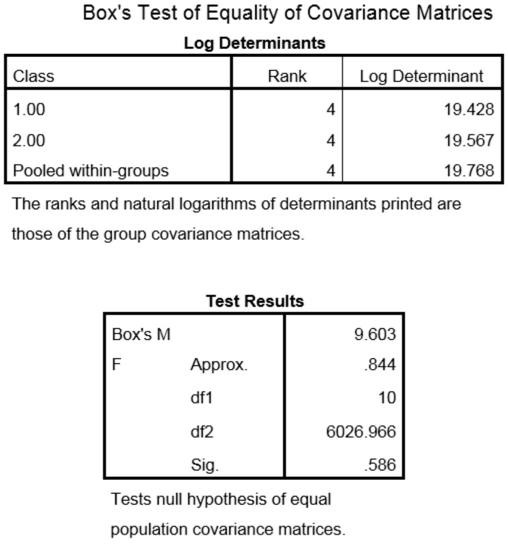
CSC334/424: Assignment #3
Solution

Total points: 65 for graduate students; 35 + 30 extra credit for undergraduate students

Note: Undergraduate students must attempt all problems and will be graded the same way as graduate students. However, a score of 35 of the 65 points will be considered full credit.

Problem 1 (20 points): Data were collected on two species of flea beetles (a) Halticus oleracea and (b) Halticus carduorum. Measures of thorax length (THORAX), elytra length (ELYTRA), length of the second antennal joint (AJ2), and length of the third antennal joint (AJ3) in microns. These data are stored under beetle.txt. Perform a linear discriminant analysis between the two beetle species assuming equal population proportions.

i. Test for the equality of the variance-covariance matrices between the two species. What are your conclusions?



Box's Test of Equality of Covariance Matrices

Log Determinants

Class	Rank	Log Determinant
1.00	4	19.428
2.00	4	19.567
Pooled within-groups	4	19.768

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

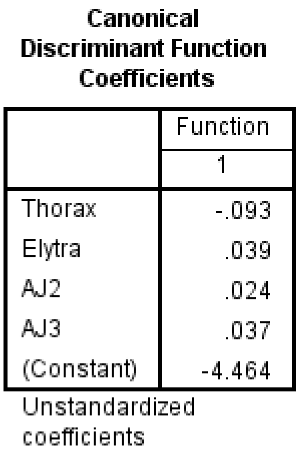
Test Results

Box's M		9.603
F	Approx.	.844
	df1	10
	df2	6026.966
	Sig.	.586

Tests null hypothesis of equal population covariance matrices.

The Box's Test of Equality of Covariance Matrices test shows a p-value of .586. In this case we are looking to fail to reject the null hypothesis, so something above .001 is good. Also note that we could double-check with the Log Determinants – they are very similar so we are in good shape to use LDA.

ii. Give the linear discriminant classification function for the two beetle species. Under what condition would an unidentified specimen be classified as Halticus oleracea?



Canonical Discriminant Function Coefficients

	Function
	1
Thorax	-.093
Elytra	.039
AJ2	.024
AJ3	.037
(Constant)	-4.464

Unstandardized coefficients

One discriminant function is needed to differentiate between the two species. A new specimen is measured for each of the original variables and then evaluated against the discriminant function specified by the coefficients in this table. For a specimen with values thorax, elytra, AJ2, AJ3, *we test:*

-*.093**thorax + *.039**elytra + *.024**AJ2 + *.037**AJ3 – *4.464 > 0*

if this is true (quantity is > 0), then we declare class 2, otherwise, class 1. Halticus oleracea is class 1.

