4 Simulation Study

4.1 Two Groups

The 500 random observations are generated in a 9-dimensional mixture. An observation $\mathbf{y}$ in this mixture is generated by $\mathbf{y} = -0.5\mathbf{d} + \mathbf{d}(\mathbf{X}) + \mathbf{Z}$ where $\mathbf{X}$ is a Bernoulli variable with P($\mathbf{X} = 1$)=0.3 (i.e the mixing proportions are 0.3 and 0.7) and $\mathbf{Z} \sim N_9(\mathbf{0},\mathbf{\Sigma})$. Let $\mathbf{d} = (d_1, d_2, \ldots, d_9)$ with $d_i = 0.05i, i = 1, 2, \ldots, 9$. The diagonal elements of $\mathbf{\Sigma}$ are set to 1 and the off diagonal elements of $\mathbf{\Sigma}$ are determined by the off diagonal elements of the matrix, $1.3\mathbf{ff}'$, where the first 4 elements of $\mathbf{f}$ are 0.8 and the last 5 elements of $\mathbf{f}$ are -0.1.

We first use Chang's method to compute the Mahalanobis distance Δ^2_{1k} computed based on each principal component PC_k, $k = 1, 2, \ldots, 9$ and obtain a selection order of principal components. The result of Chang's method shown in Table 1.

Table 1: Information (Δ^2_{1k}) Retained by Principal Component, PC_k (Factor, FA_k)

k	Δ^2_{1k}
1	1.90E-02
2	0.001636
3	0.55036
4	0.059145
5	0.030945
6	2.82E-03
7	0.0027659
8	1.91E-02
9	1.63E-04

Based on Table 1, we should first select PC_3, and then select PC_4, PC_5, ..., PC_9 as $\Delta^2_{13} > \Delta^2_{14} > \Delta^2_{15} > \cdots > \Delta^2_{19}$.

Next, we apply the method of Mardia et al. on principal component selection and factor selection. First, express the data matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_{500}]$ in terms of principal component values and principal factor scores, obtained by the principal component method. Using the procedure introduced in Section 2.2.1, we than obtain some suitable discriminators from principal components and factors, respectively. Here, we use $\alpha = 0.1$ as the significance level. Table 2 gives the results of M's statistic based on principal component values (or factor scores). Table 2 shows

Table 2: M's Statistic ($\eta_1(k)$ or $\eta_2(k)$) based on Principal Component, PC_k (or Factor, FA_k)

k	$\eta_1(k)$	$3, k$	$\eta_2(k)$
1	8.7556	3,1	1.8988
2	9.0036	3,2	2.1822
3	1.933	3,4	1.2491
4	8.187	3,5	1.7039
5	8.585	3,6	2.1627
6	8.9865	3,7	2.1637
7	8.9873	3,8	1.8969
8	8.7539	3,9	2.2064
9	9.0247		

that PC_3 and PC_4 should be selected since $\eta_1(3)$ is the smallest among $\eta_1(k)$ and is larger than $F_{8,490}(0.1) = 1.682773475$, and $\eta_2(4)$ is the smallest among $\eta_2(k)$ and is smaller than $F_{7,490}(0.1) = 1.729045051$.

We also show the scatter diagrams of PC_1, PC_2 and PC_3, PC_4 in Figure 1 and Figure 2, respectively. The sign "o" represents the principal component data from $N_p(0.5\mathbf{d}, \mathbf{\Sigma})$, and the sign "$\times$" represents the principal component data from $N_p(-0.5\mathbf{d}, \mathbf{\Sigma})$. Furthermore, the classification error rate on using PC_1 and PC_2 is 0.47936, which is worse than that of 0.32121 on using PC_3 and PC_4.

