Comparison of Discrimination Methods for High Dimensional Data

M.S.Srivastava[1] and T.Kubokawa[2]

University of Toronto and University of Tokyo

Abstract

In microarray experiments, the dimension p of the data is very large but there are only few observations N on the subjects/patients. In this article, the problem of classifying a subject into one of the two groups, when p is large, is considered. Three procedures based on Moore-Penrose inverse of the sample covariance matrix and an empirical Bayes estimate of the precision matrix are proposed and compared with the DLDA procedure.

Key Words and Phrases: Classification, discrimination analysis,minimum distance, Moore-Penrose inverse

1 Introduction

Dudoit, Fridlyand, and Speed (2002) compares several discrimination methods for the classification of tumors using gene expression data. The comparison includes the Fisher (1936)'s linear discrimination analysis methods (FLDA), classification and regression tree (CART) method of Breiman, Friedman, Olshen, and Stone (1984), aggregating classifiers of Breiman (1996) and Breiman (1998) which include "bagging" methods of Friedman(1998) and "boosting" method of Freund and Schapire (1997). The comparison also included two more methods called DQDA method and DLDA method respectively. In DQDA method, it is assumed that the population covariances are diagonal matrices, but unequal for different groups. The likelihood ratio rule is obtained assuming that the parameters are known, and then estimates are substituted in the

[1]Professor, Department of Statistics, University of Toronto, 100 St George Street, Toronto, Ontario, CANADA M5S 3G3, E-Mail: srivasta@utstat.utstat.toronto.edu phone:416-978-4450, fax:416-978-5133

[2]Professor, Faculty of Economics, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN, E-Mail: tatsuya@e.u-tokyo.ac.jpFaculty phone:+81-3-5841-5656, fax:+81-3-5841-5521

likelihood ratio rule. On the other hand, in DLDA method, it is assumed that the population covariance are not only diagonal matrices but they are all equal and the rule is obtained in the same manner as in DQDA. However, among all the preceding methods considered by Dudoit, Fridlyand, and Speed (2002), only DLDA did well. While it is not possible to give reasons as to why other methods did not perform well, the poor performance of the FLDA method may be due to the large dimension p of the data even when the degrees of freedom associated with the sample covariance $n > p$. In large dimensions, the sample covariance may become near singular with very small eigenvalues. For this reason, it may be reasonable to consider a version of the principal component method which is applicable even when $p > n$. Using the Moore-Penrose inverse, a general method based on minimum distance rule is proposed. Another method which uses an empirical Bayes estimate of the inverse of the covariance matrix along with a variation of this method are also proposed. We compare these three new methods with DLDA method of Dudoit, Fridlyand, and Speed (2002).

The organization of the paper is as follows. In Section 2 we describe the three procedures, along with DLDA, and in Section 3, we show that the estimate of Σ^{-1}, used in the two procedures, is an empirical Bayes estimate. Section 4 presents some simulation studies while Section 5 analyzes two microarray datasets which were considered by Dudoit, Fridlyand, and Speed (2002) and Dettling and Buhlmann (2002). The paper concludes in Section 6.

2 Methods of Classification

Let $\mathbf{x}_0$ be an observation vector on an individual belonging to the group C_0. It is known that $\mathbf{x}_0$ belongs to either to group C_1 or to group C_2. Independent observation vectors $\mathbf{x}_{11}, \cdots, \mathbf{x}_{1N_1}$ and $\mathbf{x}_{21}, \cdots, \mathbf{x}_{2N_1}$ are obtained from the two groups C_1 and C_2 respectively. We shall assume that $\mathbf{x}_{1i}$ are i.i.d. $N_p(\mu_{\mathbf{1}}, \Sigma)$, $i = 1, \cdots, N_1$, and $\mathbf{x}_{2i}$ are i.i.d. $N_p(\mu_{\mathbf{2}}, \Sigma)$, $i = 1, \cdots, N_2$, where $N_p(\theta, \Lambda)$ denotes the p-dimensional multivariate normal distribution with mean vector θ and covariance matrix Λ. We estimate $\mu_{\mathbf{1}}$, $\mu_{\mathbf{2}}$

