Stats 519: Final

Due on May 12, 2009

Dr. Stephen Lee 1:30

Monte J. Shaffer

Problem 1

The US News and World Report lists the 50 best cancer hospitals in the nation in 7/19/99 based on some criteria described below:

- **RANK** - The rankings were obtained by using a particular composite index that was computed using various measurements
- **NAME** - Hospital's name
- **INDX** the composite index used to do the ranking in the first column
- **REPUT** - the percentage of cancer specialists surveyed who named the hospital
- **MORTAL** - mortality rate [below 1 is good, above 1 is bad]
- **COTH** - indicting the membership in Council of Teaching Hospitals
- **TECH** - a score indicating the availability of key technologies
- **DISCH** - the number of patients discharged
- **RN_B** - the hospital-wide ratio of the number of registered nurses to the number of beds.

The highest ranked cancer hospital are listed; the data consists of the 50 best (membership in the Council of Teaching Hospitals is a requirement for consideration)

RANK	NAME	INDX	REPUT	MORTAL	COTH	TECH	DISCH	RN_B
1	Sloan-Ketterning	100.0	73.0	1.05	Yes	6.0	3544	1.85
2	M. D. Anderson	99.7	67.5	0.78	Yes	6.0	3683	1.87
3	Johns Hopkins	65.6	33.4	0.70	Yes	7.0	1278	1.33
4	Mayo Clinic	60.4	26.0	0.57	Yes	7.0	2589	1.10
5	U. Wash. Medical	39.0	9.1	0.67	Yes	6.0	606	2.03
6	Duke U Medical	38.6	10.7	0.82	Yes	7.0	2523	1.63
7	University of Chicago	37.2	6.6	0.68	Yes	7.0	1116	1.26
8	Fox Chase Cancer	35.5	5.7	0.54	Yes	4.0	872	1.88
9	U. Michigan Medical	35.2	1.3	0.56	Yes	7.0	1221	1.46

With an intent to group, we could just use the rankings (which are based on the INDX variable; REPUT is a popularity score and correlates strongly with the INDX score, suggesting the algorithm may heavily weight on that variable). A simple plot of these two variables identifies 3 clusterings: the first two hospitals, the second two hospitals, and everyone else. The intention of the grouping is important to identify the ultimate classifications.

The second plot shows some key factors related to quality cancer care. Reputation, Technology, and Nurse Care. Obviously Mortality and Discharge may normally be of interest, but generally, with cancer, expectations are high that death will occur. If caught early, you want the best technology to remove the cancer; and if late, you want the best care in the treatment of the terminal illness.

We can group in other fashions, as will be described below: for example, we could group based on just one variable (like DISCH which would capture a volume of work or size of the hospital). An exploration of the data will identify key factors to consider in choosing a grouping technique. First, I would do a simple distribution analysis on the individual variables to identify their structure and normality. Histograms of several variables align with the initial ranking results. Since the goal is not to reverse-engineer the rankings, but proceed to best classify, this approach will not be discussed further.

Since all of the top 50 have YES to COTH, this variable will not help in finding dissimilarities, so will be excluded from the remaining exploration. Also, as described above, ranking is based on the INDX and

Problem 1 continued on next page...

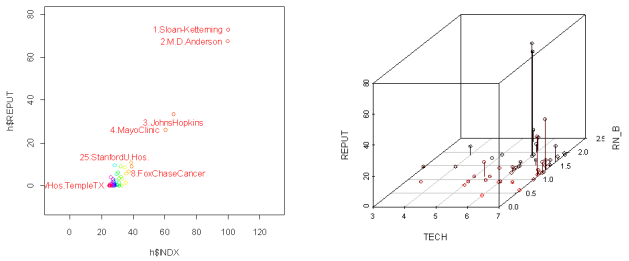

Figure 1: Simple Analysis

REPUT which are both highly (.983) correlated. Only the REPUT score will be used, since the INDX is a final score, which is then ranked.

PCA.

I like starting with PCA to identify orthogonal variables that explain the most variance.

