STA 9705: HW5

Tanay Mukherjee

 # 8.8 (a)

a
0.345249
-0.130388
-0.106434
-0.143353

8.8

(a) From the SAS output we have discriminant function coefficient vector

$$a = S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2) = \begin{pmatrix} 0.3452 \\ -0.1304 \\ -0.1064 \\ -0.1434 \end{pmatrix}$$

 # 8.8 (b)

as
4.1366401
-2.50055
-1.157705
-2.067833

(b) From the SAS output the standarized coefficients are

$$a^* = (\text{diag}(S_{pl}))^{1/2}.a = \begin{pmatrix} 4.1366 \\ -2.5006 \\ -1.1577 \\ -2.0678 \end{pmatrix}$$

8.8 (c)

T1
3.8879456

T2
-3.865239

T3
-5.691131

T4
-5.042625

(c) t-tests for individual variables is

$t_1 = 3.8879$, $t_2 = -3.8652$, $t_3 = -5.6911$, $t_4 = -5.0426$

8.8 (d)

(d) Section 8.7 in the book says we use the true value for interpretation but only the absolute value for comparing the contribution for each variable to the seperation of groups.

So, using values of $|a^*|$ we can say the contributions are ranked $\rightarrow$ $Y_1 > Y_2 > Y_4 > Y_3$

Also, using absolute values for t-stat we have the contributions ranked $\rightarrow$ $Y_3 > Y_4 > Y_1 > Y_2$

We know that in case of conflict, we go with the $|a^*|$ as the standardized coefficients takes into account the sample co-relations among variables as well as the influence of each variable in the presence of others.

8.11 (a)

Raw Canonical Coefficients		
Variable	**Can1**	**Can2**
AROMA	0.118947483	-1.822971192
FLAVOR	3.064352847	1.714018010
TEXTURE	-1.992418219	1.396730818
MOISTURE	-0.775971076	-0.150866787

8.11 (a) The eigenvectors of $E^{-1}H$ are

$$a_1 = \begin{pmatrix} 0.1189 \\ 3.0644 \\ -1.9924 \\ -0.7760 \end{pmatrix}, \quad a_2 = \begin{pmatrix} -1.8230 \\ 1.7140 \\ 1.3967 \\ -0.1509 \end{pmatrix}$$

8.11 (b)

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
					Eigenvalue	Difference	Proportion	Cumulative
1	0.864251	0.850266	0.042777	0.746930	2.9515	2.8242	0.9586	0.9586
2	0.336071	0.268316	0.149940	0.112944	0.1273		0.0414	1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero				
Likelihood Ratio	**Approximate F Value**	**Num DF**	**Den DF**	**Pr > F**
0.22448732	8.33	8	60	<.0001
0.88705614	1.32	3	31	0.2869

(b) From SAS output we have the eigen values as
$\lambda_1 = 2.9515$ and $\lambda_2 = 0.1273$

For any m^{th} step, $\Lambda_m = \prod_{i=m}^{s} 1/(1+\lambda_i)$

So, for 1^{st} step, $\Lambda_1 = \frac{1}{(1+2.9515)} \times \frac{1}{(1+0.1273)} = 0.2245$

and for 2^{nd} step, $\Lambda_2 = \frac{1}{(1+0.1273)} = 0.8871$

From the test of Ho in the table above we can compare the p-value with $\alpha = 0.05$ for significance test and confirm that for
Λ_1 we reject Ho as p-value is less than $<.001$

whereas for Λ_2 we fail to reject Ho → 1

Now, let's do it using Critical values:-

For step 1 :- $p = 4,\ K = 3,\ m = 1,\ n = 12,\ N = 36.$

$\Lambda_m = \Lambda_1 = 0.2245$ and $\Lambda_\alpha(p-m+1,\ k-m,\ N-K-m+1)$
$= \Lambda_{0.05}(4, 2, 33)$

Using table A.9 we see that
$\Lambda_{0.05}(4, 2, 33) > \Lambda_{0.05}(4, 2, 30)$
$= 0.580$

We have $\Lambda_m < \Lambda_{0.05}(4, 2, 30) < \Lambda_{0.05}(4, 2, 33)$

$\Rightarrow 0.2245 < 0.580.$

Therefore, we reject H_0.

For step 2 :- $p=4$, $K=3$, $m=2$, $n=12$, $N=36$.