and Σ by

$$\overline{\mathbf{x}}_1 = N_1^{-1} \sum_{i=1}^{N_1} \mathbf{x}_{1i}, \qquad \overline{\mathbf{x}}_2 = N_2^{-1} \sum_{i=1}^{N_2} \mathbf{x}_{2i} \tag{1}$$

and

$$S = n^{-1}[\sum_{i=1}^{N_1} (\mathbf{x}_{1i} - \overline{\mathbf{x}}_1)(\mathbf{x}_{1i} - \overline{\mathbf{x}}_1)' + (\mathbf{x}_{2i} - \overline{\mathbf{x}}_2)(\mathbf{x}_{2i} - \overline{\mathbf{x}}_2)'] \tag{2}$$

where $n = N_1 + N_2 - 2$

In the classification procedures that we will describe next, we may not need to consider all the p characteristics. We may select the variables according to the largest values of t^2-statistic, i.e. selecting $\tilde{p}$ characteristics (out of p) corresponding to the $\tilde{p}$ largest values of the statistic

$$t_i^2 = (\overline{x}_{1i} - \overline{x}_{2i})^2 / s_{ii}, \quad i = 1, \cdots, p \tag{3}$$

where

$$\overline{\mathbf{x}}_1 = (\overline{x}_{11}, \cdots, \overline{x}_{1p})', \quad \overline{\mathbf{x}}_2 = (\overline{x}_{21}, \cdots, \overline{x}_{2p})' \tag{4}$$

and s_{ii} ,$i = 1, \cdots, p$ are the diagonal elements of S.

Although no suitable criterion is yet available to select an optimum value of $\tilde{p}$, we demonstrate the performance of our discrimination procedures for several selected values of $\tilde{p}$.

Next, we describe the three new discrimination methods, along with DLDA method.

2.1 Minimum Distance Rule using Moore-Penrose Inverse(MDMP)

For any matrix A, the Moore-Penrose inverse of A is defined by A^+ satisfying the following four conditions:

(i) $AA^+A = A$

(ii)$A^+AA^+ = A^+$

(iii)$(AA^+)' = AA^+$

(iv)$(A^+A)' = A^+A$

The Moore-Penrose inverse is unique and if A is a square nonsingular matrix, $A^+ = A^{-1}$. Thus, we shall consider the Moore-Penrose inverse of the symmetric and at least

positive semi-definite sample covariance matrix S. When $n > p$, then theoretically S is positive definite with probability one and S^+ becomes S^{-1}. However, when p is large, then even when $n > p$, S behaves like a near singular matrix. Let us first consider the case when $n > p$, the $p \times p$ sample covariance matrix S can be written as

$$S = H'LH \tag{5}$$

where $HH' = I_p = H'H$, that is H is a $p \times p$ orthogonal matrix and $L = \text{diag}(l_1, \cdots, l_p)$ is a diagonal matrix with $l_1 > \cdots > l_p$ as the ordered eigenvalues of the sample covariance matrix S. We partition L and H as follows:

$$L = \begin{pmatrix} L_1 & 0 \\ 0 & L_2 \end{pmatrix} \text{ and } H' = (H_1', H_2'), \tag{6}$$

where $L_1 : r \times r, L_2 : (p-r) \times (p-r)$ and H_1' is a $p \times r$ matrix such that $H_1'H_1 = I_r$. Our L_2 will consist of approximately 5% of the smaller eigenvalues. We shall approximate S by

$$S_a = H_1'L_1H_1. \tag{7}$$

The Moore-Penrose inverse of S_a is given by

$$S_a^+ = H_1'L_1^{-1}H_1. \tag{8}$$

When $n < p$, then H is an $n \times p$ semi-orthogonal matrix $HH' = I_n$, and

$$S = H'LH, \tag{9}$$

where $L : n \times n$, $H : n \times p$. We partition H and L in the same manner as above and approximate S by S_a as given above. Note that now $L_2 : (n-r) \times (n-r)$ and H_2 is $(n-r) \times p$ matrix. We define the sample distance between the observation vector $\mathbf{x}_0$ that is to be classified, to the group C_i by

$$d_i^{MP} = (\mathbf{x}_0 - \overline{\mathbf{x}}_i)'S_a^+(\mathbf{x}_0 - \overline{\mathbf{x}}_i), \qquad i = 1, 2 \tag{10}$$

