DE GRUYTER DOI 10.1515/sagmb-2012-0017 — Statistical Applications in Genetics and Molecular Biology 2013; 12(3): 347–359

Herbert Pang, Tiejun Tong* and Michael Ng

Block-diagonal discriminant analysis and its bias-corrected rules

Abstract: High-throughput expression profiling allows simultaneous measure of tens of thousands of genes at once. These data have motivated the development of reliable biomarkers for disease subtypes identification and diagnosis. Many methods have been developed in the literature for analyzing these data, such as diagonal discriminant analysis, support vector machines, and k-nearest neighbor methods. The diagonal discriminant methods have been shown to perform well for high-dimensional data with small sample sizes. Despite its popularity, the independence assumption is unlikely to be true in practice. Recently, a gene module based linear discriminant analysis strategy has been proposed by utilizing the correlation among genes in discriminant analysis. However, the approach can be underpowered when the samples of the two classes are unbalanced. In this paper, we propose to correct the biases in the discriminant scores of block-diagonal discriminant analysis. In simulation studies, our proposed method outperforms other approaches in various settings. We also illustrate our proposed discriminant analysis method for analyzing microarray data studies.

Keywords: bias-correction; block-diagonal; classification; high-dimensional data; linear discriminant analysis.

***Corresponding author: Tiejun Tong,** Department of Mathematics, Hong Kong Baptist University, Hong Kong, e-mail: tongt@hkbu.edu.hk
Herbert Pang: Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27710, USA
Michael Ng: Department of Mathematics, Hong Kong Baptist University, Hong Kong

1 Introduction

High-throughput expression profiling allows simultaneous measure of tens of thousands of genes at once. These data have motivated the development of reliable biomarkers for disease subtypes identification and diagnosis, and the identification of novel targets for drug treatment. Many methods have been developed in the literature, such as diagonal linear discriminant analysis (DLDA) (Dudoit et al., 2002), regularized linear discriminant analysis (Guo et al., 2007), random forests (Breiman, 2001; Statnikov et al., 2008), support vector machines (SVM) (Lee et al., 2004; Vapnik and Kotz, 2006), dimension reduction methods (Antoniadis et al., 2003; Dai et al., 2006), nearest shrunken centroids methods (Tibshirani et al., 2002, 2003; Dabney and Storey, 2007; Wang and Zhu, 2007), and penalized linear discriminant analysis (Guo, 2010). These methods have been applied and reviewed by many studies, e.g., Lee et al. (2005), Huang and Zheng (2006), Statnikov et al. (2008) among others.

The three main statistical problems in cancer genomics research are (1) the identification of tumor subtypes; (2) the classification of patients into known classes; and (3) the selection of genes that distinguish between tumor subtypes. In this article, we will focus on the second and third problems. The high dimensional nature of microarray data with the number of genes much larger than the number of samples presents a challenge to classical classification methods, such as the well-known linear discriminant analysis (LDA). To overcome the singularity problem, various approaches that rely on a diagonal approximation to the covariance matrices have been proposed. This leads to the so-called diagonal discriminant analysis which have been widely used for classification in high-dimensional data (Dudoit et al., 2002; Speed, 2003; Ye et al. 2004; Lee et al., 2005; Shieh et al., 2006; Wang and Zhu, 2007; Pang et al., 2009). Due to the small sample size, DLDA performs very well in practice and usually produced lower misclassification rates than more sophisticated classifiers. DLDA is easy to implement and have been widely adopted to analyze high-dimensional data in the fields of science.

The diagonal discriminant methods have been shown to perform well for high-dimensional data with small sample sizes. Despite the popularity of the independence assumption in the literature (Dudoit et al., 2002; Bickel and Levina, 2004; Tong and Wang, 2007; Hwang et al., 2009), it is unlikely to be true in practice. Expression data consist of genes that can be highly co-expressed (Horvath and Dong, 2008; Taylor et al., 2009). In other words, the diagonal covariance matrix assumption can be too restrictive. Recently, Hu et al. (2011) have proposed a gene module based linear discriminant analysis strategy, by utilizing the correlation among genes in discriminant analysis. In essence, their method is a block-diagonal linear discriminant rule. However, this method may have lower accuracy than desired when the samples of the two classes are unbalanced. Unbalanced data sets, in which the number of samples in different subgroups are unevenly distributed, are common in biomedical studies. It is thus desired to correct the biases in the discriminant scores of block-diagonal discriminant analysis. The idea of bias correction in discriminant analysis has been around for some time (Ghurye and Own, 1969; Moran and Murphy, 1979; McLachlan, 1992). Moran and Murphy (1979) proposed several bias correction methods for the plug-in discriminant scores under the condition that the sample size for each class, n_k, is larger than p. However, the improvement of their bias-corrected rules is small (McLachlan, 1992), because the ratio of p/n_k is small. On the other hand, for high-dimensional data like microarray, the ratio p/n_k can be very large. As a result, the block-diagonal linear discriminant analysis may have low prediction accuracy when the design is fairly unbalanced. In this paper, we propose to correct the biases in the discriminant scores of block diagonal discriminant analysis when p is larger than n.

The remainder of the article is organized as follows. In the first part of Section 2, we introduce the problem and briefly review the diagonal discriminant analysis and its bias-corrected version. In the second part of Section 2, we introduce block diagonal discriminant analysis and derive the bias-corrected block diagonal estimators of the discriminant score and show that they dominate the original ones. In Section 3, we conduct realistic simulation studies to investigate the performance of the proposed methods. We then apply them to two real microarray data sets in Section 4. Finally, we conclude the paper in Section 5 with discussions and future directions.

2 Diagonal discriminant analysis and its improvement

Diagonal discriminant rules have been successfully employed for high-dimensional classification problems. However, the diagonal covariance matrix assumption can be too restrictive and it is unlikely to be true in practice. We describe below the diagonal discriminant analysis and block-diagonal discriminant analysis. In this paper, we propose bias-corrected rules for block-diagonal discriminant analysis as an improvement.

2.1 Diagonal discriminant analysis and its improvement

The main purpose of discriminant analysis is to assign an unknown subject to one of K classes on the basis of a multivariate observation $x=(x_1, ..., x_p)^T$, where p is the number of covariates. For simplicity of notation, the class labels y_i are defined to be integers ranging from 1 to K. We assume that there are n_k observations in each class k are to be i.i.d. from a multivariate normal distribution with mean μ_k and covariance matrix Σ_k, for $k=1$, ..., K. The total number of observations is $n=n_1+...+n_K$.

