Multivariate Analysis - Homework 3

1. Consider the two 2-dimensional data sets from two populations G_1 and G_2

$$\mathbf{Y}_1 = \begin{pmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{pmatrix} \quad \text{and} \quad \mathbf{Y}_2 = \begin{pmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{pmatrix}$$

   (a) Calculate the linear discriminant function.

   (b) Classify the observation $\mathbf{y}_0 = (2, 7)'$ as population G_1 or G_2.

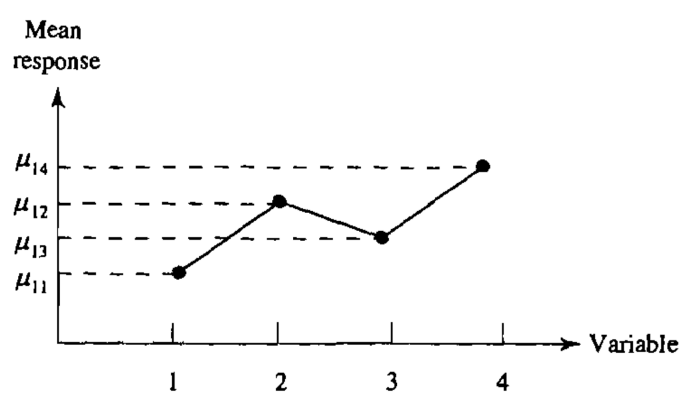
2. Show the following based on Fisher's LDA.

   (a) The maximized statistical distance between the transformed sample mean $\bar{z}_1$ and $\bar{z}_2$ is proportional to the statistical distance between the $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$.

   (b) Fisher's allocation rule is indeed to compare the statistical distance between the new observation $\mathbf{y}_0$ and $\bar{\mathbf{y}}_1$ and that between $\mathbf{y}_0$ and $\bar{\mathbf{y}}_2$.

   (c) Verify that the soluation of two-population LDA is indeed a special case of the several-population case.

3. Profile analysis pertains to situations in which a battery of p treatments are administered to two or more groups of subjects. All responses must be expressed in similar units. Further, assume that the responses for the different groups are independent of one another. Denote $\boldsymbol{\mu}_1' = (\mu_{11}, \ldots, \mu_{1p})'$ and $\boldsymbol{\mu}_2' = (\mu_{21}, \ldots, \mu_{2p})'$ as the mean responses to p treatments for populations 1 and 2. An illustration of a profile plot of population 1 (with $p = 4$) is as follows.



Formulate the following hypothesis tests, including the null hypothesis, test statistic and its null distribution, and the rejection region. Assume the two populations have common covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}$ is unknown to us.

(a) Test whether the profiles of two populations are parallel.

(b) Test whether the total measurements are the same between the two population.

(c) Assuming the profiles are parallel, test whether the profiles are linear. (Hint: First show that the linearity test can be written as H_0: $(\mu_{1j}+\mu_{2j})-(\mu_{1(j-1)}+\mu_{2(j-1)}) = (\mu_{1(j-1)}+\mu_{2(j-1)})-(\mu_{1(j-2)}+\mu_{2(j-2)})$, $j=3,\ldots,p$)

(d) Following (c), Let $n_1 = 30$, $n_2 = 30$, $\bar{\mathbf{y}}_1 = (6.4, 6.8, 7.3, 7.0)'$ and $\bar{\mathbf{y}}_2 = (4.3, 4.9, 5.3, 5.1)'$, and

$$\mathbf{S}_{\text{pooled}} = \begin{bmatrix} .61 & .26 & .07 & .16 \\ .26 & .64 & .17 & .14 \\ .07 & .17 & .81 & .03 \\ .16 & .14 & .03 & .31 \end{bmatrix}$$

Test for linear profiles, assuming that the profiles are parallel. Use $\alpha = 0.05$.

4. (R exercise) In the first phase of a study of the cost of transporting milk from farms to dairy plants, a survey was taken of firms engaged in milk transportation. Cost data on Y_1 =fuel, Y_2 =repair, and Y_3 =capital, all measured on a per-mile basis, are presented in the following table for $n_1 = 36$ gasoline and $n_2 = 23$ diesel trucks (data attached as cost.dat).

