Stats 519: HW 3 - Vector (CH 2)

Due on February 13, 2009

Dr. Stephen Lee 1:30

Monte J. Shaffer

Program 1 Useful Functions

```
> vectorLength = function (vector) {sqrt(sum(vector^2));}
> dotProduct   = function (a,b)    {t(as.matrix(a))%*%as.matrix(b);}
> calcAngle    = function (a,b)    {180*acos(dotProduct(a,b)/(vectorLength(a)*vectorLength(b)))/pi;}
> doAnalysis   = function (a,b)
                         {
                         myAnswer = numeric(5);
                              myAnswer[1]=round(vectorLength(a),digits=3);
                              myAnswer[2]=round(vectorLength(b),digits=3);
                              myAnswer[3]=round(calcAngle(a,b),digits=1);       # degrees
                              myAnswer[4]=round(calcAngle(a,b)/180,digits=2);   # radians
                              myAnswer[5]=dotProduct(a,b);
                         myAnswer;
                         }
```

Problem 1

For each of the following vectors, find the length of **a**, the length of **b**, the angle θ formed by the vectors **a** and **b**, and their scalar product:

a) **a** = (1,1); **b**=(-1,1)
b) **a** = (1,0); **b**=(1,1)
c) **a** = (4,3); **b**=(-4,-3)
d) **a** = (1,2,3); **b**=(1,1,2)

Program 2 Results

```
> a=c(1,1);     b=c(-1,1);     Answer1=doAnalysis(a,b);
> a=c(1,0);     b=c(1,1);      Answer2=doAnalysis(a,b);
> a=c(4,3);     b=c(-4,-3);    Answer3=doAnalysis(a,b);
> a=c(1,2,3);   b=c(1,1,2);    Answer4=doAnalysis(a,b);
> finalAnswer  =    rbind(Answer1,Answer2,Answer3,Answer4);
>    rownames(finalAnswer) = c("a","b","c","d");
>    colnames(finalAnswer) = c("||a||","||b||","theta","pi radians","a.b");
> finalAnswer;
     ||a||          ||b||     theta     pi radians     a.b
============================================================
a    1.414          1.414      90.0       0.50           0
b    1.000          1.414      45.0       0.25          1
c    5.000          5.000     180.0       1.00         -25
d    3.742          2.449      10.9       0.06           9
```

Problem 2

EDA for EDUC_SCORES.

Program 3 Test Scores by Gender

```
> Y=read.table('clipboard');
> X=cbind(Y$V2,Y$V3,Y$V4);
>    colnames(X)=c("x1","x2","x3");
> X=data.frame(X);
> Xbar = mean(X);
> Xd=scale(X,scale=F);
> Xs=scale(X);
> SSD =t(as.matrix(Xd))%*%as.matrix(Xd);
> n = length(X$x1);
> S = SSD / (n-1);
> SSS =t(as.matrix(Xs))%*%as.matrix(Xs);
> R = SSS / (n-1);
```

a) the centroid vector $\bar{x}' = (\bar{x}_{.1},\bar{x}_{.2},\bar{x}_{.3})$

```
   x1    x2    x3
6.625 6.125 6.625
```

b) the mean-differenced matrix X_d

```
         x1     x2     x3
[1,] -4.625 -3.125  8.375
[2,] -0.625  1.875  2.375
[3,] -1.625 -4.125  0.375
[4,]  2.375 -2.125 -3.625
[5,]  4.375  3.875 -4.625
[6,]  5.375  8.875 -5.625
[7,] -5.625 -2.125  5.375
[8,]  0.375 -3.125 -2.625
```

c) the standardized data matrix X_s

```
              x1         x2          x3
[1,] -1.16737749 -0.6917050  1.66140831
[2,] -0.15775371  0.4150230  0.47114564
[3,] -0.41015966 -0.9130506  0.07439142
[4,]  0.59946411 -0.4703594 -0.71911703
[5,]  1.10427600  0.8577142 -0.91749414
[6,]  1.35668194  1.9644422 -1.11587126
[7,] -1.41978343 -0.4703594  1.06627698
[8,]  0.09465223 -0.6917050 -0.52073992
```

d) the sum of squares matrix $X_d'X_d$

```
           x1      x2       x3
x1  109.875  90.375 -131.125
x2   90.375 142.875  -86.625
x3 -131.125 -86.625  177.875
```

e) the covariance matrix $S = \frac{1}{n-1}X_d'X_d$

```
          x1        x2        x3
x1  15.69643  12.91071 -18.73214
x2  12.91071  20.41071 -12.37500
x3 -18.73214 -12.37500  25.41071
```

f) the correlation matrix $R = \frac{1}{n-1}X_s'X_s$

```
           x1        x2         x3
x1  1.0000000  0.721308 -0.9379477
x2  0.7213080  1.000000 -0.5433850
x3 -0.9379477 -0.543385  1.0000000
```

