BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 23 2007, pages 3170–3177
doi:10.1093/bioinformatics/btm488

Gene expression

Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data

Feng Tai and Wei Pan*

Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building (MMC 303),
Minneapolis, MN 55455-0378, USA

Received on June 5, 2007; revised on August 21, 2007; accepted on September 25, 2007

Advance Access publication October 12, 2007

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Discriminant analysis for high-dimensional and low-sample-sized data has become a hot research topic in bioinformatics, mainly motivated by its importance and challenge in applications to tumor classifications for high-dimensional microarray data. Two of the popular methods are the nearest shrunken centroids, also called predictive analysis of microarray (PAM), and shrunken centroids regularized discriminant analysis (SCRDA). Both methods are modifications to the classic linear discriminant analysis (LDA) in two aspects tailored to high-dimensional and low-sample-sized data: one is the regularization of the covariance matrix, and the other is variable selection through shrinkage. In spite of their usefulness, there are potential limitations with each method. The main concern is that both PAM and SCRDA are possibly too extreme: the covariance matrix in the former is restricted to be diagonal while in the latter there is barely any restriction. Based on the biology of gene functions and given the feature of the data, it may be beneficial to estimate the covariance matrix as an intermediate between the two; furthermore, more effective shrinkage schemes may be possible.

Results: We propose modified LDA methods to integrate biological knowledge of gene functions (or variable groups) into classification of microarray data. Instead of simply treating all the genes independently or imposing no restriction on the correlations among the genes, we group the genes according to their biological functions extracted from existing biological knowledge or data, and propose regularized covariance estimators that encourages between-group gene independence and within-group gene correlations while maintaining the flexibility of any general covariance structure. Furthermore, we propose a shrinkage scheme on groups of genes that tends to retain or remove a whole group of the genes altogether, in contrast to the standard shrinkage on individual genes. We show that one of the proposed methods performed better than PAM and SCRDA in a simulation study and several real data examples.

Contact: weip@biostat.umn.edu

1 INTRODUCTION

As a classic method, linear discriminant analysis (LDA) has been well studied and widely used. It is well known for its simplicity as well as robustness. Suppose we have a class variable Y with possible values in $\mathcal{G} = \{1,2,\ldots,K\}$ and a real-valued random input vector X, the optimal decision rule based on the 0–1 loss is the Bayes rule: $\hat{Y}(X) = \text{argmax}_{k\in\mathcal{G}} P(Y = k|X = x)$. In the context of sample classifications with microarray gene expression data, Y represents one of K sample groups (e.g. tumors or normal tissues) and X represents the gene expression profile of a patient. According to the Bayes theorem, we have

$$P(Y = k|X = x) \propto P(X = x|Y = k)P(G = k).$$

In LDA, $X|Y = k$ is assumed to have a multivariate normal distribution MVN(μ_k, Σ). By some simple calculations, we have

$$\hat{Y}(X) = \text{argmax}_k \delta_k(x),$$

where

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log(\pi_k)$$

is a linear discriminant function in x. Thus, the classification problem reduces to estimating the parameters in the distribution $f(X|Y = k)$. Traditionally, the maximum-likelihood estimators (MLEs) of μ_k and Σ, the sample mean and sample covariance, are used:

$$\hat{\mu_k} = (\hat{\mu}_{1k}, \ldots, \hat{\mu}_{pk})^T,\ \ \hat{\mu}_{ik} = \frac{1}{n_k}\sum_{y_j=k} x_{ij},$$

$$\hat{\Sigma} = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{y_j=k}(x_j - \hat{\mu}_k)(x_j - \hat{\mu}_k)^T,$$

along with $\pi_k = n_k/n$, where n_k and n are the sample sizes for class k and the pooled samples, respectively. However, with high-dimensional and low-sample-sized data, as arising in sample classification with microarray gene expression data, LDA suffers from the singularity of the sample covariance matrix $\hat{\Sigma}$ due to the 'large p, small n' problem, and the lack of the capability of conducting variable selection. In order to remedy these two weaknesses, Tibshirani *et al.* (2003) proposed a simple modification to LDA, the nearest shrunken centroid (also known as Predictive analysis of microarray (PAM)) method, which assumes the independence among the variables to sidestep the singularity problem, and uses a

*To whom correspondence should be addressed.

shrinkage estimator, instead of MLE, to conduct gene selection. PAM has gained much popularity because of its simplicity and superior performance in practice. On the other hand, completely ignoring possible correlations among the genes as in PAM may be too extreme and thus degrade classification performance. Guo *et al*. (2007) proposed another modified version of LDA, shrunken centroids regularized discriminant analysis (SCRDA), which aims to estimate the covariance matrix in a more general way through regularization, and then adopt the same technique of shrinkage as in PAM for estimate regularization and variable selection, though as to be discussed later, it cannot really realize variable selection. SCRDA was shown to slightly outperform PAM in some occasions.

We feel that both PAM and SCRDA are at the ends of the two extremes: the covariance matrix in the former is restricted to be diagonal while in the latter there is barely any restriction. Based on the biology of gene functions, we aim to estimate the covariance matrix as an intermediate between the two. The idea is in line with recent efforts of integrating biological information of genes, as available in the Gene Ontology (GO) annotations (Ashburner *et al*., 2000), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa, 1996) or constructed gene networks, into the process of model building. Several methods in this line have been proposed. Lottaz and Spang (2005) proposed a structured analysis of microarray data (StAM), while Wei and Li (2007) proposed a modified boosting method called non-parametric pathway-based regression (NPR). More recently, Tai and Pan (2007a) proposed a group penalization method that handles the genes within different functional groups with different penalty terms. The common rationale behind all these methods is simple: many genes are known to have the same function or involved in the same pathway as that of some known/putative cancer-related genes, and the genes in the same functional group or pathway are more likely to work together. The methods aim to borrow information from existing biological knowledge to improve both predictive accuracy and interpretability of a resulting classifier.

