations are observed, suggesting the sensibility of adopting the network structure.

For each dataset, each gene expression is normalized to zero mean and unit variance. Genes identified by iLDA with network N.1 and N.2 are listed in Table 7. We see that the two iLDA models present considerable similarity. They identify the same 20 genes across the three datasets. Among these datasets, as expected, considerable overlaps are presented. For example, D1 and D2 share 19 and 24 identified genes under iLDA with N.1 and N.2, respectively. In addition, datasets have dataset-specific genes. For instance, under network N.2, MMP7 is specific for D1. For comparison, three single-dataset classification methods including NSC, l_1PLD, and DSDA are applied, and meta-analysis is conducted. We summary the numbers of genes and their overlaps identified by different methods. To better comprehend their similarity/difference, we also compute the modified RV-coefficients (Smilde et al., 2009) between the identified gene sets of two approaches. This coefficient measures the common information of two matrices (observation matrices of genes identified by two different approaches in our study), with a larger value indicating higher similarity. The summary comparison results are presented in Table 8. We observe that the numbers of genes identified by NSC and l_1PLD are far greater than that of iLDA. This is consistent with the finding in simulation. We also see that moderate to high modified RV-coefficients exist between iLDA and alternatives. This indicates that, our method, with just over 20 discriminative genes, can cover most of the data information that is contained in the alternatives with hundreds of genes.

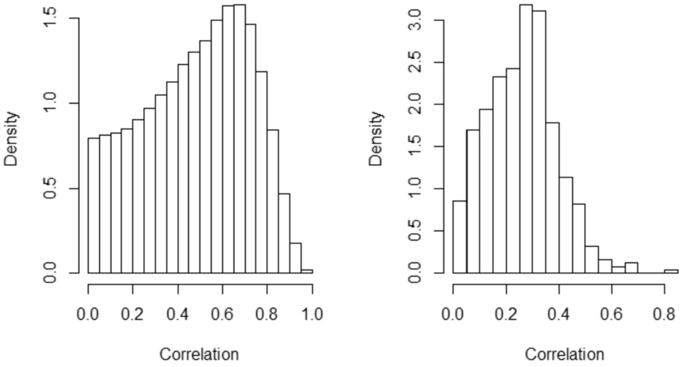

Figure 1. Data analysis: absolute values of Pearson Correlations (the left panel is for all genes; the right one is for a randomly selected gene (HLA-DRB1)).

We also present the network structures of selected genes under N.1 and N.2 in Figures 2 and 3, respectively. The eigenvector centrality of a gene is indicated by the size of its node, and gene communities are colorized differently.

Table 7. Breast cancer studies: gene identification and parameter estimation.

Gene	N.1			N.2		
	D1	D2	D3	D1	D2	D3
LOC101928826				-0.519	-0.473	-0.028
RPL13A	-0.114	-0.136	-0.087	-0.027		-0.525
RPS3	0.394	0.433		0.270	0.144	-0.119
IGH				-0.246	-0.057	-0.135
CD24(216379_x_at)	0.252	-0.753		0.354	-0.513	
RPL23A	0.505	-0.909		0.621	-0.885	-0.253
CD24(209771_x_at)			-0.650	-0.112	-0.034	-0.341
EEF1G			0.632	0.100	0.657	0.020
UBB	-0.500		0.184	-0.583		0.306
RPL27A	0.093	-0.675		0.447	-1.277	
HMGN1	0.171	0.309		0.291	0.036	0.077
IER2	0.131		-0.925	0.250	-0.001	-1.244
LOC100508408				-0.391	-0.142	0.593
PTN	0.386	0.090	0.101	0.319	0.555	-0.187
CSN1S1	-0.344	-0.056	-0.092	-0.243	0.388	0.617
PIK3R1	0.032		0.502	0.022	0.226	0.560
RPS16	-0.470		-0.099	-0.494	-0.068	-0.476
ZNF721				0.196	-0.106	0.630
AKR1C2	-0.548	0.204		-0.697	0.185	-0.066
LPL	-0.463	0.209	0.194	-0.396	0.191	-0.396
JUN	-0.372	0.117	-0.042	-0.323	0.191	
MMP7				0.017		
HNRNPD	0.294	0.365	0.219	0.445	0.339	0.735
RGS1	0.441	-0.003	-0.104	0.365	-0.369	0.146
HAUS2	0.285	0.410	0.303	0.271	0.462	0.210
ACTN1	0.231	-0.697		0.258	-0.769	0.036
CAT				0.238	-0.708	0.715
IGK	-0.140	-0.467	0.116			
EEF2	-0.260	-0.020	-0.125			
MYL12B	0.219	0.039	0.430			
RBP4	0.060	-0.281	-0.433			

Table 8. Numbers of genes, overlaps, and the modified RV-coefficients between different approaches.

