Math 760

Chapter 3 HW

Gabrielle Salamanca
March 10, 2024

2. Given the data matrix

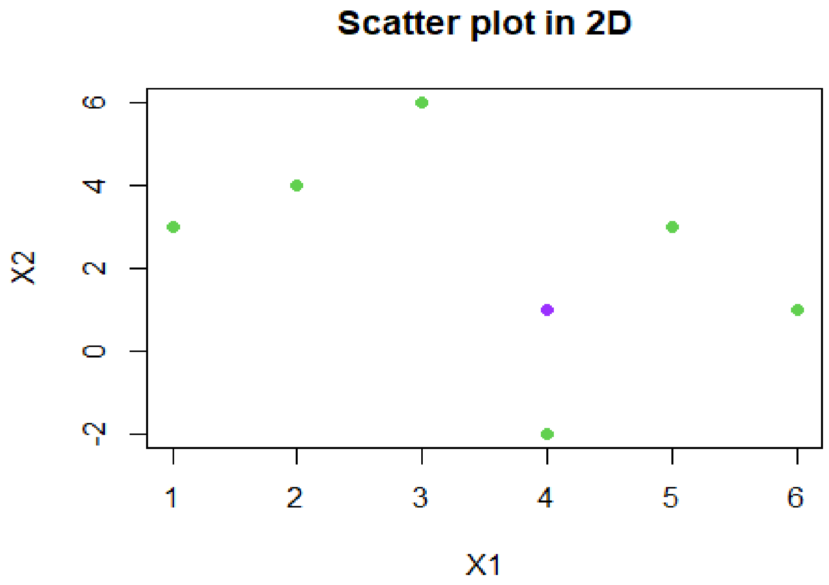
$$X = \begin{bmatrix} 3 & 4 \\ 6 & -2 \\ 3 & 1 \end{bmatrix}$$

(a) Graph the scatter plot in p = 2 dimensions, and locate the sample mean on your diagram.

The sample mean matrix is:

```
##      [,1]
## [1,]    4
## [2,]    1
```

Now, we can graph the scatter plot, and note that the sample mean is colored in purple.



(b) Sketch the n = 3-space representation of the data, and plot the deviation vectors $y_1 - \bar{x}_1{}'1$ and $y_2 - \bar{x}_2{}'1$.

Let's first find the deviation vectors $y_1 - \bar{x}_1{}'1$ and $y_2 - \bar{x}_2{}'1$.

Our sample mean matrix is:

```
##      [,1]
## [1,]    4
## [2,]    1
```

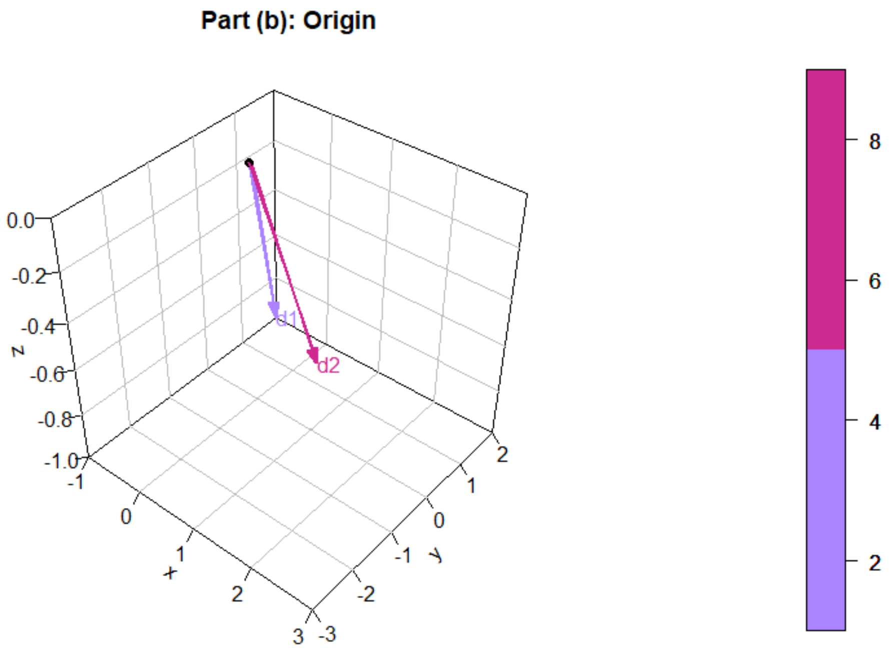
But because X is a 3x2 matrix, we will need to make the sample mean as such. Then, we can subtract the X and mean matrix to get:

```
##      d1 d2
## [1,] -1  3
## [2,]  2 -3
## [3,] -1  0
```

Now, we can plot them in a 3D space.