In this article, we propose several versions of a modified LDA, group regularized discriminant analysis (GRDA) that aims to take advantage of existing gene functional groups. Specifically, we lean on, but do not require, the assumption that the genes within the same group are correlated with each other, but are independent of the genes from other groups, leading to a block-diagonal covariance structure. In addition, rather than shrinking individually for each gene as in PAM and SCRDA, again by taking advantage of known gene groups, we propose a group shrinkage scheme to identify biologically significant gene functional groups or pathways.

The rest of the article is organized as follows. We first review the PAM and SCRDA. Then, we introduce our new methods GRDA with three combinations of choosing between two regularized covariance matrices and between two shrinkage schemes. We also discuss some computational issues such that an efficient implementation makes it feasible to handle very high-dimensional data. Results from simulation studies and analyses of four public cancer data sets are presented to evaluate the proposed methods, demonstrating their potential gains over PAM and SCRDA. We end with a short summary and discussion.

2 METHODS

2.1 PAM

The nearest shrunken centroids (PAM) method assumes the independence among the genes, ignoring any possible correlation among the genes. It uses a soft-thresholding rule to shrink $\hat{\mu}_{ik}$ towards 0, thus eliminating noise genes. Suppose we have n samples and p genes. Let x_{ij} be the expression level of gene i in sample j. To simplify the discussion, we assume that x_{ij}'s across j have been centered at 0; that is, $\bar{x}_i = \sum_j x_{ij}/n = 0$ for all genes.

The basic idea of PAM is to shrink the class centroids or class-specific sample means $\hat{x}_{ik}$ towards the overall centroid $\bar{x}_i$. Let

$$d_{ik} = \frac{\bar{x}_{ik}}{m_k(s_i + s_0)}$$

where s_i is the pooled within-class SD defined by

$$s_i^2 = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{j\in C_I}(x_{ij} - \bar{x}_{ik})^2$$

with s_0 being a positive constant, usually set as the median of $\{s_i : i = 1, \dots, p\}$ and $m_k = \sqrt{1/n_k - 1/n}$. Basically, d_{ik} is a modified t-statistic for gene i. We can rewrite

$$\bar{x}_{ik} = m_k(s_i + s_0)d_{ik}$$

and shrink d_{ik} towards zero by an amount $\lambda \geq 0$ by soft thresholding:

$$d'_{ik} = \text{sign}(d_{ik})(|d_{ik}| - \lambda)_+ \tag{1}$$

where λ is a tuning parameter that has to be decided, usually by cross-validation (CV). We thus obtain a new shrunken centroid

$$\bar{x}'_{ik} = m_k(s_i + s_0)d'_{ik}$$

and define a discriminant score for class k as

$$\delta_k(\tilde{x}) = \sum_{i=1}^{p}\frac{(\tilde{x}_i - \bar{x}'_{ik})^2}{s_i^2} - 2\log(\pi_k)$$

where $\tilde{x} = (\tilde{x}_1, \dots, \tilde{x}_p)$ is a new test sample and $\pi_k = n_k/n$ is the estimated class prior probability.

It is noted that, if the class centroids $\bar{x}_{ik} = 0$ for all k, then gene i does not contribute to classification and is regarded as a noise gene.

2.2 SCRDA

Instead of completely ignoring the correlations between genes, the SCRDA aims to estimate the covariance matrix in a general way:

$$\tilde{\Sigma} = \alpha\hat{\Sigma} + (1-\alpha)\mathbf{I}\ ,\ \ \alpha \in [0, 1], \tag{2}$$

where $\hat{\Sigma}$ is the sample covariance matrix (i.e. MLE). It aims to adaptively find an optimal intermediate between the unstructured and the independence covariance structures. By regularization, $\tilde{\Sigma}$ will typically be non-singular.

Next, SCRDA shrinks the transformed class mean $\hat{\mu}_k^*$ towards 0:

$$\hat{\mu}_{k(s)}^* = \text{sign}(\hat{\mu}_k^*)(|\hat{\mu}_k^*| - \lambda)_+,\ \ \hat{\mu}_k^* = \tilde{\Sigma}^{-1}\hat{\mu}_k. \tag{3}$$

Finally, SCRDA transforms $\hat{\mu}_{k(s)}^*$ back to get $\hat{\mu}_{k(s)} = \tilde{\Sigma}\hat{\mu}_{k(s)}^*$, and classifies a new sample $\tilde{x}$ using discriminant score

$$\begin{aligned}\delta_k(\tilde{x}) &= \tilde{x}^T\tilde{\Sigma}^{-1}\hat{\mu}_{k(s)} - \frac{1}{2}\hat{\mu}_{k(s)}^T\tilde{\Sigma}^{-1}\hat{\mu}_{k(s)} + \log(\pi_k)\\ &= \tilde{x}^T\hat{\mu}_{k(s)}^* - \frac{1}{2}\hat{\mu}_{k(s)}^T\hat{\mu}_{k(s)}^* + \log(\pi_k)\end{aligned}$$

One problem with SCRDA is that in fact it cannot realize gene selection, which has not been pointed out previously in the literature. The reason is the following. First, each $\hat{\mu}_{ik}^*$ is a linear combination of

$\hat{\mu}_{ik}$'s. Second, if $\hat{\mu}^*_{ik(s)} = 0$ for $k = 1,\ldots,K$, according to the decision rule, gene i in new sample $\tilde{x}$ will not contribute to classification; however, it still contributes to constructing the decision rule since other $\hat{\mu}^*_{jk(s)}$ for $j \neq i$ depends on gene i via $\hat{\mu}_{ik}$ and $\tilde{\Sigma}$.