	Overlap					Modified RV-coefficient			
	N.1	N.2	NSC	l_1PLD	DSDA	N.2	NSC	l_1PLD	DSDA
N.1	25	20	12	11	10	0.96	0.65	0.72	0.75
N.2		27	11	10	9		0.63	0.70	0.73
NSC			111	105	19			0.93	0.83
l_1PLD				207	15				0.79
DSDA					19				

We see that these two network structures share a considerable number of important nodes, including ACTN1, HNRNPD, RPS3, RPS16, UBB, RPS16, EEF1G, RPL27A, and RPL13A. In addition, we can obtain some information on genes' interconnections from the community detection result. For instance, RPL27A, RPL13A, UBB, EEF1G, and RPS16 are clustered into the same community under both N.1 and N.2. It is of interest to note that these genes have similar groups of coefficients in the sense of 1-norm. For example, for genes clustered into the purple (black in printed version) community in Figure 2, most of their coefficients' 1-norms range from 0.5 to 0.7, while for most green (dark grey in printed version) genes, the norms are between 0.7 to 0.9. Similar results can be obtained under N.2.

To evaluate prediction performance, we adopt a random splitting approach. Each dataset is randomly split into a training set and a testing set with sizes 3:1. To avoid an extreme split, this process is repeated 100 times. Three criteria, namely sensitivity, specificity, and prediction error, are calculated to measure prediction performance. Sensitivity and specificity measure the prediction accuracy of case and control samples, respectively. It is shown that, iLDA

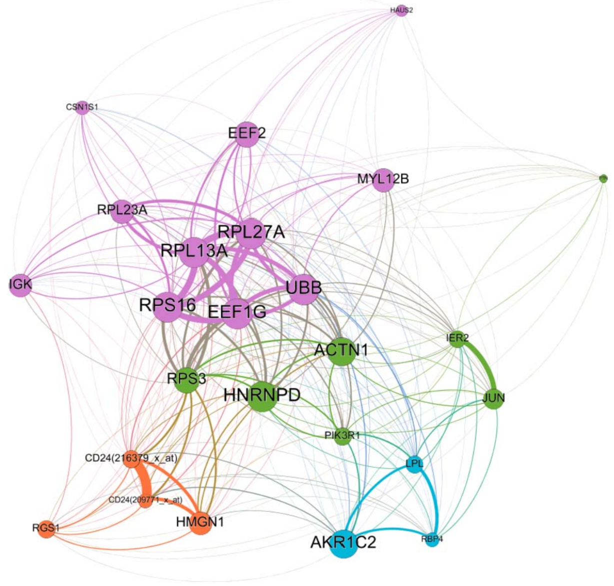

Figure 2. Network structure for genes identified by iLDA under N.1: each gene community (cluster) is colorized uniquely, and the eigenvector centrality of a gene is represented by the size of its node.

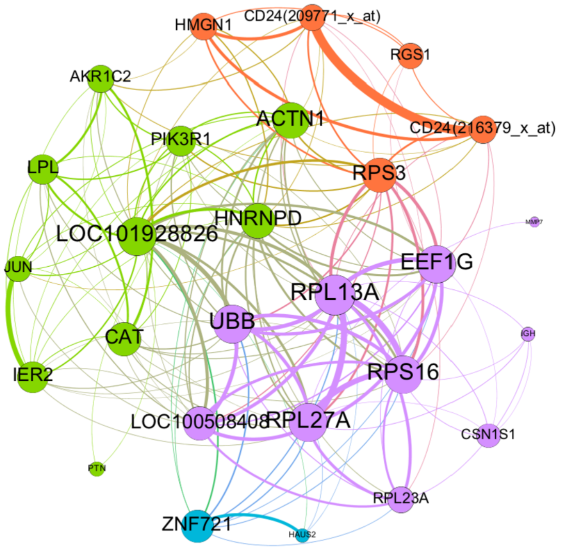

Figure 3. Network structure for genes identified by iLDA under N.2: each gene community (cluster) is colorized uniquely, and the eigenvector centrality of a gene is represented by the size of its node.

has smaller prediction errors (0.143 under both networks) than benchmarks (0.200 for all three benchmarks). iLDA also shows higher sensitivity, with 0.89 and 0.90 under N.1 and N.2, respectively, while the results for NSC, l_1PLD, and DSDA are 0.87, 0.79, and 0.83, respectively. Specificities are calculated as 0.82 (N.1), 0.84 (N.2), 0.79 (NSC), 0.82 (l_1PLD), and 0.78 (DSDA).

