Biostatistics (2010), **11**, 4, *pp.* 599–608
doi:10.1093/biostatistics/kxq023
Advance Access publication on May 26, 2010

Simultaneous variable selection and class fusion for high-dimensional linear discriminant analysis

JIAN GUO

Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA
guojian@umich.edu

SUMMARY

In many high-dimensional microarray classification problems, an important task is to identify subsets of genes that best discriminate the classes. Nevertheless, existing gene selection methods for microarray classification cannot identify which classes are discriminable by these selected genes. In this paper, we propose an improved linear discriminant analysis (LDA) method that simultaneously selects important genes and identifies the discriminable classes. Specifically, a pairwise fusion penalty for LDA was used to shrink the differences of the class centroids in pairs for each variable and fuse the centroids of indiscriminable classes altogether. The numerical results in analyzing 2 gene expression profiles demonstrate the proposed approach help improve the interpretation of important genes in microarray classification problems.

Keywords: Class fusion; Linear discriminant analysis; Microarray; Shrunken centroid estimator; Variable selection.

1. INTRODUCTION

It has been considered important to predict the clinical class of a sample based on its gene expression profiles from microarray experiments. However, this is a challenging task due to the huge number of genes. Linear discriminant analysis (LDA), originally introduced by Fisher (1936), is a classification technique which has been successfully applied in microarray classification problems (Tibshirani *and others*, 2002; Guo *and others*, 2007; Tai and Pan, 2007; Wang and Zhu, 2007, and the references therein). LDA assumes that the observations in each class come from a specific Gaussian-distributed component, and it also assumes that these Gaussian components have different means but equal covariance matrices. In a prediction procedure, the label of a new observation is determined by the Bayes rule (Hastie *and others*, 2001). LDA performs well for low-dimensional data. In particular, it has some nice properties, such as the robustness to deviations from model assumptions and the almost-"Bayes" optimality (Guo *and others*, 2007). Nevertheless, the performance of LDA is far from optimal in high-dimensional cases, especially when the number of the variables is much larger than the sample size ($p \gg n$) (Di Pillo, 1976, 1977). There are 2 major limitations here. First, the sample covariance matrix is singular and cannot be inverted when $p > n$. To address this problem, Friedman (1977) proposed a method to regularize the common covariance matrix of the Gaussian components in LDA. Second, it is a common assumption that only a small proportion of variables contribute to classification in high-dimensional data. Nevertheless, it is challenging to identify such important variables in practice. Tibshirani *and others* (2002) proposed a modified

LDA, namely shrunken centroids estimator. By assuming the diagonal shape of the covariance matrix, it shrinks the class centroids toward the global centroid by using soft thresholding and thus removes unimportant variables (i.e. the centroids of all classes are shrunken together) from the model. Wang and Zhu (2007) reformulated the shrunken centroids estimator as a Lasso-type problem (Tibshirani, 1996) and proposed 2 new penalties to improve the effectiveness of variable selection. Tai and Pan (2007) improves the shrunken centroids estimator by incorporating group structures among the variables. Guo *and others* (2007) extended the idea of shrunken centroids estimator to LDA with general covariance matrix.

In existing variable selection methods for multiclass LDA (Tibshirani *and others*, 2002; Wang and Zhu, 2007), the important variables are those effectively discriminate at least 2 out of all classes. In many real problems, however, people are also interested in identifying which specific classes can be discriminated by an important variable. Imagining, for example, a disease with 3 subtypes (denoted as types I, II, and III). By observing the gene expression profiles, we may see that some genes can discriminate types I and II but cannot discriminate types II and III; on the other hand, some other genes can discriminate types II and III but cannot discriminate types I and II. Such scenarios often appear in high-dimensional gene expression profiles, and thus it is necessary to identify these class-specific information. For this aim, the paper proposes a penalized LDA method that simultaneously selects important variables and identifies specific classes that can be discriminated by these variables. Specifically, a pairwise fusion penalty was used in the proposed model to fuse the class centroids for each variable. Two classes are considered indiscriminable if their class centroids are fused together. Moreover, if all class centroids associated with a variable are fused, this variable is regarded as unimportant to all classes and removed from the model.