If $d_1^{MP} < d_2^{MP}$ then x_0 is classified into group C_1, otherwise into group C_2.This is our proposed minimum distance rule. We shall refer to this rule as minimum distance

Moore-Penrose rule(MDMP). In terms of eigenvalues and eigenvectors, the above distance defined in (10) can be written as

$$d_i^{MP} = (\mathbf{x}_0 - \overline{\mathbf{x}}_i)' H_1' L_1^{-1} H_1 (\mathbf{x}_0 - \overline{\mathbf{x}}_i), \qquad i = 1, 2 \tag{11}$$

2.2 Minimum Distance Rule using Empirical Bayes Estimate of Σ

In Section 3 we show that an empirical Bayes estimator of Σ^{-1} is given by

$$\hat{\Sigma}_\lambda^{-1} = c(S + \frac{\text{tr}(S)}{\min(n, p)} I)^{-1} \tag{12}$$

where c is a constant whose value is not important in our investigation. It may be noted that $\hat{\Sigma}_\lambda^{-1}$ exists irrespecture of whether $n < p$ or $n > p$. Thus we define a sample distance between $\mathbf{x}_0$ and the group C_i by

$$d_i^{EB} = (\mathbf{x}_0 - \overline{\mathbf{x}}_i)' \hat{\Sigma}_\lambda^{-1} (\mathbf{x}_0 - \overline{\mathbf{x}}_i), \qquad i = 1, 2 \tag{13}$$

We shall refer to this rule as Minimum Distance Empirical Bayes Rule(MDEB).

2.3 Minimum Distance Rule Using Modified Empirical Bayes Estimate of Σ

In this section we consider the distance function defined in equation (13) using the modified values of L. That is, we define the sample distance between $\mathbf{x}_0$ and group C_i by

$$d_i^{MEB} = (\mathbf{x}_0 - \overline{\mathbf{x}}_i)' H_1' (L_1 + \frac{\text{tr}(L_1)}{r} I)^{-1} H_1 (\mathbf{x}_0 - \overline{\mathbf{x}}_i), \qquad i = 1, 2 \tag{14}$$

We shall refer to this rule as Minimum Distance Modified Empirical Bayes Rule(MDMEB).

2.4 Diagonal Linear discrimination Analysis Method

Finally, we describe the DLDA method of Dudoit, Fridlyand, and Speed (2002) in terms of our notations defined in above equations. In DLDA method, the distance function is defined by

$$d_i^{DLDA} = \sum_{j=1}^{p} \frac{(x_0 - \overline{x}_{ij})^2}{s_{jj}}, \qquad i = 1, 2 \tag{15}$$

3 Empirical Bayes Estimator of Σ^{-1}

In this section we derive an empirical Bayes estimate of Σ^{-1}. Let

$$V = nS \tag{16}$$

where $n = N_1 + N_2 - 2$. Then it can be shown by using an orthogonal transformation (see,e.g. Page 78 of Srivastava and Khatri (1979)) that

$$V = YY' \tag{17}$$

where $Y = (\mathbf{y}_1, \cdots, \mathbf{y}_n)$ and $\mathbf{y}_i$ are i.i.d. $N_p(\mathbf{0}, \Sigma)$. Thus, the joint p.d.f. of $\mathbf{y}_1, \cdots, \mathbf{y}_n$ given Σ^{-1} is given by

$$f(Y \mid \Sigma^{-1}) = (2\pi)^{-\frac{1}{2}pn}|\Sigma^{-1}|^{\frac{1}{2}n}\text{etr}(-\frac{1}{2}\Sigma^{-1}YY') \tag{18}$$

where $\text{etr}(A)$ stands for the exponential of the trace of the matrix A. For the prior distribution of Σ^{-1}, we assume that Σ^{-1} has a Wishart distribution with mean $\lambda^{-1}I$, $\lambda > 0$, and degree of freedom $l \geq p$, i.e.

$$g(\Sigma^{-1}) = c(p,l)|\lambda^{-1}I|^{-\frac{l}{2}}|\Sigma^{-1}|^{\frac{l-p-1}{2}}\text{etr}(-\frac{1}{2}\lambda\Sigma^{-1}) \tag{19}$$

where

$$c(p,l) = \left[2^{\frac{pr}{2}}\Gamma_p(\frac{1}{2}l)\right]^{-1}, \quad \Gamma_p(\frac{1}{2}l) = \pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\Gamma(\frac{l-i+1}{2}) \tag{20}$$