(c) Sketch the deviation vectors in (b) emanating from the origin. Calculate their lengths and the cosine of the angle between them. Relate these quantities to S_n and R.

Let's actually put the deviation vectors into a proper 3D graph.



With this in mind, let's calculate their lengths and the cosine of the angle between them. From (3-5), we get the squared lengths, so we'll square root them.

$$L_{d_i}^2 = d'_i d_i \Rightarrow \sqrt{L_{d_i}^2} = L_{d_i}$$

```
## The length of d1 is 2.44949
## The length of d2 is 4.242641
```

With this in mind, we can find the S_n matrix. The formulas are:
$3s_{11} = d_1{'}d_1$
$3s_{22} = d_2{'}d_2$
$3s_{12} = d_1{'}d_2$

```
## s11 is 2
## s22 is 6
## s12 is -3
##      [,1] [,2]
## [1,]    2   -3
## [2,]   -3    6
```

Let θ be the angle between d_1 and d_2, then, $\cos(\theta)$ is:

```
## [1] -0.8660254
```

$\cos(\theta)$ is also known as the sample correlation coefficient; thus, **R** is:

```
##            [,1]       [,2]
## [1,]  1.0000000 -0.8660254
## [2,] -0.8660254  1.0000000
```

6. Consider the data matrix

$$X = \begin{bmatrix} -1 & 3 & -2 \\ 2 & 4 & 2 \\ 5 & 2 & 3 \end{bmatrix}$$

(a) Calculate the matrix deviations (residuals), $X - 1\bar{x}'$. Is this matrix of full rank? Explain.

We will need to first find the $\bar{x}$, which is:

```
##      [,1]
## [1,]    2
## [2,]    3
## [3,]    1
```

Then, we can calculate the residuals, which are:

```
##      [,1] [,2] [,3]
## [1,]   -3    0   -3
## [2,]    0    1    1
## [3,]    3   -1    2
```

The matrix deviations are as follows:

```
## d1 is [ -3 0 -3 ]
## d2 is [ 0 1 1 ]
## d3 is [ 3 -1 2 ]
```

And because they aren’t the same, the matrix is not of full rank.

(b) Determine S and calculate the generalized sample variance |S|. Interpret the latter geometrically.

To find **S**, we can multiply the residuals twice, then divide it by 2.

```
##      [,1] [,2] [,3]
## [1,]  9.0 -1.5 -7.5
## [2,] -1.5  1.0  0.5
## [3,] -7.5  0.5  7.0
```

Now we’ll calculate |**S**|, which is:

```
## [1] 5.995204e-15
```

This means that the 3 deviation vectors lie in a 2D subspace, and the 3D volume enclosed by the deviation vectors is 0.

(c) Using the results in (b), calculate the total sample variance. [See (3-23)]

According to (3-23), the total sample variance = $s_{11} + s_{22} + \ldots + s_{pp}$. So, the total sample variance is:

```
## [1] 17
```

10. When the generalized variance is zero, it is the columns of the mean corrected data matrix $X_c = X - 1\bar{x}'$ that are linearly dependent, not necessarily those of the data matrix itself. Given the data

$$\begin{bmatrix} 3 & 1 & 0 \\ 6 & 4 & 6 \\ 4 & 2 & 2 \\ 7 & 0 & 3 \\ 5 & 3 & 4 \end{bmatrix}$$

(a) Obtain the mean corrected data matrix, and verify that the columns are linearly dependent. Specify an $a' = [a_1 \quad a_2 \quad a_3]$ vector that establishes the dependence.