The remainder of the paper is organized as follows: Section 2 introduces the methodology of proposed method and discusses algorithmic issues. Section 3 illustrates the performance of the proposed method with 2 simulated examples, and Section 4 applies this method to 2 microarray data sets, respectively. Finally, some concluding remarks are drawn in Section 5.

2. METHODOLOGY

2.1 *High-dimensional LDA*

Suppose the data matrix $\boldsymbol{X} = (x_{i,j})_{n\times p}$ consists of n observations and p variables. Without loss of generality, we assume $\boldsymbol{X}$ is centered along each column, that is, $\sum_{i=1}^{n} x_{i,j} = 0$, $1 \leqslant j \leqslant p$. In a K-class LDA problem, the observations in the kth class ($1 \leqslant k \leqslant K$) are assumed to be i.i.d. generated from a Gaussian distribution with mean $\boldsymbol{\mu}_k = (\mu_{k,1}, \ldots, \mu_{k,p})$ and the common covariance matrix $\boldsymbol{\Sigma}$. In addition, it is a common assumption in high-dimensional settings that the covariance matrix is diagonal, that is, $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$. This assumption significantly reduces the number of parameters to be estimated and its advantages are theoretically justified by Bickel and Levina (2004).

The parameters of high-dimensional LDA can be estimated by solving the following criterion:

$$\min_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \frac{1}{2} \sum_{k=1}^{K} \sum_{i\in S_k} \sum_{j=1}^{p} \left[\frac{(x_{i,j} - \mu_{k,j})^2}{\sigma_j^2} - \log \sigma_j^2 \right], \tag{2.1}$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)^T$ and S_k is the index set of the kth class. Let $\widehat{\pi}_k = n_k/n$ be the estimate of the prior of the kth class, whose sample size is n_k. The prediction procedure is based on the Bayes rule (Hastie *and others*, 2001). Specifically, given the estimate $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ from (2.1), a new observation $\boldsymbol{x}^* = (x_1^*, \ldots, x_p^*)$ is assigned to the class which achieves

$$\arg \max_{1\leqslant k\leqslant K} \widehat{\pi}_k \phi(\boldsymbol{x}^*; \widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}), \tag{2.2}$$

where ϕ is the density function of p-variate Gaussian distribution.

2.2 *The pairwise fusion penalty*

To fuse indiscriminable classes for each important variable, we use the following pairwise fusion penalty to regularize criterion (2.1):

$$\min_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{j=1}^{p}\left[\frac{(x_{i,j}-\mu_{k,j})^2}{\sigma_j^2}-\log\sigma_j^2\right]+\lambda\sum_{j=1}^{p}\sum_{1\leqslant k<k'\leqslant K} w_{k,k'}^{(j)}|\mu_{k,j}-\mu_{k',j}|, \tag{2.3}$$

where λ is a tuning parameter. The penalty aims at shrinking the differences between every pair of class centroids for each variable. Similar to the scenario in Lasso (Tibshirani, 1996), the ℓ_1-norm in the penalty shrinks some differences to be exactly zero, resulting in some class centroids $\hat{\mu}_{k,j}$'s having identical values. If $\widehat{\mu}_{k,j}=\widehat{\mu}_{k',j}$, for some $1\leqslant k<k'\leqslant K$, then variable j cannot discriminate class k and class k', though it may be effective to discriminate other classes. Moreover, if all class centroids for some variable are fused together, that is, $\widehat{\mu}_{1,j}=\widehat{\mu}_{2,j}=\cdots=\widehat{\mu}_{p,j}$, then this variable is considered unimportant to the classification task and can be removed from the model. We borrow the idea from Zou (2006) and define the adaptive weights $w_{k,k'}^{(j)}=|\tilde{\mu}_{k,j}-\tilde{\mu}_{k',j}|^{-1}$, $1\leqslant k<k'\leqslant K$, $1\leqslant j\leqslant p$, where $\tilde{\mu}_{k,j}$ is the estimate of $\mu_{k,j}$ from criterion (2.1). With these adaptive weights, the pairwise fusion penalty tends to lightly fuse classes k and k' ($1\leqslant k<k'\leqslant K$) if variable j is effective to discriminate them and heavily fuses them otherwise. Note that the pairwise fusion penalty has been applied in Guo *and others* (2010) for clustering purpose.