2.3 Group Regularized Discriminant Analysis (GRDA)

Many studies have shown that genes in the same functional group or involved in the same pathway are more likely to co-express, hence their expression levels tend to be correlated. In this article, we aim to incorporate such biological information into the development of a regularized covariance matrix estimator and a grouped shrinkage scheme. In contrast, neither PAM nor SCRDA takes advantage of such biological information.

2.3.1 Regularized covariance matrix Instead of shrinking the sample covariance matrix to an independence structure, we shrink it to a *between-group* independence structure:

$$\tilde{\Sigma} = \alpha_1 \hat{\Sigma} + \alpha_2 \hat{\Sigma}^* + (1 - \alpha_1 - \alpha_2)\hat{\mathbf{D}}, \tag{4}$$

where α_1,α_2 and $\alpha_1 = \alpha_2 \in [0,1]$ are some tuning parameters to be determined; as a simpler alternative, we also consider using

$$\tilde{\Sigma} = \alpha \hat{\Sigma}^* + (1 - \alpha)\hat{\mathbf{D}}\ ,\ \ \alpha \in [0, 1], \tag{5}$$

where $\hat{\Sigma}$ is the sample covariance matrix, $\hat{\mathbf{D}} = \mathrm{diag}(\hat{\Sigma})$ is a diagonal matrix with the sample variances as diagonal entries and $\hat{\Sigma}^* = \mathrm{diag}(\hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_G)$ is block-diagonal with $\hat{\Sigma}_i$ being a $p_i \times p_i$ sample covariance matrix for genes in group i. We call the resulting discriminant analysis with the regularized covariance estimate in (4) or (5) GRDA, and denote them as GRDA-1 and GRDA-2 respectively.

2.3.2 Shrunken centroids GRDA (SCGRDA) Next we consider two shrinkage schemes. The first one is exactly the same as in SCRDA; we call it individual shrinkage as opposed to the second one, called group shrinkage. The first one works as follows,

$$\hat{\mu}^*_{k(s)} = \mathrm{sign}(\hat{\mu}^*_k)(|\hat{\mu}^*_k| - \lambda)_+,\ \ \hat{\mu}^*_k = \tilde{\Sigma}^{-1}\hat{\mu}_k.$$

As pointed out before, this shrinkage scheme with a non-diagonal covariance matrix cannot realize gene selection.

The second one is a group shrinkage that tends to retain or remove all the variables or genes in a group (Cai, 1999; Yuan and Lin, 2006). Instead of shrinking each $\hat{\mu}^*_{ik}$ individually, we shrink them as a group. With regularized covariance matrix (5), that assumes between-group independence, we can actually perform gene selection at group level. Specifically,

$$\hat{\mu}^*_{kg(s)} = \hat{\mu}^*_{kg}\left(1 - \frac{\lambda\sqrt{p_g}}{||\hat{\mu}^*_{kg}||}\right)_+,$$

where $||\hat{\mu}^*_{kg}|| = \sqrt{\sum_{l \in G_g} \hat{\mu}^2_{lk}}$ is the L_2 norm and $p_g = |G_g|$ is the group size. If $||\hat{\mu}^*_{kg}||$ for group g is larger than the threshold $\lambda\sqrt{p_g}$, then all the $\hat{\mu}^*_{ik}$'s in group g are retained and only shrunken towards 0; otherwise, all the $\hat{\mu}^*_{ik}$ in this group are shrunken to be exactly zero, not contributing to classification, thus realizing gene selection at a group level.

With the two choices of a regularized covariance matrix and two choices of a shrinkage scheme, we have three possible methods:

(1) ISCGRDA-1: GRDA-1 with individual shrinkage

(2) ISCGRDA-2: GRDA-2 with individual shrinkage

(3) GSCGRDA: GRDA-2 with group shrinkage.

2.3.3 Tuning parameters In GRDA-1, we have three tuning parameters whose values need to be determined by CV: α_1, α_2 and λ, and two tuning parameters for GRDA-2: α and λ. We perform a grid search in a tuning parameter space. The grids for α_1, α_2 or α range from 0 to 0.99 with a equally spaced grids and a was 10 in our study. The grids for λ conventionally range from 0 to the maximum of the absolute values of the parameter that needs to be shrunken; e.g. in PAM, it ranges from 0 to $\max(|d_k|)$. However, in SCRDA and our method SCGRDA, $\hat{\mu}^*_k$, the parameter that needs to be shrunken, depends on $\tilde{\Sigma}$, which changes with α. SCRDA fixes the range of λ to simplify the computation.

Instead of directly searching in grids of λ, we introduce another parameter θ, which is a proportion of the total number of genes or gene groups remaining in the model. The range of θ was fixed from 0 to 30 in our study. In CV, given α_1, α_2 or α, we can obtain $\hat{\mu}^*_{ik}$. Then we calculate $\hat{\mu}^*_i = \max_k(|\hat{\mu}^*_{ik}|)$ for each gene. Suppose the order statistic $\hat{\mu}^*_{(i)}$ is the $100(1-\theta)$th percentile of $\{\hat{\mu}^*_i, i = 1, \ldots, p\}$, we let $\lambda = \hat{\mu}^*_{(i-1)}$ and use it to shrink. Similarly, for group shrinkage, we replace $\hat{\mu}^*_{ik}$ by $||\hat{\mu}^*_{kg}||$.

The best combination of the shrinkage parameters were selected based on the smallest number of test errors. When there are two or more combinations giving the same smallest test error rates, to break ties, our strategy is to first use as a small number of genes as possible, which means to choose the smallest θ; if still tied, we choose the smallest α_1 to decrease the weight of the sample covariance matrix; if still tied, we choose the group independence over the individual independence by choosing largest α_2 or α.