The $\bar{x}$ matrix is:

```
##      [,1]
## [1,]    5
## [2,]    2
## [3,]    3
```

Now, we can obtain the mean corrected data matrix, which is:

```
##      [,1] [,2] [,3]
## [1,]   -2   -1   -3
## [2,]    1    2    3
## [3,]   -1    0   -1
## [4,]    2   -2    0
## [5,]    0    1    1
```

And notice, if we added up the first 2 columns of the mean corrected data matrix, they equal the third column. This means, if we want $X_c a = 0$, then matrix a will have to be $a' = [1 \quad 1 \quad -1]$. Therefore, the columns are linearly dependent.

```
##      [,1] [,2] [,3]
## [1,]    1    1   -1
##      [,1] [,2] [,3] [,4] [,5]
## [1,]   -2    1   -1    2    0
## [2,]   -1    2    0   -2    1
## [3,]   -3    3   -1    0    1
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    0    0    0    0    0
```

(b) Obtain the sample covariance S, and verify that the generalized variance is zero.

The sample covariance matrix is:

```
##      [,1] [,2] [,3]
## [1,]  2.5  0.0  2.5
## [2,]  0.0  2.5  2.5
## [3,]  2.5  2.5  5.0
```

And using vector a from (a), Sa = 0. Thus, the generalized variance is zero.

```
##      [,1] [,2] [,3]
## [1,]    0    0    0
```

(c) Show that the columns of the data matrix are linearly independent in this case.

Let’s start with this formula: Xa = 0.

We have our data matrix:

```
##      [,1] [,2] [,3]
## [1,]    3    1    0
## [2,]    6    4    6
## [3,]    4    2    2
## [4,]    7    0    3
## [5,]    5    3    4
```

And our A matrix:

$$a' = [a_1 \quad a_2 \quad a_3]$$

And when multiplied, we have this:
$3a_1 + a_2 = 0$
$6a_1 + 4a_2 + 6a_3 = 0$
$4a_1 + 2a_2 + 2a_3 = 0$
$7a_1 + 3a_3 = 0$
$5a_1 + 3a_2 + 4a_3 = 0$

With these equations, we can use them to show that all a's equal 0.
$3a_1 + a_2 = 0 \Rightarrow a_2 = -3a_1$
$7a_1 + 3a_3 = 0 \Rightarrow 3a_3 = -7a_1 \Rightarrow a_3 = -\frac{7}{3}a_1$

$$4a_1 + 2a_2 + 2a_3 = 0 \Rightarrow 4a_1 + 2(-3a_1) + 2\left(-\frac{7}{3}a_1\right) = 0$$
$$\Rightarrow 4a_1 - 6a_1 - \frac{7}{3}a_1 = 0$$
$$\Rightarrow \frac{12 - 18 - 7}{3}a_1 = 0$$
$$\Rightarrow -\frac{13}{3}a_1 = 0$$
$$\Rightarrow a_1 = 0$$

We have proven $a_1 = 0$, then plugging it into the first two equations:
$a_2 = -3a_1 = 0$
$a_3 = -\frac{7}{3}a_1 = 0$

Therefore, the columns of the data matrix are linearly independent.

15. Repeat Exercise 3.14 using the data matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & 4 & 3 \\ 6 & 2 & 6 \\ 8 & 3 & 3 \end{bmatrix}$$

and the linear combinations

$$\boldsymbol{b'X} = [1 \quad 1 \quad 1]\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \qquad \boldsymbol{c'X} = [1 \quad 2 \quad -3]\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

(a) Evaluate the sample means, variances, and covariance b'X and c'X from first principles. That is, calculate the observed values of b'X and c'X, and then use the sample mean, variance, and covariance formulas.