Thus, the joint p.d.f. of Y and Σ^{-1} is given by

$$(2\pi)^{-\frac{pn}{2}}c(p,l)\lambda^{\frac{lp}{2}}|\Sigma^{-1}|^{\frac{n+l-p-1}{2}}\text{etr}(-\frac{1}{2}\Sigma^{-1}(\lambda I + YY')) \tag{21}$$

Hence, the marginal distribution of Y is given by

$$C_1\lambda^{\frac{lp}{2}}|\lambda I + YY'|^{-\frac{n+l}{2}} \tag{22}$$

where

$$C_1 = (2\pi)^{-\frac{pn}{2}}c(p,l)/c(p,n+l) \tag{23}$$

Thus the conditional distribution of Σ^{-1} given Y is given by

$$c(p,n+l)|\lambda I + YY'|^{\frac{n+l}{2}}|\Sigma^{-1}|^{\frac{n+l-p-1}{2}}\text{etr}(-\frac{1}{2}\Sigma^{-1}(\lambda I + YY')) \tag{24}$$

which is the p.d.f. of a Wishart distribution with mean $(\lambda I + YY')^{-1}$ and degree of freedom $n + l$. From this the Bayes estimator is obtained as

$$\begin{aligned} E[\Sigma^{-1} \mid Y] &= (n+l)(\lambda I + YY')^{-1} \quad (25) \\ &= \frac{n+l}{n}(n^{-1}\lambda I + S)^{-1} \quad (26) \end{aligned}$$

To obtain an empirical Bayes estimator, we need an estimator of λ which can be obtained from the marginal distribution of Y given in (22). We may use any resonable method to obtain an estimate of λ, such as the maximum likelihood method, and the method of moments.

When $n > p$, then from the marginal p.d.f. of Y, given in (22),and Lemma 3.2.3 of Srivastava and Khatri (1979), we get the p.d.f. of $V = YY'$ as

$$\begin{aligned} & C_1 \frac{\pi^{\frac{1}{2}pn}}{\Gamma_p(n/2)} \lambda^{\frac{lp}{2}} |V|^{\frac{n-p-1}{2}} \ |\lambda I + V|^{-\frac{n+l}{2}} \\ = \ & \frac{c(p,n)c(p,l)}{c(n,n+l)} |\lambda^{-1}V|^{\frac{n-p-1}{2}} \ |I + \lambda^{-1}V|^{-\frac{n+l}{2}} \lambda^{-\frac{p(p+1)}{2}} \quad (27) \end{aligned}$$

Hence, from page 92 of Srivastava and Khatri (1979), we get

$$\begin{aligned} E[|\lambda^{-1}V|] &= \frac{c(p,n)c(p,l)}{c(p,n+l)} \frac{c(p,n+l}{c(p,n+2)c(p,l-2)} \\ &= \frac{\prod_{i=1}^{p} \Gamma(\frac{n-i+1}{2})}{\prod_{i=1}^{p} \Gamma(\frac{l-2+i+1}{2})} \quad (28) \end{aligned}$$

which is equal to one if $l = n + 2$. In any case, its exact value is not important in our investigation and we choose it equal to one. Thus, using the method of moments, we can estimate λ by

$$\hat{\lambda} = |V|^{1/p} \quad (29)$$

that is, by the geometric mean of the eigenvalues of nS. For convenience, however, we use the arithmetic mean of the non-zero eigenvalues of nS, which can be written as $\frac{n\text{tr}(S)}{p}$.