REMARK 2.1 It is of interest to compare the method defined in (2.3) with the ℓ_1-regularized high-dimensional LDA as follows:

$$\min_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{j=1}^{p}\left[\frac{(x_{i,j}-\mu_{k,j})^2}{\sigma_j^2}-\log\sigma_j^2\right]+\lambda\sum_{j=1}^{p}\sum_{k=1}^{K}\xi_{k,j}|\mu_{k,j}|, \tag{2.4}$$

where $\xi_{k,j}$'s are adaptive weights defined as $\xi_{k,j}=1/|\tilde{\mu}_{k,j}|$. The ℓ_1-penalty in (2.4) shrinks the individual $\mu_{k,j}$'s toward zero (which is the global centroid of the entire centered data) and removes variable j from the model if all $\widehat{\mu}_{k,j}$, $1\leqslant k\leqslant K$, are set to zeros. However, it cannot correctly identify which specific classes are discriminable by each important variable. Following Wang and Zhu (2007), we can show that (2.4) is actually equivalent to the shrunken centroids estimator (Tibshirani *and others*, 2002) if we set $\xi_{k,j}=\sqrt{1/n_k-1/n}$ instead. For clarification, we denote the estimators defined by criterions (2.3) and (2.4) as LDA-PF and LDA-L1, respectively.

2.3 *Parameter estimation*

Notice that criterion (2.3) can be decomposed into p individual minimization problems, where the jth one is

$$\min_{\boldsymbol{\mu}_{(j)},\boldsymbol{\Sigma}} \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in S_k}\left[\frac{(x_{i,j}-\mu_{k,j})^2}{\sigma_j^2}-\log\sigma_j^2\right]+\lambda\sum_{1\leqslant k<k'\leqslant K} w_{k,k'}^{(j)}|\mu_{k,j}-\mu_{k',j}|, \tag{2.5}$$

where $\boldsymbol{\mu}_{(j)}$ is the jth column of $\boldsymbol{\mu}$. By taking the first derivative of objective function (2.3) with respect to σ_j^2's, we can obtain the closed-form solution $\widehat{\sigma}_j^2=1/n\sum_{k=1}^{K}\sum_{i\in S_k}(x_{i,j}-\tilde{\mu}_{k,j})^2$, where $\tilde{\mu}_{k,j}=1/n_k\sum_{i\in S_k}x_{i,j}$. The estimation of $\mu_{k,j}$'s is nontrivial. When σ_j^2's are replaced by their

estimates, objective function (2.3) can be transformed into a quadratic programming problem. We propose an efficient iterative algorithm based on the standard local quadratic approximation algorithm (Fan and Li, 2001), which has been used in a number of variable selection procedures and whose convergence properties have been studied by Fan and Li (2001) and Hunter and Li (2005). Specifically, let $\widehat{\mu}_{k,j}^{(t)}$ be the estimates from the tth iteration ($t = 1, 2, \ldots$), we approximate

$$|\mu_{k,j}^{(t+1)} - \mu_{k',j}^{(t+1)}| \approx \frac{(\mu_{k,j}^{(t+1)} - \mu_{k',j}^{(t+1)})^2}{2|\widehat{\mu}_{k,j}^{(t)} - \widehat{\mu}_{k',j}^{(t)}|} + \frac{1}{2}|\widehat{\mu}_{k,j}^{(t)} - \widehat{\mu}_{k',j}^{(t)}|\,, \tag{2.6}$$

which results in an approximation to (2.5):

$$\min_{\boldsymbol{\mu}_{(j)}} \frac{1}{2(\widehat{\sigma}_j^{(t)})^2} \sum_{k=1}^{K} \sum_{i \in S_k} (x_{i,j} - \mu_{k,j})^2 + \lambda \sum_{1 \leqslant k < k' \leqslant K} w_{k,k'}^{(j)} \frac{(\mu_{k,j} - \mu_{k',j})^2}{2|\widehat{\mu}_{k,j}^{(t)} - \widehat{\mu}_{k',j}^{(t)}|}\,. \tag{2.7}$$