Note – R does not automatically generate that constant term as part of the LDA process. That term represents the value that is used to separate the classes on the projected line. There are multiple ways to calculate it. The simplest thing to do is take the mean of the each class group in the projected space (i.e. using scores on the LDA vector), and then use the weighted average of those two means as the splitting criteria. For weighting, use the number of points in each class.

iii. Suppose that an unidentified specimen with the following measurement is obtained:

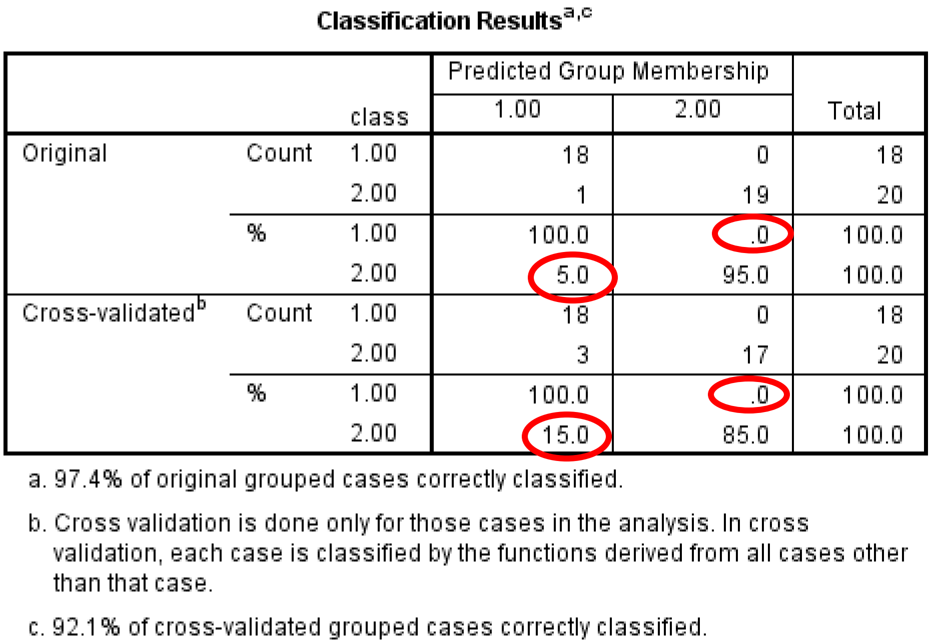
Variable	**Measurement**
Thorax	184
Elytra	275
AJ2	143
AJ3	192

Which species would you classify this specimen into?

-.093(184) + .039*(275) + .024*(143) + .037*(192) – 4.464 = -.315*

Because -.315 < 0, we classify this new specimen as Halticus oleracea.

iv. Give the apparent confusion matrix for the data. Estimate the percentage of beetles of each species that will be misclassified under the linear discriminant rule.



Classification Results[a,c]

		class	Predicted Group Membership		Total
			1.00	2.00	
Original	Count	1.00	18	0	18
		2.00	1	19	20
	%	1.00	100.0	.0	100.0
		2.00	5.0	95.0	100.0
Cross-validated[b]	Count	1.00	18	0	18
		2.00	3	17	20
	%	1.00	100.0	.0	100.0
		2.00	15.0	85.0	100.0

a. 97.4% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 92.1% of cross-validated grouped cases correctly classified.

Looking at the above confusion matrix, we can see that for the non-cross-validated data, we expect 5% of class 2 to be misclassified, but none of class 1. In the leave-one-out-cross-validated case, the estimate is 15% of class 2 will be misclassified. The cross-validated version is a better predictor of future classification performance.

Problem 2 (15 points):

(a) Both LDA and multiple linear regression can learn a function mapping multiple variables to predictions of values for another variable. What is the main difference in when they are used?

LDA is used when the dependent variable is binary as opposed to numeric (scalar/metric).

(b) Both LDA and PCA can be thought of as finding an optimal vector onto which to project data. What is the difference between the techniques in terms of what criteria the vector is chosen to optimize?

The PCA vector is chosen to optimize accounting for data variance. The vector chosen in LDA is optimized for discriminating power between specified classes of data points.

(c) Briefly describe what is being optimized in Fisher's Linear Discriminant (the optimization criterion function).

Fisher's Linear Discriminant optimizes the ratio of between-class scatter to within-class scatter.

(d) Hierarchical clustering can allow you to determine the number of clusters after the clustering has already been completed – how?

