See Quiz 6 for classification discriminant analysis

Math 585 Exam 2, Fall 2012 Name_____

sum to 4.0001

```
> rprcomp(z)
Standard deviations: = √λ
[1] 1.8203714 0.6191696 0.4608657 0.3007983
Rotation:
            PC1        PC2        PC3        PC4
syct  0.4797222  0.7041879 -0.4295802 -0.2990767
mmin -0.5045870 -0.2572312 -0.7548543 -0.3307855
mmax -0.5102727  0.3811138  0.4767305 -0.6058894
perf -0.5048567  0.5410210 -0.1355871  0.6588111
```

3.3138 .3834 .2124 .0905

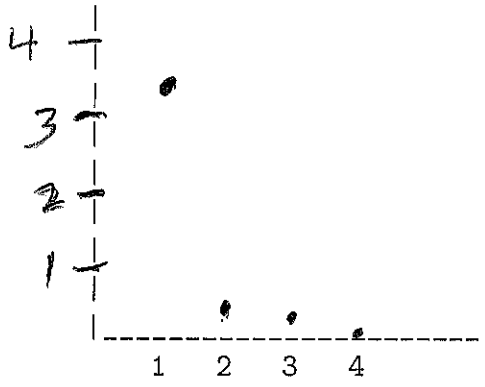
λ

3.3138 0.383 0.212 0.090

sum to 3.9982

1) Shown above is PCA output using the correlation matrix for Venables and Ripley (2003) CPU data. The variables were syct = log(cycle time + 1), mmin = log(minimum main memory + 1), mmax = log(maximum main memory + 1) and perf = log(published performance + 1).

a) Make a scree plot.




b) What proportion of the trace is explained by the first principal component?

$\frac{3.3138}{4} = 0.8284$

c) Interpret the first principal component. (sum of + term) − (sum of − terms) since magnitudes are about the same

syct − mmin − mmax − perf

(or mmin + mmax + perf − syct)

21

2) The $R$ output below is for a canonical correlation analysis on Venables and Ripley (2003) CPU data. The variables were syct = log(cycle time + 1),
mmin = log(minimum main memory + 1),
chmin = log(minimum number of channels + 1),
chmax = log(maximum number of channels + 1),
perf = log(published performance + 1) and
estperf = $20/\sqrt{(\text{estimated performance} + 1)}$. These six variables had a linear scatterplot matrix and DD plot and similar variances. Want to compare the two performance variables with the four remaining variables.

a) What is the first canonical correlation $\hat{\rho}_1$? 0.8769

b) What is $\hat{\boldsymbol{a}}_1$? $\begin{pmatrix} .02536 \\ -.0412 \end{pmatrix}$

c) What is $\hat{\boldsymbol{b}}_1$? $\begin{pmatrix} -.0136 \\ .0375 \\ .0069 \\ .0200 \end{pmatrix}$

.155 and .143 ≈ .149

d) Interpret the second canonical variable $U_2 = \hat{\boldsymbol{a}}_2^T \boldsymbol{w}$.

an average of perf and estperf

```
> cancor(w,y)
$cor
[1] 0.8769433 0.2278554

$xcoef
                [,1]       [,2]
perf      0.02536432 0.1558717
estperf -0.04121870 0.1431100

$ycoef
               [,1]         [,2]         [,3]         [,4]
syct  -0.013613254  0.05700360  0.089757416 -0.011423664
mmin   0.037485282 -0.01874858  0.084442460  0.005859654
chmin  0.006932264  0.09843612 -0.021782624  0.090756713
chmax  0.019998948  0.01159728  0.007855559 -0.094198608
```

not starts

3) What two attractors are used by the FCH estimator?

MB and DGK

4) SAS output for PCA using the correlation matrix is shown below. The Khattree and Naik (1999, p. 11) cork data gives the weights of cork borings in four directions for 28 trees in a block of plantations.

a) What is the variance explained by the first two principal components?

0.9626

b) Interpret the first principal component.

(an average of the 4 variables

(coefs ≈ −.5)

Eigenvalues of the Covariance Matrix

| | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| 1 | 3.5967 | 3.3431 | 0.8992 | 0.8992 |
| 2 | 0.2536 | 0.1735 | 0.0634 | 0.9626 |
| 3 | 0.0801 | 0.0107 | 0.0200 | 0.9826 |
| 4 | 0.0694 | | 0.0174 | 1.0000 |

Eigenvectors

| | Prin1 | Prin2 | Prin3 | Prin4 |
|---|---|---|---|---|
| north | -0.5108992 | 0.1267234 | 0.803287920 | 0.2786606 |
| east | -0.4829921 | 0.7604818 | -0.328918253 | -0.2831940 |
| south | -0.5082783 | -0.3006659 | -0.496526386 | 0.6361719 |
| west | -0.4973468 | -0.5614345 | 0.001687729 | -0.6613884 |