Problem 3

Linear weights: The researcher is interested in finding a linear combination to summarize the test performance of each of the eight students. He proposes giving a weight of 25 percent to the language aptitude score (x_1), a weight of 25 percent to the analogical reasoning score (x_2), and a weight of 50 percent to the geometric reasoning score (x_3).

a) Find the vector $w' = (w_1, w_2, w_3)$ with unit length that achieves this relative weighting scheme. ($w^* = (0.25, 0.25, 0.50)$ scaled to unit length $\frac{w^*}{||w^*||}$.)

```
> w_star=c(.25,.25,.50);

[1] 0.25 0.25 0.50

> vectorLength(w_star);

[1] 0.6123724

> w=w_star/vectorLength(w_star);

[1] 0.4082483 0.4082483 0.8164966
```

Problem 3 continued on next page. . .

b) Find the linear combination using the raw data ($z_1 = Xw$) and the standardized data ($z_2 = X_s w$).

```
> z_1=as.matrix(X)%*%as.matrix(w);
```

```
          [,1]
[1,] 14.288690
[2,] 13.063945
[3,]  8.573214
[4,]  7.756718
[5,] 10.206207
[6,] 11.839200
[7,] 11.839200
[8,]  7.348469
```

```
> z_2=as.matrix(Xs)%*%as.matrix(w);
```

```
            [,1]
[1,]  0.59756696
[2,]  0.48971855
[3,] -0.47945799
[4,] -0.53444982
[5,]  0.05184832
[6,]  0.44473820
[7,]  0.09896393
[8,] -0.66892814
```

c) Compare the scores of students C and D. Which student is higher on z_1? Which student is higher on z_2?

```
> c=cbind(z_1[3],z_2[3]);
> d=cbind(z_1[4],z_2[4]);
> myAnswer = round(rbind(c,d),digits=2);
>    colnames(myAnswer)=c("Raw Data","Standardized");
>    rownames(myAnswer)=c("Student C","Student D");
> myAnswer;
```

```
          Raw Data       Standardized
          ===========================
Student C     8.57            -0.48
Student D     7.76            -0.53
```

Student C is higher using Raw Data; Student C is also very slightly higher using Standardized scores (less negative in number of standard deviations).

Problem 4

Gender Differences: Divide the data in the file into two groups: males and females. For each group, form the matrix $X = [x_1\ x_2\ x_3]$ and compute the following:

Program 4 Gender Differences

```
> F=rbind(X[3,],X[5,],X[6,],X[8,]);
> Fbar = mean(F);
> n = length(F$x1);
> Fd=scale(F,scale=FALSE);
> Fs=scale(F);
> S_F = t(as.matrix(Fd))%*%as.matrix(Fd)/(n-1);
> M=rbind(X[1,],X[2,],X[4,],X[7,]);
> Mbar = mean(M);
> n = length(M$x1);
> Md=scale(M,scale=FALSE);
> Ms=scale(M);
> S_M = t(as.matrix(Md))%*%as.matrix(Md)/(n-1);
```

a) Compare Mean Vectors

```
> myAnswerMean = rbind(Fbar,Mbar);
>    rownames(myAnswerMean)=c("Females","Males");

          x1   x2   x3
Females 8.75 7.50 3.50
Males   4.50 4.75 9.75
```

b) Compare Covariance

```
> S_F;

         x1        x2        x3
x1 10.91667  19.50000  -8.50000
x2 19.50000  37.66667 -14.66667
x3 -8.50000 -14.66667   7.00000

> S_M;

           x1        x2     x3
x1  13.666667  2.833333 -17.50
x2   2.833333  4.916667  -2.75
x3 -17.500000 -2.750000  26.25
```

Problem 4 continued on next page. . .