5. DISCUSSION

In high-dimensional classification, single dataset analysis may be unsatisfactory owing to the small sample size. Integrative analysis pools and analyses raw data from multiple datasets, and can effectively increase sample size and improve estimation and selection result. However, the existing integrative analysis studies are focused on regression. In this study we have developed the iLDA method for the integrative analysis of LDA. Advancing from the published studies on classification, we have considered the interconnections among variables by constructing a network structure. To achieve variable selection and simultaneous estimation, we have adopted the 1-norm group MCP method in which the effect of one covariate across all datasets is represented by a group of coefficients. In addition, a Laplacian penalty has been adopted to incorporate the network information. For computation, a local linear approximation has been conducted, and based on this approximation, we have further adopted the coordinate descent algorithm to estimate parameters. Simulation study has been conducted under different models. Compared with several alternatives, iLDA demonstrates superior performance in terms of true positives, false positives, and prediction accuracy. In an application with three breast cancer datasets, iLDA shows satisfactory prediction results, and the interconnections among genes are properly accommodated.

In practical data analysis, the proposed approach can provide a competitive solution to multi-dataset classification. This is supported by our extensive simulations and data analysis. The three alternatives compared in our numerical study are originally designed for single-dataset analysis. Their relative advantages (and disadvantages) have been discussed in Mai et al. (2012) and other published studies and will not be reiterated here. Developing their integrative analysis counterparts (and comparing with iLDA) is beyond the scope of this article. It is also noted that it is possible to “reduce” the proposed iLDA to single-dataset analysis. This would lead to a LDA approach that accommodates the network structure of covariates. With the iLDA numerical results presented in this article and single-dataset regression studies that accommodate the covariate network structures in the published literature, it may be reasonable to conjecture that such a LDA approach would have competitive performance. However, a detailed examination of this approach is also beyond our scope.

This study can be potentially extended in multiple directions. Besides LDA, integrative analysis can be conducted based on other classification methods. For variable selection and estimation, 1-norm gMCP penalization is adopted. Other penalties are also expected to be applicable, such as the group Bridge, 1-norm group SCAD, adaptive group Lasso, and composite group MCP. We postpone such research to the future.

APPENDIX

Table 9. Simulation 5 under Nonnorm.B: the first row is the number of true positives (standard deviation), the second row is the false positives (standard deviation), and the third row is prediction mean classification error (standard deviation).