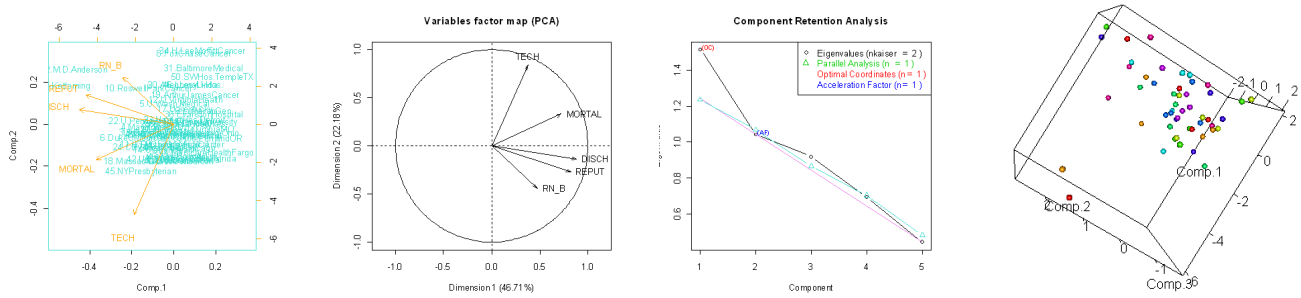

Figure 2: PCA

EFA.

EFA can reduce the variables, but doesn't really help much with grouping. The first graph identifies EFA compared to the second with PCA. The last is a transpose of the data and a multi-stage, multi-step bootstrapping of the hierarchical clustering using the WARD's technique, which is an alternative way to consider the reducing of variables `pvclust`.

Clustering.

`pvclust` was calculated which is a hierarchical clustering using WARD; `kmeans` was also run, with up to 5 choices. A comparison of these clusters to classification will be reported in the next section.

Centers for 5 groups is reported below (kmeans):

```
          REPUT      MORTAL        TECH       DISCH       RN_B
1  4.383358094   1.5530211 -0.04228496   3.5131314  1.1986300
2 -0.266982095 -0.2057749 -0.12059044 -0.5404056  1.2883967
3 -0.301686126 -1.0723971 -2.43339876 -0.4961541  0.1336708
4 -0.002693571  0.6401352  0.56636219  0.2802295 -0.1111475
5 -0.392144175 -0.8498041 -0.11972994 -0.4879266 -0.9294053
```