Figure 1: Plot of PC 1 and PC 2

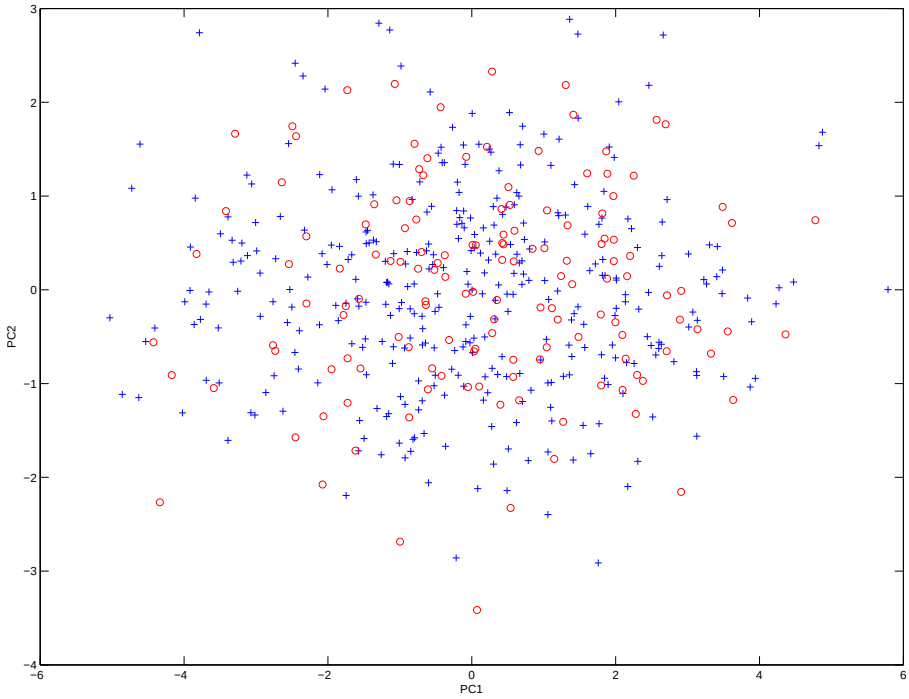
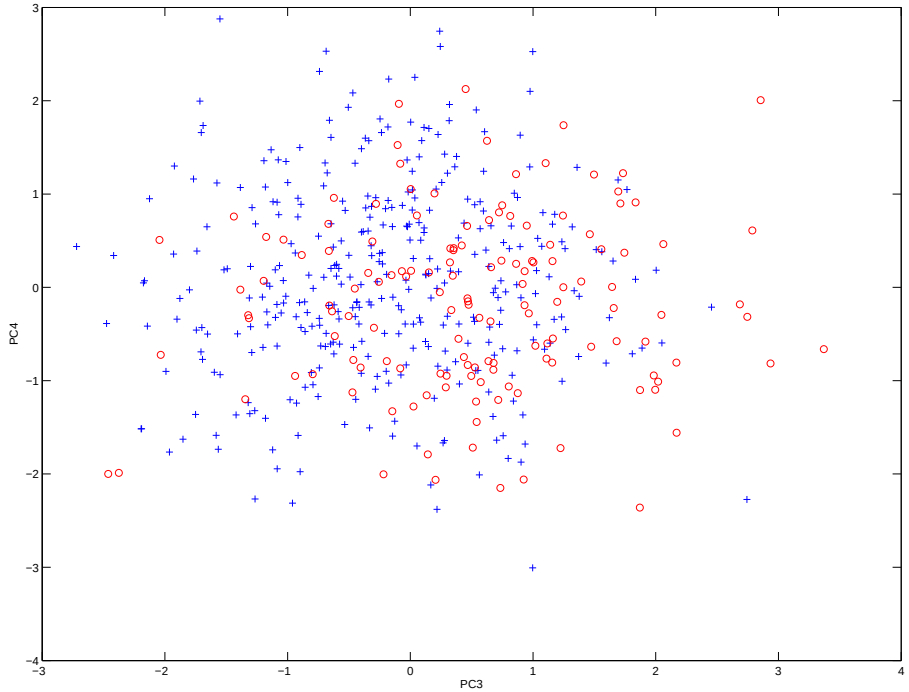


Figure 2: Plot of PC 3 and PC 4



4.2 Three Groups

In this section, we introduce how to use the method of Mardia et al. when our data is from three populations. Let $\mathbf{f} = (-0.95, -0.95, -0.95, 0.15, 0.15)'$ and $\mathbf{d} = (0.8, 0.75, 0.7, 0.65, 0.6)'$. Define $\mathbf{\Sigma} = -0.15(\mathbf{ff}')$ with diagonal elements 1. We simulate 1000 random vectors from three populations :

1. Simulate 250 random vectors $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{250}$ from the first population π_1, where $\mathbf{z}_i \sim N_5(-0.5\mathbf{d}, \mathbf{\Sigma}), i = 1, 2, \ldots, 250$.

