Biology 680, Data Analysis in Biology
Notes on grading of Final Exam

19 December 1996

1. circle **column** (vs. row) in response to query about type of vector.

Use the Leslie Matrix as the premultiplier and the Age vector as the postmultiplier (as stated in instructions) to find:

\[
\begin{array}{cc}
nt+1 & nt+2 \\
127.5 & 108.75 \\
30 & 51 \\
24 & 12 \\
0 & 12
\end{array}
\]

2. Principal Component Analysis on Correlation vs. V-CV matrix:

<table>
<thead>
<tr>
<th>lambda interpretation</th>
<th>R matrix</th>
<th>V-CV matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC I</td>
<td>PC II</td>
<td>PC I</td>
</tr>
<tr>
<td>50%</td>
<td>15%</td>
<td>94%</td>
</tr>
<tr>
<td>many variables with tail (-)</td>
<td>head (+) vs. total 0.87 tail 0.87</td>
<td>tail total -0.48: length of tail</td>
</tr>
<tr>
<td>weak positive loadings: size</td>
<td>contrast of extaudmea length</td>
<td>vs. length of core</td>
</tr>
</tbody>
</table>

Note in the V-CV matrix that the variances of total length and tail length are gigantic compared to variances of all other variables (scan through diagonal elements of the matrix). Also note from the correlation matrix that total length and tail length are highly positively correlated (r=0.8662). Thus, these variables drive the PCA on the V-CV matrix. Both these points should be discussed when comparing the outputs.

With the ANOVA on PC I scores, PC I from the correlation matrix differed significantly among groups overall and multiple comparisons (Tukey’s) showed that all but 6 pairwise differences were significant. Slightly better separation of groups occurred using PC I from the V-CV matrix (F statistic is higher and significance for all but 4 pairwise comparisons).

**Note that Discriminant analysis will always perform better** at distinguishing among groups than the 2-step procedure (PCA followed by ANOVA on PC scores) because Discriminant Analysis maximizes the among-group component of overall variability, while PCA examines overall variability without respect to group identity. If you compared the results above to the results of the Discriminant Analysis earlier in the semester, you were comparing apples and oranges—the skin variables were not included in the DFA earlier in the semester. Inclusion of the skin variables in a DFA for comparison with the PCA-ANOVA 2-step shows that DFA results in more
significant among-group differences and all but 2 pairwise comparisons are significant.

3. I used the variables loglay, eggvol,dwg1,pw,fw,clsize,pos,nhatch,nfledg--if you used a somewhat different set you'd get somewhat different results than those noted below. If you used hatch date instead of lay date, you would have different results, for example. In the Stepwise Discriminant Analysis, FW (P=0.0310) entered first, followed by loglay (P=0.0336) and POS (P=0.0277). The classification results aren't great: 77% of resites are classified correctly but only 48% of the non-resites are correctly classified. The crossvalidation classification shows we'd do about as well by flipping a coin.

The logistic regression enters the same three independent variables in the same order. At a cutpoint of 0.660 in the classification table, the point at which misclassification errors are most nearly balanced for both groups, about 62% of both groups are classified correctly. Comparing other lines of this table to the DFA classification shows that Logistic Regression does a slightly better job at correctly classifying the 2 groups. Perhaps this is because the underlying assumption of equality of within-group V-CV matrices is violated (see DISCRIM output).

You can conduct 2-way ANOVA's on loglay and FW, but not POS, since it is only binary (you could do a 3-way contingency table analysis, but nobody got too carried away on that). For loglay, there is a significant year effect and a significant resite effect, meaning that laying dates varied among years and were earlier for resites than non-resites (based on looking at means). There was no interaction between year and resite, indicating that the difference between resite and non-resite groups was consistent across years. For FW, neither the year effect (P=0.0631) nor the resite effect (P=0.1019) was significant, a counter-intuitive result based on the above analyses. Note that the ANOVA apportions considerable variability to each year and resite codes but rather equally between them. Thus, adding year shows that some of the variability attributed to resite in the Discriminant Analysis and Logistic Regression actually should be attributed to year, a categorical variable we couldn't include in those analyses. Thus the significance of a particular variable can depend on which other variables are in the model--a point we'd talked about, but this example shows it well.