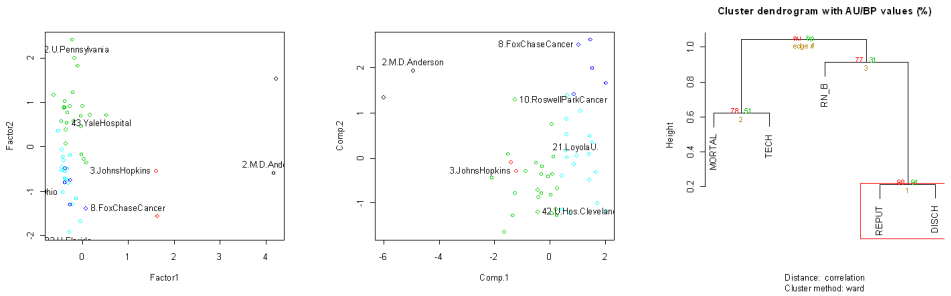

Figure 3: EFA

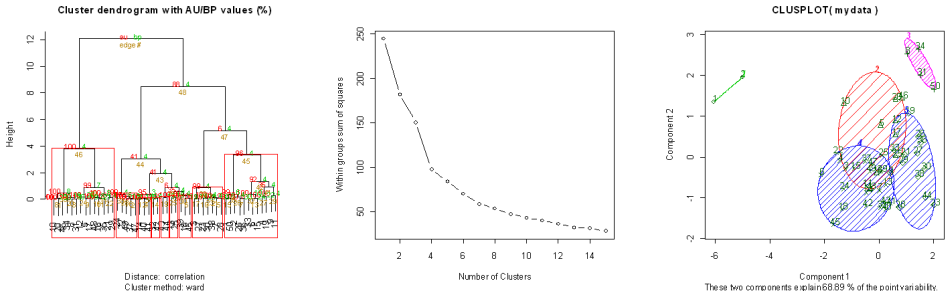

Figure 4: Clustering

`Mclust` using library `mclust02` was used to model a mixed Model-based clustering technique. Centers for 5 groups is reported below (classification):

```
           REPUT     MORTAL       TECH      DISCH        RN_B
[1,]  4.38335809  1.6150005 -0.1406084 -0.3271366 -0.38590612
[2,]  1.55302115 -0.5933062  0.7191128 -1.0070256 -0.67370069
[3,] -0.04228496  0.9644998  0.4650800 -2.0878737 -0.04696275
[4,]  3.51313145  1.0600313  0.1285541 -0.4445441 -0.52054393
[5,]  1.19863004 -0.3804475  0.2176944  0.2032965 -0.41029725
```

Correlations between the two results (kmeans and classification) are shown in the following table:

```
       Mclust
kmeans  1  2  3  4  5
     1  2  0  0  0  0
     2  0  0  4  1  4
     3  0  0  0  4  0
     4  0  2 19  0  1
     5  0  0  0  0 13
```

Based on this information, with an intent to capture the most number of classifications (personal choice), and based on the initial discussion of rankings, I would conclude with the Mclust classification results. Although there may be limitations to the Gaussian assumption, the probabilistic assignment to groups is of value.

Since the scaled scores are reported, refer to the centroids under the classifications. The first cluster is high reputation, high mortality, good level of Nurse care, medium sized (medium level of DISCH), and

rather low-tech. This may suggest that this cluster is a "quality care for terminal illness" where the service is a key feature. The second cluster is similar in some aspects with two notable exceptions: one, it uses high technology to fight agressively, and there is some success. The third cluster is noteworthy for being small (based on DISCH). The fourth cluster is similar to the first, with slightly higher technology and lower reputation. The fifth cluster is similar to the second cluster with lower technology and higher volume (DISCH).

Classification.

Discriminant analysis

Conclusion.

Problem 2

Consider the following 7 (n=food) by 5 (p=attributes) data set:

```
         energy protein fat calcium iron
beef        180      22  10      17  3.7
chicken     170      25   7      12  1.5
mackerel    155      16   9     157  1.8
salmon      120      17   5     159  0.7
sardines    180      22   9     367  2.5
tuna        170      25   7       7  1.2
shrimp      110      23   1      98  2.6
```

All of the analyses (PCA, EFA, Clustering, Classification) should be conducted on the standardized data.

- **a)** Find the first two PCA directions and show that they are perpendicular to each other.

```
Xs = scale(f);
R = t(as.matrix(Xs))%*%as.matrix(Xs)/(n-1)
R.e = eigen(R);
     Lambda    = R.e$values;
     U    = R.e$vectors;
     (u1 = U[1,]);  ## FIRST DIRECTION
[1]  0.66919290 -0.00476882  0.23590335 -0.18474602  0.67998282
     (u2 = U[2,]);  ## SECOND DIRECTION
[1]  0.21785297 -0.68199720 -0.08697995 -0.59707317 -0.35122360
     round((finalAnswer = doAnalysis(u1,u2)),2);
          ||u1||  ||u2||   theta   pi radians   u1.u2
u1 and u2   1.0     1.0     90.0       0.5       0.0
```

Orthogonal is demonstrated in the angle (90 degrees) and the dot product.

- **b)** Write the first principal component in terms of the 5 variables energy, protein, fat, calcium, and iron.