2. Simulate 400 random vectors $\mathbf{z}_{251}, \mathbf{z}_{252}, \ldots, \mathbf{z}_{650}$ from the second population π_2, where $\mathbf{z}_i \sim N_5(\mathbf{d}, \mathbf{\Sigma}), i = 251, 252, \ldots, 650$.

3. Simulate 350 random vectors $\mathbf{z}_{651}, \mathbf{z}_{652}, \ldots, \mathbf{z}_{1000}$ from the third population π_3, where $\mathbf{z}_i \sim N_5(0.5\mathbf{d}, \mathbf{\Sigma}), i = 651, 652, \ldots, 1000$.

We first transform the data into principal component values. We'll repeat the same procedure as in Section 4.1 three times on π_1 and π_2, π_1 and π_3, π_2 and π_3, respectively. Now, we will use the method of Mardia et al. to obtain a set S of principal components having discriminating power.

Step 1. Apply the method on the random vectors from π_1 and π_2. Obtain a set S_1 of principal components having discriminating power.

Step 2. Apply the method on the random vectors from π_1 and π_3. Obtain a set S_2 of principal components having discriminating power.

Step 3. Apply the method on the random vectors from π_2 and π_3. Obtain a set S_3 of principal components having discriminating power.

Step 4. Take $S = S_1 \cup S_2 \cup S_3$.

Tables $3 \sim 5$ give the M's statistics on discriminating π_1 and π_2, π_1 and π_3, π_2 and π_3, respectively. Table 3 suggests PC_1 as the best variable to discriminate π_1 and π_2, Table 4 suggests (PC_1,PC_5) as the best pair of variables to discriminate π_1 and π_3, and Table 5 suggests that (PC_1,PC_5) as the best pair of variables to discriminate π_2 and π_3. Hence, we would take PC_1 and PC_5 (since $\{PC_1, PC_5\}=\{PC_1\}\cup \{PC_1, PC_5\} \cup \{PC_1, PC_5\}$) as the variables to discriminate π_1, π_2 and π_3.

Table 3: M's Statistic ($\eta_1(k)$) based on Principal Component, PC_k for Discriminating Populations 1 and 2

k	$\eta_1(k)$
1	2.3386
2	105.52
3	105.52
4	105
5	105.4

The critical value is $F_{4,644}(0.05) = 2.385768028$.

Table 4: M's Statistic ($\eta_1(k)$ or $\eta_2(k)$) based on Principal Component, PC_k for Discriminating Populations 1 and 3

k	$\eta_1(k)$	k	$\eta_2(k)$
1	5.9136	1,2	7.4104
2	381.69	1,3	7.525
3	373.26	1,4	8.085
4	376.85	1,5	1.1841
5	376.72		

The critical values are $F_{4,594}(0.05) = 2.386933318$ and $F_{3,594}(0.05) = 2.61990607$. Moreover, we will find all the possible classification error rates using 1, 2, 3, 4 or 5 variables as discriminant variables, respectively. The results are given in Table 6.

From the Tables $3 \sim 5$, we would use PC_1 and PC_5 as our discriminant variables, as explained in Section 4.2. However, Table 6 shows that the error rate is the smallest when using PC_1, PC_3 and PC_5 as discriminant variables. Although the classification error rate (0.2479) corresponding to the case using PC_1 and PC_5 as discriminant variables is not the smallest, it is not much different from the smallest error rate even it is not the smallest. Moreover, our simulations also show that the

error rate using the method in Section 4.2 is significantly improved than that using PC_1 and PC_2.