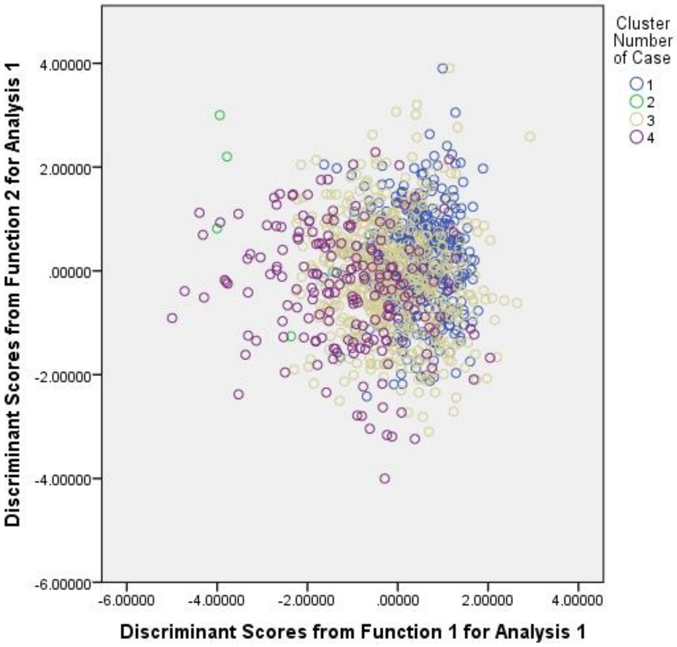
With hierarchical clustering, a full hierarchical structure of the data point relationships is created. At the end, deciding the number of clusters just means picking the subtrees that are considered clusters based on a threshold. Think of this like the dendogram visualization – the result of hierarchical clustering looks like a tree. You can just draw a line across it to cut the boundary of how many clusters you want. For k-means, you would have to rerun the algorithm if you want to change the number of clusters.

(e) Describe one advantage of DBSCAN over k-means and one advantage of k-means of DBSCAN.

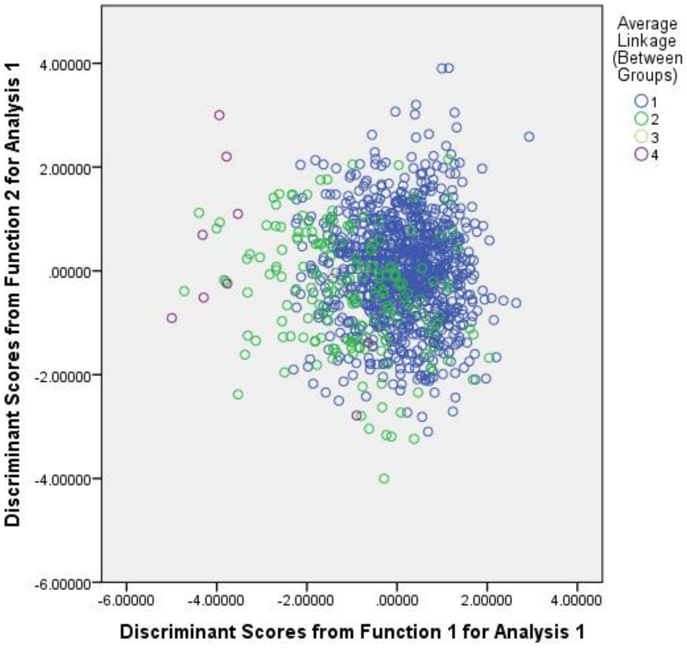
DBSCAN can find clusters of arbitrary shape and can label noise, while k-means can only discover convex shapes defined by a center point and is sensitive to noise. DBSCAN requires only one pass through the data, as opposed to multiple iterations, but it's complexity class is O(n^2) whereas k-means is O(tkn), which can be much faster because generally t, k << n.

Problem 3 (30 points):
We will revisit the data on faculty from the second lab, now called *faculty.sav*. Recall that the PCA projection was not very helpful for visualizing the data. We will now consider a projection using LDA. Recall that you can get more than one discriminant vector. Run LDA with the same data as before (variables *item13* through *item24*) and use *faculty rank* as the dependent variable. Keep two discriminant vectors and save the scores of all data points on these discriminants. Plot the LDA projection by plotting those new score variables.

a. Run k-means clustering and use the cluster assignment to color the points.