Scenario	Bayes	iLDA		NSC	l_1PLD	DSDA
		N.1	N.2			
			$\chi^2\,(5)$			
Homo.A		17(1.75)	17(1.83)	13(2.32)	17(2.44)	10(0.72)
		20(9.69)	16(5.46)	10.5(88.94)	0(0.69)	0.5(1.76)
	17.61(0.02)	17.67(0.02)	17.00(0.02)	15.00(0.02)	20.00(0.02)	15.01(0.02)
Hetero.B		20(3.14)	22.5(2.48)	19(2.94)	26.5(2.91)	14(1.28)
		12(13.24)	11(10.49)	68.5(86.29)	1(1.19)	4(5.04)
	4.67(0.01)	8.00(0.01)	7.44(0.01)	10.78(0.01)	15.39(0.02)	8.33(0.01)
			Log-normal			
Homo.A		20(0.91)	20(0.81)	27(2.58)	30(0)	11(1.64)
		19.5(5.64)	14(4.89)	141(188.92)	14(18.92)	1.5(4.29)
	13.11(0.01)	15.33(0.01)	14.56(0.01)	7.83(0.01)	19.33(0.02)	14.73(0.01)
Hetero.B		25(2.54)	28(2.02)	29.5(2.59)	30(0)	15(1.71)
		11.5(13.52)	6(3.42)	50(95.06)	10(5.82)	4(5.22)
	6.56(0.01)	7.39(0.01)	6.89(0.01)	7.83(0.01)	19.55(0.02)	8.67(0.01)
			Two-component normal mixture			
Homo.A		15(1.78)	15.5(1.53)	12(1.29)	12(0.92)	10.5(0.53)
		24.5(4.56)	17.5(3.29)	11(12.38)	15(20.91)	0.5(4.42)
	14.33(0.01)	15.89(0.02)	15.33(0.02)	14.39(0.01)	15.28(0.01)	14.39(0.01)
Hetero.B		15(1.65)	16(1.97)	17(2.04)	30(0.32)	14(0.74)
		27(2.47)	25(2.45)	18(32.40)	8(64.20)	4(2.98)
	9.17(0.02)	9.11(0.01)	8.56(0.01)	8.22(0.01)	28.33(0.03)	8.89(0.01)
			Uniform			
Homo.A		21(0.65)	21(0.61)	18(1.46)	21(1.51)	17(1.50)
		0(2.94)	0(1.23)	39.5(73.49)	58.5(51.23)	1(4.87)
	7.72(0.01)	7.33(0.01)	7.44(0.01)	10.56(0.02)	11.50(0.01)	8.57(0.01)
Hetero.B		25(2.58)	26(0.47)	28(1.57)	29(1.36)	22(1.78)
		13(14.17)	12(1.67)	101(64.53)	101(87.20)	10(7.59)
	11.72(0.01)	11.44(0.01)	10.61(0.01)	13.17(0.01)	13.00(0.01)	12.00(0.01)
			$t\,(5)$			
Homo.A		21(0.85)	21(0.85)	18(2.11)	19(1.14)	15(1.38)
		12(5.16)	7(1.29)	34.5(66.63)	33(58.20)	4(12.87)
	11.78(0.01)	12.89(0.02)	12.33(0.01)	12.44(0.01)	12.27(0.01)	12.83(0.02)
Hetero.B		26(3.31)	29(1.41)	29(2.87)	30(0.95)	22(3.03)
		26(16.83)	17(5.29)	149.5(115.40)	107(97.04)	8(12.10)
	6.22(0.01)	6.44(0.01)	6.22(0.01)	8.11(0.01)	6.78(0.01)	7.56(0.01)
			$t\,(30)$			
Homo.A		21(0.91)	21(0.94)	18(1.91)	18(1.46)	15(1.28)
		9.5(5.31)	7(5.24)	37(62.39)	40(91.29)	1(5.56)
	10.89(0.01)	12.22(0.02)	11.39(0.02)	12.78(0.02)	12.22(0.02)	12.33(0.01)
Hetero.B		26(2.70)	30(0.86)	30(1.58)	30(1.26)	26(2.51)
		25(21.64)	23.5(8.59)	89(90.39)	107(87.20)	12.5(10.58)
	11.28(0.01)	11.78(0.01)	11.56(0.01)	8.22(0.01)	7.61(0.01)	13.33(0.01)

Table 10. Simulation 5 under Nonnorm.C: the first row is the number of true positives (standard deviation), the second row is the false positives (standard deviation), and the third row is prediction mean classification error (standard deviation).