Table 5: M's Statistic ($\eta_1(k)$ or $\eta_2(k)$) based on Principal Component, PC_k for Discriminating Populations 2 and 3

k	$\eta_1(k)$	k	$\eta_2(k)$
1	7.8445	1,2	10.212
2	143.94	1,3	10.335
3	141.07	1,4	10.704
4	142.11	1,5	0.6012
5	140.72		

The critical values are $F_{4,744}(0.05) = 2.3838993$ and $F_{3,744}(0.05) = 2.616872052$.

Table 6: Error Rates (E.R.) using $1, 2, \ldots,$ or 5 Principal Components as Discriminating Variables

PC	E.R.	PC	E.R.	PC	E.R.	PC	E.R.	PC	E.R.
1	0.24931	1,2	0.25026	1,2,3	0.24931	1,2,3,4	0.25181	1,2,3,4,5	0.2474
2	0.65776	1,3	0.24657	1,2,4	0.25181	1,2,3,5	0.2454		
3	0.6435	1,4	0.25026	1,2,5	0.2464	1,2,4,5	0.24669		
4	0.65779	1,5	0.2479	1,3,4	0.24907	1,3,4,5	0.2444		
5	0.63952	2,3	0.64698	1,3,5	0.24107	2,3,4,5	0.63107		
		2,4	0.65436	1,4,5	0.24824				
		2,5	0.63393	2,3,4	0.64007				
		3,4	0.6541	2,3,5	0.6375				
		3,5	0.6465	2,4,5	0.6269				
		4,5	0.6251	3,4,5	0.62783				

We shall simulate 500 random observations in a 9-dimensional mixture distribution for 10 times. Define $\mathbf{d} = (0.05, 0.1, 0.15, 0.2, 0.25)'$ and a 9-dimensional

random vector $\mathbf{f}$ with component f_i, $-2 \leq f_i \leq 3, i = 1, 2, \ldots, 9$. First, the diagonal elements of $\mathbf{\Sigma}$ are 1's and the off diagonal elements of $\mathbf{\Sigma}$ are determined by the off diagonal elements of the matrix, $1.3\mathbf{ff}'$. Using this $\mathbf{\Sigma}$, 500 observation $\mathbf{y}'$s are generated by $\mathbf{y} = -0.5\mathbf{d} + \mathbf{dX} + \mathbf{Z}$ where $\mathbf{X}$ is a Bernoulli variable with P($\mathbf{X}$=1)=0.3, and $\mathbf{Z}$ has a 9-dimensional normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$. Another set of 500 observations are generated again for 9 more times corresponding to different $\mathbf{\Sigma}$ (i.e random vector $\mathbf{f}$). We apply the method of Mardia et al. on principal components to select the appropriate discriminating variables. Using the selected discriminating variables by Mardia et al., we find the classification error rate, denoted by ER(M). We also compute the classification error rate, denoted by ER(PC), using PC_1 and PC_2 as discriminating variables. We then compute the relative error rate (RER)with RER=(ER(PC)-ER(M))/ER(PC) . The results are shown in Table 7.

Table 7: Relative Error Rate

Simulation number	ER(M)	ER(PC)	RER
1	0.3277 (PC 3, PC 6)	0.4322	24.2%
2	0.3415(PC 4, PC 6)	0.4272	20.1%
3	0.379(PC 3, PC 6)	0.415	8.7%
4	0.3509(PC 3, PC 5)	0.4064	13.7%
5	0.34(PC 4, PC 5)	0.4447	23.5%
6	0.3907(PC 1, PC 6)	0.4451	12.2%
7	0.3239(PC 3, PC 9)	0.433	25.2%
8	0.3687(PC 7, PC 9)	0.4183	11.9%
9	0.3857(PC 4, PC 5)	0.4371	11.8%
10	0.31306(PC 2, PC 7)	0.3831	18.3%

From Table 7, we can see that the classification error rate using the method of Mardia et al. is significantly improved than that using PC_1 and PC_2.