Under the normal distribution assumption, we assign a new subject x to class k which minimizes the following discriminant score $d_k^Q(x)=(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)+\ln|\Sigma_k|-2\ln\pi_k$, where π_k denotes the prior probability of observing a class k member, i.e., we assign x to class $\operatorname{argmin}_k d_k^Q(x)$. This is the so-called quadratic discriminant analysis (QDA). When the covariance matrices are all the same, the $\ln|\Sigma_k|$ term disappears and $\Sigma_k=\Sigma$. We have LDA. Classification rules based on QDA are known to require larger samples than those based on LDA (Wald and Kronmal, 1977) and seem to be more sensitive to violations of the basic assumptions. These so-called "plug-in" estimates are straightforward to compute, but enjoy no optimality properties (Anderson, 1958; Friedman, 1989).

QDA requires that $\min\{n_1, ..., n_K\} \geq p$ to make $\hat{\Sigma}_k$ non-singular. LDA requires that $n \geq p$ to make $\hat{\Sigma}$ non-singular. To overcome the singularity problem, Dudoit et al. (2002) introduced two simplified discrimination rules by assuming independence of components and replacing off-diagonal elements of the sample covariance matrices with zeros. The first rule is called the diagonal quadratic discriminant analysis (DQDA). Specifically, they estimate $\hat{\Sigma}_k = \mathrm{diag}(\hat{\sigma}_{1k}^2, \ldots, \hat{\sigma}_{pk}^2)$, which simplifies the QDA rule to

$$\mathscr{C}(x) = \underset{k}{\operatorname{argmin}} \left(\hat{d}_k^Q(x) = \sum_{i=1}^{p} (x_i - \hat{\mu}_{ik})^2 / \hat{\sigma}_{ik}^2 + \sum_{i=1}^{p} \ln \hat{\sigma}_{ik}^2 - 2 \ln \hat{\pi}_k \right), \tag{1}$$

where $\hat{\pi}_k = n_k / n$. The second rule is DLDA where a common diagonal covariance matrix is assumed. Specifically, they estimate $\hat{\Sigma} = diag(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2)$, which simplifies the LDA rule to

$$\mathscr{C}(x) = \underset{k}{\operatorname{argmin}} \left(\hat{d}_k^L(x) = \sum_{i=1}^{p} (x_i - \hat{\mu}_{ik})^2 / \hat{\sigma}_i^2 - 2 \ln \hat{\pi}_k \right). \tag{2}$$

Due to the small sample size, DLDA and DQDA, perform very well in practice and usually produced lower misclassification rates than more sophisticated classifiers. Both DQDA and DLDA are easy to implement and have been widely adopted to analyze high-dimensional data in the fields of science. See for example Speed (2003), Ye et al. (2004), Noushath et al. (2006), Asyali et al. (2006), Son and Lee (2006), Heilemann and Schuhr (2008), Natowicz et al. (2008), among many others.

DLDA is also called "naive Bayes" classifier because it arises in a Bayesian setting. For more theory on why the "naive Bayes" classifiers which assume independent covariates work well when $p > n$, see Bickel and Levina (2004). Although DLDA and DQDA have been shown to perform well for high-dimensional data with small sample sizes, they suffer from a serious drawback as their discriminant scores are biased. As a consequence, DLDA and DQDA, may result in low prediction accuracy, especially when the design is fairly unbalanced. Huang et al. (2010) proposed bias-correction to improve DLDA and DQDA.

2.2 Block-diagonal discriminant analysis and its improvement

In the setting of microarray data analysis, Storey and Tibshirani (2001) suggested that the clumpy dependence (i.e., the block diagonal matrix) is a likely form of dependence, where the clumpy dependence means that the genes are dependent within groups and independent among groups. This is also mentioned in Langaas et al. (2005); Pang et al. (2009), etc. In recent development, Hu et al. (2011) proposed a clumpy dependence structure linear discriminant analysis method. Inspired by Huang et al. (2010), in this paper we propose bias-corrected rules for block-diagonal discriminant analysis as an improvement.

2.2.1 Block-diagonal discriminant analysis

As before, we assume that there are n_k observations in each class k with

$$x_{k,1}, \ldots, x_{k,n_k} \overset{i.i.d}{\sim} \mathrm{MVN}(\mu_k, \Sigma_k), k = 1, \ldots, K,$$

where μ_k and Σ_k are the corresponding mean vector and covariance matrix. The total number of observations is $n = n_1 + \ldots + n_K$.

Given the block-diagonal nature, we assume that all the genes can be assembled to a total of H groups (e.g., H pathways) with each group size p_h, where $h = 1, \ldots, H$ and $p_1 + \ldots + p_H = p$. Let $\mu_k(h)$ and $\Sigma_k(h)$ be the mean vector and covariance matrix for group h, respectively. Note that the diagonal discriminant analysis is a special case of this general setting with $p_1 = \ldots = p_H = 1$ and $H = p$.

Without loss of generality, we can relabel the genes so that the first p_1 genes are from group 1 (referred to as block 1), the subsequent p_2 genes are from group 2 (referred to as block 2), and so on. Then we have

$$\mu_k=(\mu_k^T(1),\ldots,\mu_k^T(H))^T,$$
$$\Sigma_k=\mathrm{diag}(\Sigma_k(1),\ldots,\Sigma_k(H)).$$

When the covariance matrices Σ_k are all the same, we denote it by Σ which has the form

$$\Sigma=\mathrm{diag}(\Sigma(1),\ldots,\Sigma(H)).$$

Let the new observation be $x=(x^T(1),\ldots,x^T(H))^T$. According to the above notations, the quadratic rule assigns x to class k which minimizes the following discriminant score

$$d_k^{BQ}(x)=\sum_{h=1}^{H}\left(x(h)-\mu_k(h)\right)^T[\Sigma_k(h)]^{-1}\left(x(h)-\mu_k(h)\right)+\sum_{h=1}^{H}\ln|\Sigma_k(h)|-2\ln\pi_k.$$