Thus, when $n > p$, an empirical Bayes estimator of Σ^{-1} is given by

$$\hat{\Sigma}^{-1} = c(S + \frac{\text{tr}S}{p})^{-1} \quad (30)$$

where c is a constant which depends on (n, l). When $n < p$, from the marginal distribution of Y given in (22), we obtain the p.d.f. of $W = Y'Y$. This is given by

$$\begin{aligned} & \frac{c_1 \pi^{(pn)/2}}{\Gamma_n(\frac{p}{2})} \lambda^{-\frac{1}{2}np} |W|^{\frac{p-n-1}{2}} \ |I + \lambda^{-1} W|^{-\frac{n+l}{2}} \\ = \ & \frac{c_1 \pi^{(pn)/2}}{\Gamma_n(\frac{p}{2})} |\lambda^{-1} W|^{\frac{p-n-1}{2}} \ |I + \lambda^{-1} W|^{-\frac{n+l}{2}} \ \lambda^{-\frac{n(n+1)}{2}} \end{aligned} \tag{31}$$

Hence,

$$E[\ |\lambda^{-1} W|\ ] = D_2(l, n, p) \tag{32}$$

where D_2 is a constant whose exact evaluation is not necessary in our investigation. Thus λ can be estimated by

$$\hat{\lambda} = |W|^{\frac{1}{n}} / (D_2(l, n, p))^{\frac{1}{n}} \tag{33}$$

that is, by the geometric mean of the eigenvalue of YY'(equivalently of $Y'Y$). For simplicity, however, we consider the arithmetic mean of the non-zero eigenvalues of nS, which is written as $\frac{n\text{tr}(S)}{n}$.

Thus, when $n < p$, an empirical Bayes estimator of Σ^{-1} is given by

$$\hat{\Sigma}^{-1} = c(S + \frac{\text{tr}S}{n})^{-1} \tag{34}$$

Putting them together,

$$\hat{\lambda} = \frac{n\text{tr}(S)}{\min(n, p)} \tag{35}$$

Thus, an empirical Bayes estimate of Σ^{-1} is given by

$$\hat{\Sigma}^{-1} = c(S + \frac{1}{n}\hat{\lambda} I)^{-1} \tag{36}$$

$$= c(S + \frac{\text{tr}S}{\min(n, p)})^{-1} \tag{37}$$

4 Simulation Studies

In this section we use the simulated datasets to compare the performance of our three proposed discrimination methods with DLDA of Dudoit, Fridlyand, and Speed (2002) who have shown that the DLDA method is simple and performs better than most

other procedures considered in the literature. The simulation results show that MDEB performs better than DLDA in all cases. When the correlation between variables are not negligible , all the three proposed methods are superior over DLDA.

4.1 Generation of Datasets

In order to include the correlation between variables into the dataset, we generate the datasets as follows:

$$\begin{aligned} \mathbf{x}_{1i} \text{ i.i.d. } &\sim \mathrm{N}_p(\mu_{\mathbf{1}}, \Sigma), i = 1, 2, \cdots, N_1 \\ \mathbf{x}_{2i} \text{ i.i.d. } &\sim \mathrm{N}_p(\mu_{\mathbf{2}}, \Sigma), i = 1, 2, \cdots, N_2 \end{aligned} \tag{38}$$

where

$$\Sigma = \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_p \end{pmatrix} \begin{pmatrix} \rho^{|1-1|^{\frac{1}{7}}} & \rho^{|1-2|^{\frac{1}{7}}} & \cdots & \rho^{|1-p|^{\frac{1}{7}}} \\ \rho^{|2-1|^{\frac{1}{7}}} & \rho^{|2-2|^{\frac{1}{7}}} & \cdots & \rho^{|2-p|^{\frac{1}{7}}} \\ & & \cdots & \\ \rho^{|p-1|^{\frac{1}{7}}} & \rho^{|p-2|^{\frac{1}{7}}} & \cdots & \rho^{|p-p|^{\frac{1}{7}}} \end{pmatrix} \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_p \end{pmatrix} \tag{39}$$

We chose $p = 2000$ and generated three datasets with $\rho = 0.0, 0.2, 0.5$ respectively. This represents different levels of correlations between the p variables. Without loss of generality, we chose the 2000 dimensional mean vector of the second group as zero vector, i.e. $\mu_{\mathbf{2}} = (0, 0, \cdots, 0)'$. The mean vector $\mu_{\mathbf{1}}$ of the first group was chosen as $\mu_{\mathbf{1}} = (\mathbf{m_1}, 0, \cdots, 0)$, where $\mathbf{m_1}$ is a 100-dimensional vector generated from Uniform(0.5,1.5) and then ordered decreasingly,then assigned "+" or "−" signs with probability 0.5. $\mathbf{m_1}$ may be visualized as follows:

$$\mathbf{m_1} = (-1.50, -1.49, 1.46, -1.46, -1.44, \cdots, -0.54, 0.53, -0.52, 0.52, -0.50)$$