2.4 Connection to penalized likelihoods

Let X be centered raw expression data, i.e. $x_{ij} = x^{\mathrm{raw}}_{ij} - \bar{x}^{\mathrm{raw}}_j$. Each gene expression sample is denoted by a $p \times 1$ vector $X_i = (x_{i1},\ldots,x_{ip})^T$ and mean of class k is denoted by $\mu_k = (\mu_{1k},\ldots,\mu_{pk})^T$. In LDA, μ_k and Σ are estimated by MLEs. In this section, we show the connections between our method and penalized log-likelihood methods; for derivations, see Tai and Pan (2007b). The penalized log-likelihood can be expressed as

$$L_\lambda = L(\mu_k, \Sigma) - P_\lambda(\mu_k, \Sigma) \tag{6}$$

where

$$L(\mu_k, \Sigma) = -\frac{n}{2}\log\det(\Sigma) - \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_i - \mu_k)^T\Sigma^{-1}(X_i - \mu_k)$$

is a multivariate normal log-likelihood and $P\lambda(\mu_k, \Sigma)$ is a penalty function for parameters μ and Σ.

2.4.1 PAM As pointed out by Wu (2006) and Wang and Zhu (2007), if we let $Z_i = X_i/m_k$ for gene i in class k and $\Sigma = \mathrm{diag}((s_1 + s_0)^2,\ldots,(s_p + s_0)^2)$, and use an L_1 norm penalty $P_\lambda(\mu_k, \Sigma) = \lambda\sum_{k=1}^{K}\sum_{j=1}^{p} n_k|\mu_{kj}|$ on μ_k, the nearest shrunken centroids estimators are the maximizer of L_λ with X_i replaced by Z_i.

2.4.2 GSCGRDA For GSCGRDA, we assume a between-group independence covariance matrix $\Sigma = \mathrm{diag}(\Sigma_1,\ldots,\Sigma_G)$, where Σ_g is the covariance for group g ($g = 1,\ldots,G$). Accordingly, we apply a group penalty on μ_k (Yuan and Lin, 2006), resulting in the following penalized log-likelihood

$$L_\lambda = L(\mu_k, \Sigma) - \lambda\sum_{k=1}^{K}\sum_{g=1}^{G} n_k\sqrt{p_g}||\mu_{kg}|| \tag{7}$$

where $||\mu_{kg}|| = \sqrt{\sum_{l \in G_g}\mu^2_{kl}}$ is the L_2 norm of mean vector and p_g is the group size for group g. It can be shown that a sufficient and necessary condition for μ_k to be a maximizer of (7) is

$$\Sigma_g^{-1}\mu_{kg} + \frac{\lambda\sqrt{p_i}}{||\mu_{kg}||}\mu_{kg} = \Sigma_g^{-1}\bar{x}_{kg},\ \ \forall\mu_{kg} \neq \mathbf{0} \tag{8}$$

and

$$||\Sigma_g^{-1}\bar{x}_{kg}|| \leq \lambda\sqrt{p_i},\ \ \mu_{kg} = \mathbf{0} \tag{9}$$

where $\bar{x}_{kg}$ is a $p_g \times 1$ vector whose elements are average gene expressions over class k for each gene that belongs to group g. Equation (8) is a non-linear system and has no closed-form solution. In GSCGRDA, we use

$$\hat{\mu}_{kg} = \left(1 - \frac{\lambda\sqrt{p_i}}{||\Sigma_g^{-1}\bar{x}_{kg}||}\right)_+ \bar{x}_{kg} \tag{10}$$

as an approximate solution to (8). If

$$||\bar{x}_{kg}||\Sigma_g^{-1}\bar{x}_{kg} = \bar{x}_{kg}||\Sigma_g^{-1}\bar{x}_{kg}|| \tag{11}$$

or in other words, if $\bar{x}_{kg}$ is an eigenvector of Σ_g, then solution (10) becomes exact for (8).

The covariance matrix estimator $\tilde{\Sigma} = \alpha\hat{\Sigma}^* + (1-\alpha)\mathbf{D}$ used in GSCGRDA can be regarded as a maximizer of penalized likelihood

$$L_\lambda = L(\mu_k, \Sigma) - \lambda\sum_{g=1}^{G}\mathrm{tr}(D_g\Sigma_g^{-1})$$

with μ_k estimated by $\bar{x}_k$, and $\mathbf{D} = \mathrm{diag}(D_1,\ldots,D_G)$ and $\Sigma = \mathrm{diag}(\Sigma_1,\ldots,\Sigma_G)$. The penalty term $\mathrm{tr}(D_g\Sigma_g^{-1})$ corresponds to prior distribution of Σ_g as an inverse Wishart distribution with mean D_g. A similar idea of deriving $\tilde{\Sigma}$ in (2) as an empirical Bayes estimator was discussed by Srivastava and Kubokawa (2007). In this article we use $D = \hat{\mathrm{D}}$.

