# STAT 530/J530
# September 13th, 2005

Instructor: Brian Habing
Department of Statistics
LeConte 203
Telephone: 803-777-3578
E-mail: habing@stat.sc.edu



## Homework 2 – Question 1

1) Imagine that someone wanted to come up with a total score to summarize each persons view of the oil crisis (Q1-Q20).

a) Explain why it doesn't make sense to just add up all of the numbers.

b) Find the correlation matrix for Q1-Q20 data set and suggest two separate groups of questions that might be added separately.

c) How could these two scores be combined to form a single score?



| | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Q1 | 1.0 | -0.3 | -0.1 | 0.3 | 0.2 | 0.0 | 0.1 | 0.0 | 0.2 | 0.0 |
| Q2 | -0.3 | 1.0 | 0.3 | -0.1 | -0.3 | 0.0 | -0.1 | 0.1 | -0.2 | 0.0 |
| Q3 | -0.1 | 0.3 | 1.0 | -0.1 | -0.3 | 0.1 | 0.0 | 0.1 | -0.1 | -0.1 |
| Q4 | 0.3 | -0.1 | -0.1 | 1.0 | 0.2 | 0.1 | 0.0 | 0.0 | 0.2 | 0.0 |
| Q5 | 0.2 | -0.3 | -0.3 | 0.2 | 1.0 | 0.1 | 0.0 | -0.1 | 0.3 | 0.3 |
| Q6 | 0.0 | 0.0 | 0.1 | 0.1 | 0.1 | 1.0 | 0.1 | -0.1 | 0.0 | 0.1 |
| Q7 | 0.1 | -0.1 | 0.0 | 0.0 | 0.0 | 0.1 | 1.0 | -0.1 | 0.0 | 0.0 |
| Q8 | 0.0 | 0.1 | 0.1 | 0.0 | -0.1 | -0.1 | -0.1 | 1.0 | 0.1 | 0.0 |
| Q9 | 0.2 | -0.2 | -0.1 | 0.2 | 0.3 | 0.0 | 0.0 | 0.1 | 1.0 | 0.1 |
| Q10 | 0.0 | 0.0 | -0.1 | 0.0 | 0.3 | 0.1 | 0.0 | 0.0 | 0.1 | 1.0 |
| Q11 | 0.0 | 0.0 | 0.1 | 0.1 | 0.1 | 0.1 | -0.3 | 0.3 | 0.1 | 0.1 |
| Q12 | 0.0 | 0.2 | 0.2 | 0.0 | -0.3 | 0.0 | 0.1 | 0.0 | -0.1 | -0.5 |
| Q13 | 0.0 | 0.2 | 0.2 | 0.0 | -0.3 | 0.0 | 0.0 | 0.1 | -0.1 | -0.3 |
| Q14 | 0.0 | 0.1 | 0.0 | 0.1 | 0.1 | 0.0 | -0.2 | 0.4 | 0.2 | 0.1 |
| Q15 | 0.2 | -0.2 | -0.2 | 0.1 | 0.3 | -0.1 | 0.1 | 0.0 | 0.3 | 0.1 |
| Q16 | -0.1 | -0.1 | -0.1 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.1 | 0.2 |
| Q17 | 0.0 | 0.0 | 0.1 | 0.1 | -0.1 | 0.5 | 0.1 | -0.1 | -0.1 | 0.0 |
| Q18 | 0.0 | 0.0 | 0.1 | 0.0 | -0.1 | 0.0 | 0.0 | 0.0 | -0.1 | -0.2 |
| Q19 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.6 | 0.1 | -0.1 | -0.1 | 0.0 |
| Q20 | 0.0 | 0.0 | 0.0 | -0.1 | -0.2 | 0.0 | 0.3 | -0.2 | 0.0 | -0.2 |