From first principles, we will start calculating the observed values of **b'X**.

```
## The observed value of bX1 is: 8
## The observed value of bX2 is: 14
## The observed value of bX3 is: 14
```

Then, we can find the sample mean:

$$\bar{x} = \frac{x_1 + x_2 + x_3}{3}$$

```
## The sample mean of bX is: 12
```

And finally, the sample variance:

$$s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2}{2}$$

```
## The sample variance of bX is: 12
```

Now, let’s do the observed values, sample mean, and sample variance of **c’X**.

```
## The observed value of cX1 is: 44
## The observed value of cX2 is: 68
## The observed value of cX3 is: 47
## The sample mean of cX is: 53
## The sample variance of cX is: 171
```

Finally, let’s find the sample covariance of **b’X** and **c’X**.

$$Cov(b'X, c'X) = \frac{\left(b'x_1 - \overline{b'x}\right)\left(c'x_1 - \overline{c'x}\right) + \left(b'x_2 - \overline{b'x}\right)\left(c'x_2 - \overline{c'x}\right) + \left(b'x_3 - \overline{b'x}\right)\left(c'x_3 - \overline{c'x}\right)}{2}$$

```
## The sample covariance of b'X and c'X is: 27
```

(b) Calculate the sample means, variances, and covariances of b’X and c’X using (3-36). Compare the results in (a) and (b).

From (3-36):
Sample mean of **b’X** = $b'\bar{x}$
Sample mean of **c’X** = $c'\bar{x}$
Sample variance of **b’X** = **b’Sb**
Sample variance of **c’X** = **c’Sc**
Sample covariance of **b’X** and **c’X** = **b’Sc**

We will need to find the $\bar{x}$ and S matrices before we find the sample matrices.

The $\bar{x}$ matrix is:

```
##      [,1]
## [1,]    5
## [2,]    3
## [3,]    4
```

The S matrix is:

```
##      [,1] [,2] [,3]
## [1,] 13.0 -2.5  1.5
## [2,] -2.5  1.0 -1.5
## [3,]  1.5 -1.5  3.0
```

Now, we can calculate the sample summaries as according to (3-36).

```
## The sample mean of b'X is 12
## The sample mean of c'X is -1
## The sample variance of b'X is 12
## The sample variance of c'X is 43
## The sample covariance of b'X and c'X is -3
```

Let’s compare these results to the results from (a).

```
## The sample mean of bX is: 12
## The sample mean of cX is: 53
## The sample variance of bX is: 12
## The sample variance of cX is: 171
## The sample covariance of b'X and c'X is: 27
```

We see that the **b’X** sample mean and variance are the same in both (a) and (b). However, everything else is different. Everything in (a) is positive, while (b) has two negative values. There may have been some error in calculating in the sample summaries on my part in part (a), even if I was following the formulas provided in the book.

18. Energy consumption in 2001, by state, from the major sources

x_1 = petroleum

x_2 = natural gas

x_3 = hydroelectric power

x_4 = nuclear electric power

is recorded in quadrillions (10^{15}) in BTUs (Source: *Statistcal Abstract of the United States 2006*). The resulting mean and covariance matrix are

$$\bar{x} = \begin{bmatrix} 0.766 \\ 0.508 \\ 0.438 \\ 0.161 \end{bmatrix} \qquad S = \begin{bmatrix} 0.856 & 0.635 & 0.173 & 0.096 \\ 0.635 & 0.568 & 0.128 & 0.067 \\ 0.173 & 0.127 & 0.171 & 0.039 \\ 0.096 & 0.067 & 0.039 & 0.043 \end{bmatrix}$$

(a) Using the summary statistics, determine the sample mean and variance of a state’s total energy consumption for these major sources.

Let $y = x_1 + x_2 + x_3 + x_4$ be the total energy consumption.

Then, $\bar{y}$ is:

```
##       [,1]
## [1,] 1.873
```

Then s_y^2 is:

```
##        [,1]
## [1,] 3.913
```

(b) Determine the sample mean and variance of the excess of petroleum consumption over natural gas consumption. Also find the sample covariance of this variable with the total variable in part a.

The excess of petroleum consumption over natural gas consumption will be: $y = x_1 - x_2$.