2.5 Computational issues

With high-dimensional microarray data it takes too much memory space or even infeasible to invert a $p \times p$ covariance matrix, where p is in the order of thousands or tens of thousands. In order to efficiently and stably invert such a large and potentially sparse matrix, we use the Woodbury formula so that the memory requirement is reduced from inverting a $p \times p$ matrix to an $n \times n$ matrix; the latter is quite small for microarray data with n less than hundreds. The Woodbury formula can be expressed as

$$(\mathbf{A} + \mathbf{U}\mathbf{V}')^{-1} = \mathbf{A}^{-1} - [\mathbf{A}^{-1}\mathbf{U}(\mathbf{I} + \mathbf{V}'\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}'\mathbf{A}^{-1}]$$

For simplicity of the discussion, we replace $\hat{\mathbf{D}}$ with $\mathbf{I}$ in formulas (4) and (5). Let $A = \alpha_2\hat{\Sigma}^* + (1 - \alpha_1 - \alpha_2)\mathbf{I}$, $U = \alpha_1 X/(n - K)$ and $V = X$, then

$$\tilde{\Sigma}^{-1} = (A + UV')^{-1}.$$

Hence, if we have A^{-1}, we can calculate $\tilde{\Sigma}$ by applying the Woodbury formula. Notice that $A = \alpha_2\hat{\Sigma}^* + (1 - \alpha_1 - \alpha_2)\mathbf{I}$ is a block diagonal matrix with each block denoted as A_i and $A^{-1} = \mathrm{Diag}(A_1^{-1}, \ldots, A_G^{-1})$. For each A_i, if $p_i \leq n$, we invert it by the Cholesky decomposition; otherwise, we apply the Woodbury formula to A_i,

$$\begin{aligned} A_i^{-1} &= [\alpha_2\hat{\Sigma}_i + (1 - \alpha_1 - \alpha_2)\mathbf{I}_{p_i}]^{-1} \\ &= [(1 - \alpha_1 - \alpha_2)\mathbf{I}_{p_i} + \frac{\alpha_2}{n - K}X_iX_i')]^{-1} \\ &= a\mathbf{I}_{p_i} - b[X_i(\mathbf{I}_n + bX_i'X_i)^{-1}X_i'] \end{aligned}$$

where $a = 1/(1 - \alpha_1 - \alpha_2)$ and $b = \alpha_2/(n - K)(1 - \alpha_1 - \alpha_2)^2$. In this way, the largest matrix we need to invert is the $n \times n$ matrix $\mathbf{I}_n + bX_i'X_i$, which is computationally affordable, considering n is usually small in microarray data analysis. In the context of discriminant analysis, our final goal is to compute the discriminant score $\delta_k(x)$. If we can obtain $\tilde{\Sigma}^{-1}\mu_k$, which is $p \times K$, then we can compute the discriminant scores for classification.

$$\begin{aligned} \tilde{\Sigma}^{-1}\mu_k &= (A + \frac{\alpha_1}{n - K}XX')^{-1}\mu_k \\ &= A^{-1}\mu_k - c[A^{-1}X(\mathbf{I}_n + cX'A^{-1}X)^{-1}X'A^{-1}\mu_k] \end{aligned}$$

where $c = \alpha_1/(n - K)$, $A^{-1}\mu_k = (A_1^{-1}\mu_{1k}, \ldots, A_G^{-1}\mu_{Gk})^T$ and $A^{-1}\mu_k = (A_1^{-1}X_1, \ldots, A_G^{-1}X_G)^T$. Thus we only need to store $A^{-1}X$ and $A^{-1}\mu_k$, instead of $\tilde{\Sigma}^{-1}$. The above computational strategy proves to be efficient and robust in practice.

3 RESULTS

3.1 Simulation

For evaluation, we compared our methods to PAM and SCRDA on simulated data. As discussed above, the major difference between these methods is the assumption on the form of the covariance matrix. Hence, we considered simulation set-ups with four different covariance matrices. For each simulated dataset in each case, we had two classes and $p = 1000$ variables; there were 50 training samples and 500 test samples for each class; the small training samples reflected typical microarray data sizes while the large test samples were used for obtaining accurate test error rates for the purpose of evaluations only. Gene expression levels X were generated from two multivariate normal distributions with different mean vectors but the same covariance structure. Specifically,

$$X_1 \sim \mathrm{MVN}(\mu_1, \Sigma), X_2 \sim \mathrm{MVN}(\mu_2, \Sigma)$$

μ_1 and μ_2 were $p \times 1$ vector. All the elements in μ_1 were 0. The first 100 elements in μ_2 were randomly drawn from uniform (0,1), $\mu_{2,1},\ldots,\mu_{2,100} \sim U(0,1)$ and the remaining ones were 0. The choices of covariance matrices were respectively,

(1) an identity matrix

(2) a compound symmetric (CS) matrix with $\rho = 0.2$: the correlation between any two genes was ρ

(3) a block diagonal matrix with each block as CS: the block size was 50×50, resulting in a total of 20 blocks. The within-block-wise correlations were $\rho_1,\ldots,\rho_{20} \sim U(0,1)$

(4) a block diagonal matrix plus a weak CS correlation for other off-block elements: the blocking structure was the same as in case 3 with the within-block correlations $\rho_1,\ldots,\rho_{20} \sim U(0.5,1)$, and the correlation for off-block elements was $\rho \sim U(0,0.1)$.

For the purpose of comparison, we also included results for a weighted PAM (wPAM) method (Tai and Pan, 2007a) and support vector machines (SVMs) (Vapnik, 1998). The wPAM is a modification to PAM with multiple shrinkage parameters to accommodate gene groups. We used binary SVMs implemented in the `svm` function in R package e1071; we used the default radial kernel; for each simulated dataset, 10-fold CV was used to select two tuning parameters, the gamma (a parameter with the radial kernel) and the cost parameter C; both parameters were searched on a grid (10^{-5}, 10^{-4},…,10^4, 10^5); because the SVM results were added in the revision, the generated datasets might be different from the ones used for other methods, but we do not expect that the conclusion would change otherwise. Note that SVM was very slow while could not conduct variable selection.

For the grouped methods, we grouped variables according to the blocking structure in case 3 and used this grouping scheme throughout all four simulation set-ups when applying any group-based method. More specifically, we grouped 1000 variables into 20 groups with 50 variables in each group, the first 50 variables in group 1, the next 50 variables in group 2, etc.