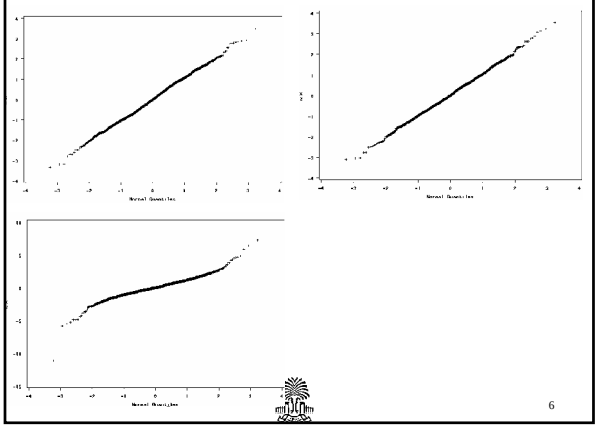
| | Q11 | Q12 | Q13 | Q14 | Q15 | Q16 | Q17 | Q18 | Q19 | Q20 |
|---|---|---|---|---|---|---|---|---|---|---|
| Q1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | -0.1 | 0.0 | 0.0 | 0.0 | 0.0 |
| Q2 | 0.0 | 0.2 | 0.2 | 0.1 | -0.2 | -0.1 | 0.0 | 0.0 | 0.0 | 0.0 |
| Q3 | 0.1 | 0.2 | 0.2 | 0.0 | -0.2 | -0.1 | 0.1 | 0.1 | 0.1 | 0.0 |
| Q4 | 0.1 | 0.0 | 0.0 | 0.1 | 0.1 | 0.0 | 0.1 | 0.0 | 0.0 | -0.1 |
| Q5 | 0.1 | -0.3 | -0.3 | 0.1 | 0.3 | 0.2 | -0.1 | -0.1 | 0.0 | -0.1 |
| Q6 | 0.1 | 0.0 | 0.0 | 0.0 | -0.1 | 0.0 | 0.5 | 0.0 | 0.6 | 0.0 |
| Q7 | -0.3 | 0.1 | 0.0 | -0.2 | 0.1 | 0.0 | 0.1 | 0.0 | 0.1 | 0.3 |
| Q8 | 0.3 | 0.0 | 0.1 | 0.4 | 0.0 | 0.0 | -0.1 | 0.0 | -0.1 | -0.2 |
| Q9 | 0.1 | -0.1 | -0.1 | 0.2 | 0.3 | 0.1 | -0.1 | -0.1 | -0.1 | 0.0 |
| Q10 | 0.1 | -0.5 | -0.3 | 0.1 | 0.1 | 0.2 | 0.0 | -0.2 | 0.0 | -0.2 |
| Q11 | 1.0 | 0.0 | 0.1 | 0.3 | 0.0 | 0.1 | 0.1 | 0.0 | 0.1 | -0.3 |
| Q12 | 0.0 | 1.0 | 0.5 | 0.0 | -0.2 | -0.1 | 0.1 | 0.3 | 0.1 | 0.2 |
| Q13 | 0.1 | 0.5 | 1.0 | 0.0 | -0.2 | -0.3 | 0.1 | 0.4 | 0.1 | 0.1 |
| Q14 | 0.3 | 0.0 | 0.0 | 1.0 | 0.0 | 0.1 | 0.0 | -0.1 | 0.0 | -0.1 |
| Q15 | 0.0 | -0.2 | -0.2 | 0.0 | 1.0 | 0.1 | -0.1 | -0.2 | -0.1 | 0.0 |
| Q16 | 0.1 | -0.1 | -0.3 | 0.1 | 0.1 | 1.0 | 0.0 | -0.1 | 0.0 | -0.1 |
| Q17 | 0.1 | 0.1 | 0.1 | 0.0 | -0.1 | 0.0 | 1.0 | 0.0 | 0.5 | 0.0 |
| Q18 | 0.0 | 0.3 | 0.4 | -0.1 | -0.2 | -0.1 | 0.0 | 1.0 | 0.0 | 0.0 |
| Q19 | 0.1 | 0.1 | 0.1 | 0.0 | -0 | 0.0 | 0.5 | 0.0 | 1.0 | 0.1 |
| Q20 | -0.3 | 0.2 | 0.1 | -0.1 | 0 | -0.1 | 0.0 | 0.0 | 0.1 | 1.0 |