Scenario	Bayes	iLDA		NSC	l_1PLD	DSDA
		N.1	N.2			
			$\chi^2(5)$			
Homo.A		16(1.76)	16(1.95)	8.5(2.26)	19(2.32)	12(1.05)
		47.5(9.71)	39.5(14.35)	30.5(43.29)	0(1.10)	1(10.79)
	24.67(0.03)	28.56(0.03)	28.11(0.03)	27.67(0.03)	35.39(0.02)	30.29(0.04)
Hetero.B		20.5(3.51)	21.5(3.79)	14.5(3.88)	27(1.70)	16(0.86)
		11.5(9.37)	8.5(5.49)	39.5(59.52)	0(0.31)	6(5.98)
	11.77(0.01)	12.72(0.01)	12.67(0.01)	13.39(0.01)	37.78(0.02)	13.72(0.01)
			Log-normal			
Homo.A		20(0.97)	20(1.22)	20(3.24)	21(0)	16(1.89)
		32.5(7.82)	25.5(9.91)	47(56.49)	5(19.50)	13(6.05)
	19.22(0.01)	24.28(0.02)	23.39(0.02)	19.61(0.01)	26.78(0.02)	23.44(0.02)
Hetero.B		26(1.77)	27(1.71)	29.5(0.72)	30(0.37)	19(2.06)
		8(15.12)	17(7.43)	71.5(100.89)	6.5(2.45)	5(5.41)
	10.09(0.01)	10.33(0.01)	10.17(0.01)	11.11(0.01)	20.11(0.01)	13.33(0.01)
			Two-component normal mixture			
Homo.A		12(2.46)	11(2.63)	6(1.90)	21(2.21)	4(0.31)
		48(50.29)	39(20.41)	31.5(81.29)	5(32.44)	0(6.78)
	23.11(0.02)	28.61(0.02)	27.72(0.02)	28.27(0.02)	42.81(0.02)	23.77(0.02)
Hetero.B		13(1.65)	14(2.11)	10(3.42)	30(0)	7(1.03)
		23(10.91)	18(5.92)	13(23.40)	5(5.60)	6(3.29)
	15.17(0.01)	14.01(0.02)	13.50(0.02)	14.17(0.01)	36.50(0.02)	13.94(0.02)
			Uniform			
Homo.A		21(0.18)	21(0.18)	21(1.10)	21(0)	21(0.49)
		0(3.12)	0(0.98)	46(43.25)	41(39.10)	5(7.87)
	7.83(0.01)	7.55(0.01)	7.61(0.01)	12.83(0.02)	13.28(0.01)	8.78(0.01)
Hetero.B		28(2.01)	29(1.17)	30(0)	30(0)	30(0.50)
		28(9.76)	33(10.29)	93(90.23)	63(81.23)	5(6.40)
	2.61(0.01)	2.89(0.01)	2.78(0.01)	5.28(0.01)	7.22(0.01)	2.94(0.01)
			$t(5)$			
Homo.A		20(1.41)	21(0.89)	10.5(2.05)	19(0.93)	19(0.95)
		31(12.92)	25.5(8.93)	17.5(28.98)	29(82.54)	23(5.25)
	22.50(0.02)	25.11(0.02)	24.72(0.02)	20.56(0.02)	41.22(0.02)	27.89(0.02)
Hetero.B		26(2.05)	29(1.42)	30(0.94)	30(0.18)	23(2.92)
		18(13.61)	17(8.14)	80.5(57.17)	55(56.48)	12(10.72)
	8.00(0.01)	8.56(0.01)	7.61(0.01)	10.89(0.01)	9.33(0.01)	10.56(0.01)
			$t(30)$			
Homo.A		21(1.14)	21(0.67)	10(1.09)	18(0)	10(0.56)
		32.5(46.82)	26.5(30.22)	15.5(45.48)	26.5(85.34)	5.5(8.37)
	23.01(0.05)	26.22(0.04)	26.17(0.05)	24.39(0.04)	40.44(0.02)	26.11(0.02)
Hetero.B		25(2.52)	30(0.91)	30(1.15)	30(0)	25(2.79)
		21(20.31)	16(5.92)	89(80.08)	104(47.92)	13(10.78)
	7.67(0.01)	8.11(0.01)	6.89(0.01)	9.22(0.01)	8.17(0.01)	9.11(0.01)

Table 11. Simulation 5 under Nonnorm.D: the first row is the number of true positives (standard deviation), the second row is the false positives (standard deviation), and the third row is prediction mean classification error (standard deviation).