Denote $\boldsymbol{\mathcal{Y}}$ as the jth of $\boldsymbol{X}$ and $\boldsymbol{\mathcal{X}}$ as an $n \times K$ matrix whose kth ($1 \leqslant k \leqslant K$) column is composed of ones for those components in S_k and zeros for those outside S_k^c. Let $\boldsymbol{\beta} = \boldsymbol{\mu}_{(j)}$. We also denote $\boldsymbol{G} = (g_{k,k'})_{K \times K}$ as a $K \times K$ matrix whose off-diagonal element $g_{k,k'} = -w_{k,k'}^{(j)}/|\widehat{\mu}_{k,j}^{(t)} - \widehat{\mu}_{k',j}^{(t)}|$ and whose diagonal element $g_{k,k} = \sum_{1 \leqslant k' \leqslant K; k' \neq k} w_{k,k'}^{(j)}/|\widehat{\mu}_{k,j}^{(t)} - \widehat{\mu}_{k',j}^{(t)}|$. Then the following proposition shows that (2.7) has a closed-form solution.

PROPOSITION 2.2 Objective function (2.7) is equivalent to the following generalized ridge regression problem:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}\boldsymbol{\beta}\|^2 + \lambda\sigma_j^2 \boldsymbol{\beta}^T \boldsymbol{G} \boldsymbol{\beta} \tag{2.8}$$

with a closed-form solution

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{\mathcal{X}}^T \boldsymbol{\mathcal{X}} + \lambda\sigma_j^2 \boldsymbol{G})^{-1} (\boldsymbol{\mathcal{X}}^T \boldsymbol{\mathcal{Y}})\,. \tag{2.9}$$

This procedure was repeated over $t = 1, 2, \ldots$ until convergence. We list the proposed algorithm as follows:

Step 1. Initialize $\widehat{\mu}_{k,j}^{(1)} = \widetilde{\mu}_{k,j}^{(1)} = 1/n_k \sum_{i \in S_k} x_{i,j}$, $1 \leqslant k \leqslant K$, $1 \leqslant j \leqslant p$;
Step 2. In the tth iteration, update $\widehat{\mu}_{k,j}^{(t+1)}$, $1 \leqslant k \leqslant K$, $1 \leqslant j \leqslant p$, with (2.9);
Step 3. Repeat Step 2 until some stopping criterion achieves.

REMARK 2.3 In this work, the stopping criterion is defined as $\sum_{1 \leqslant k \leqslant K} \sum_{1 \leqslant j \leqslant p} |\widehat{\mu}_{k,j}^{(t+1)} - \widehat{\mu}_{k,j}^{(t)}| / \sum_{1 \leqslant k \leqslant K} \sum_{1 \leqslant j \leqslant p} |\widehat{\mu}_{k,j}^{(t)}| < 10^5$.

REMARK 2.4 For numerical stability, we threshold the absolute value of $\widehat{\mu}_{k,j}^{(t)} - \widehat{\mu}_{k',j}^{(t)}$ at a lower bound of 10^{-10}, and at the end of the iterations, set all estimates less than 10^{-10} to zero.

3. SIMULATION STUDY

To evaluate the performance of the proposed method, we modified the simulated examples in Guo *and others* (2010).

In this section, we evaluate the performance of LDA-PF on 2 simulated examples. In each example, we generate 50 data sets, each consisting of a training set, an independent validation set, and an independent test set, with 20, 20, and 2000 observations, respectively. The model is estimated on the training set, and the tuning parameter is selected on the validation set by minimizing the corresponding prediction error rate. We repeat this procedure on 50 data sets for each simulation and recorded the test error rates, the false-negative rates (the proportions of incorrectly removed important variables), and the false-positive rates (the proportions of incorrectly selected unimportant variables), averaged on the 50 data sets, respectively.