Similarly, when Σ_k are all the same, it reduces to a linear rule with the discriminant score as following:

$$d_k^{BL}(x)=\sum_{h=1}^{H}(x(h)-\mu_k(h))^T[\Sigma(h)]^{-1}(x(h)-\mu_k(h))-2\ln\pi_k.$$

Note that both $d_k^{BQ}(x)$ and $d_k^{BL}(x)$ involve unknown quantities. To get the sample version rules, we need to have the estimates of π_k, $\mu_k(h)$, $\Sigma_k(h)$ and $\Sigma(h)$, respectively. The same as before, we estimate π_k by $\hat{\pi}_k=n_k/n$. The quantities $\mu_k(h)$, $\Sigma_k(h)$ and $\Sigma(h)$ are estimated by the following sample mean and sample covariance matrices:

$$\hat{\mu}_k(h)=\frac{1}{n_k}\sum_{l=1}^{n_k}x_{k,l}(h),\ h=1,\ldots,H,$$
$$\hat{\Sigma}_k(h)=\frac{1}{n_k-1}\sum_{l=1}^{n_k}(x_{k,l}(h)-\hat{\mu}_k(h))(x_{k,l}(h)-\hat{\mu}_k(h))^T,$$
$$\hat{\Sigma}(h)=\frac{1}{n-K}\sum_{k=1}^{K}(n_k-1)\hat{\Sigma}_k(h),$$

where $x_{k,l}=\left(x_{k,l}^T(1),\ldots,x_{k,l}^T(H)\right)^T$ is the l-th observation from class k.

With the above sample estimates, the quadratic discriminant rule is given as $\mathscr{C}(x)=\underset{k}{\mathrm{argmin}}\,\hat{d}_k^{BQ}(x)$, where

$$\hat{d}_k^{BQ}(x)=\sum_{h=1}^{H}(x(h)-\hat{\mu}_k(h))^T[\hat{\Sigma}_k(h)]^{-1}(x(h)-\hat{\mu}_k(h))+\sum_{h=1}^{H}\ln|\hat{\Sigma}_k(h)|-2\ln\hat{\pi}_k. \tag{3}$$

We refer to it as the block-diagonal quadratic discriminant analysis (BD-QDA). Similarly, the sample version rule for the linear discriminant analysis is $\mathscr{C}(x)=\underset{k}{\mathit{argmin}}\,\hat{d}_k^{BL}(x)$, where

$$\hat{d}_k^{BL}(x)=\sum_{h=1}^{H}(x(h)-\hat{\mu}_k(h))^T[\hat{\Sigma}(h)]^{-1}(x(h)-\hat{\mu}_k(h))-2\ln\hat{\pi}_k. \tag{4}$$

We refer to it as the block-diagonal linear discriminant analysis (BD-LDA).

Finally, we mention that to avoid the singularity of the sample covariance matrices, to make BD-LDA work we require that $\max\{p_1,\ldots,p_H\}\le n$. Similarly, to make BD-QDA work we require that $\max\{p_1,\ldots,p_H\}\le\min\{n_1,\ldots,n_k\}$, a stronger requirement compared to that for BD-LDA especially when the data are fairly unbalanced. This makes BD-LDA a more applicable method for classifying high-dimensional data than BD-QDA.

2.2.2 Bias-corrected block-diagonal discriminant rules

Similarly as in Huang et al. (2010), the proposed block-diagonal discriminant rules, can be shown to suffer from the serious drawback of biased discriminant scores. With this insight, we propose bias-corrected rules

for BD-QDA and BD-LDA, respectively, by removing the biases in the discriminant scores in this section. Technical details for the bias-corrected rules will also be presented.

2.2.2.1 Bias-corrected rules for BD-QDA

For ease of notation, let

$$Q_k(h) = (x(h)-\mu_k(h))^T\,[\Sigma_k(h)]^{-1}(x(h)-\mu_k(h)),$$
$$\hat{Q}_k(h)=(x(h)-\hat{\mu}_k(h))^T[\hat{\Sigma}_k(h)]^{-1}(x(h)-\hat{\mu}_k(h)),$$

where $\hat{Q}_k(h)$ is the estimates of $Q_k(h)$ with h=1, ..., H and k=1, ..., K. Then for BD-QDA, the discriminant score (3) can be rewritten as

$$\hat{d}_k^{BQ}(x)=\sum_{h=1}^{H}\hat{Q}_k(h)+\sum_{h=1}^{H}\ln|\hat{\Sigma}_k(h)|-2\ln\hat{\pi}_k.$$

To calculate the expectations of $\hat{Q}_k(h)$ and $\ln|\hat{\Sigma}_k(h)|$, we employ the following well-known results (Das Gupta, 1968; McLachlan, 1992),

$$E([\hat{\Sigma}_k(h)]^{-1})=\frac{n_k-1}{n_k-p_h-2}([\Sigma_k(h)]^{-1}), \tag{5}$$

$$E(\ln|\hat{\Sigma}_k(h)|)=\ln|\Sigma_k(h)|-p_h\ln(n_k-1)+\sum_{i=1}^{p_h}\Psi\left(\frac{1}{2}(n_k-i)\right), \tag{6}$$

where $\Psi(\cdot)$ is the digamma function Abramowitz and Stegun (1972).

By (5), together with the fact that $\hat{\mu}_k(h)$ and $\hat{\Sigma}_k(h)$ are independent of each other, we have

$$\begin{aligned}E[\hat{Q}_k(h)]&=E\left\{\mathrm{tr}\left[(x(h)-\hat{\mu}_k(h))^T[\hat{\Sigma}_k(h)]^{-1}(x(h)-\hat{\mu}_k(h))\right]\right\}\\&=\mathrm{tr}\left\{E\left[(x(h)-\hat{\mu}_k(h))\,(x(h)-\hat{\mu}_k(h))^T\right]E([\hat{\Sigma}_k(h)]^{-1})\right\}\\&=\frac{n_k-1}{n_k-p_h-2}\left\{\frac{p_h}{n_k}+Q_k(h)\right\}.\end{aligned}$$

where "tr" is the trace of a matrix. This implies that

$$\tilde{Q}_k(h)=\frac{n_k-p_h-2}{n_k-1}\hat{Q}_k(h)-\frac{p_h}{n_k} \tag{7}$$

is an unbiased estimator of $Q_k(h)$ for any h=1, ..., H and k=1, ..., K.