c) Does there appear to be a difference across the two groups in either the level or dispersion of their test scores?

```
> myAnswerMean = rbind(Fbar,Mbar);
>    rownames(myAnswerMean)=c("Females","Males");

          x1   x2   x3
Females 8.75 7.50 3.50
Males   4.50 4.75 9.75
> myAnswerSE=round(rbind(sqrt(diag(S_F)/n),sqrt(diag(S_M)/n)),digits=2);
>    rownames(myAnswerSE)=c("Females","Males");

          x1   x2   x3
Females 1.65 3.07 1.32
Males   1.85 1.11 2.56
```

Yes, there appears to be differences in means and variances. Pairwise t-tests (e.g. Welch's) could verify the statistical significance of these differences. The graph plots the means and standard error of the means for each gender ($se = \frac{\sigma}{\sqrt{n}}$). Specifically, it appears that Language Aptitude (x_1) scores are higher for females and Geometric Reasoning (x_3) scores are higher for males.

```
> doSegments = function(myAnswerMean,myAnswerSE)
     {
     myMean    = as.numeric(myAnswerMean);
     mySE = as.numeric(myAnswerSE);
     for(i in 1:6)
          {
          myExtra = 0;if(i > 2){myExtra = 1;}if(i > 4){myExtra = 2;}
          yTop = myMean[i]+mySE[i]; yBottom  = myMean[i]-mySE[i];
          par(new=T);
          segments(i+0.5+myExtra,yTop,i+0.5+myExtra,yBottom,lwd=5);
          }
     }
> barplot(myAnswerMean,beside=T,ylim=c(0,12),col=c("pink","blue"));
> doSegments(myAnswerMean,myAnswerSE);
```

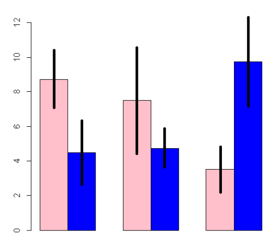

Figure 1: Gender Differences of Test Results

Problem 5

Program 5 Results

```
> RAND = read.table('clipboard');
> # errata, not really normalized, so must use scale()
> x4=0.8*scale(RAND$V1)+0.6*scale(RAND$V2);
> x5=0.8*scale(RAND$V1)+0.6*scale(RAND$V3);
> X=cbind(x4,x5);
> Xd=scale(X,scale=FALSE);
> Xs=scale(X);
> W=matrix(c(0.866,-0.500,0.500,0.866),byrow=T,nrow=2,ncol=2); # ORTHOGONAL
> Z=X%*%W;
```

The file RANDOM1 contains three variables (x_1, x_2, x_3) that were created using a random number generator. Each consists of $n = 100$ observations drawn independently from a unit normal distribution and each has been standardized (errata). Using the data in RANDOM1, form the following linear combinations:

$$\begin{aligned} x_4 &= 0.80x_1 + 0.60x_2 \\ x_5 &= 0.80x_1 + 0.60x_3 \end{aligned} \tag{1}$$

Form the matrix $X = [x_4\ x_5]$ and perform the following matrix multiplication $Z = XW$ where $W = [w_1\ w_2]$ such that $||w_1|| = ||w_2|| = 1$:

$$W = \begin{bmatrix} 0.866 & -0.500 \\ 0.500 & 0.866 \end{bmatrix}$$

a) Create a scatterplot of the new variables, z_1 and z_2.

```
> plot(x4,x5);
>    abline(v=0,col="gray");
>    abline(h=0,col="gray");
> plot(Z[,1],Z[,2]);
>    abline(v=0,col="gray");
>    abline(h=0,col="gray");
```

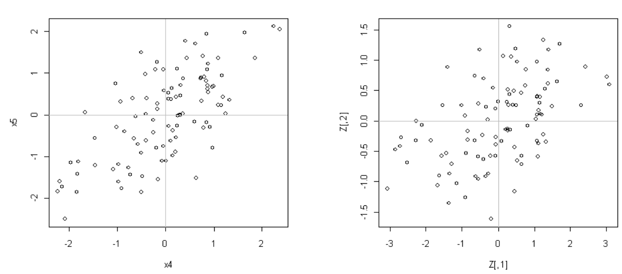

Figure 2: Scatter Plots of X and Z

b) Calculate the covariance matrix $\frac{1}{n-1}Z'Z$ (already mean centered).

```
> n=100;
> S=t(Z)%*%Z / (n-1);

          [,1]      [,2]
[1,] 1.5911760 0.4145034
[2,] 0.4145034 0.4629700
```

c) What is the determinant of the covariance matrix of Z? How does it compare to the determinate of the covariance matrix of X?

```
> cov(X);

          [,1]      [,2]
[1,] 0.9501781 0.6958079
[2,] 0.6958079 1.1040582

> det(S);

[1] 0.5648536

> det(cov(X));

[1] 0.5649033
```