5) Edited SAS output for SAS Institute (1985, p. 146) Fitness Club Data is given below for CCA. Three physiological and three exercise variables measured on 20 middle aged men at a fitness club.

a) What is the first canonical correlation $\hat{\rho}_1$? 0.7956

b) What is $\hat{\boldsymbol{a}}_1$? $\begin{pmatrix} -0.0314 \\ 0.0493 \\ -0.0082 \end{pmatrix}$

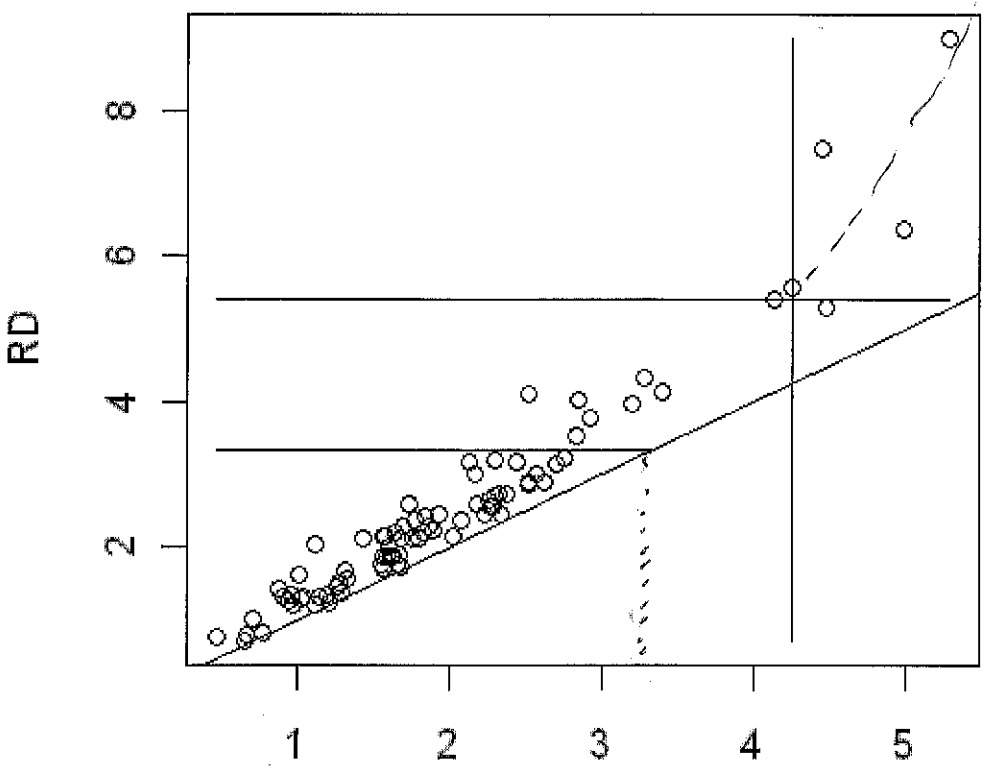
c) What is $\hat{\boldsymbol{b}}_1$? $\begin{pmatrix} -0.0661 \\ -0.0168 \\ 0.0140 \end{pmatrix}$

21

```
 Canonical
Correlation
0.7956
0.2006
0.0726

Raw Canonical Coefficients for the Physiological Variables
        PHYS1    PHYS2    PHYS3
weight -0.0314  -0.0763  -0.0077
waist   0.0493   0.3687   0.1580
pulse  -0.0082  -0.0321   0.1457

Raw Canonical Coefficients for the Exercise Variables
         Exer1    Exer2   Exer3
chinups -0.0661  -0.0714  -0.2428
situps  -0.0168   0.0020   0.0198
jumps    0.0140   0.0207  -0.0082
```




≃ intersects on line plotted points follow

note! MVN region has volume that is too small

draw vertical line to get "MVN" region using $(\bar{x}, S)$

> ddplot4(log(mussels))

$cuplim 95% 4.245233 $ruplim 95% 5.398374 $mvnlim[1] 3.327236

90% covers 95% of the training data

(6) The mussels data has n = 87 and variables log(length), log(width), log(height), log(shell mass) and log(muscle mass). The 90% prediction regions use cuplim for the nonparametric region that uses the classical estimator, ruplim for the semiparametric region and mvnlim for the parametric MVN region that use the robust RMVN estimator. a) What is the the semiparametric region in the DD plot shown above? b) What does the DD plot suggest about the distribution of the data?

a) RD < 5.398

b) EC but not MVN