Note that

$$\mathrm{var}(\tilde{Q}_k(h))=\left(\frac{n_k-p_h-2}{n_k-1}\right)^2\mathrm{var}(\hat{Q}_k(h))<\mathrm{var}(\hat{Q}_k(h)).$$

This shows that $\tilde{Q}_k(h)$ has a smaller variance than $\hat{Q}_k(h)$. Then together with the fact that $\tilde{Q}_k(h)$ is also an unbiased estimator of $Q_k(h)$, $\tilde{Q}_k(h)$ is a better estimator as it has a smaller mean squared error (MSE) than $\hat{Q}_k(h)$. Or equivalently, we say that $\tilde{Q}_k(h)$ dominates $\hat{Q}_k(h)$ under the quadratic loss function $L(\theta,\hat{\theta})=(\hat{\theta}-\theta)^2$.

By (6), an unbiased estimator of $\ln|\Sigma_k(h)|$ is given as

$$\ln|\hat{\Sigma}_k(h)|+p_h\ln(n_k-1)-\sum_{i=1}^{p_h}\Psi\left(\frac{1}{2}(n_k-i)\right). \tag{8}$$

The same as above, the new estimator (8) dominates $\ln|\hat{\Sigma}_k(h)|$ under the quadratic loss function.

By (7) and (8), we define the biased-corrected discriminant score of BD-QDA as

$$\tilde{d}_k^{BQ}(x)=\sum_{h=1}^{H}\left[\tilde{Q}_k(h)+\ln|\hat{\Sigma}_k(h)|+p_h\ln(n_k-1)-\sum_{i=1}^{p_h}\Psi\left(\frac{1}{2}(n_k-i)\right)\right]-2\ln\hat{\pi}_k.$$

And correspondingly, we classify the new subject x to class $\underset{k}{\operatorname{argmin}}\,\tilde{d}_k^{BQ}(x)$.

2.2.2.2 Bias-corrected rules for BD-LDA

For ease of notation, let

$$L_k(h)=(x(h)-\mu_k(h))^T\,[\Sigma(h)]^{-1}\,(x(h)-\mu_k(h)),$$
$$\hat{L}_k(h)=(x(h)-\hat{\mu}_k(h))^T[\hat{\Sigma}(h)]^{-1}(x(h)-\hat{\mu}_k(h)),$$

where $\hat{L}_k(h)$ is the estimates of $L_k(h)$ with h=1, ..., H and k=1, ..., K. Then for BD-LDA, the discriminant score (4) can be rewritten as

$$\hat{d}_k^{BL}(x)=\sum_{h=1}^{H}\hat{L}_k(h)-2\ln\hat{\pi}_k.$$

Note that $\hat{\Sigma}(h)=1n-K\sum_{k=1}^{K}(n_k-1)\hat{\Sigma}_k(h)$. It is known that $(n-K)\hat{\Sigma}(h)$ follows a Wishart distribution with $n-K$ degrees of freedom (Das Gupta, 1968; McLachlan, 1992),. That is

$$(n-K)\hat{\Sigma}(h)\sim W_{p_h}(\Sigma(h),n-K),\ h=1,\ldots,H.$$

This leads to

$$E([\hat{\Sigma}(h)]^{-1})=\frac{n-K}{n-K-p_h-1}[\Sigma(h)]^{-1}. \tag{9}$$

By (9), together with the fact that $\hat{\mu}_k(h)$ and $\hat{\Sigma}_k(h)$ are independent of each other, we have

$$\begin{aligned}E[\hat{L}_k(h)]&=E\left\{\mathrm{tr}\left[(x(h)-\hat{\mu}_k(h))^T[\hat{\Sigma}(h)]^{-1}(x(h)-\hat{\mu}_k(h))\right]\right\}\\&=\mathrm{tr}\left\{E\left[(x(h)-\hat{\mu}_k(h))(x(h)-\hat{\mu}_k(h))^T\right]E([\hat{\Sigma}(h)]^{-1})\right\}\\&=\frac{n-K}{n-K-p_h-1}\left\{\frac{p_h}{n_k}+L_k(h)\right\}.\end{aligned}$$

This implies that

$$\tilde{L}_k(h)=\frac{n-K-p_h-1}{n-K}\hat{L}_k(h)-\frac{p_h}{n_k} \tag{10}$$

is an unbiased estimator of $L_k(h)$ for any h=1, ..., H and k=1, ..., K. Noting that

$$\mathrm{var}(\tilde{L}_k(h))=\left(\frac{n-K-p_h-1}{n-K}\right)^2\mathrm{var}(\hat{L}_k(h))<\mathrm{var}(\hat{L}_k(h)),$$

This shows that $\tilde{L}_k(h)$ has a smaller variance than $\hat{L}_k(h)$. Then together with the fact that $\tilde{L}_k(h)$ is also an unbiased estimator of $L_k(h)$, we can conclude that $\tilde{L}_k(h)$ provides to be a better estimator as it has a smaller MSE than $\hat{L}_k(h)$. Or equivalently, we say that $\tilde{L}_k(h)$ dominates $\hat{L}_k(h)$ under the quadratic loss function.

Finally, we define the biased-corrected discriminant score of BD-LDA as

$$\tilde{d}_k^{BL}(x)=\sum_{h=1}^{H}\tilde{L}_k(h)-2\ln\hat{\pi}_k,\ k=1,\ldots,K.$$

And correspondingly, we classify the new subject x to class $\underset{k}{\operatorname{argmin}}\,\tilde{d}_k^{BL}(x)$.

To illustrate how the bias-corrected rules may perform better than the original rules, we consider BD-LDA with K=2. Also assume that n and p_h are given. (i) When the data are balanced, i.e., when n_1=n_2=n/2, by (10) it is easy to verify that

$$\tilde{L}_1(h)>\tilde{L}_2(h) \quad \textit{is equivalent to} \quad \hat{L}_1(h)>\hat{L}_2(h).$$

This indicates that the bias-corrected BD-LDA always provides a similar decision as the original BD-LDA, though $\tilde{L}_k(h)$ is usually a better estimator of $L_k(h)$ than $\hat{L}_k(h)$. (ii) When the data are unbalanced, i.e., when $n_1\neq n_2$, however,

$$\hat{L}_1(h)>\hat{L}_2(h) \quad \textit{does not necessarily lead to} \quad \tilde{L}_1(h)>\tilde{L}_2(h).$$

Therefore, the bias-corrected BD-LDA is likely to provide a different decision compared with the original BD-LDA for unbalanced designs. Such discrepancy may increase when the data are fairly unbalanced, i.e., when n_1 and n_2 are far different from each other.