The first PCA is a linear combination of the original Xs: $PCA_1 = \sum_{i=1}^{p} c_i \cdot X_i = X \cdot u_1$

```
Xs.PCA=princomp(Xs);
          rownames(Xs.PCA$loadings)=myAttributes;
          rownames(Xs.PCA$scores)=myFoods;
     summary(Xs.PCA);
     round(Xs.PCA$scores[,1],3);

    beef  chicken mackerel   salmon sardines     tuna   shrimp
   1.864    0.409    0.018   -1.930    1.221    0.298   -1.881
```

- **c)** Find the correlation of first principal component and protein.

```
     round(cor(Xs.PCA$scores[,1],Xs[,2]),3);
[1] 0.314
```

- **d)** Write down the second principal component scores, and then compute the variance of this set of scores.

  $\lambda_2 = var(PCA_2)$, so:

```
     round(Xs.PCA$scores[,2],3);

    beef  chicken mackerel   salmon sardines     tuna   shrimp
  -0.459   -1.154    1.467    0.992    1.335   -1.153   -1.027

     round(var(Xs.PCA$scores[,2]),3);
[1] 1.474
     round(Lambda[2],3);
[1] 1.474
```

- **e)** Produce a biplot of the PCA result

```
# PCA Variable Factor Map
     library(FactoMineR)
     result = PCA(Xs) # graphs generated automatically

     biplot(Xs.PCA);
```

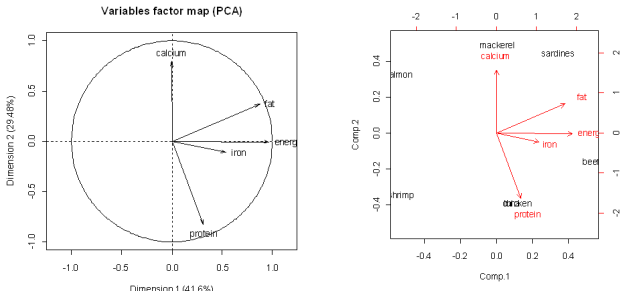

Figure 5: PCA

- **f)** Fit the smallest sufficient factor analysis model to the data. Write down this model in all details: model specifications, assumptions, parameters, parameter estimates, and the decomposition of the covariance matrix of the data.

```
multifactanal(factors=1:5, covmat=R, n.obs=n);

   n items factors total.df rest.df model.df        LL      AIC      AICc      BIC
1  7     5       1       15       5       10 -4.360927 28.72185 -26.27815 28.18096
2  7     5       2       15       1       14 -1.066844 30.13369 -22.36631 29.37643
3 NA    NA      NA       NA      NA       NA        NA       NA        NA       NA
4 NA    NA      NA       NA      NA       NA        NA       NA        NA       NA
5 NA    NA      NA       NA      NA       NA        NA       NA        NA       NA

factanal(factors = 2, covmat = R, n.obs=n, rotation = "varimax");
```

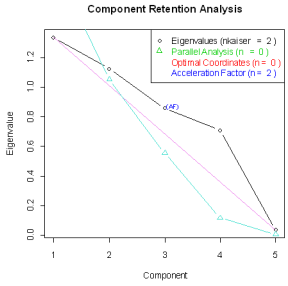

Figure 6: EFA

There is only enough data to fit a one- or two- factor model (degrees of freedom run out). According to above, the two-factor model fits slightly better. The goodness of fit (χ^2) verifies that we can retain the H_0: Good Fit on the two- but not one- factor model.

```
Call:
factanal(factors = 2, covmat = R, n.obs = n, rotation = "varimax")

Uniquenesses:
 energy protein     fat calcium    iron
  0.005   0.005   0.005   0.838   0.889

Loadings:
        Factor1 Factor2
energy   0.986   0.155
protein  0.206   0.976
fat      0.941  -0.331
calcium         -0.402
iron     0.305   0.137

               Factor1 Factor2
SS loadings      1.993   1.266
Proportion Var   0.399   0.253
Cumulative Var   0.399   0.652

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 2.31 on 1 degree of freedom.
The p-value is 0.128
```

The model is a derivative from PCA ($R = F'F$) becomes $R \approx F_c'F_c = \Lambda_c'\Lambda_c + \Omega$ where Λ_c is analagous to F_c and Ω contains the uniqueness (don't want uniqueness for a variable to be close to one: e.g., calcium and iron).