To investigate the robustness of the proposed methods to group misspecification, we also considered three scenarios:

- Mixed groups (mix). We randomly chose $m \sim U(0{,}40)$ genes from each group, then randomly reassigned them to one of the 20 groups; in this way, ~40% genes' groups were misspecified.
- New groups (new). We randomly chose $m \sim U(0{,}40)$ genes from each group, and then treated them as separate groups.
- Divided groups (div). We randomly divided each of four groups (1 informative and 3 non-informative ones) into two groups of nearly equal size.

The results are summarized in Table 1. As expected, in general, the method with the correct assumption on the underlying covariance matrix outperformed other methods with incorrect assumptions. There was no big difference among all the methods for the true independence model, perhaps due to the RDA-based methods' flexibility of including the independence model as a special case; surprisingly, GSCGRDA performed even slightly better than PAM. Albeit not really practical for microarray data, the CS case was most suitable for SCRDA, which outperformed other methods; though ISCGRDA-1 had the flexibility of modeling any general covariance structure, it performed slightly worse than SCRDA due to the cost of former's estimating one more tuning parameter in its regularized covariance estimator. The block diagonal-CS model was ideal for the group-based RDA methods, which outperformed SCRDA, PAM and wPAM by significant margins; GSCGRDA performed best in this cases. For the block diagonal CS plus a weak off-block CS case, which was perhaps most representative for real microarray data with stronger within-group correlations and weaker between-group correlations, GSCGRDA was a clear winner, followed by our other two proposed methods; the three new methods, even under largely mis-specified groups, performed much better than PAM and SCRDA. In general, the new methods were robust to group mis-specifications.

Table 1. Simulation results

	Classifier	# errors	# Info	# Non-info
Identity	ISCGRDA-1	22.60	36.43	11.47
	ISCGRDA-2	20.18	38.66	12.64
	GSCGRDA	12.86	90.50	4.00
	SCRDA	18.27	1.47	114.48
	PAM	13.67	54.55	74.39
	wPAM	9.17	62.26	23.23
	SVM	40.81	100	900
CS	ISCGRDA-1	58.45	68.52	189.78
	ISCGRDA-2	148.82	33.49	46.71
	GSCGRDA	113.87	94.00	46.50
	SCRDA	26.49	2.78	831.41
	PAM	144.72	42.11	101.46
	wPAM	145.46	51.64	133.53
	SVM	34.30	100	900
Block CS	ISCGRDA-1	83.39	38.85	34.25
	ISCGRDA-1 (mix)	117.83	23.82	44.68
	ISCGRDA-1 (new)	108.13	27.67	41.73
	ISCGRDA-1 (div)	92.28	26.66	39.34
	ISCGRDA-2	82.38	40.19	35.21
	ISCGRDA-2 (mix)	123.66	21.9	36.94
	ISCGRDA-2 (new)	110.04	26.43	34.67
	ISCGRDA-2 (div)	91.24	29.11	45.59
	GSCGRDA	37.62	82.00	15.50
	GSCGRDA (mix)	87.02	78.59	49.39
	GSCGRDA (new)	78.29	56.90	69.94
	GSCGRDA (div)	38.90	60.58	41.15
	SCRDA	165.22	1.48	229.71
	PAM	221.46	29.27	23.08
	wPAM	232.54	40.20	32.82
	SVM	222.15	100	900
Block CS + CS	ISCGRDA-1	43.72	41.91	30.79
	ISCGRDA-1 (mix)	75.68	26.38	41.22
	ISCGRDA-1 (new)	70.25	33.49	49.41
	ISCGRDA-1 (div)	43.49	33.31	53.79
	ISCGRDA-2	40.89	44.72	31.28
	ISCGRDA-2 (mix)	95.60	26.14	39.46
	ISCGRDA-2 (new)	81.94	31.44	39.46
	ISCGRDA-2 (div)	38.94	33.81	44.89
	GSCGRDA	10.50	74.50	5.50
	GSCGRDA (mix)	47.82	73.67	34.97
	GSCGRDA (non)	45.53	62.99	60.28
	GSCGRDA (div)	12.69	52.28	22.94
	SCRDA	109.84	2.01	273.28
	PAM	280.98	24.54	13.40
	wPAM	298.92	36.43	40.38
	SVM	176.17	100	900

The average numbers of test errors and selected informative genes (# Info) and non-informative genes (# Non-info) for the methods are listed.

PAM and wPAM performed pretty well under the independence case, but suffered severely for other cases because of their ignoring existing correlations. SCRDA performed well in general for all cases, especially when correlations among the genes were fairly strong, but it tended to use more genes than other methods probably because it was unable to capture the sparseness of an underlying covariance matrix. ISCGRDA-1 performed well in all cases due to its generality. However, it suffered from its high computational cost. ISCGRDA-2 performed similarly to ISCGRDA-1 except for the CS case. In terms of prediction error, GSCGRDA outperformed all the other methods except in the CS case, in which it was ranked the second only behind SCRDA. For variable selection, it selected a much higher proportion of informative genes while eliminating much more noise genes. In addition, by using a block covariance structure along with the group shrinkage, GSCGRDA had the capability of genuine gene selection as compared to other RDA-based methods. In particular, GSCGRDA performed better than SVM, which was not really surprising because the former method (along with other

LDA-type methods) used the normality assumption on the predictors.

3.2 Real data

We also applied all methods to four public microarray gene expression datasets for tumor classifications.

(1) Breast cancer data (Huang *et al.*, 2003). There were in total $n = 52$ samples (18 with recurrence of tumor and 34 without). Each sample contained $p = 12\,625$ genes.

(2) Lung cancer data (Gordon *et al.*, 2002). The goal was to discriminate between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. There were $n = 181$ tissue samples (31 MPM and 150 ADCA). Each sample contained $p = 12\,533$ genes.