$\Lambda_m = \Lambda_2 = 0.8871$ and $\Lambda_\alpha(p-m+1, K-m, N-K-m+1)$

$= \Lambda_{0.05}(3, 1, 32)$

Using table A.9 we see that

$\Lambda_{0.05}(3, 1, 32) > \Lambda_{0.05}(3, 1, 40)$

$= 0.816.$

Now, the confusion could be whether to ~~test~~ for $V_E = 30$ or $V_E = 40$. When we fail to reject H_0 for lower bound, always test for higher bound next and if that's fails to reject we can conclude.

So, we have $\Lambda_m > \Lambda_{0.05}(3, 1, 32) > \Lambda_{0.05}(3, 1, 40)$

$\Rightarrow 0.8871 > 0.816.$

Therefore, we fail to reject H_0. $\rightarrow$ 2

Both 1 and 2 give us the same conclusion that is - first discriminant function is significant but the second discriminant fuction is not significant.

8.11 (c)

Pooled Within-Class Standardized Canonical Coefficients		
Variable	**Can1**	**Can2**
AROMA	0.075820332	-1.162010988
FLAVOR	1.553387218	0.868873071
TEXTURE	-1.181660941	0.828371392
MOISTURE	-0.439076751	-0.085366711

(c) The standarized discriminant fuction coefficients are:-

$$a_1^* = \begin{pmatrix} 0.0758 \\ 1.5534 \\ -1.1817 \\ -0.4391 \end{pmatrix}, \quad a_2^* = \begin{pmatrix} -1.1620 \\ 0.8689 \\ 0.8284 \\ -0.0854 \end{pmatrix}$$

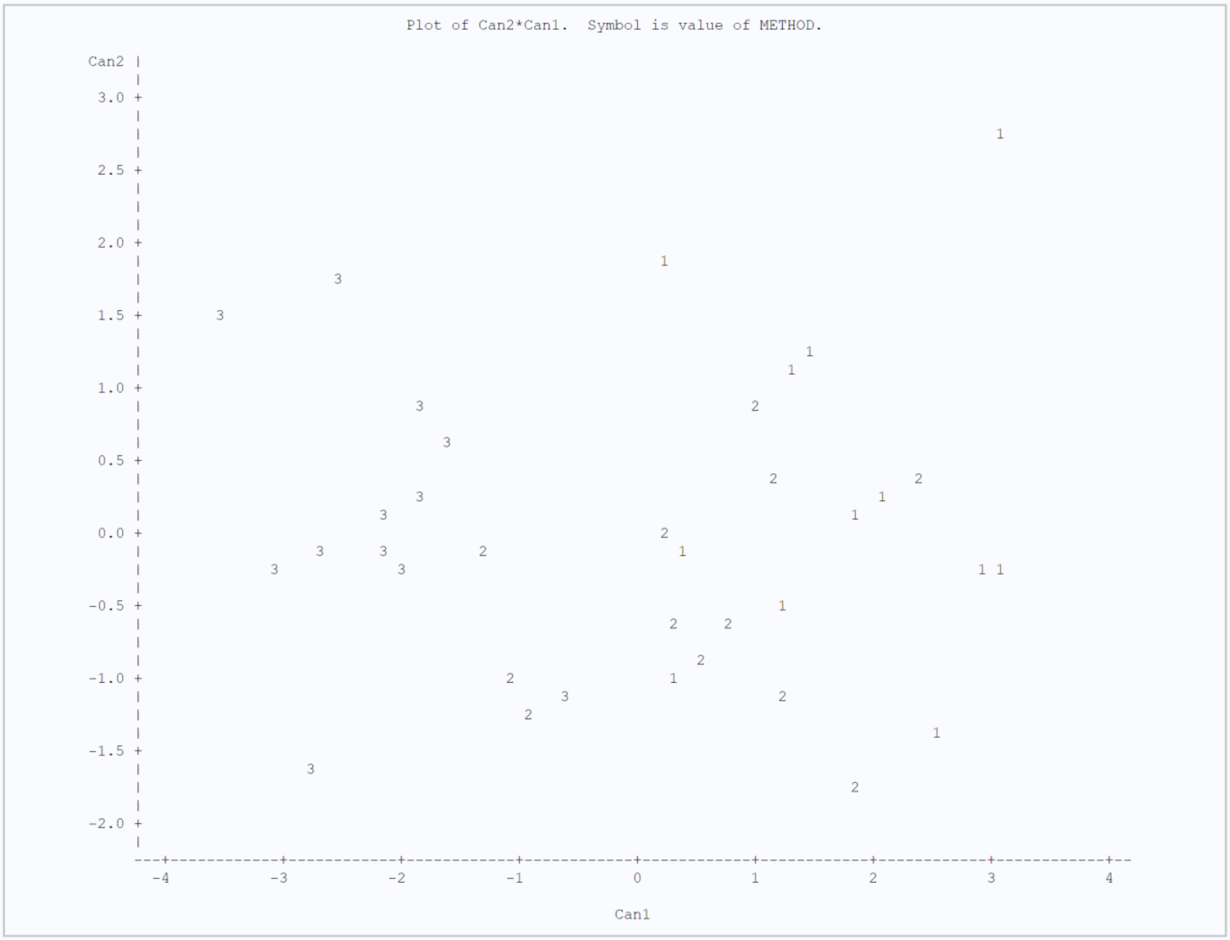
We know for contribution we use absolute value where as we use true value only for interpretatio