Except for slight variation in rounding ($\frac{\sqrt{3}}{2} \approx 0.866$), these are identical. Considering the intent of the projection (to make a random scatter), each axis should be orthogonal, such that $cov(z_i, z_j) = 0 \; \forall \; z_i, z_j$ where $i \neq j$. If the columns of matrix W are orthogonal, this should be expected.

```
> myAnswer5 = doAnalysis(W[,1],W[,2]);
[1]  1.0  1.0 90.0  0.5  0.0
```

Problem 6

Program 6 Results

```
> RAND = read.table('clipboard');
> # errata, not really normalized, so must use scale()
> x1=scale(RAND$V1);
> x2=scale(RAND$V2);
> X=cbind(x1,x2);
> Xd=scale(X,scale=FALSE);
> Xs=scale(X);
> W=matrix(c(0.866,0.500,0.500,0.866),byrow=T,nrow=2,ncol=2);  # ORTHOGONAL ?!?
> Z=X%*%W;
```

Using the data in RANDOM1, form the matrix $X = [x_1\ x_2]$ and perform the following matrix multiplication $Z = XW$ where $W = [w_1\ w_2]$ such that $||w_1|| = ||w_2|| = 1$:

$$W = \begin{bmatrix} 0.866 & 0.500 \\ 0.500 & 0.866 \end{bmatrix}$$

a) Create a scatterplot of the new variables, z_1 and z_2.

```
> plot(x1,x2);
>    abline(v=0,col="gray");
>    abline(h=0,col="gray");
> plot(Z[,1],Z[,2]);
>    abline(v=0,col="gray");
>    abline(h=0,col="gray");
```

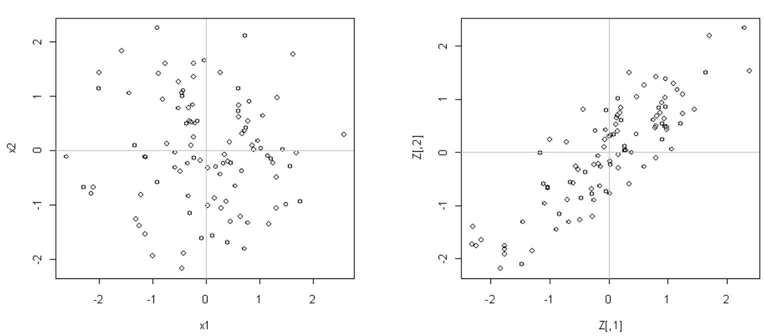

Figure 3: Scatter Plots of X and Z

b) Calculate the covariance matrix $\frac{1}{n-1}Z'Z$ (already mean centered).

```
> n=100;
> S=t(Z)%*%Z / (n-1);


             [,1]       [,2]
```

```
[1,] 0.9550125 0.8141045
[2,] 0.8141045 0.9550125
```

c) What is the determinant of the covariance matrix Z? How does it compare to the determinate of the covariance matrix of X?

```
> cov(X);

          [,1]       [,2]
[1,]  1.0000000 -0.0518978
[2,] -0.0518978  1.0000000

> det(S);

[1] 0.2492828

> det(cov(X));

[1] 0.9973066
```

Clearly not equal. The reason is that the projection vectors (w_1 and w_2 are not orthogonal to each other). As a result the stretching and shrinking that is occurring is being applied more to one projection dimension than the other.

```
> myAnswer6 = doAnalysis(W[,1],W[,2]);
[1]  1.000  1.000 30.000  0.170  0.866
```

Comparing Problem 5 to Problem 6, we can see the differences

```
> finalAnswer = rbind(myAnswer5,myAnswer6);
> colnames(finalAnswer) = c("||a||","||b||","theta","pi radians","a.b");
> rownames(finalAnswer) = c("#5","#6");

   ||a|| ||b|| theta pi radians   a.b
#5     1     1    90       0.50 0.000
#6     1     1    30       0.17 0.866
```