Scenario	Bayes	iLDA		NSC	l_1PLD	DSDA
		N.1	N.2			
$\chi^2\,(5)$						
Homo.A		19(1.31)	19(0.97)	19(0.97)	19(1.23)	17(0.90)
		15(14.30)	14(14.30)	14.5(28.90)	45(51.39)	2(2.74)
	8.94(0.01)	9.00(0.01)	9.11(0.01)	10.64(0.02)	14.17(0.01)	9.67(0.01)
Hetero.B		23(3.52)	23(1.49)	30(0)	0(0)	25(1.67)
		11(32.13)	9(24.42)	21(18.14)	150(89.82)	1(4.23)
	4.11(0.01)	4.56(0.01)	4.33(0.01)	5.83(0.01)	45.28(0.02)	4.22(0.01)
Log-normal						
Homo.A		21(2.02)	21(1.66)	20(0.82)	0(2.02)	17(1.19)
		21(19.83)	18(13.16)	26(48.07)	131(283.23)	0(2.04)
	9.44(0.01)	9.39(0.01)	8.72(0.01)	10.56(0.01)	49.78(0.02)	9.89(0.01)
Hetero.B		24(3.19)	28(2.43)	29(0.92)	8(6.13)	26(1.40)
		30(39.69)	26(21.32)	26(121.63)	100(132.19)	0(2.87)
	4.44(0.01)	4.50(0.01)	4.22(0.01)	6.44(0.01)	50.33(0.02)	4.72(0.01)
Two-component normal mixture						
Homo.A		21(0)	21(0.32)	20(1.05)	0(0)	17.5(0.67)
		110(67.70)	112(87.19)	19.5(37.28)	149(189.02)	1(3.12)
	9.38(0.01)	11.39(0.01)	11.29(0.01)	10.33(0.01)	50.5(0.02)	10.00(0.01)
Hetero.B		19(2.28)	23(2.05)	30(0.98)	2(3.29)	26.5(1.64)
		91.5(66.25)	90(62.29)	44(88.93)	280(230.21)	1(1.33)
	4.89(0.01)	6.89(0.01)	6.44(0.01)	5.61(0.01)	39.93(0.02)	5.11(0.01)
Uniform						
Homo.A		21(1.76)	20(1.49)	20(1.13)	21(0.31)	17(1.39)
		0(1.43)	1(2.47)	69(150.96)	37(82.22)	1(3.99)
	9.28(0.01)	9.22(0.01)	8.94(0.01)	9.56(0.01)	8.94(0.01)	9.89(0.01)
Hetero.B		25(2.16)	27(3.23)	30(0.84)	30(0)	26(1.72)
		18(8.70)	6(11.12)	5(63.47)	60(50.38)	1(2.44)
	4.39(0.01)	4.56(0.01)	4.22(0.01)	4.00(0.01)	4.00(0.01)	4.72(0.01)
$t\,(5)$						
Homo.A		21(0.31)	21(0.43)	19(1.35)	21(0.48)	17(1.39)
		16(4.66)	16(4.63)	2(32.38)	33(80.59)	1(3.10)
	9.50(0.01)	10.39(0.01)	10.00(0.01)	10.44(0.01)	9.78(0.01)	9.83(0.01)
Hetero.B		25(2.74)	28(2.01)	29(0.52)	30(0)	26(1.79)
		9(16.60)	2(14.87)	46(213.65)	33(122.19)	1(4.89)
	4.17(0.01)	4.33(0.01)	4.00(0.01)	5.06(0.01)	4.78(0.01)	4.44(0.01)
$t\,(30)$						
Homo.A		21(0.31)	21(0.73)	20(1.41)	21(0.38)	17(1.04)
		7(4.62)	6(4.28)	6(131.06)	47(87.08)	0(1.78)
	9.33(0.01)	10.00(0.01)	9.78(0.01)	9.33(0.01)	9.11(0.01)	9.78(0.01)
Hetero.B		27(2.32)	29(2.65)	29(1.16)	30(0)	26(1.57)
		3(10.58)	5(12.94)	29(113.50)	56(80.97)	0(12.96)
	4.28(0.01)	4.28(0.01)	4.06(0.01)	4.39(0.01)	4.11(0.01)	4.78(0.01)

Table 12. Simulation 5 under Nonnorm.E: the first row is the number of true positives (standard deviation), the second row is the false positives (standard deviation), and the third row is prediction mean classification error (standard deviation)