Therefore, contribution of a_1^* is based on $|a_1^*|$ values and it is ranked as $Y_2 > Y_3 > Y_4 > Y_1$

Similarly, contribution of a_2^* is based on $|a_2^*|$ values and it is ranked as $Y_1 > Y_2 > Y_3 > Y_4$.

8.11 (e)

(e) The first discriminant function separates groups 1 and group 2 from group 3 but the second discriminant function fails to separate group 1 from group 2. From the graph below:-



The first discriminant function "Can 1" (horizontal axis) separates group 1 & group 2 from 3.

However, second discriminant function "Can 2" (vertical axis) fails to separate group 1 from group 2.

9.10 (a)

Linear Discriminant Function for METHOD			
Variable	**1**	**2**	**3**
Constant	-72.76878	-65.18045	-68.56609
AROMA	0.80819	2.12237	0.67639
FLAVOR	15.15136	10.11279	2.79198
TEXTURE	-1.03021	0.23934	6.54334
MOISTURE	10.01533	11.06496	13.09289

9.10(a) We know linear classification function is given by:

$$L_i(y_{new}) = c_i' y_{new} + c_{0i}$$

where $c_i = S_{pl}^{-1}\bar{y}_i$ and $c_{0i} = -\frac{1}{2}\bar{y}_i' S_{pl}^{-1}\bar{y}_i$

For the given dataset we have

$$L_1(y) = -72.77 + 0.81y_1 + 15.15y_2 - 1.03y_3 + 10.02y_4$$

$$L_2(y) = -65.18 + 2.12y_1 + 10.11y_2 - 0.24y_3 + 11.06y_4$$

$$L_3(y) = -68.57 + 0.68y_1 + 2.79y_2 + 6.54y_3 + 13.09y_4$$

9.10 (b)

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.FISH
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into METHOD				
From METHOD	1	2	3	Total
1	9 75.00	3 25.00	0 0.00	12 100.00
2	3 25.00	7 58.33	2 16.67	12 100.00
3	0 0.00	1 8.33	11 91.67	12 100.00
Total	12 33.33	11 30.56	13 36.11	36 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for METHOD				
	1	2	3	Total
Rate	0.2500	0.4167	0.0833	0.2500
Priors	0.3333	0.3333	0.3333	

(b) Error rate = 1 - correct classification rate

$$= 1 - \frac{n_{11} + n_{22} + n_{33}}{n_1 + n_2 + n_3}$$

$$= 1 - \left[(9+7+11)/36\right] = 9/36 = 0.25$$

9.10 (c)

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.FISH
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into METHOD				
From METHOD	**1**	**2**	**3**	**Total**
1	10 83.33	2 16.67	0 0.00	12 100.00
2	2 16.67	8 66.67	2 16.67	12 100.00
3	0 0.00	1 8.33	11 91.67	12 100.00
Total	12 33.33	11 30.56	13 36.11	36 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for METHOD				
	1	**2**	**3**	**Total**
Rate	0.1667	0.3333	0.0833	0.1944
Priors	0.3333	0.3333	0.3333	

(c) Error rate = 1 - Correct Classification rate

$$= 1 - \frac{n_{11} + n_{12} + n_{33}}{n_1 + n_2 + n_3}$$

$$= 1 - [(10 + 8 + 11)/36] = 7/36 = 0.1944$$

9.10 (d)