- **g)** Explain in matrix terms why fitting a FA model to the original data is the same as fitting it to the standardized data.

  The model is fit in terms of the correlation matrix R which mathematically is derived from the original data (cov $S = \frac{X_d'X_d}{n-1}$) or the scaled data (cor $R = \frac{X_s'X_s}{n-1}$). In order to fit the data the S will be scaled to unit variance to create R.

- **h)** Explain why the chi square statistics has only 1 degree of freedom for the 2-factor model.

  The degrees of freedom are calculated as follows: $df(R) - df(F_1) - df(F_2)$. The correlation matrix is symmetric, and the diagonals are one, so the only variation is the upper triangle (excluding the diagonal): for five parameters that is 4+3+2+1 = 10. The first factor eats up 5 degrees of freedom; the second factor eats up 4 degrees of freedom: $10 - 5 - 4 = 1$. In general p parameters eats up p,$p-1$,$p-2$, etc. for each factor [Notes from March 11]

- **i)** Do you know how much percent of the variation of iron is not explained by the common factors in the smallest sufficient factor analysis model?

```
Uniquenesses:
 energy protein     fat calcium    iron
  0.005   0.005   0.005   0.838   0.889
```

  Yes, 88.9% of iron is due to uniqueness and not to the two common factors.

- **j)** Can you find the first factor direction? Why/why not?

  Sure, any matrix can be "eigen-ized." In this case $R - \Theta = \Lambda_c' \Lambda_c'$ is analagous to $R = F'F$ of PCA. Instead of finding the eigen values and vectors of R, I can find them for $R - \Theta$.

- **k)** Perform a hierarchical clustering analysis on the data using the Ward's criterion, and plot the dendrogram.

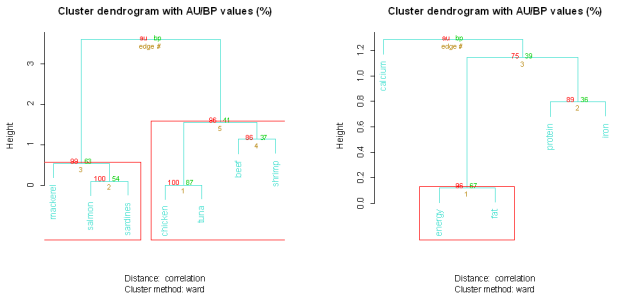

Figure 7: Ward Hclust with Bootstrapping `pvclust`

  Both the data and the transposed data are shown. The transposed data would be similar to a classification of the facors (e.g., factor analysis).

- **l)** Suppose we would like to discriminant between land and sea species of the data set. Find the Fisher direction and standardized the direction to unit length, and the related Fisher ratio.

```
mySea = c(0,0,1,1,1,1,1);
library(MASS);
(Xs.LDA = lda(Xs,mySea));
(Xs.PRED = predict(Xs.LDA,Xs));
table(Actual = mySea, Predicted = Xs.PRED$class);
```

```
      Predicted
Actual 0 1
     0 1 1
     1 0 5
```

```
     (w_star=Xs.LDA$scaling);

                LD1
energy   7.3123568
protein -4.0959138
fat     -7.6167388
calcium  0.4331651
iron    -0.3174329

     (w=w_star/vectorLength(w_star));

                 LD1
energy   0.64494255
protein -0.36125550
fat     -0.67178874
calcium  0.03820473
iron    -0.02799726

     (ratio_F = Xs.LDA$svd);
[1] 3.118029
```

The unit length vector is found by scaling the given vector (w^*) by its vector lenght: $w = \frac{w^*}{||w^*||}$.

- **m)** Find the value of the Fisher ratio along the first PCA direction.

  The ratio is defined as the between over within variance. The $var(PC_1) = \lambda_1 = 2.08$. In the scale model, the within is predefined to be 1. The Fisher ratio of PCA_1 is 2.08; the variance. The total variance of the model is the standardized variance times the parameters $p = 5$. [Notes April 20]

```
round(var(Xs.PCA$scores[,1]),3);
[1] 2.08
```

- **n)** Compare the Fisher ratio values from the previous two parts and comment on your observations.