Gasoline trucks			Diesel trucks		
x_1	x_2	x_3	x_1	x_2	x_3
16.44	12.43	11.23	8.50	12.26	9.11
7.19	2.70	3.92	7.42	5.13	17.15
9.92	1.35	9.75	10.28	3.32	11.23
4.24	5.78	7.78	10.16	14.72	5.99
11.20	5.05	10.67	12.79	4.17	29.28
14.25	5.78	9.88	9.60	12.72	11.00
13.50	10.98	10.60	6.47	8.89	19.00
13.32	14.27	9.45	11.35	9.95	14.53
29.11	15.09	3.28	9.15	2.94	13.68
12.68	7.61	10.23	9.70	5.06	20.84
7.51	5.80	8.13	9.77	17.86	35.18
9.90	3.63	9.13	11.61	11.75	17.00
10.25	5.07	10.17	9.09	13.25	20.66
11.11	6.15	7.61	8.53	10.14	17.45
12.17	14.26	14.39	8.29	6.22	16.38
10.24	2.59	6.09	15.90	12.90	19.09
10.18	6.05	12.14	11.94	5.69	14.77
8.88	2.70	12.23	9.54	16.77	22.66
12.34	7.73	11.68	10.43	17.65	10.66
8.51	14.02	12.01	10.87	21.52	28.47
26.16	17.44	16.89	7.13	13.22	19.44
12.95	8.24	7.18	11.88	12.18	21.20
16.93	13.37	17.59	12.03	9.22	23.09
14.70	10.78	14.58			
10.32	5.16	17.00			
8.98	4.49	4.26			
9.70	11.59	6.83			
12.72	8.63	5.59			
9.49	2.16	6.23			
8.22	7.95	6.72			
13.70	11.22	4.91			
8.21	9.85	8.17			
15.86	11.42	13.06			
9.18	9.18	9.49			
12.49	4.67	11.94			
17.32	6.86	4.44			

(a) Test for differences in the mean cost vectors at the significance level 0.01.

(b) If the hypothesis of equal cost vectors is rejected, conduct the univariate tests at the same significance level. What is your conclusion?

(c) If the hypothesis of equal cost vectors is rejected, find the linear combination of mean components most responsible for the rejection. Interpret the coefficients in the linear combination.

(d) Now only consider the first 23 gasoline trucks and the 23 diesel trucks. Suppose the ith gasoline truck and the ith diesel truck are from the same farm to the same dairy plant. Redo (a)-(c).

5. (R exercise) The tail lengths in millimeters X_1 and wing lengths in millimeters X_2 for 45 male hook-billed kites are (data attached as male.dat):

x_1 (Tail length)	x_2 (Wing length)	x_1 (Tail length)	x_2 (Wing length)	x_1 (Tail length)	x_2 (Wing length)
180	278	185	282	284	277
186	277	195	285	176	281
206	308	183	276	185	287
184	290	202	308	191	295
177	273	177	254	177	267
177	284	177	268	197	310
176	267	170	260	199	299
200	281	186	274	190	273
191	287	177	272	180	278
193	271	178	266	189	280
212	302	192	281	194	290
181	254	204	276	186	287
195	297	191	290	191	286
187	281	178	265	187	288
190	284	177	275	186	275

Similar measurements for female hook-billed kites are (data attached as female.dat):

x_1 (Tail length)	x_2 (Wing length)	x_1 (Tail length)	x_2 (Wing length)	x_1 (Tail length)	x_2 (Wing length)
191	284	186	266	173	271
197	285	197	285	194	280
208	288	201	295	198	300
180	273	190	282	180	272
180	275	209	305	190	292
188	280	187	285	191	286
210	283	207	297	196	285
196	288	178	268	207	286
191	271	202	271	209	303
179	257	205	285	179	261
208	289	190	280	186	262
202	285	189	277	174	245
200	272	211	310	181	250
192	282	216	305	189	262
199	280	189	274	188	258

(a) Plot the male hook-billed kite data as a scatterplot, and visually check for outliers.

(b) Test for equality of mean vectors for the populations of male and female hook-billed kites. Set $\alpha = 0.05$. If $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$ is rejected, find the linear combination most responsible for the rejection of H_0. You may want to eliminate any outlier found in Part (a) for the male hook-billed kite data before conducting this test. Alternatively, you may try to interpret the outlier as a misprint and conduct the test with a more reasonable imputation/substitute. Does it make any difference in this case how outliers for the male hook-billed kite data are treated?

(c) Determine and draw the 95% confidence region for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

(d) Are male or female birds generally larger?