2.2.3 Gene and network module selection

The selection of the top genes for all the methods was done according to the ratio of between-group to within-group sums of squares (Dudoit et al., 2002). Specifically, the ratio for gene j is:

$$\frac{BSS(j)}{WSS(j)}=\frac{\sum_i\sum_k I(y_i=k)(\bar{x}_{kj}-\bar{x}_{\star j})^2}{\sum_i\sum_k I(y_i=k)(x_{ij}-\bar{x}_{kj})^2}, \tag{11}$$

where $\bar{x}_{kj}$ denotes the average expression level of gene j across samples being in class k, and $\bar{x}_{\star j}$ denotes the average expression level of gene j across all the samples.

Therefore, the same number of genes as the number selected in network modules will be used in the non-block diagonal, support vector machines and k-Nearest Neighbor (k-nn). By doing this, we ensure that the differences in the results is not due to differences in the genes used for classification. For finding network modules, the top genes are then partitioned into groups based on affinity propagation clustering (Frey and Dueck, 2007; Bodenhofer et al. 2011). Affinity propagation clustering determines for each cluster an exemplar, a sample that is most representative for this cluster. Unlike most clustering algorithms (e.g., k-means), this clustering technique allow exemplars to be chosen among the observed data samples without pre-specifying the number of clusters. The affinity propagation clustering requires a similarity matrix. The similarity of data point a to data point b, called $s(a,b)$, is the suitability of point b to serve as the exemplar for data point a. Let $x(a)$ and $x(b)$ are vector measures of points a and b, respectively. If the data are real-valued, a common choice for similarity is negative squared Euclidean distance: $s(a,b)=-(x(a)-(x(b))^T\,(x(a)-x(b))$, where T represents the transpose. A squared negative distances similarity measure was used in accordance with the examples in Frey and Dueck (2007). Another popular distance measure, Gaussian kernel, was included as comparison. Gene network modules are more biologically interpretable and are suitable for the identification of drug targets (Suthram et al., 2010). The use of a non-parametric clustering method based on local shrinking was also investigated (Wang et al., 2007). However, the performance is less desirable than the Affinity propagation method. Therefore, those results are not presented here.

2.2.4 Prediction accuracy

Prediction accuracy is widely used to assess the performance of classifiers. However, when the number of samples in the different classes is fairly unbalanced, commonly used classification methods that favor the majority class may appear to perform better (Qiao and Liu, 2009; Huang et al., 2010). In this article, we apply the class-weighted accuracy (CWA) criterion (Cohen et al., 2006) which is defined as $CWA=\sum_{k=1}^{K} w_k a_k$, where a_k are the per-class predication accuracies and w_k are nonnegative weights with $\sum_{k=1}^{K} w_k = 1$. For simplicity, we assume equal weights, that is, $w_k=1/K$, and set the prior probability $\pi_k=1/K$ as well. This is the same measure used by Huang et al. (2010). Qiao and Liu (2009) referred to the CWA criterion as the mean within group error with one-step fixed weights criterion.

3 Simulation study

We conduct simulation studies in this section to evaluate the proposed bias-corrected BD-LDA. The simulation study is to investigate the misclassification rates by using the above discriminant scores. We examine the misclassification rates using CWA. Different simulation setups were considered to investigate the behavior of the proposed BD-LDA in a controlled manner. We chose SVM with radial basis and k-nn with k=3, two well known classification schemes for comparison. These settings are chosen as they were found to be best in Pang et al. (2009) and Huang et al. (2010).

Realistic covariance matrix structures were used for simulations. We consider two classes of multivariate normal distributions: MVN(μ_1, Σ) and MVN(μ_2, Σ). The covariance matrix Σ is a block-diagonal matrix of size 1000×1000 with each of the $b\times b$ diagonal blocks the compound symmetric matrix Σ_p, and the rest of the matrix is zero, where

$$\Sigma=\begin{pmatrix} \sigma_1^2\Sigma_\rho & 0 & \cdots & \cdots & \cdots & \vdots \\ 0 & \sigma_2^2\Sigma_\rho & 0 & \ddots & \ddots & \vdots \\ \vdots & 0 & \sigma_3^2\Sigma_\rho & 0 & \ddots & \vdots \\ \vdots & \ddots & 0 & \sigma_4^2\Sigma_\rho & 0 & \vdots \\ \vdots & \ddots & \ddots & 0 & \sigma_5^2\Sigma_\rho & \vdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}_{1000\times 1000},$$

with Σ_p as

$$\Sigma_\rho=\begin{pmatrix} 1 & \rho & \cdots & \rho & \rho \\ \rho & 1 & \ddots & \cdots & \rho \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho & \cdots & \ddots & 1 & \rho \\ \rho & \rho & \cdots & \rho & 1 \end{pmatrix}_{b\times b}.$$

Since having some signal and non-signal genes in a pathway is reasonable, block matrices of size $b\times b$ with b=5, 10, and 20 for the respective number of signal genes of (2 or 3), (2 or 3), and (3 or 5) were chosen for different runs. For each gene i, we let μ_i=0 if it is a non-signal gene and μ_i=0.5 if it is a signal gene. Matrices with a compound symmetric structure with ρ=0, 0.25, 0.5, 0.75 and 0.9 were chosen for this setup. Two scenarios of σs were considered. In one case, we set all of the σs to be 1. Therefore, when ρ=0, it reduces to the identity covariance matrix. In the other case, σs are generated from Uniform[0.5, 1.5]. For each simulation, a training set of size n was generated as described above, and a validation set of size $2n$ was generated with the identical setup in order to assess the error rate. At each iteration of simulation, top 50 genes were selected according to the ratio of between-group to within-group sums of squares and are then partitioned based on affinity

propagation described in Section 2.2.3. The mean error rates for each method were obtained by running 300 simulations and taking an average over them. For each setup, we generated training sets of n_1=40 and n_2=10 for the respective validation set of sizes $2n_1$ and $2n_2$.