The square root of the diagonal elements of Σ, $(\sigma_1, \cdots, \sigma_{2000})$, were generated from Uniform(2,3), which may be visualized as follows:

$$(\sigma_1, \cdots, \sigma_{2000}) = (2.23, 2.56, 2.17, 2.95, 2.89, \cdots, 2.16, 2.40, 2.59, 2.05, 2.97)$$

4.2 Simulation Results

For $\rho = 0.0, 0.2, 0.5$, we generated three datasets as described above. Each dataset consists of a training set which comprises of 50 cases from each group, and a testing

set which comprises of 300 cases from each group. The classifiers are built with the parameters estimated with the training dataset with 100 cases. We then perform the classification procedure on the testing set with 600 cases. The methods are compared in terms of the correct classification rates. We test the performance of the four discrimination procedures on several selected values of $\tilde{p}$. The correct classification rates are presented in Table 1. Several conclusions can be drawn from this table.

First, MDEB performs consistently better than all the other methods considered in this paper. It works better even when the variables are independent. Thus, it is the most worthy method we have seen in the literature. Secondly, MDEB and DLDA are more stable with respect to the choice of $\tilde{p}$ than MDMP and MDMEB. That is, with the increase in the number of selected variables which may include noise variables, the performance of MDEB and DLDA methods remain unaffected.

Secondly, when the correlation between variables becomes stronger, DLDA gets worse substantially. This is reasonable since DLDA assumes the variables are completely independent, and therefore it has worse performance when the correlation is not negligible. All of the three proposed methods which accounts for the correlation are superior over DLDA when $\rho = 0.5$ consistently and substantially.

5 Two examples of microarray datasets

In this section we test the performance of all the four discrimination methods on two publicly available datasets with Leave-One-Out cross validation. We analyze only these two datasets because we concentrate in this paper only on two samples problems.

5.1 Method of Comparison

In the context of microarray data, the number of observations are usually very small. Therefore it is not appropriate to extract a subset from the original dataset to form a testing dataset. Instead, a widely used way to test a discrimination method is the Leave-One-Out cross validation. For $i = 1, 2, \cdots, N$, the x_i is treated as the testing case

$\tilde{p}$	20	40	60	80	100	120	140	160	200	300
					$\rho = 0$					
DLDA	0.82	0.86	0.83	0.84	0.83	0.82	0.84	0.81	0.79	0.80
MDEB	0.82	0.86	0.88	0.87	0.85	0.84	0.86	0.84	0.85	0.86
MDMEB	0.82	0.86	0.84	0.80	0.80	0.77	0.78	0.75	0.74	0.66
MDMP	0.80	0.85	0.81	0.78	0.78	0.77	0.74	0.74	0.71	0.62
					$\rho = 0.2$					
DLDA	0.78	0.79	0.82	0.80	0.77	0.75	0.77	0.80	0.77	0.75
MDEB	0.80	0.82	0.86	0.85	0.84	0.82	0.83	0.82	0.83	0.82
MDMEB	0.80	0.84	0.82	0.80	0.77	0.71	0.73	0.72	0.67	0.65
MDMP	0.77	0.79	0.73	0.73	0.73	0.69	0.62	0.61	0.61	0.61
					$\rho = 0.5$					
DLDA	0.75	0.66	0.72	0.70	0.70	0.70	0.67	0.67	0.66	0.62
MDEB	0.86	0.86	0.90	0.92	0.93	0.92	0.92	0.89	0.87	0.84
MDMEB	0.86	0.87	0.85	0.85	0.80	0.80	0.79	0.76	0.73	0.70
MDMP	0.84	0.76	0.79	0.83	0.78	0.78	0.74	0.72	0.70	0.67

Table 1: Correct Classification Rates for DLDA, MDEB, MDMEB and MDMP given by predicting 600 testing cases with 300 from each group based on 100 training cases with 50 from each group. The standard deviation of each value in this table is estimated to be 0.0145, by taking 0.85 as the approximate correct classification rates for all cells

and the remaining cases are used as training set, i.e. used to estimate the parameters $\mu_{\mathbf{1}}, \mu_{\mathbf{2}}, \Sigma$. The discrimination methods are compared in terms of correct classification rate.

