In the homework assignments 8 and 9, you analyzed the fish data using MANOVA and Profile analysis. For this homework you will analyze it using discriminant analysis.

1. **It is reasonable to perform a discriminant analysis on the fish data to describe the differences between cooking methods? Why or why not.**

   Yes it is reasonable because the test that the cooking methods do not different on the four measures (i.e., $\mu_1 = \mu_2 = \mu_3$) was rejected in the MANOVA:

   \[
   \text{Wilks } \lambda^* = 0.22448732, \quad F_{8,60} = 8.33, \quad p < .0001
   \]

2. **It is reasonable to assume that the covariance matrices are equal? Why or why not. If reasonable what estimated covariance matrix should be used to computer the vector of weights that defined Fisher’s linear discriminant function.**

   It is reasonable to assume that the covariance matrices are equal because the test statistic for testing $H_0 : \Sigma_1 = \Sigma_2 = \Sigma_3$ equals $X^2 = 14.221958, df = 20$ and $p = 0.82$.

   The best choice is the pooled covariance matrix.

3. **How many linear discriminant are needed to capture the differences between groups (report null hypotheses, test statistic, p-value).**

   Only one linear discriminant is needed, because

   **Test 1:** The null hypothesis could be stated as any one of these:...
   
   - $H_0 : \rho_1 = \rho_2 = 0$ versus not $H_0$.
   - $H_0 :$ The two canonical correlations equal 0.
   - There is no relationship between the $y_1$–$y_4$ set of variables and the cooking methods
   - Neither discriminant function will distinguish between the methods in terms of the $y_1$–$y_4$ measures.
• Since eigenvalue_{i} = \rho_{i}^{2} / (1 - \rho_{i}^{2}), testing whether \rho_{1} = \rho_{2} = 0, is the same as testing whether the eigenvalues that are measures of differences between methods equals 0; that is, the ratio of between versus within squared distances equals 0.

\[ F_{8,60} = 8.33 \quad p < .01 \]

Reject \( H_{0} \) and conclude at least one canonical correlation is not zero or at least one discriminant provides information about the differences between methods in terms of the measures \( y1 - -y4 \).

**Test 2:** The null hypothesis could be

- \( \rho_{2} = 0 \) versus \( \rho_{2} \neq 0 \).
- The second eigenvalue equals 0.
- etc.

\[ F_{3,31} = 1.32 \quad p = .29 \]

Retain \( H_{0} \).

Conclusion: only one linear discriminant is needed (to adequately capture the nature of the differences between the cooking methods).

4. **Report the weights used to define the significant functions found in part 4. Interpret (i.e., how are the cooking methods different?)**

\[ y_{da} = 0.1189(\text{aroma}) + 3.0644(\text{flavor}) - 1.9924(\text{texture}) - 0.7760(\text{moisture}) \]

These are the “raw canonical coefficients” and should be OK to interpret since the variables are on comparable scale and have similar variances.

The major difference between the methods appears to be in terms of flavor and a bit of a contrast between texture and moisture versus flavor. A large value on the discriminant function would indicate a method has a high score on flavor and lower scores on texture and moisture. A small value on \( y_{da} \) would indicate a method leads to a fish that has a low score on flavor but high scores on texture and moisture.

5. **Report the means on the (significant) discriminant function(s) and comment.**

The CANDISC Procedure

Class Means on Canonical Variables
method     Can1

1  1.709200429
2  0.511874865
3  -2.221075294

SAS Note: These were computed used the weights given above however the values of the variables were mean centered; that is,

\[ y_{dak} = 0.1189(\bar{y}_{1k} - \bar{y}_1) + 3.0644(\bar{y}_{2k} - \bar{y}_2) - 1.9924(\bar{y}_{3k} - \bar{y}_3) - 0.7760(\bar{y}_{4k} - \bar{y}_4) \]

for methods \( k = 1, 2, 4 \).