Scenario	Bayes	iLDA		NSC	l_1PLD	DSDA
		N.1	N.2			
			$\chi^2\,(5)$			
Homo.A		18(2.76)	18(2.75)	20(0.97)	18(0.83)	17(1.09)
		16(18.35)	16(5.92)	35.5(34.91)	75.38(65.39)	0(3.39)
	9.44(0.01)	9.67(0.01)	9.72(0.01)	13.00(0.03)	15.49(0.01)	10.39(0.01)
Hetero.B		26(3.21)	24.5(2.69)	30(0.43)	0(0)	26(1.59)
		29(47.73)	23(4.29)	22(28.53)	280(158.29)	0(4.04)
	4.62(0.01)	5.67(0.01)	5.28(0.01)	5.78(0.01)	49.39(0.02)	4.89(0.01)
			Log-normal			
Homo.A		21(0.98)	20(1.25)	19(1.29)	3(1.99)	17(1.23)
		21(19.47)	20(9.90)	17(42.51)	180(292.39)	0(2.92)
	8.94(0.01)	8.72(0.01)	8.61(0.01)	9.81(0.01)	50.17(0.02)	9.56(0.01)
Hetero.B		25(3.60)	28(1.74)	29(1.26)	0(4.48)	27(1.89)
		36(37.89)	34(29.31)	16.5(77.92)	210(165.23)	1(1.03)
	5.00(0.01)	5.50(0.01)	5.11(0.01)	3.67(0.01)	49.44(0.02)	5.33(0.01)
			Two-component normal mixture			
Homo.A		21(0)	21(0.42)	21(1.10)	0(1.75)	18(0.99)
		92(6.89)	91(6.97)	23(41.26)	298(144.24)	1(0.69)
	10.33(0.01)	11.33(0.01)	11.44(0.01)	10.50(0.01)	50.89(0.02)	10.67(0.01)
Hetero.B		21(1.67)	23(1.10)	30(0.71)	0(3.86)	25(2.00)
		78(9.91)	70(9.82)	47(120.54)	300(346.41)	2(1.98)
	4.78(0.01)	6.28(0.01)	5.72(0.01)	5.17(0.01)	5.01(0.02)	5.11(0.01)
			Uniform			
Homo.A		21(0.86)	20(1.29)	20(1.57)	21(0.18)	17(1.59)
		0(2.71)	0.5(2.31)	33(52.30)	13(45.49)	0(3.59)
	8.83(0.01)	9.33(0.01)	9.33(0.01)	9.33(0.01)	9.11(0.01)	10.22(0.01)
Hetero.B		28(2.01)	29(1.17)	30(0.85)	30(0)	30(0.50)
		28(9.75)	33(11.87)	18(90.23)	23(30.52)	6(6.52)
	2.61(0.01)	2.89(0.01)	2.78(0.01)	4.50(0.01)	4.44(0.01)	2.94(0.01)
			$t\,(5)$			
Homo.A		12(0.84)	12(1.20)	11(0.52)	11(0.52)	10(0.71)
		21(5.93)	16(4.39)	7(29.18)	22(28.32)	1(3.16)
	14.72(0.01)	16.89(0.01)	15.89(0.01)	14.89(0.01)	14.22(0.01)	16.50(0.01)
Hetero.B		26(3.13)	28(1.51)	30(0.42)	30(0)	27(1.51)
		14(10.38)	3(0.52)	80(99.68)	117(138.73)	2(1.58)
	4.44(0.01)	4.50(0.01)	4.44(0.01)	4.06(0.01)	4.44(0.01)	4.56(0.01)
			$t\,(30)$			
Homo.A		12(0.45)	12(0.78)	11(0.88)	11(0.53)	10(0.63)
		20(5.58)	15.5(5.17)	10(66.49)	19(43.20)	0(3.29)
	14.78(0.01)	16.27(0.02)	15.89(0.02)	15.44(0.02)	14.72(0.01)	14.89(0.02)
Hetero.B		26(3.12)	29(1.89)	29(1.19)	30(0.19)	26(1.70)
		6(8.85)	6(11.95)	94(179.39)	43.5(78.21)	1(4.96)
	5.44(0.01)	5.78(0.01)	5.44(0.01)	5.50(0.01)	5.67(0.01)	6.00(0.01)

ACKNOWLEDGEMENTS

We would like to thank the editor and reviewers for their useful comments and suggestions, which have led to a significant improvement of this study. Wang's work was supported by the National Natural Science Foundation of China (71601076), Humanity and Social Science Youth Foundation of Ministry of Education of China (16YJCZH104), and Social Science Foundation of Hunan Province (15YBA085). Zhang's work was supported by the Fundamental Research Funds for the Central Universities (20720171064, 20720181003). Ma's work was supported by the National Bureau of Statistics of China (2016LD01).

Received 3 February 2018

REFERENCES

[1] Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics.* **20(18)** 3583–3593.

[2] Fan, J., Fan, Y. (2008). High Dimensional Classification Using Features Annealed Independence Rules. *Annals of Statistics.* **36(6)** 2605–2637. MR2485009

[3] Fan, J., Feng,Y., Tong, X. (2012). A road to classification in high dimensional space: The regularized optimal affine discriminant. *Journal of the Royal Statistical Society.* **74(4)** 745–771. MR2965958

[4] Guerra, R., Goldstein, D. R. (2009). *Meta-analysis and combining information in genetics and genomics.* CRC Press, New York. MR2569308

[5] Guo, Y., Hastie, T., Tibshirani, R. (2006). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics.* **8(1)** 86–100.

[6] Hastie, T., Tibshirani, R., Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition. Springer, New York. MR2722294

[7] Huang, J., Breheny, P., Ma, S. (2012(a)). A Selective Review of Group Selection in High-Dimensional Models. *Statistical Science.* **27(4)** 481–499. MR3025130

[8] Huang, Y., Huang, J., Shia, B. C., Ma, S. (2012(b)). Identification of cancer genomic markers via integrative sparse boosting. *Biostatistics.* **13(3)** 509–522.