3.1 Simulation results

Denote BcBD-LDA as the bias-corrected block-diagonal LDA. We see that BcBD-LDA outperforms all other methods when the correlation gets large which closely resembles real microarray data, see Table 1. When ρ is small the results are similar for BcBD-LDA and Bc-DLDA. The pattern of DLDA and BD-LDA is similar to the two we have just discussed. As ρ gets large, BD-LDA outperforms DLDA. SVM and *k*-nn do not perform well under this setting. Their performance also decline when ρ becomes large. Overall, when there is a high correlation among genes, BcBD-LDA should be chosen over Bc-DLDA. Results of the 1000 genes data simulations with 5 signal genes for each of the 50 blocks and with 2 signal genes for each of the 100 blocks with varying ρ values can be found in Figures 1 and 2, respectively. Additional figures of simulations results can be found in Supplementary Materials, Figures S1–10.

We also performed simulations with expression data generated using simulator from the "boost" R package (Dettling, 2004). This program allows us to retain the mean and correlation structure from real data. For this simulation, we used the Shipp et al. (2002) data set. Results are shown in supplementary table S1. Again, the results demonstrate that the superior performance of BcBD-LDA other methods under this setting. We also investigated the settings when half of the genes in each block are signal genes, e.g., Figure S20s block size of 20 has 10 signal genes per block, and Figure S21s block size of 10 has 5 signal genes per block. Under these settings, our approach still performs well when the correlation is high among genes within each block.

Table 1 CWA for simulated data sets 100 blocks with 3 signal genes in each block.

Method	ρ=0	ρ=0.25	ρ=0.5	ρ=0.75	ρ=0.9
DLDA	0.731	0.732	0.732	0.728	0.721
Bc-DLDA	0.774	0.773	0.765	0.755	0.738
k-nn	0.652	0.653	0.645	0.661	0.637
SVM	0.613	0.622	0.623	0.634	0.628
BD-LDA	0.719	0.711	0.719	0.748	0.777
BcBD-LDA	0.752	0.743	0.757	0.799	0.825

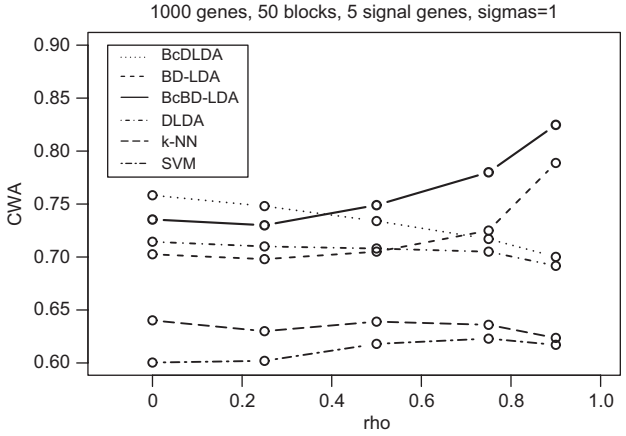

Figure 1 Results of a simulation study with 1000 genes, 50 blocks and 5 signal genes per block, σ s=1, with varying ρ values.

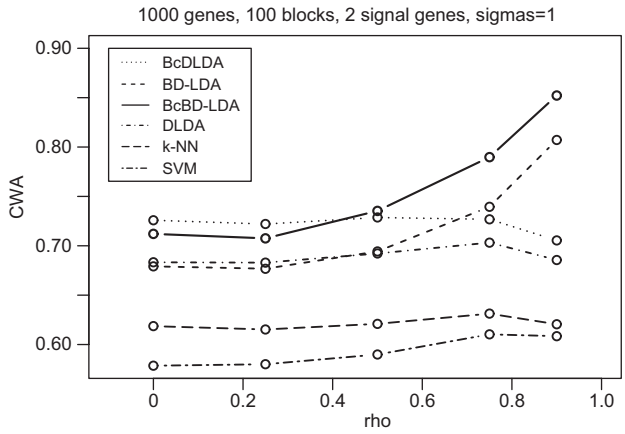

Figure 2 Results of a simulation study with 1000 genes, 100 blocks and 2 signal genes per block, σs=1, with varying ρ values.

4 Applications to microarray data

In this section, we applied the proposed bias-corrected BD-LDA to real data sets and compare them with several other popular classification methods as we did in the simulations studies: 1) Bc-DLDA (Huang et al., 2010), 2) BD-LDA, 3) SVM, and 4) *k*-nn methods. In order to assess the performance of the different methods, we randomly divide the data into a training set and a validation set. Approximately 60% of the samples are assigned to the training set. The rest, about 40%, is used as the validation set to assess the error rate. This process is repeated 100 times. For every training set, gene selection is performed as outlined in Section 2.2.3.

We presented two unbalanced microarray data sets with binary outcomes. Shipp et al. (2002) is a lymphoma study and Wong et al. (2009) is a pediatric critic care study, see Table 2. The Shipp et al. (2002) data set contains 6817 probe sets. It is a study consisting of 58 diffuse large B-cell lymphoma and 19 follicular lymphoma pretreatment biopsy samples. The goal of their study is to differentiate between the two lymphoma subtypes. Although the two types of lymphoma have different responses to cancer therapy, they share similar morphologic and clinical features over time. The Wong et al. (2009) data set is based on the Affymetrix HGU-133plus2 chip with over 55,000 probe sets. It is an unbalanced design study with 18 normal control and 67 pediatric patients with septic shock on the first day to understand pediatric septic shock of critically ill children with sepsis and septic shock.

When we consider the top 50 genes for building the network modules, BD-LDA performed worst in Shipp et al. (2002) that was used as illustrations in Huang et al. (2010). The non-network module based methods, DLDA, Bc-DLDA, SVM, *k*-nn will have the same genes as what we used in network modules. For Shipp et al. (2002), we see that BcBD-LDA outperforms all of the other methods across including bias-corrected DLDA. An example of the network modules for this data set, can be found in Figure 3. For the Wong et al. (2009) data set, similarly, top 50 genes were used for building the network modules. Again, we see that BcBD-LDA performs

Table 2 CWA for two real microarray data sets.