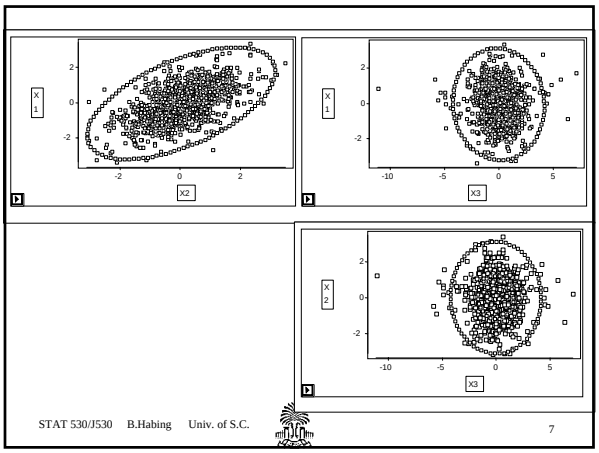


## Homework 2 – Question 2

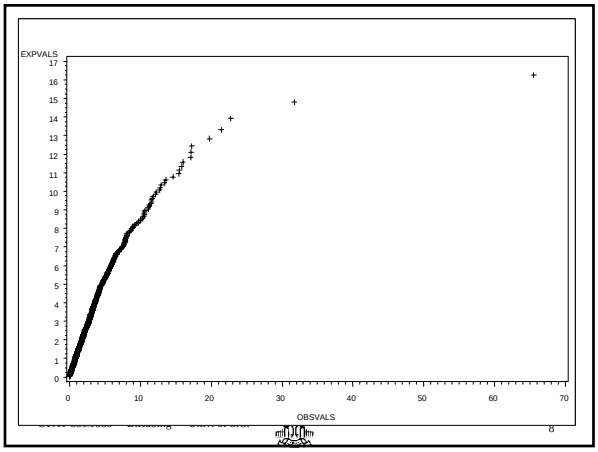
2) Check whether the data set normsamp.txt is actually multivariate normal.









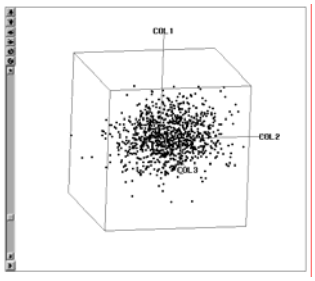



## Principal Components Analysis

The main idea of Principal Components Analysis is that we would like to come up with new combinations of the original variables that are “easier to work with”.

## 3-D Example

```
PROC IML;
sigma = {1 .2 2,
         .2 1 2,
         2 2 10};
mu = {0 0 0};
n = 1000;
seed = 91505;
q=NROW(sigma);
MUMAT=REPEAT(mu,n,1);
SROOT=ROOT(sigma);
Z=NORMAL(REPEAT(seed,n,q));
x=Z*SROOT+MUMAT;
CREATE mvnormdata FROM x;
APPEND FROM x;
QUIT;
```






## Goal

Find the coefficients (a's) of the x's so that:

$$Y=a_1X_1+a_2X_2+\cdots+a_qX_q$$

has the largest possible variance subject to the condition that the length of the coefficient vector is 1.



## Example Cont.

```
library(MASS)
mu<-c(5,0,-1)
sigma<-
  matrix(c(1,0.2,2,0.2,1,2,2,2,10),
ncol=3,byrow=T)
x<-mvrnorm(n=1000,mu,sigma)

coef<-princomp(x,cor=F)$loadings[,1]

pc1<-princomp(x,cor=F)$scores[,1]
```

## Oildata Q's

```
coef<-princomp(sect3,cor=F)$loadings[,1]

        Q1          Q2          Q3          Q4
 0.14749703 -0.18642164 -0.24432644  0.09980284
        Q5          Q6          Q7          Q8
 0.35633250 -0.17468260 -0.04253903  0.01630308
        Q9         Q10         Q11         Q12
 0.25808813  0.29008088  0.01900861 -0.41185841
       Q13         Q14         Q15         Q16
-0.24074574  0.07216813  0.38062551  0.16292494
       Q17         Q18         Q19         Q20
-0.23470978 -0.23005868 -0.19195023 -0.11866190
```



## What Else?

Imagine that you are giving directions… but instead of using North and East, you used North and North/East.



## Dot Product

The dot product of two vectors a and b is $a_1b_1+a_2b_2 +\cdots+a_qb_q$

In vector notation this is:

It relates to distance by:

It relates to correlation by:

## The Next Principal Component

The coefficient vector should be length 1.

The coefficient vector should be orthogonal to the previous one(s).

It should explain the largest possible amount of variance.



## Eigen Values and Vectors

The Eigen Vectors of the Covariance matrix are exactly the coefficients that do this!

And the Eigen Values are the variances of the principal components!



## Example Cont.

```
coeffs<-eigen(cov(x))$vectors
vars<-eigen(cov(x))$values

t(coeffs[,1])%*%coeffs[,1]

scores<-x%*%coeffs

round(var(scores),2)
```