(3) Prostate cancer data (Singh *et al.*, 2002). The goal was to classify between 77 tumor samples and 59 normal samples. Each sample contained $p = 12\,600$ genes.

(4) Leukemia data (Armstrong *et al.*, 2001). The goal was to classify each of 62 leukemia samples, among which there were 24 acute lymphoblastic leukemia (ALL) samples, 20 acute leukemia samples carrying chromosomal translocations involving the mixed-lineage leukemia gene (MLL), and 28 acute myelogenousleukemia (AML) samples, into one of the three subtypes. There were $p = 12\,582$ genes.

Gene groups were formed based on the KEGG pathways (Kanehisa, 1996): the genes in a KEGG pathway formed a group, while each of the genes that was not annotated in any KEGG pathway formed its own group with group size one. To deal with the genes annotated in multiple pathways, we applied two methods: (1) (the default): we kept a gene to the pathway with the smallest ID while ignoring its belonging to other pathways (2) (dup): we duplicated the expression profile of the gene in each pathway to which it belonged to.

As a pre-screening, we only used various subsets of the genes in each dataset: we used the genes in KEGG pathways along with the top 3000 genes with the largest sample variances. Results in Table 2 were based on 10 independent repeats of 10-fold CV; within each CV, a second-level CV was used to select tuning parameters.

Table 2. Real data results from 10 repeats of 10-fold double CV

	Classifier	Number of errors	Number of genes[a]
Breast cancer ($p = 6116$)	ISCGRDA-1	9.7/52	342.68
	ISCGRDA-2	9.9/52	461.89
	GSCGRDA	9.8/52	2586.52
	GSCGRDA (dup)	9.3/52	1676.56
	SCRDA	12.3/52	5690.21
	PAM	11.2/52	1830.34
	wPAM	11.4/52	259.69
	SVM	10.9/52	–
Lung cancer ($p = 6013$)	ISCGRDA-1	2.8/181	1008.47
	ISCGRDA-2	2.7/181	1153.95
	GSCGRDA	1.8/181	1822.88
	GSCGRDA (dup)	2.2/181	1993.20
	SCRDA	3.8/181	5700.57
	PAM	1.4/181	142.46
	wPAM	0.8/181	6.57
	SVM	1.6/181	–
Prostate cancer ($p = 6163$)	ISCGRDA-1	37.3/136	871.63
	ISCGRDA-2	36.5/136	1349.19
	GSCGRDA	19.3/136	4132.16
	GSCGRDA (dup)	19.3/136	2052.34
	SCRDA	15.1/136	6151.19
	PAM	59.9/136	15.18
	wPAM	53.7/136	10.89
	SVM	10.8/136	–
Leukemia ($p = 6117$)	ISCGRDA-1	6.8/62	1188.07
	ISCGRDA-2	6.6/62	1165.38
	GSCGRDA	2.4/62	2709.51
	GSCGRDA (dup)	2.5/62	2039.53
	SCRDA[b]	–	–
	PAM	6.3/62	3454.83
	wPAM	7.4/62	2458.56
	SVM	4.0/62	–

[a] ISCGRDA-1, ISCGRDA-2 and SCRDA do not have the capability of gene selection. Their given gene numbers were that for non-zero elements of $\hat{\mu}^*_{k(s)}$.
[b] R-package rda failed with error message: 'error code 1 from Lapack routine 'dgesdd' Execution halted'.

The RDA-based methods tended to use more genes than PAM or wPAM; in particular, SCRDA used almost all the genes for each dataset. However, RDA-based methods performed significantly better than PAM and wPAM for the prostate cancer data. The GSCGRDA performed consistently well for all datasets: it was the best for the breast cancer data and leukemia data, the third best for the prostate cancer data and performed closely to the winner (wPAM) for the lung cancer data. It also consistently outperformed other two GRDA-based methods. In addition, the genes selected from GSCGRDA had a biological interpretation: any group of more than one genes selected at the end corresponded to a KEGG pathway. In Table 3, we show the top 10 most frequently selected pathways by GSCGRDA based on 100 models from 10 repeats of 10-fold CV. Since GSCGRDA selected almost all of the genes for the prostate cancer data, we do not list the selected pathways for the data.

It was somewhat reassuring that the two treatments of the genes with multiple annotations in GSCGRDA gave almost equal performance in terms of misclassifications, while the duplication method seemed to select an almost equal number of or fewer unique genes as compared to the first treatment; more work is needed on how to handle the genes with multiple annotations for grouped methods.

To empirically verify the 'stronger within-group and weaker between-group correlations' assumption, based on the lung cancer data, we calculated pairwise Pearson's correlations among the genes within three pathways (with IDs 04514, 04020 and 04360) and the collection of all other genes not in any pathway. For the three pathways, the median correlations were 0.092, 0.095 and 0.097, and the 75th percentiles were 0.169, 0.167, 0.168, respectively, larger than 0.090 and 0.159 for the non-annotated genes.

Table 3. Top 10 frequently selected pathways by GSCGRDA

	Pathway ID	Description	Freq
Breast cancer	04010	MAPK signaling pathway	100
	04360	Axon guidance	100
	04060	Cytokine-cytokine receptor interaction	100
	01430	Cell Communication	100
	04080	Neuroactive ligand-receptor interaction	100
	04730	Long-term depression	100
	04020	Calcium signaling pathway	100
	04510	Focal adhesion	99
	04740	Olfactory transduction	99
	02010	ABC transporters	97
Lung cancer	04514	Cell adhesion molecules	100
	04010	MAPK signaling pathway	96
	04020	Calcium signaling pathway	95
	04360	Axon guidance	90
	04060	Cytokine-cytokine receptor interaction	90
	04140	Regulation of autophagy	87
	03022	Basal transcription factors	85
	04950	Maturity onset diabetes of the young	81
	05120	Epithelial cell signaling in Helicobacter pylori infection	79
	00603	Glycosphingolipid biosynthesis	78
Leukemia	04080	Neuroactive ligand-receptor interaction	100
	04610	Complement and coagulation cascades	100
	04514	Cell adhesion molecules	100
	01430	Cell Communication	100
	04060	Cytokine-cytokine receptor interaction	100
	04010	MAPK signaling pathway	100
	04020	Calcium signaling pathway	100
	00230	Purine metabolism	99
	04360	Axon guidance	99
	02010	ABC transporters	97