  The LDA Fisher ratio is larger than the PCA Fisher ratio. This is intuitive since the PCA is a simple ratio with unit variance, but the LDA is the maximization of the ratio. The LDA maximizes between variance while simultaneously minimizing within variance.

- **o)** Compute the angle (in degree) between the Fisher direction and the first PCA direction.

```
round((finalAnswer = doAnalysis(u1,w)),2);
          ||u1||  ||w||    theta   pi radians   u1.w
u1 and w    1.00   1.00    75.60      0.42       0.25
```

About 75 degrees.

Problem 3

Consider just the two species `versicolor` and `virginica` in the iris data set which can be obtained using the R commands and use this data to answer the following questions:

```
ir=iris[-c(1:50),];
ir;

     n1=n2=50;
     N = n1+n2;

versi = ir[c(1:50),1:4];
     x1_bar = mean(versi);
     s1 = var(versi);
     w1 = s1*(n1-1);
virgi = ir[-c(1:50),1:4];
     x2_bar = mean(virgi);
     s2 = var(virgi);
     w2 = s2*(n2-1);

G=2; # number of groups
p=4; # number of variables

q1=n1/(N); q2=n2/(N);

Sp = Cw = (w1+w2)/(n1+n2-2); # pooled
Cw_inv = solve(Cw);

#### Box's test of covariance matrix equality. ####
sum1 = 1/(n1-1) + 1/(n2-1);
sum2 = (n1+n2-2);
     (c = (sum1-1/sum2)*(2*p^2+3*p-1)/(6*(p+1)*(G-1)));
sum3 = (n1-1)*log(det(s1)) + (n2-1)*log(det(s2));
     (B = (1-c)*(sum2*log(det(Cw))-sum3));
(df = 0.5*p*(p+1)*(G-1));
pchisq(B,df);
```

- **a)** Find the two within group sum of squares matrices and group centroids.

```
w1;
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length      13.0552       4.174        8.962      2.7332
Sepal.Width        4.1740       4.825        4.050      2.0190
Petal.Length       8.9620       4.050       10.820      3.5820
Petal.Width        2.7332       2.019        3.582      1.9162

x1_bar;
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
       5.936        2.770        4.260        1.326
```

```
w2;
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length      19.8128      4.5944      14.8612      2.4056
Sepal.Width        4.5944      5.0962       3.4976      2.3338
Petal.Length      14.8612      3.4976      14.9248      2.3924
Petal.Width        2.4056      2.3338       2.3924      3.6962

x2_bar;
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
       6.588        2.974        5.552        2.026
```

- **b)** Find the two within group variance covariance matrices.

```
s1;

             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length   0.26643265  0.08518367   0.18289796  0.05577959
Sepal.Width    0.08518367  0.09846939   0.08265306  0.04120408
Petal.Length   0.18289796  0.08265306   0.22081633  0.07310204
Petal.Width    0.05577959  0.04120408   0.07310204  0.03910612

s2;

             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length   0.40434286  0.09376327   0.30328980  0.04909388
Sepal.Width    0.09376327  0.10400408   0.07137959  0.04762857
Petal.Length   0.30328980  0.07137959   0.30458776  0.04882449
Petal.Width    0.04909388  0.04762857   0.04882449  0.07543265
```

- **c)** Assume multivariate normality for the groups, write down the probability density function (pdf) of the 2 groups with suitable parameter estimates.