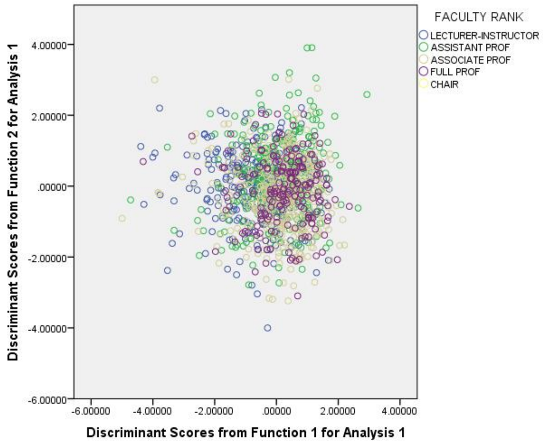


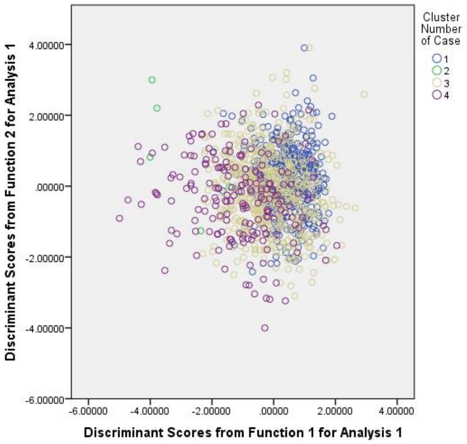
b. Run hierarchical clustering and use the cluster assignment to color the points.



c. Compare the k-means clustering to the correct labels.

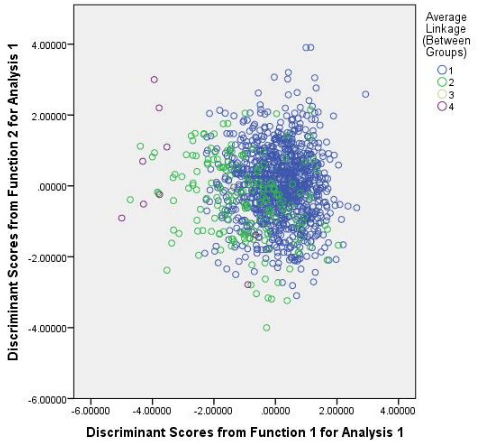
Here is the plot with the correct labels used for colors:

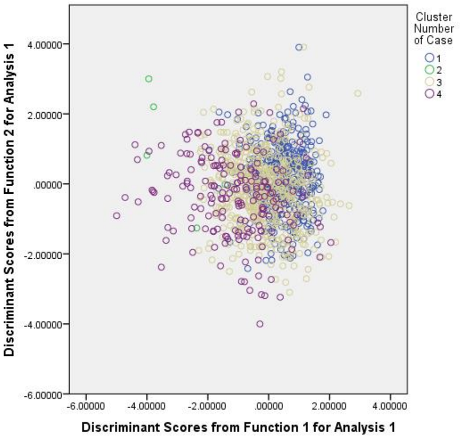




Keep in mind that k-means cannot produce clusters that overlap. The overlap we see is an artifact of the projection to 2D space. This question was intended as an exercise in using available tools for visualization. What we can see from this visual inspection is that the k-means clusters did not capture the same pattern as our chosen class. There may look to be some areas they have in common, like in the middle or upper right or lower left, but the projection and not-ideal color palette may be misleading. Use this type of comparison to guide further study, e.g. with a confusion matrix of labels against cluster values, or a stacked histogram. We could try to see if other choices of label match up better with the clusters or try different clustering or rerunning k-means because it's non-deterministic. However, since there are labels, what we probably want to do is see if we can build a classifier for faculty rank or some of the other label variables.

d. Compare the k-means clustering to the hierarchical clustering. What part of the process of hierarchical clustering accounts for the sparse outlying cluster?





These look much more similar than k-means does to the faculty rank. It looks like hierarchical clustering swallowed up 3 and 1 from k-means into 1. Cluster 2 from hierarchical looks similar to cluster 4 from k-means. On the left in hierarchical *clustering, there is a sparse cluster (#4) that got bigger than its counterpart in k-means. These were points that are far away from most of the data but close to each other. In this case, hierarchical clustering, because it builds from the most similar pairs up, can pick up these points into their own cluster.*