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.FISH
Cross-validation Summary using Linear Discriminant Function

Number of Observations and Percent Classified into METHOD				
From METHOD	1	2	3	Total
1	7 58.33	5 41.67	0 0.00	12 100.00
2	4 33.33	5 41.67	3 25.00	12 100.00
3	0 0.00	1 8.33	11 91.67	12 100.00
Total	11 30.56	11 30.56	14 38.89	36 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for METHOD				
	1	2	3	Total
Rate	0.4167	0.5833	0.0833	0.3611
Priors	0.3333	0.3333	0.3333	

(d) Error rate = 1 - Correct classification rate

$$= 1 - \frac{n_{11} + n_{12} + n_{33}}{n_1 + n_2 + n_3}$$

$$= 1 - [(7+5+11)/36] = 13/36 = 0.3611$$

APPENDIX:

This section will have the entire SAS code.

 # 8.8

Code:

```
DATA work.FBEETLES;

  INFILE "/folders/myfolders/data/T5_5_FBEETLES.dat";
  INPUT NUM TYPE Y1 Y2 Y3 Y4;

TITLE "HW5 Q-8.8";

PROC IML;
  USE work.FBEETLES;
  READ ALL VAR {Y1 Y2 Y3 Y4} INTO X;
  X1 = X[1:19,];
  X2 = X[20:39,];

  RESET PRINT;
  N1 = NROW(X1);
  N2 = NROW(X2);

  X1BAR = 1/N1*X1`*J(N1,1);
  X2BAR = 1/N2*X2`*J(N2,1);

  S1 = 1/(N1-1)*X1`*(I(N1)-1/N1*J(N1))*X1;
  S2 = 1/(N2-1)*X2`*(I(N2)-1/N2*J(N2))*X2;
```

```
  Spl = 1/(N1+N2-2)*((N1-1)*S1+(N2-1)*S2);

  T1 = (X1BAR[1]-X2BAR[1])/SQRT(Spl[1,1]*(1/n1+1/n2));
  T2 = (X1BAR[2]-X2BAR[2])/SQRT(Spl[2,2]*(1/n1+1/n2));
  T3 = (X1BAR[3]-X2BAR[3])/SQRT(Spl[3,3]*(1/n1+1/n2));
  T4 = (X1BAR[4]-X2BAR[4])/SQRT(Spl[4,4]*(1/n1+1/n2));

  a = INV(Spl)*(X1BAR-X2BAR);
  as=J(4,1);

  as[1]=SQRT(Spl[1,1])*a[1];
  as[2]=SQRT(Spl[2,2])*a[2];
  as[3]=SQRT(Spl[3,3])*a[3];
  as[4]=SQRT(Spl[4,4])*a[4];

  z1 = a`*X1`;
  z1 = z1`;
  z2 = a`*X2`;
  z2 = z2`;
  PRINT X1BAR,X2BAR,Spl,T1,T2,T3,T4,a,as,z1,z2;

RUN;
```

 # 8.11

Code:

```
DATA work.FISH;

  INFILE "/folders/myfolders/data/T6_17_FISH.dat";
  INPUT METHOD AROMA FLAVOR TEXTURE MOISTURE;
RUN;

TITLE "HW5 Q-8.11";

PROC FORMAT;
  VALUE METHOD 1='METHOD 1' 2='METHOD 2' 3='METHOD 3';
RUN;

PROC CANDISC OUT=CAND;
  CLASS METHOD;
RUN;

PROC PRINT DATA=CAND;
RUN;

PROC PLOT DATA=CAND;
  PLOT CAN2*CAN1=METHOD;
RUN;
```

 # 9.10

Code:

```
DATA work.FISH;

  INFILE "/folders/myfolders/data/T6_17_FISH.dat";
  INPUT METHOD AROMA FLAVOR TEXTURE MOISTURE;
RUN;

TITLE "HW5 Q-9.10";

proc discrim data=FISH outstat=ftstat
method=NORMAL pool=yes list crossvalidate;
class METHOD;
var AROMA FLAVOR TEXTURE MOISTURE;

proc discrim data=FISH outstat=ftstat
method=NORMAL pool=no list crossvalidate;
class METHOD;
var AROMA FLAVOR TEXTURE MOISTURE;

proc discrim data=FISH outstat=ftstat
method=npar k=5 pool=yes list crossvalidate;
class METHOD;
var AROMA FLAVOR TEXTURE MOISTURE;
```