Method	Shipp et al. (2002)	Wong et al. (2009)
DLDA	0.828	0.869
Bc-DLDA	0.832	0.870
k-nn	0.835	0.891
SVM	0.824	0.884
BD-LDA	0.812	0.884
BcBD-LDA	0.868	0.894

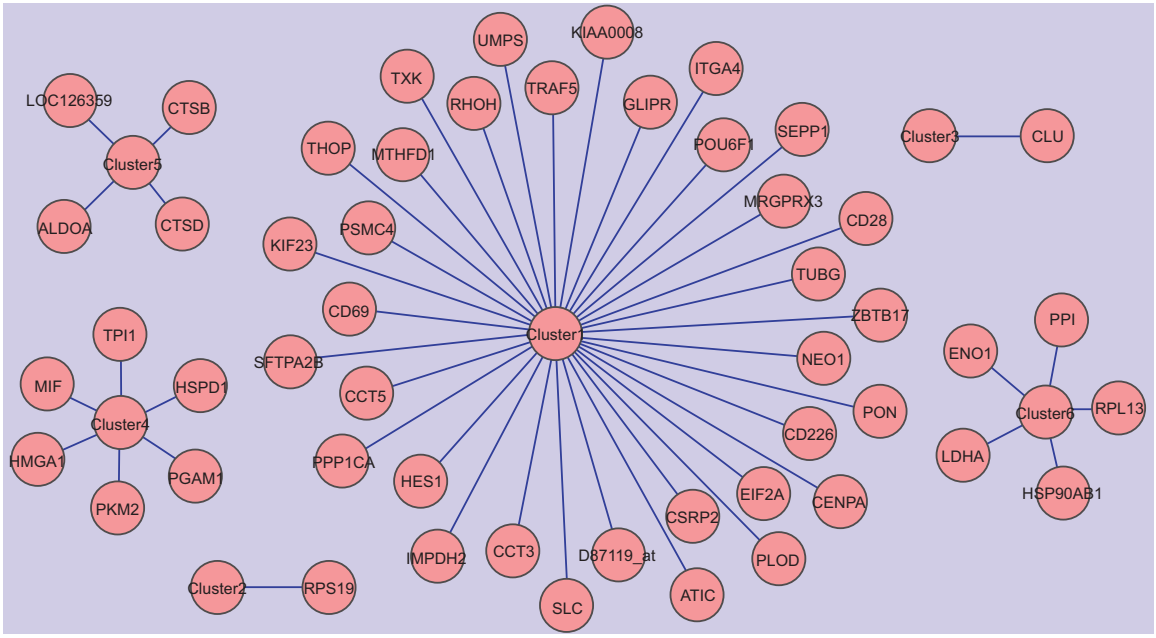

Figure 3 Gene network modules for Shipp et al. Lymphoma study.

best among all the methods. DLDA did worst in this data set. However, the *k*-nn faired quite well compared with BcBD-LDA. Overall, BcBD-LDA has smaller misclassification rates than all the other methods and clearly beats the discriminant analysis-based counterparts. Network modules for this data set, can be found in Supplementary Materials Figure S11.

5 Discussion

In this paper, we proposed new discriminant analysis rules that aim to better classify subjects and provide more precise prediction in gene expression data for unbalanced data sets than existing methods. This is done through the development of bias-corrected discriminant rules that allows non-independence blocks structure of genes. This approach overcomes the strong independence assumptions of diagonal discriminant analysis and corrects for bias in the case of imbalance subjects in different subgroups. It aids the interpretation of genes in microarray data as they tend to work in pathways rather than individually.

We have illustrated in simulations that under moderate correlation, our bias-corrected block-diagonal discriminant rules outperform existing methods. The simulation studies do attempt to model real situation in experimental gene expression data. In addition, we demonstrate the advantages of using our new method on two real microarray data.

Further work can be considered. For example, Pang et al. (2009) demonstrated good performance of the shrinkage-based discriminant rules. Their method's performance is enhanced mainly due to reduced variances in the discriminant scores. However, the bias terms still remain for unbalanced data. Therefore, one may propose to correct the biases for the shrinkage-based block-diagonal discriminant rules.

Acknowledgments: Herbert Pang's research was supported in part by National Institute of Health under Award P01CA142538 and funds from DUMC. Tiejun Tong's research was supported in part by Hong Kong Research Grants Council under Grant 202711 and HKBU FRGs. Michael Ng's research was supported in part by Hong Kong Research Grants Council under Grant 201508 and HKBU FRGs. The authors are grateful to the editor, the associate editor, and two reviewers for their constructive comments and suggestions that have led to a substantial improvement in the article.