It is worth mentioning that we need to perform the variable selection each time with the remaining cases other than x_i, rather than pre-select the variables with the complete dataset then perform the cross validation with the subset. This is because that the variable selection should be based on only the training set and should not use any information from the testing cases. Performing the cross validation with the pre-selected subset may lead to misleading results. An extreme example is when $\mu_{\mathbf{1}} = \mu_{\mathbf{2}}$, i.e. all variables are actually indistinguishable across the two groups. If we pre-select $\tilde{p}$ variables based on the complete dataset by some methods, for example, the t^2-statistic as described earlier, the selected $\tilde{p}$ variables may show difference across the two groups, even though there is no actual difference. Later when we use Leave-One-Out cross validation to test the discrimination method a high correct classification rate may be obtained. This is avoided by selecting variables again when we change the testing case.

5.2 Description of the Datasets

Colon

In this dataset, expression levels of 40 tumor and 22 normal colon tissues for 6500 human genes are measured using the Affymetrix technology. A selection of 2000 genes with highest minimal intensity across the samples has been made by Alon, Barkai, Motterman, Gish, Mack, and Levine (1999). Thus $p = 2000$, and the degrees of freedom available to estimate the covariance matrix is only 60.

These data are publicly available at "http://www.molbio.princeton.edu/ colondata". A base 10 logarithmic transformation is applied.

Leukemia

This dataset contains gene expression levels of 72 patients either suffering from acute lymphoblastic leukemia(ALL, 47 cases) or acute myeloid leukemia(AML 25 cases) and was obtained from Affymetrix oligonucleotide microarrays. More information can be

found in Golub, Slonim, Tamayo, Huard, Gaasenbeek, Mesirov, Coller, Loh, and Downing (1999); Following the protocol in Dudoit, Fridlyand, and Speed (2002), we preprocess them by thresholding, filtering, a logarithmic transformation and standardization, so that the data finally comprise the expression values of $p = 3571$ genes, and the degrees of freedom available for estimating the covariance is only 70.

The description of the above datasets and preprocessing are due to Dettling and Buhlmann (2002), except that we do not process the datasets such that each tissue sample has zero mean and unit variance across genes, which is not explainable in our framework. We roughly check the normality assumption by QQ-plotting around 50 genes selected randomly. The results are nearly satisfactory.

For the Colon dataset, the first 10 elements of the mean vectors for the two groups and the corresponding pooled estimate of the variances that are placed in descending order w.r.t t^2-statistics are listed in Table 2.

$\mu_\mathbf{1}$	1.207	0.741	2.092	-0.070	1.121	0.289	1.583	-1.061	1.239	-1.319	$\cdots$
$\mu_\mathbf{2}$	0.0745	-0.0097	0.725	-1.098	-0.380	1.138	0.581	-0.262	-0.0012	-0.607	$\cdots$
σ^2	0.280	0.128	0.487	0.289	0.742	0.240	0.388	0.250	0.626	0.207	$\cdots$

Table 2: Visualization of Colon Dataset

The Hotelling's T^2-statistic calculated with the first $\tilde{p}$= 60 variables that have the largest t^2-statistics is 170 which leads to a p-value of 0; the same is true for $p = 30, 40$ and 50. Such simple calculation as well as the above layout of mean vectors and variances indicate that the two groups are greatly separated in at least 60 variables. We may note that $N_1 + N_2 - 2 = n = 60$.

The same simple analysis was done on Leukemia dataset. The first 10 elements of the mean vectors and the variances are listed in Table 3.

$\mu_\mathbf{1}$	-1.216	-1.121	-0.702	-0.809	-0.292	1.002	1.945	-0.899	1.662	1.221	$\cdots$
$\mu_\mathbf{2}$	1.705	0.860	2.261	0.526	1.729	-0.315	0.902	1.139	-0.740	2.198	$\cdots$
σ^2	0.702	0.336	0.759	0.168	0.475	0.205	0.148	0.569	0.892	0.149	$\cdots$

Table 3: Visualization of Leukemia Dataset

The Hotelling's T^2-statistic calculated with the first 70 variables that have the

largest t^2-statistics is 844, corresponding to a p-value of 0; the same is true for $p = 30, 40, 50$ and 60.. From such analysis we can see that the two dataset of the two groups in Colon and Leukemia are statistically very distinct, and thus it is very unlikely that any reasonable method will not do well. It may be pointed out again that the FLDA did not do well because there were a few very small eigenvalues making the sample covariance matrix almost singular. None of the four methods suffer from this problem.