EXAMPLE 3.1 In this scenario, there are $K = 4$ classes and $p = 202$ variables with the first 2 being important and the remaining ones unimportant. The variables were generated as follows: the first variable follows distributions $N(2.5, 1)$, $N(0, 1)$, $N(0, 1)$, and $N(-2.5, 1)$ in the 4 classes, respectively; the second variable follows distributions $N(1.5, 1)$, $N(1.5, 1)$, $N(-1.5, 1)$, and $N(-1.5, 1)$ in the 4 classes, respectively. All remaining 200 variables are i.i.d. $N(0, 1)$ for all 4 classes. In this simulation setting, variable 1 cannot discriminate classes 2 and 3, while variable 2 cannot discriminate classes 1 and 2 (as well as classes 3 and 4).

EXAMPLE 3.2 This example considers a 5-class scenario. There are a total of $p = 203$ variables with the first 3 important and the other 200 unimportant. Similarly to simulation 1, the important variables follow normal distributions with unit variances but different means in the 5 classes. Specifically, the means of variable 1 are 2.5, 2.5, 0, 0, -2.5, the means of variable 2 are -2.5, 0, 0, 0, 2.5, and the means of variable 3 is 2.5, 0, 0, -2.5, -2.5. In this scenario, variable 1 cannot discriminate classes 1 and 2, as well as classes 3 and 4; variable 2 cannot discriminate classes 2, 3, and 4; and variable 3 cannot discriminate classes 2 and 3, as well as classes 4 and 5.

The results over 50 replications for both examples are summarized in Table 1. We can see that in both examples, LDA-PF exhibit similar performance to LDA-L1 in terms of false-negative rate and false-positive rate. and it achieves slightly lower error.

Table 2 summarizes the results of identifying indiscriminable classes for those important variables. Specifically, each row in the table gives the average proportion of the important variables that correctly identify the corresponding indiscriminable pair of classes. For example, the first row shows that for LDA-PF, on average 96.0% of the 50 replications, variable 1 can correctly fuse classes 2 and 3. It is also clear that LDA-PF dominates LDA-L1 in terms of correctly fusing the indiscriminable classes. It should also be pointed out that although LDA-L1 correctly fuses some class centroids, respectively (e.g. in the first row), these results are artifacts. For example, in Example 3.1, the centroids of classes 2 and 3 for variable 1 are all equal to zero, which happens to be the value that the ℓ_1-penalty shrinks to. The same reasoning also applies to classes 2, 3, and 4 for variable 2 in Example 3.2.

Table 1. *Prediction and variable selection results for Examples* 3.1 *and* 3.2. *Each table cell exhibits the result averaged over* 50 *repetitions and the associated standard deviation (in the parentheses). "ER" is the average prediction error rate on the test set, "FN" is the average false-negative rate, that is, the average proportion of incorrectly removed important variables, and "FP' is the false-positive rate, that is, the average proportion of incorrectly selected unimportant variables*

Example	Method	ER (%)	FN (%)	FP (%)
1	LDA-L1	15.6 (1.3)	0 (0)	0.2 (0.4)
	LDA-PF	15.1 (1.4)	0 (0)	0.2 (0.5)
2	LDA-L1	13.4 (1.1)	0 (0)	0.5 (0.8)
	LDA-PF	12.9 (1.2)	0 (0)	0.5 (1.3)

Table 2. *Pairwise class fusion results for Examples* 3.1–3.2. *"Pair" corresponds to indiscriminable class pairs for the variables in the corresponding row. For example, the first row indicates that variable 1 is unimportant for discriminating classes* 2 *and* 3. *The numbers in the following columns give the proportions of the important variables that correctly identify the corresponding indiscriminable pair of classes. All results are averaged over* 50 *repetitions with the corresponding standard deviations in the parentheses*

Example	Variable	Pair	LDA-L1 (%)	LDA-PF (%)
1	1	2/3	96.0 (19.8)	96.0 (19.8)
	2	1/2	0 (0)	96.0 (19.8)
		3/4	4.0 (19.8)	92.0 (27.4)
2	1	1/2	6.0 (24.0)	96.0 (19.8)
		3/4	42.0 (49.9)	94.0 (24.0)
	2	2/3	100 (0)	100 (0)
		2/4	98.0 (14.1)	98.0 (14.1)
		3/4	98.0 (14.1)	98.0 (14.1)
	3	2/3	44.0 (50.1)	90.0 (30.3)
		4/5	0 (0)	90.0 (30.3)