References

Abramowitz, M. and I. Stegun (1972): Handbook of mathematical functions. New York: Dover.
Anderson, T. W. (1958): An Introduction to multivariate analysis. New York: John Wiley.
Antoniadis, A., S. Lambert-Lacroix and F. Leblanc (2003): "Effective dimension reduction methods for tumor classification using gene expression data," Bioinformatics, 19, 563–570.
Asyali, M. H., D. Colak, O. Demirkaya and M. S. Inan (2006): "Gene expression profile classification: a review," Curr. Bioinformatics, 1, 55–73.
Bickel, P. J. and E. Levina (2004): "Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations," Bernoulli, 10, 989–1010.
Bodenhofer, U., A. Kothmeier and S. Hochreiter (2011): "APCluster: an R package for affinity propagation clustering," Bioinformatics, 27, 2463–3464.
Breiman, L. (2001): "Random forests," Mach. Learn., 45, 5–32.
Cohen, G., M. Hilario, H. Sax, S. Hugonnet and A. Geissbuhler (2006): "Learning from imbalanced data in surveillance of nosocomial infection," Artif. Intell. Med., 37, 718.
Dabney, A. R. and J. D. Storey (2007): "Optimality driven nearest centroid classification from genomic data," PLoS ONE, 2, e1002.
Dai, J., L. Lieu and D. Rocke (2006): "Dimension reduction for classification with gene expression microarray data," Stat. Appl. Genetics Mol. Biol., 5, 6.
Das Gupta, S. (1968): "Some aspects of discrimination function coefficients," Sankhya, 30, 387–400.
Dettling, M. (2004): "Bagboosting for tumor classification with gene expression data," Bioinformatics, 20, 3583–3593.
Dudoit, S., J. Fridlyand and T. P. Speed (2002): "Comparison of discrimination methods for the classification of tumors using gene expression data," J. Am. Stat. Assoc., 97, 77–87.
Frey, B. and D. Dueck (2007): "Clustering by passing messages between data points," Science, 315, 972–976.
Friedman, J. H. (1989): "Regularized discriminant analysis," J. Am. Stat. Assoc., 84, 165–175.
Ghurye, S. G. and I. Own (1969): "Unbiased estimation of some multivariate probability densities and related functions," Ann. Math. Stat., 40, 1261–1271.
Guo, J. (2010): "Simultaneous variable selection and class fusion for high-dimensional linear discriminant analysis," Biostatistics, 11, 599–608.
Guo, Y., T. Hastie and R. Tibshirani (2007): "Regularized linear discriminant analysis and its application in microarrays," Biostatistics, 8, 86–100.
Heilemann, U. and R. Schuhr (2008): "On the evolution of German business cycles 1958–2004," Jahrbucher fur Nationalokonomie und Statistik, 228, 84–109.
Horvath, S. and J. Dong (2008): "Geometric interpretation of gene coexpression network analysis," PLoS Comput. Biol., 4, e1000117.
Hu, P., S. Bull and H. Jiang (2011): "Gene network modules-based liner discriminant analysis of microarray gene expression data," Lect. Notes Comput. Sci., 6674, 286–296.
Huang, D. and C. Zheng (2006): "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data," Bioinformatics, 22, 1855–1862.
Huang, S., T. Tong and H. Zhao (2010): "Bias-corrected diagonal discriminant rules for high-dimensional classification," Biometrics, 66, 1096–1106.
Hwang, J. T. G., J. Qiu and Z. Zhao (2009): "Empirical Bayes confidence intervals shrinking both means and variances," J. Roy. Stat. Soc. B, 71, 265–285.
Langaas, M., B. H. Lindqvist and E. Ferkingstad (2005): "Estimating the proportion of true null hypotheses, with application to DNA microarray data," J. Roy. Stat. Soc., B, 67, 555–572.
Lee, J. W., J. B. Lee, M. Park and S. H. Song (2005): "An extensive comparison of recent classification tools applied to microarray data," Comput. Stat. Data An., 48, 869–885.
Lee, Y. K., Y. Lin and G. Wahba (2004): "Multicategory support vector machines: theory and application to the classification of microarray data and satellite radiance data," J. Am. Stat. Assoc., 99, 67–81.
McLachlan, G. J. (1992): Discriminant Analysis and Statistical Pattern Recognition. New York: Wiley-Interscience, John Wiley & Sons.
Moran, M. A. and B. J. Murphy (1979): "A closer look at two alternative methods of statistical discrimination," Appl. Stat., 28, 223–232.
Natowicz, R., R. Incitti, E. G. Horta, B. Charles, P. Guinot, K. Yan, C. Coutant, F. Andre, L. Pusztai and R. Rouzier (2008): "Prediction of the outcome of preoperative chemotherapy in breast cancer using DNA probes that provide information on both complete and incomplete responses," BMC Bioinformatics, 9, 149.
Noushath, S., G. H. Kumar and P. Shivakumara (2006): "Diagonal Fisher linear discriminant analysis for efficient face recognition," Neurocomputing, 69, 1711–1716.

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

Qiao, X. and Y. Liu (2009): "Adaptive weighted learning for unbalanced multicategory classification," Biometrics, 65, 159–168.

Shieh, G. S., Y. C. Jiang and Y. S. Shih (2006): "Comparison of support vector machines to other classifiers using gene expression data," Commun. Stat. Simul. C., 35, 241–256.

Shipp, M. A., K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C. Aster and T. R. Golub (2002): "Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," Nat. Med., 8, 68–74.

Son, B. and Y. Lee (2006): "The fusion of two user-friendly biometric modalities: iris and face," IEICE T. Inf. Syst., e89- d, 372–376.

Speed, R. (2003): Statistical analysis of gene expression microarray data. London: Chapman and Hall.

Statnikov, A., L. Wang and C. F. Aliferis (2008): "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification," BMC Bioinformatics, 9, 319.

Storey, J. D. and R Tibshirani (2001): Estimating the positive false discovery rate under dependence, with applications to DNA microarrays. *Technical Report 2001–28,* Department of Statistics, Stanford University.

Suthram, S., J. Dudley, A. Chiang, R. Chen, T. Hastie and A. Butte (2010): "Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets," PLoS Comput. Biol., 6, e1000662.

Taylor I, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris and J. Wrana (2009): "Dynamic modularity in protein interaction networks predicts breast cancer outcome," Nat. Biotechnol., 27, 199–204.

Tibshirani, R., T. Hastie, B. Narasimhan and G. Chu (2002): "Diagnosis of multiple cancer types by shrunken centroids of gene expression," Proc. Natl. Acad. Sci., 99, 6567–6572.

Tibshirani, R., T. Hastie, B. Narasimhan and G. Chu (2003): "Class prediction by nearest shrunken centroids, with applications to DNA microarrays," Stat. Sci., 18, 104–117.

Tong, T. and Y. Wang (2007): "Optimal shrinkage estimation of variances with applications to microarray data analysis," J. Am. Stat. Assoc., 102, 113–122.

Vapnik, V. and S. Kotz (2006): Estimation of Dependences Based on Empirical Data. New York: Springer.

Wald, P. M. and R. A. Kronmal (1977): "Discriminant functions when covariates are unequal and sample sizes are moderate," Biometrics, 33, 479–484.

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

Wang, S., W. Qiu and R. Zamar (2007): "Clues: a non-parametric clustering method based on local shrinking," Comput. Stat. Data An., 52, 286–298.

Wong, H., N. Cvijanovich, G. Allen, R. Lin, N. Anas, K. Meyer, R. Freishtat, M. Monaco, K. Odoms, B. Sakthivel, T. Shanley and Genomics of Pediatric SIRS/Septic Shock Investigators (2009): "Genomic expression profiling across the pediatric systemic inflammatory response syndrome, sepsis, and septic shock spectrum," Crit. Care Med., 37, 1558–1566.

Ye, J., T. Li, T. Xiong and R. Janardan (2004): "Using uncorrelated discriminant analysis for tissue classification with gene expression data," IEEE/ACM Trans. Comput. Biol. Bioinform., 1, 181–190.