[9] Huang, J., Ma, S., Li, H., Zhang, C. H. (2011). The sparse Laplacian shrinkage estimator for high-dimensional regression. *Annals of Statistics.* **39(4)** 2021–2046. MR2893860

[10] Knudsen, S. (2006). *Cancer Diagonostics with DNA Microarrays.* Wiley, New Jersey.

[11] Li, C., Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics.* **24(21)** 1175–1182.

[12] Li, C., Li, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *Annals of Applied Statistics.* **4(3)** 1498–1516. MR2758338

[13] Liu, J., Huang, J., Ma, S. (2014). Integrative Analysis of Cancer Diagnosis Studies with Composite Penalization. *Scandinavian Journal of Statistics.* **41(1)** 87–103. MR3181134

[14] Liu, J., Huang, J., Ma, S. (2013(a)). Incorporating Network Structure in Integrative Analysis of Cancer Prognosis Data. *Genetic Epidemiology.* **37(2)** 173–183.

[15] Liu, J., Huang, J., Ma, S., Wang, K. (2013(b)). Incorporating group correlations in genome-wide association studies using smoothed group Lasso. *Biostatistics.* **14(2)** 205–219.

[16] Ma, S., Huang, Y., Huang, J., Fang, K. (2012). Gene network-based cancer prognosis analysis with sparse boosting. *Genetics Research.* **94(4)** 205–221.

[17] Ma, S., Huang, J., Song, X. (2011(a)). Integrative analysis and variable selection with multiple high-dimensional data sets. *Biostatistics.* **12(4)** 763–775.

[18] Ma, S., Huang, J., Wei, F., Xie, Y., Fang, K. (2011(b)). Integrative analysis of multiple cancer prognosis studies with gene expression measurements. *Statistics in Medicine.* **30(28)** 3361–3371. MR2861619

[19] Mai, Q., Zou, H., Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika.* **99(1)** 29–42. MR2899661

[20] Ni, I. B. P., Zakaria Z., Muhanmmad, R., et al. (2010). Gene expression patterns distinguish breast carcinomas from normal breast tissues: the Malaysian context. *Pathology-Research and Practice.* **206(4)** 223–228.

[21] Shi, X., Liu, J., Huang, J., Zhou, Y., Shia, B., Ma, S. (2013). Integrative analysis of high-throughput cancer studies with contrasted penalization. *Genetic Epidemiology.* **38(2)** 144–151.

[22] Smilde, A. K., Kiers, H. A., Bijlsma, S. (2009). Matrix correlations for high-dimensional data: the modified RV-coefficient. *Bioinformatics.* **25(3)** 401–405.

[23] Tibshirani, R., Hastie, T., Narasimhan, B., Chu G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences.* **99(10)** 6567–6572.

[24] Tripathi, A., King, C., Morenas, A., et al. (2008). Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients. *International Journal of Cancer.* **122(7)** 1557–1566.

[25] Tseng, G., Ghosh, D., Zhou, X. (2015). *Integrating Omics Data.* Cambridge University Press, New York.

[26] Wu, M. C., Zhang, L., Wang, Z., Christiani, D. C., Lin, X. (2009). Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics.* **25(9)** 1145–1151.

[27] Witten, D. M., Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society.* **73(5)** 753–772. MR2867457

[28] Zhao, Q., Shi, X., Huang, J., Liu, J., Li, Y., Ma, S. (2015). Integrative analysis of '-omics' data using penalty functions. *Wiley Interdisciplinary Reviews: Computational Statistics.* **7(1)** 99–108. MR3348725

[29] Zou, H., Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics.* **36(4)** 1509–1533. MR2435443

Xiaoyan Wang
College of Finance and Statistics
Hunan University
Changsha, Hunan 410079
China
Department of Biostatistics
Yale University
New Haven, CT 06511
USA
E-mail address: xywang@hnu.edu.cn

Kuangnan Fang
School of Economics
Xiamen University
Xiamen, Fujian 361005
China
Fujian Key Lab of Science
Xiamen University
Xiamen, Fujian 361005
China
E-mail address: xmufkn@163.com

Qingzhao Zhang
School of Economics
Wang Yanan Institute for Studies in Economics
MOE Key Lab of Ecomonics
Fujian Key Lab of Statistics
Xiamen University
Xiamen, Fujian 361005
China
E-mail address: qzzhang@xmu.edu.cn

Shuangge Ma
Department of Biostatistics
Yale University
New Haven, CT 06511
USA
E-mail address: shuangge.ma@yale.edu