It is still correct is you use took the linear combination of the non-centered data in which case the means on the first discriminant will differ by \( a'\bar{y} = 1.01512 \), that is you would have gotten

method     Mean on DA

1  2.7243209
2  1.5269953
3  -1.205955

Method 1 has the largest value on the discriminant function which implies that fish cooked using this method tend to be high on flavor but lower on texture (and a bit lower on moisture). What seems to be “driving” the significant difference is a contrast between flavor and texture. Method 3 has the smallest value and implies that fish cooked using method 3 tend to be lower on flavor and higher on texture (and moisture). Methods 2 is in between but closer to method 1.

6. Plot the observation’s values on the 2 discriminant function and identify which observation corresponds to which cooking method. Comment on how well the groups are separated.
Methods 1 and 3 are well separated; however, values for fish cooked by method 2 overlap with methods 1 and 3. Looking at this figure I expect there would be miss-classification errors mostly for fish cooked by method 2.

7. In terms of classification, how well do the discriminant functions perform?

Using equal priors (since there are the same number of fish per condition), yields the following

<table>
<thead>
<tr>
<th>From method</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Total</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>12</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>75.00</td>
<td>25.00</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>12</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>25.00</td>
<td>58.33</td>
<td>16.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>11</td>
<td>12</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>8.33</td>
<td>91.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>11</td>
<td>13</td>
<td>36</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>33.33</td>
<td>30.56</td>
<td>36.11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Priors 0.33333 0.33333 0.33333
As expected fish cooked by method 2 were miss-classified most often (i.e., 3 + 2 = 5 or 5/12 × 100% = 42%) than were those cooked by method 1 (i.e., 3 or 3/12 × 100% = 20%) or method 3 (i.e., 1/12 × 100% = 8%).

Note: You can get much better (perfect) classification if you use “method=npar r=.5”. The formula for this:

The DISCRIM Procedure

Posterior Probability of Membership in Each method

\[ m(X) = \text{Proportion of obs in group } k \text{ within the radius } R \text{ of } X_k \]

\[ Pr(j|X) = \frac{m(X) \text{ PRIOR }}{\sum (m(X) \text{ PRIOR })} \]

8. Based on just the first discriminant function and given a new observation with value \( y_o = (5, 6, 6, 6)' \), what cooking method would you predict was used for this fish?

To answer this problem, I used equation (5) and computed \( y_o \)'s value on the linear discriminant and got \( y_o = 1.355 \). Since this value is closest to the mean on the discriminant for method 1, I would predict that it was cooked by method 1.

Can get PROC DISCRIM to do the classification for you:

data newobs;
  input y1 y2 y3 y4;
datalines;
  5 6 6 6
run;

proc discrim data=fish testdata=newobs testlist testout=scored ;
  class method;
  var y1 y2 y3 y4;
  priors equal;
  title 'Discriminant Analysis via PROC DISCRIM';
run;
You can use a different method (e.g., npar).

9. Summarize your analyses on this data set considering your answer here and from previous homework assignments.

**Homework 8:** From the MANOVA, I found that the methods were significantly different ($F_{8,60} = 8.33, p < .0001$). Multivariate contrasts indicated that the methods 1 and 2 differed and 1 & 2 differed from method 3. Simultaneous $T^2$ intervals indicated that methods 1 and 2 differed in terms of flavor but were the same on the other three methods. Methods 1 & 2 differed from method 3 in terms of both flavor and texture where Methods 1 & 2 had lower values on texture but higher values on flavor than method 3. Lastly, these results were somewhat consistent with ANOVAs and subsequent pairwise comparisons using Tukey’s method. To summarize the pairwise comparisons

- Flavor: Methods 1 & 3 differed but methods 1 & 2 and methods 2 & 3 did not differ.
- Texture: Methods 2 & 3 differed but not methods 1 & 2 or methods 1 & 3.
- Moisture and aroma: no differences among methods.

**Homework 9:** Profile analysis was conducted on only methods 1 & 2. These were found to be parallel, coincident and flat; in other words, there were not differences between methods and no differences over variables.

**Homework 10:** The biggest difference is between methods 1 and 3 and the difference is mostly in terms of a contrast between flavor and texture where Method 1 has better flavor than 3 but Method 3 has better texture. Method 2 is “in between” Methods 1 and 2 on these variables.

**Conclusion:** Homework 10 conclusion pretty much sums up the findings.