4 DISCUSSION

In this article, we have proposed a class of modified LDA to incorporate prior knowledge on gene functions into building a classifier. A main difference from other modifications of LDA is that we regularize the covariance matrix by considering group relationships among variables. Unlike most standard classifiers, which treat all the genes equally a priori, our methods assume that the genes in the same group are more likely to function similarly and thus have correlated expressions while the genes from different functional groups or pathways are more likely to be independent or only weakly correlated. We introduce a between-group independence (i.e. block-diagonal) covariance structure into regularization and put more weight on it to account for the biological belief of higher within-group but lower between-group gene correlations. Another main difference is our consideration of a group shrinkage scheme that tends to retain or remove a whole group of the genes altogether. When gene groups are formed informatively, it may not only improve predictive performance, but also facilitate interpretation of results. Although the genes within such selected pathways may be useful for the purpose of clinical outcome classifications, we cannot determine whether differential expressions of these genes are the cause or only manifestation of different outcomes; in fact, even for the purpose of diagnosis or prognosis, more studies may be still needed to evaluate whether identified genes are really trustworthy predictors.

Among the methods studied, in general, GSCGRDA performed best for our simulated and real data. In addition, an advantage of GSCGRDA over other RDA-based methods is that it can realize gene selection while the others cannot. In particular, gene selection is accomplished at the group level, thus naturally associating selected gene groups to their biological interpretations, e.g. pathways. Furthermore, compared to ISGRDA-1 and SCRDA, GSCGRDA is much less computationally intensive by excluding the use of the unrestrictive sample covariance estimate.

Although we only used the KEGG pathways to form gene groups in the real data examples, other sources of biological knowledge for gene functions or pathways, such as GO, can be also utilized in our proposed methods. However, how to take advantage of the hierarchical structure of GO annotations, or known or predicted gene networks, is unclear. Furthermore, how to handle genes with multiple annotations warrants more research. Finally, it may be productive to combine the idea proposed here with other improved PAM methods (Wang and Zhu, 2007). These are interesting topics to be studied.

ACKNOWLEDGEMENTS

This research was partially supported by NIH grant HL65462 and a UM AHC Faculty Research Development grant. The authors thank the three reviewers and the associate editor for helpful and constructive comments.

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Armstrong, *et al*. (2001) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia'. *Nat. Genet*., **30**, 41–47.

Cai,T. (1999) Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Stat*., **27**, 898–924.

Gordon,J. *et al*. (2002) Translation of microarray data into clinically relevant cancer diagnostic tests using gege expression ratios in lung cancer and mesothelioma. *Cancer Res*., **62**, 4963–4967.

Gui,J. and Li,H. (2005) Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, **21**, 3001–3008.

Guo,Y. *et al*. (2007) Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, **8**, 86–100.

Hastie,T. *et al*. (2001) *The Elements of Statistical Learning. Data mining, Inference, and Prediction*. Springer, New York, USA.

Huang,X. and Pan,W. (2003) Linear regression and two-class classification with gene expression data. *Bioinformatics*, **19**, 2072–2078.

Huang,E. *et al*. (2003) Gene expression predictors of breast cancer outcomes. *Lancet*, **361**, 1590–1596.

Huang,D. and Pan,W. (2006) Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. *Bioinformatics*, **22**, 1259–1268.

Kanehisa,M. (1996) Toward pathway engineering: a new database of genetic and molecular pathway. *Sci. Tech. Jpn.*, **59**, 34–38.

Lottaz,C. and Spang,R. (2005) Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data. *Bioinformatics*, **21**, 1971–1978.

Pan,W. (2005) Incorporating biological information as a prior in an empirical Bayes approach to analyzing microarray data. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article 12.

Pan,W. (2006) Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics*, **22**, 795–801.

Pang,W. *et al.* (2006) Pathway analysis using random forests classification and regression. *Bioinformatics*, **22**, 2028–2036.

Singh,D. *et al.* (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

Srivastava,M.S. and Kubokawa,T. (2007) Comparison of discrimination methods for high dimensional data. *J. Jpn. Stat. Soc.*, **37**, 123–134.

Tai,F. and Pan,W. (2007a) Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms. *Bioinformatics*, **23**, 1775–1782.

Tai,F. and Pan,W. (2007b) Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. *Research report 2008–020*, Division of Biostatisitics, University of Minnesota. Available at http://www.biostat.umn.edu./rrs.php

Tibshirani,R. (1996) Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc.,B*, **58**, 267–288.

Tibshirani,R. *et al.* (2003) Class prediction by nearest shrunken centroids with applications to DNA Microarrays. *Stat. Sci.*, **18**, 104–117.

Vapnik,V. (1998) *Statistical Learning Theory*. Wiley, New York, USA.

Wang,S. and Zhu,J. (2007) Improved centroids estimation for the nearest shrunken centroid classifier. *Bioinformatics*, **23**, 972–979.

Wang,Y. *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671–679.

Wei,Z. and Li,H. (2007) Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics*, **8**, 265–284.

Wu,B. (2006) Differential gene expression detection and sample classification using penalized linear regression models. *Bioinformatics*, **22**, 472–476.

Yuan,M. and Lin,Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B*, **68**, 49–67.