The four discrimination methods are applied to Colon and Leukemia datasets. The Leave-One-Out cross validation is used to test their performance by selecting a variety of number of genes. The results are shown in Table 4. From this table we see that our proposed methods work as well as DLDA. The reason that our methods do not show superiority may be due to the fact that the two groups are far apart in terms of Mahalanobis squared distance. It is surprising that "boosting","bagging" and "CART" methods did not do well despite the fact that the two groups are far apart. When the two groups are far apart it makes little difference whether correlation between variables are taken into account or not.

$\tilde{p}$	20	40	60	80	100	120	140	160	200	300
Colon										
DLDA	0.89	0.86	0.86	0.86	0.87	0.87	0.87	0.87	0.87	0.87
MDEB	0.89	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
MDMEB	0.89	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
MDMP	0.87	0.87	0.86	0.81	0.84	0.86	0.86	0.87	0.87	0.87
Leukemia										
DLDA	0.93	0.95	0.93	0.97	0.97	0.97	0.96	0.96	0.96	0.97
MDEB	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
MDMEB	0.95	0.97	0.96	0.97	0.97	0.97	0.96	0.96	0.96	0.97
MDMP	0.95	0.96	0.96	0.96	0.97	0.96	0.96	0.96	0.96	0.97

Table 4: Correct Classification Rates of DLDA,MDEB,MDMEB, MDMP given by Leave-One-Out cross validation on the two real datasets. The standard deviation of each value for Colon data is estimated to be 0.043, by taking 0.87 as the approximate correct classification rates, and for Leukemia data, the standard deviations are estimated to be 0.019, by taking 0.97 as the approximate correct classification rate

The classification results reported in Table 4 for the Colon data are better substantially than most of those reported by Dettling and Buhlmann (2002). For the Leukemia data, our results are also better, though not by as much as in the Colon data, than most of those given by Dettling and Buhlmann (2002).

6 Concluding Remarks

In this paper we propose three discrimination methods, namely MDEB,MDMEB and MDMP. Our simulation results show that MDEB performs better than all the other methods under all circumstances, and all of the proposed methods, are superior to DLDA when the correlation between variables are not negligible. These methods work well in two microarray datasets.

Acknowledgements

We wish to thank Dr. Marcel Dettling for providing the two preprocessed microrray datasets analyzed here. We would like to express our sincere thanks to Longhai Li for deligently carrying out the computation of the paper. The research was supported by the Natural Sciences and Engineering Research Council of Canada.

References

Alon, U., N. Barkai, D. Motterman, K. Gish, S. Mack, and J. Levine, 1999, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *PNAS*, pp. 6745–6750.

Breiman, L., 1996, "Out-of-bag estimation," Discussion paper, Statistics Department, U.C. Berkeley.

Breiman, L., 1998, "Arcing the classifiers," *Annals of Statistics*, pp. 801–824.

Breiman, L., J. Friedman, R. Olshen, and C. Stone, 1984, *Classification and regression trees*, Wadsworth International Group.

Dettling, M., and P. Buhlmann, 2002, "Boosting for tumor classification with gene expression data," *Bioinfomatics*, pp. 1–9.

Dudoit, S., J. Fridlyand, and T. P. Speed, 2002, "Comparison of Discimination Methods for the Classification of Tumors Using Gene Expression Data," *Journal of American Statistical Association*, pp. 77–87.

Fisher, R., 1936, "The use of Multiple Measurement in Taxonomic Problems," *Annals of Eugenics*, pp. 179–188.

Freund, Y., and R. Schapire, 1997, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, pp. 119–139.

Golub, T., D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, and J. Downing, 1999, "Molecular classification of cancer:class discovery and class prediction by gene expression monitoring," *Science*, pp. 531–537.

Srivastava, M., and C. Khatri, 1979, *An introduction to multivariate statistics*, New York:North-Holland.