4. REAL DATA ANALYSIS

In this section, we apply LDA-PF to 2 microarray data sets: SRBCT and PALL, whose descriptions are listed below:

- SRBCT data set: This data set contains the expression profiles of 2308 genes, obtained from 83 tissue samples of small round blue cell tumors (SRBCT) of childhood cancer (Khan *and others*, 2001). The 83 samples are classified into 4 tumor subtypes: Ewing's sarcoma, rhabdomyosarcoma (RMS), neuroblastoma, and Burkitt's lymphoma.
- PALL data set: This data set contains gene expression profiles for 12 625 genes from 248 patients (samples) with pediatric acute lymphoblastic leukemia (PALL) (see Yeoh *and others*, 2002, for more details). The samples are classified into 6 tumor subtypes: T-ALL (43 cases), E2A-PBX1 (27 cases), TEL-AML (79 cases), Hyperdiploid $>$ 50 (64 cases), BCR-ABL (15 cases), and MLL (20 cases). The original data had a large number of missing intensities and the following preprocessing was applied. All intensity values less than one were set to one; then all intensities were transformed to log-scale. Further, all genes with log-intensities equal to zero for more than 80% of the samples were discarded, thus leaving 12 083 genes for further consideration.

In each data set, all observations were randomly split into 2 groups: a training set (70% of all observations) and a test set (30% of all observations). LDA-PF was estimated on the training set and its performance was evaluated on the test set. Note that both test error rate and number of selected genes depend on the choice of the tuning parameter λ. Figure 1 illustrates the test error rate with respect to the number of selected genes when varying λ over different values. In both data sets, we can see that the lowest test error rate is achieved when number of selected genes varies in a large range. The optimal tuning parameter was selected on the training set by 5-fold cross-validation. Since there may be multiple tuning parameters corresponding to the same error rate, we choose the largest one among them. Table 3 shows the test error rates for both LDA-L1 and LDA-PF. We can see that both methods produce the same error rate in SRBCT and PALL data sets.

LDA-PF selected 8 and 124 genes in SRBCT and PALL data sets, respectively. Figures 2 and 3 illustrate the centroids of the selected genes estimated by LDA-PF in these 2 data sets using heatmaps.

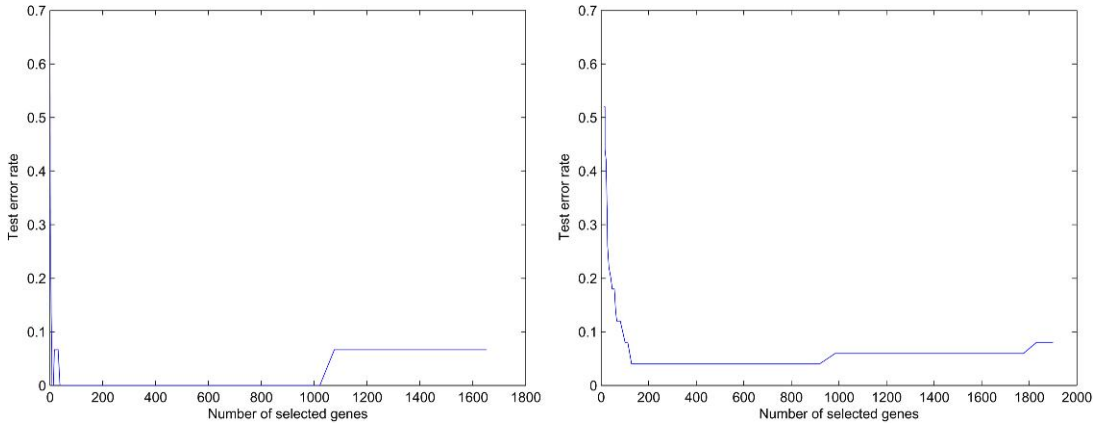

Fig. 1. The curves of the test error rates with respect to the number of selected variables. The figure in the left panel is about SRBCT data set, and the figure in the right panel is about PALL data set.

