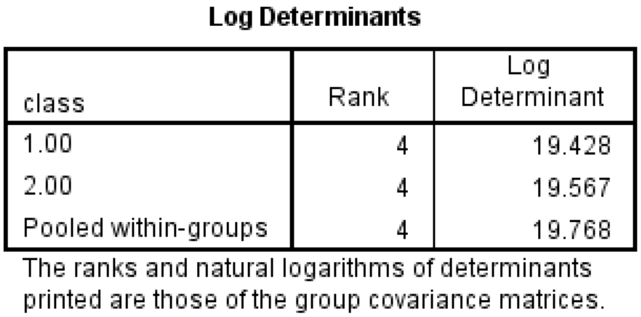
CSC 424 Assignment 3/4
Sarah Cummings

Problem 1- Perform LDA on beetle data assuming equal population proportions

a) Test for the equality of covariance matrix for the two species— conclusions?

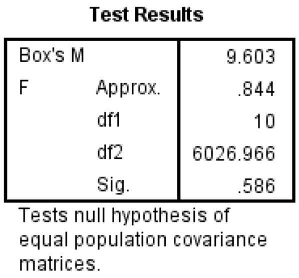


Log Determinants

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.

Box's Test of Equality of Covariance Matrices:



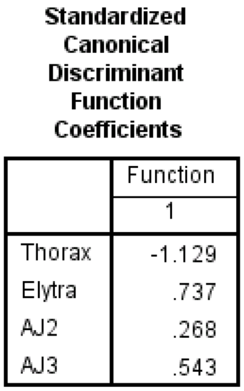
Test Results

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

Tests null hypothesis of equal population covariance matrices.

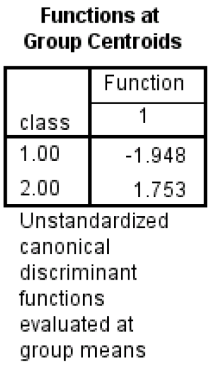
With an F statistic of 0.844 and p value 0.566, we fail to reject the null hypothesis that the covariance matrix for the two species is equal. Thus we can assume that "the vector of the dependent variables follow a multivariate normal distribution, and the variance-covariance matrices are equal across the cells formed by the between-subjects effects" according to the SPSS help document. The variance across the groups is equal.

b) Give LD classification function for each of the two types of beetles. Under what condition would an unidentified beetle be classified as a Halticus olercea?



Standardized Canonical Discriminant Function Coefficients

	Function
	1
Thorax	-1.129
Elytra	.737
AJ2	.268
AJ3	.543



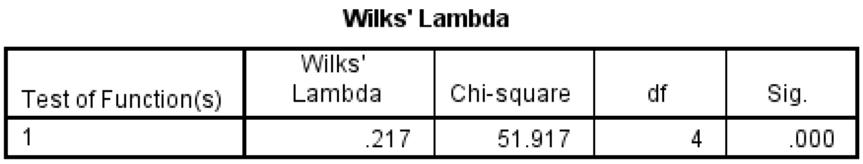
Functions at Group Centroids

class	Function
	1
1.00	-1.948
2.00	1.753

Unstandardized canonical discriminant functions evaluated at group means

As seen in the output at the left,
DF= -1.29(thorax) +0.737(elytra)+0.268(AJ2) + 0.543(AJ3)

If the DF value for a given case is closer to -1.948, the beetle belongs to class 1 (Halticus olercea) and if the value is closer to 1.753, the beetle belongs to class 2 (Halticus carduorum). A beetle is thus classified as Halticus olercea if its DF value is closer to -1.948 than it is to 1.753.



Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.217	51.917	4	.000

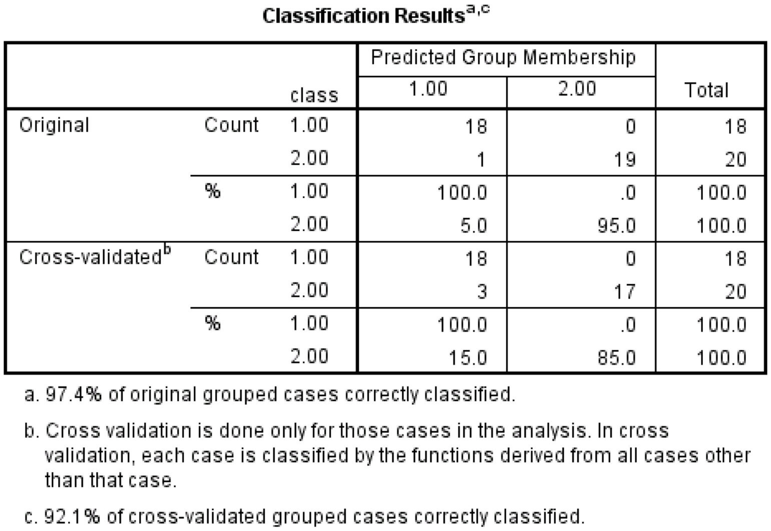
Also note that Wilks Lambda is significant, meaning we have a significant model for predicting group membership.

c) Unidentified beetle has following measurements: Thorax (184), Elytra (275), AJ2 (143), and AJ3(192). What type of beetle would it be classified as?

Given the function from part b, we calculate its DF value as:
DF= -1.29(184) +0.737(275)+0.268(143) + 0.543(192)= 69.571
and thus this beetle would be classified as class 2.

d) Apparent confusion matrix and estimate of percentage of each type of beetle that will be misclassified under the LD rule

The apparent confection matrix is the top half of the matrix below. With this matrix, we estimate the 97.4% of beetles are correctly classified and thus 2.6% of the beetles are misclassified.



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.

e) Cross validation matrix and estimate of percentage of each type of beetle that will be misclassified under the LD rule

The cross validation matrix is the bottom half of the matrix above. With this matrix, we estimate that 92.1% of the cross validated grouped cases are correctly classified and thus 7.9% are misclassified.

Problem 2

a) What is the main difference in when LDA and multiple linear regression are used?
LDA is a form of feature extraction used for classification, whereas Multiple Linear Regression is a form of feature selection used for prediction with a quantitative dependent variable.

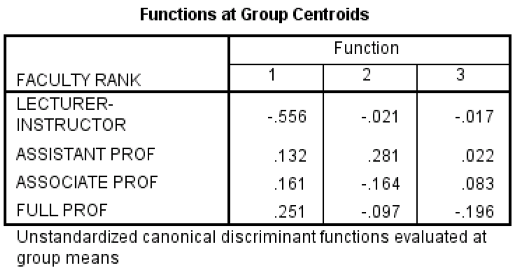
b) What is the difference between the techniques in terms of what criteria the vector is chosen to optimize in LDA versus PCA?
LDA is a form of supervised learning whereas PCA is unsupervised. LDA focuses on modeling the difference between classes, and projects the data in that there is maximum separation between groups. PCA, on the other hand, does not focus on group separation and rather aims to capture the most amount of variation in the data.

c) Briefly describe what is being optimized in Fischer's Linear Discriminant.
In FDA, our goal is to find a low-dimensional space that maximizing the separation between classes when the data is projected. We want to maximize $J(W)= |Y_1-Y_2|^2 / (s_{y1}^2 + s_{y2}^2)$

d) Hierarchical clustering can allow you to determine the number of clusters after the clustering has already been completed – how?
By using a dissimilarity measure, we can provide a certain distance or dissimilarity criteria for our algorithm that will find what number of clusters makes we will have to fit in to the criteria. In this way, the number of clusters is determined at the end of the clustering process.

e) Describe one advantage of DBSCAN over k-means and one advantage of k-means.
DBSCAN is more advantageous in that it is resistant to noise and it also can form clusters of different sizes and shapes. K-means is more advantageous because it works well with high dimensionality data and clustering with varying densities.