  Bivariate normal PDF with vector $\mu = (\mu_1, \mu_2)'$ where μ_i is the centroid (vector) for the group i; under assumed equality of variance (which is a valid assumption based on Box's test), $\Sigma = \Sigma_i$ is the covariance matrix approximated by the pooled sample variance as C_w.

$$f_x(x_1, x_2) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}$$

- **d)** A new observation is drawn with x*=(Sepal.Length=6.0, Sepal.Width=3.0, Petal.Length=4.0, Petal.Width=1.5), compute the posterior probabilities P(C—x) using the Bayes formula and the normality pdf's. [Hint: the sum of these posterior probabilities is 1.0]

  This technique simplifies to include Mahalanobis' distance and priors; since they are equal priors (50 in each classification), there should not be any difference between this technique and basic Mahalanobis on the classification.

```
x_new = c(6.0,3.0,4.0,1.5);
```

```
Dsq_1 = t(x_new-x1_bar)%*%Cw_inv%*%(x_new-x1_bar);
Dsq_2 = t(x_new-x2_bar)%*%Cw_inv%*%(x_new-x2_bar);

(logRatio = log(q1/q2)-((Dsq_1-Dsq_2)/2)); if(logRatio > 0) myG = 2 else myG = 3;
myG;
```

- **e)** Based on these posterior probabilities, which group will you classify x* to, and why? Group 2, the first group of this subset of data, `versicolor`.
- **f)** Use Mahalanobis approach to classify x*.

```
Dsq_1 = t(x_new-x1_bar)%*%Cw_inv%*%(x_new-x1_bar);
Dsq_2 = t(x_new-x2_bar)%*%Cw_inv%*%(x_new-x2_bar);

if(Dsq_1 < Dsq_2) myG = 2 else myG = 3;
myG;
```

- **g)** Comment on the similarity and differences of the posterior probability approach and the Mahalanobis approach in classification.

  In this case, identical because equal priors. The freqentist approach (Mahalanobis) is also the Maximum Likelihood approach; under normality, the Fisher approach is equivalent to the Maximum Likelihood approach for two groups.

- **h)** Use Fisher's approach (also known as Linear Discriminant) to classify x*.

```
     group = numeric(N);
newIr = cbind(ir[,1:4],group);
     newIr[1:50,5]=2;
     newIr[51:100,5]=3;


library(MASS);
     (newIr.LDA = lda(newIr[,1:4],newIr[,5]));
     (newIr.PRED = predict(newIr.LDA,newIr[,1:4]));
     table(Actual = newIr[,5], Predicted = newIr.PRED$class);

      Predicted
Actual  2  3
     2 48  2
     3  1 49

     (a=newIr.LDA$scaling);
proj_new = x_new%*%a;

diff1 = abs(proj_new - x1_bar%*%a);     diff2 = abs(proj_new - x2_bar%*%a);
if(diff1 < diff2) myG = 2 else myG = 3;
myG;
```

- **i)** Fit a 2-means to this 2-species iris data set without class labels.

```
newIr.kmeans2 = kmeans(newIr[,1:4],2);

table(Actual = 4-newIr[,5], Predicted = newIr.kmeans2$cluster);

      Predicted
Actual  1  2
     1 36 14
     2  2 48
```

- **j)** Explain how you would use this 2-means model to classify x*.

  x_{new} is nearest which centroid (from `centers` of `kmeans`)?

```
myDist = function(x,y)
          {
          # simple Euclidean
          tDist = 0;

          for(d in 1:length(x))
               {
               tDist = tDist + (x[d]-y[d])^2;
               }
          sqrt(tDist);
          }


dist1 = myDist(x_new,newIr.kmeans2$centers[2,]); # or x1_bar
dist2 = myDist(x_new,newIr.kmeans2$centers[1,]); # or x2_bar
if(dist1 < dist2) myG = 2 else myG = 3;
myG;
```

  All approaches, even `kmeans` gives the same grouping. The centers need to appropriately match the same groups (reversed centers).

- **k)** Conduct an appropriate hypothesis test to test for difference between the means of these two iris species `versicolor` and `virginica` at 5% level of significance?

```
d=x2_bar-x1_bar;

#### Hotelling's T^2                      ####
T2 = as.numeric(n1*n2/(N)*t(d)%*%Cw_inv%*%d);
1-pf((N-p-1)/(p*(N-2))*T2,p,N-p-1);
```

  The null hypothesis is that the two centroids are equal; the p-value is clearly smaller than the given $\alpha = .05$, so statistical significance will allow me to reject the null and conclude that the two centroids are NOT equal.