Table 3. *Classification results for the SRBCT and PALL data sets. "ER" is the prediction error rate on the test set*

Example	Method	ER (%)
SRBCT	LDA-L1	0
	LDA-PF	0
PALL	LDA-L1	4.0
	LDA-PF	4.0

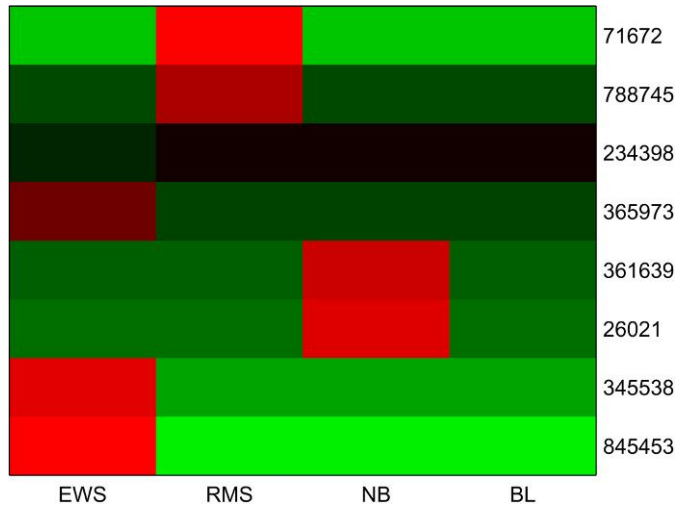

Fig. 2. The heatmap of the estimated centroids for the 8 genes selected by LDA-PF.

In each figure, the columns correspond to classes and rows to genes. The red (green) spots represent positive (negative) values in estimated centroids. It is easy to read the discriminable/indiscriminable classes from these heatmaps. For example, gene "71 672" in Figure 2 can discriminate class RMS from