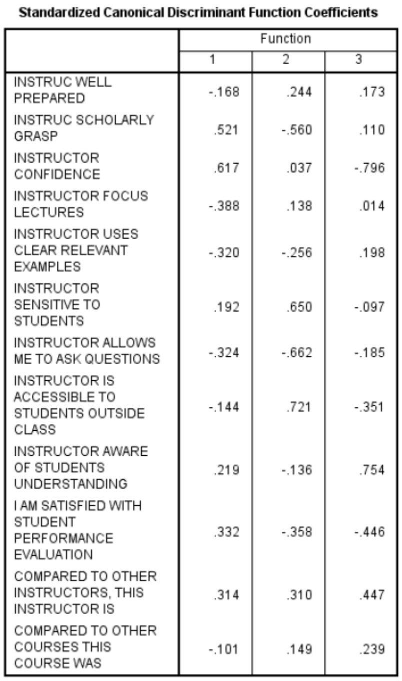
Problem 3- K means with faculty data, with faculty rank as the dependent variable and item13-item24 as independent variables. First, run LDA with those variables Keep two discriminant vectors and save the scores of all data points on these discriminants. Plot the LDA projection by plotting those new score variables.



Functions at Group Centroids

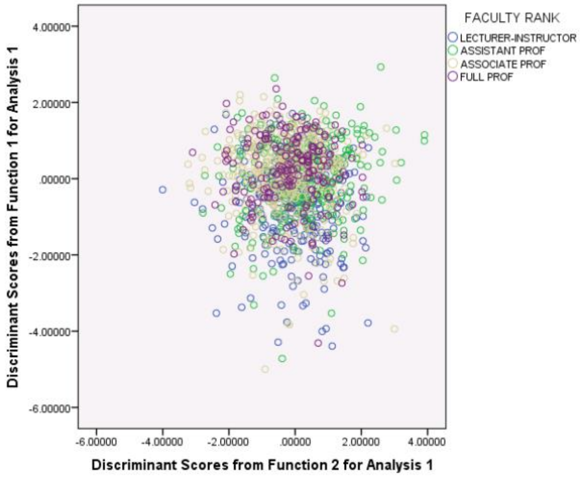
FACULTY RANK	Function 1	2	3
LECTURER-INSTRUCTOR	-.556	-.021	-.017
ASSISTANT PROF	.132	.281	.022
ASSOCIATE PROF	.161	-.164	.083
FULL PROF	.251	-.097	-.196

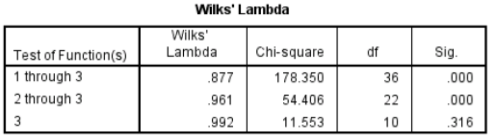
Unstandardized canonical discriminant functions evaluated at group means



Standardized Canonical Discriminant Function Coefficients

	Function 1	2	3
INSTRUC WELL PREPARED	-.168	.244	.173
INSTRUC SCHOLARLY GRASP	.521	-.560	.110
INSTRUCTOR CONFIDENCE	.617	.037	-.796
INSTRUCTOR FOCUS LECTURES	-.388	.138	.014
INSTRUCTOR USES CLEAR RELEVANT EXAMPLES	-.320	-.256	.198
INSTRUCTOR SENSITIVE TO STUDENTS	.192	.650	-.097
INSTRUCTOR ALLOWS ME TO ASK QUESTIONS	-.324	-.662	-.185
INSTRUCTOR IS ACCESSIBLE TO STUDENTS OUTSIDE CLASS	-.144	.721	-.351
INSTRUCTOR AWARE OF STUDENTS UNDERSTANDING	.219	-.136	.754
I AM SATISFIED WITH STUDENT PERFORMANCE EVALUATION	.332	-.358	-.446
COMPARED TO OTHER INSTRUCTORS, THIS INSTRUCTOR IS	.314	.310	.447
COMPARED TO OTHER COURSES THIS COURSE WAS	-.101	.149	.239





Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 3	.877	178.350	36	.000
2 through 3	.961	54.406	22	.000
3	.992	11.553	10	.316

a) Run k-means clustering and use cluster assignment to color the points
b) Run hierarchical clustering and use cluster assignment to color the points
c) Compare k-means clustering to the correct labels
d) Compare k-means to hierarchical clustering. What part of the hierarchical clustering process account for sparse outlying cluster?