6. (R exercise.) The admission officer of a business school has used an "index" of X_1 - undergraduate grade point average (GPA) and X_2 - graduate management aptitude test (GMAT) scores to help decide which applicants should be admitted to the school's graduate programs. The data are listed in the following table (attached as gpa-gmat.dat) and categorized to three groups.

(a) Calculate the group means, overall means, and pooled sample covariance matrix.

(b) Obtain the scatterplot between GPA and GMAT, and label the three groups. Comment.

(c) Assuming equal covariance matrices, conduct Fisher's LDA. DO NOT use R package. Start with $W^{-1}B$. Give the discriminant functions, and obtain the scatterplot in the discriminant space. Compare it to (b).

(d) Further assuming bivariate normality, and assume the Admit, DO not admit, and Borderline groups are predetermined to be proportional to 3:6:1. Classify the new observation $(3.21, 497)'$ into one of the three groups. DO NOT use R package.

(e) Conduct (d) using "lda" function in R. Compare the results.

Applicant no.	GPA (x_1)	GMAT (x_2)	Applicant no.	GPA (x_1)	GMAT (x_2)	Applicant no.	GPA (x_1)	GMAT (x_2)
1	2.96	596	32	2.54	446	60	2.86	494
2	3.14	473	33	2.43	425	61	2.85	496
3	3.22	482	34	2.20	474	62	3.14	419
4	3.29	527	35	2.36	531	63	3.28	371
5	3.69	505	36	2.57	542	64	2.89	447
6	3.46	693	37	2.35	406	65	3.15	313
7	3.03	626	38	2.51	412	66	3.50	402
8	3.19	663	39	2.51	458	67	2.89	485
9	3.63	447	40	2.36	399	68	2.80	444
10	3.59	588	41	2.36	482	69	3.13	416
11	3.30	563	42	2.66	420	70	3.01	471
12	3.40	553	43	2.68	414	71	2.79	490
13	3.50	572	44	2.48	533	72	2.89	431
14	3.78	591	45	2.46	509	73	2.91	446
15	3.44	692	46	2.63	504	74	2.75	546
16	3.48	528	47	2.44	336	75	2.73	467
17	3.47	552	48	2.13	408	76	3.12	463
18	3.35	520	49	2.41	469	77	3.08	440
19	3.39	543	50	2.55	538	78	3.03	419
20	3.28	523	51	2.31	505	79	3.00	509
21	3.21	530	52	2.41	489	80	3.03	438
22	3.58	564	53	2.19	411	81	3.05	399
23	3.33	565	54	2.35	321	82	2.85	483
24	3.40	431	55	2.60	394	83	3.01	453
25	3.38	605	56	2.55	528	84	3.03	414
26	3.26	664	57	2.72	399	85	3.04	446
27	3.60	609	58	2.85	381			
28	3.37	559	59	2.90	384			
29	3.80	521						
30	3.76	646						
31	3.24	467						

7. (R exercise.) Use the beetle data in the following (data attached as T5_5_FBEETLES.DAT):

Table 5.5 Four Measurements on Two Species of Flea Beetles

Haltica oleracea					*Haltica carduorum*				
Experiment Number	y_1	y_2	y_3	y_4	Experiment Number	y_1	y_2	y_3	y_4
1	189	245	137	163	1	181	305	184	209
2	192	260	132	217	2	158	237	133	188
3	217	276	141	192	3	184	300	166	231
4	221	299	142	213	4	171	273	162	213
5	171	239	128	158	5	181	297	163	224
6	192	262	147	173	6	181	308	160	223
7	213	278	136	201	7	177	301	166	221
8	192	255	128	185	8	198	308	141	197
9	170	244	128	192	9	180	286	146	214
10	201	276	146	186	10	177	299	171	192
11	195	242	128	192	11	176	317	166	213
12	205	263	147	192	12	192	312	166	209
13	180	252	121	167	13	176	285	141	200
14	192	283	138	183	14	169	287	162	214
15	200	294	138	188	15	164	265	147	192
16	192	277	150	177	16	181	308	157	204
17	200	287	136	173	17	192	276	154	209
18	181	255	146	183	18	181	278	149	235
19	192	287	141	198	19	175	271	140	192
					20	197	303	170	205

(a) Find the discriminant function coefficient vector. Obtain the transformed univariate observations.

(b) Find the discriminant coefficient vector based on the individually standardized observations. Obtain the transformed univariate observations.

(c) Compare the results from (a) and (b). Comment.

(d) Calculate t-tests for individual variables.

(e) Compare the results of (a), (b) and (d) as to the contribution of each variable to separation of the groups.