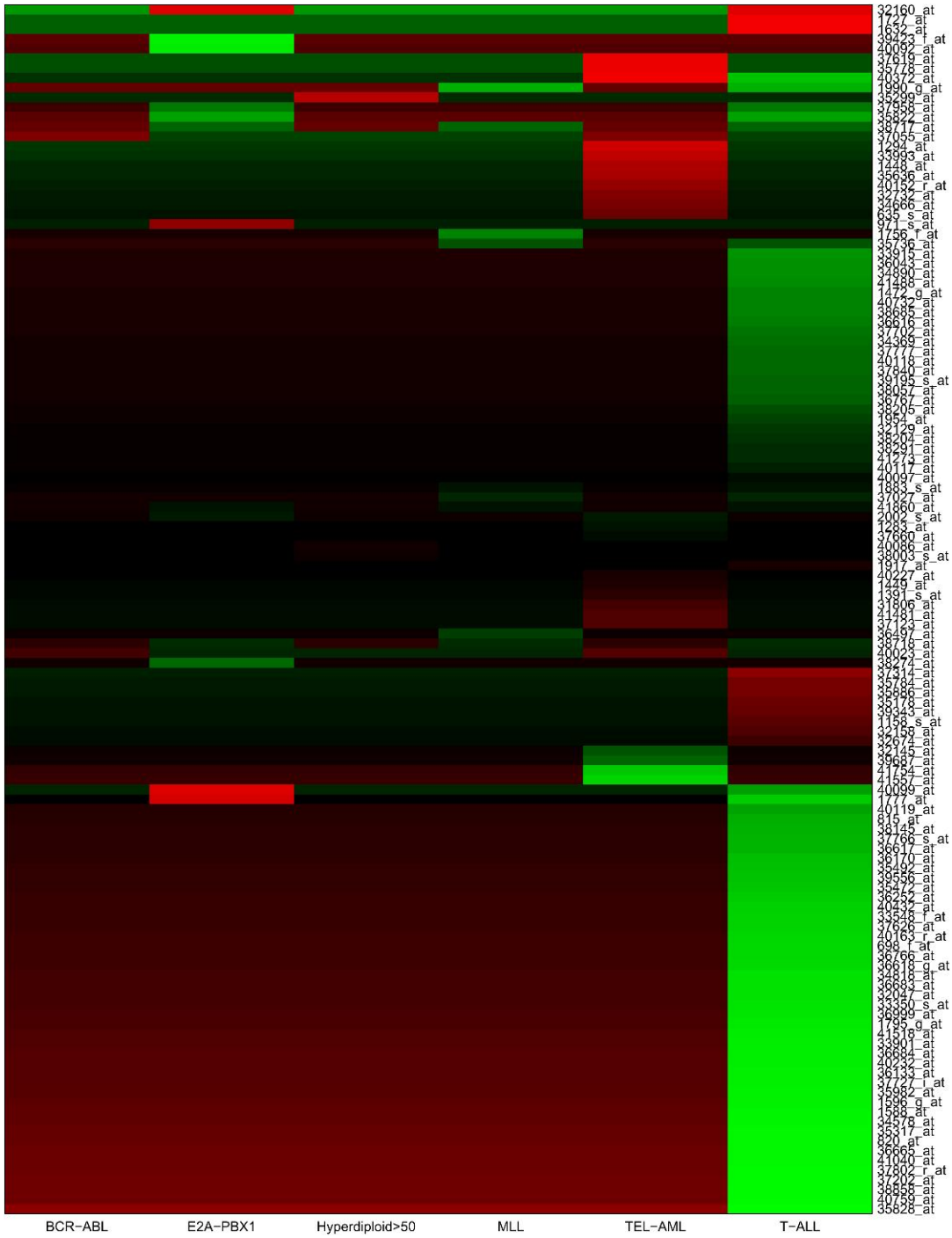

Fig. 3. The heatmap of the estimated centroids for the 124 genes selected by LDA-PF.

the remaining classes; gene "1727" in Figure 3 can discriminate class "T-ALL" from the remaining classes.

5. Conclusions

We have developed a penalized LDA method for simultaneously selecting important genes and identify the corresponding discriminable classes from expression profiles and it help improve the interpretation for the functions of particular genes in different classes. The pairwise fusion penalty introduced here can also be applied to other classification techniques such as quadratic discriminant analysis, logistic regression and (linear) support vector machines.

Acknowledgments

The author thanks Elizaveta Levina, George Michailidis, and Ji Zhu for their helpful suggestions. *Conflict of Interest:* None declared.

References

Bickel, P. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. *Bernoulli* **10**, 989–1010.

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

Di Pillo, P. (1977). Further applications of bias to discriminant analysis. *Communications in Statistics - Theory and Methods* **6**, 933–943.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Asscociation* **96**, 1348–1360.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.

Friedman, J. (1977). Regularized discriminant analysis. *Journal of the American Statistical Asscociation* **84**, 165–175.

Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2010). Pairwise variable selection for high-dimensional model-based clustering. *Biometrics* doi: 10.1111/j.1541-0420.2009.01341.x.

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

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer.

Hunter, D. and Li, R. (2005). Variable selection using MM algorithms. *Annals of Statistics* **33**, 1617–1642.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C. *and others* (2001). Classification and diagnostic prediction of cancers using gene expression profiling andartificial neural networks. *Nature Medicine* **7**, 673–679.

Tai, F. and Pan, W. (2007). Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. *Bioinformatics* **23**, 3170–3177.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B. AND CHU, G. (2002). Diagnosis of multiple cancer types by shrunken centriods of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6567–6572.

WANG, S. AND ZHU, J. (2007). Improved centroids estimation for the nearest shrunken centroid classifier. *Bioinformatics* **23**, 972–979.

YEOH, E.-J., ROSS, M., SHURTLEFF, S., WILLIAMS, W., PATEL, D., MAHFOUZ, R., BEHM, F., RAIMONDI, S., RELLING, M., PATEL, A. *and others* (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* **1**, 133–143.

ZOU, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Asscociation* **101**, 1418–1429.

[*Received August 10, 2009; revised December 30, 2009; accepted for publication March 2, 2010*]