Then, $\bar{y}$ is:

```
##        [,1]
## [1,] 0.258
```

Then, s_y^2 is:

```
##        [,1]
## [1,] 0.154
```

Code Appendix

```r
knitr::opts_chunk$set(echo = FALSE)
library(ggplot2)
library(ggExtra)
library(ggpubr)
library(matlib)
library(plotly)
library(plot3D)
library(rgl)
library(rglwidget)
library(webshot)
library(webshot2)
Xmat <- c(3,4,6,-2,3,1)
X <- matrix(Xmat, nrow = 3, ncol = 2, byrow = TRUE)
one <- as.matrix(rep(1, dim(X)[1]))
n <- dim(X)[1]
sample_mean <- 1/n*t(X) %*% one
print(sample_mean)
plot(Xmat, pch = 19, col = 3, main = "Scatter plot in 2D", xlab = "X1", ylab
= "X2")
points(4,1, col = "purple1", pch = 19)
sample_mean
mean <- matrix(c(4,4,4,1,1,1), ncol = 2)
d <- X - mean
colnames(d) <- c("d1","d2")
vector <- rbind(5*diag(3),t(d))
rownames(vector) <- c("X", "Y", "Z", "d1","d2")
#col <- c(rep("black",3))
d
d1 <- t(d[,1])
d2 <- t(d[,2])
rownames(d1) <- c("d1")
rownames(d2) <- c("d2")
# 3D
vectors3d(d1, headlength = 0.05, ref.length = NULL, radius = 0.05,
          labels = TRUE, cex.lab = 1.2, adj.lab = 0.5, frac.lab = 1.1)
vectors3d(d2, headlength = 0.05, ref.length = NULL, radius = 0.05,
          labels = TRUE, cex.lab = 1.2, adj.lab = 0.5, frac.lab = 1.1)
x0 <- c(0, 0)
y0 <- c(0, 0)
z0 <- c(0, 0)
x1 <- c(-1, 3)
y1 <- c(2, -3)
z1 <- c(-1, 0)
cols <- c("mediumpurple1","maroon3")
# 3D
arrows3D(x0, y0, z0, x1, y1, z1, colvar = x1^2, col = cols,
         lwd = 2, d = 3, main = "Part (b): Origin", bty ="b2", ticktype =
"detailed")
points3D(x0, y0, z0, add = TRUE, col="black",
          colkey = FALSE, pch = 19, cex = 1)
text3D(x1, y1, z1, c("d1", "d2"), colvar = x1^2, col = cols, add = TRUE,
colkey = FALSE)
sum1 <- rowSums(d1^2)
sum2 <- rowSums(d2^2)
l1 <- sqrt(sum1)
l2 <- sqrt(sum2)
# cat
cat("The length of d1 is", l1, "\n")
cat("The length of d2 is", l2)
s11 <- (d1 %*% t(d1))/3
s22 <- (d2 %*% t(d2))/3
s12 <- (d1 %*% t(d2))/3
# cat
cat("s11 is", s11, "\n")
cat("s22 is", s22, "\n")
cat("s12 is", s12, "\n")
# matrix
sMat <- c(s11, s12, s12, s22)
S <- matrix(sMat, nrow = 2, ncol = 2, byrow = TRUE)
S
theta <- acos( sum(d1*d2) / ( sqrt(sum(d1 * d1)) * sqrt(sum(d2 * d2)) ) )
angle <- cos(theta)
angle
rMat <- c(1, angle, angle, 1)
R <- matrix(rMat, nrow = 2, ncol = 2, byrow = TRUE)
R
xMat <- c(-1, 3, -2,
          2, 4, 2,
          5, 2, 3)
X <- matrix(xMat, nrow = 3, ncol = 3, byrow = TRUE)
one <- as.matrix(rep(1, dim(X)[1]))
n <- dim(X)[1]
xbar <- 1/n*t(X) %*% one
xbar
dev <- X - one %*% t(xbar)
dev
cat("d1 is [", dev[1, 1:3], "] \n")
cat("d2 is [", dev[2, 1:3], "] \n")
cat("d3 is [", dev[3, 1:3], "]")
double <- dev %*% t(dev)
S <- double/2
S
det(S)
S[1,1] + S[2,2] + S[3,3]
dataMat <- c(3, 1, 0,
             6, 4, 6,
             4, 2, 2,
```

```r
             7, 0, 3,
             5, 3, 4)
data <- matrix(dataMat, nrow = 5, ncol = 3, byrow = TRUE)
one <- as.matrix(rep(1, dim(data)[1]))
n <- dim(data)[1]
xbar <- 1/n*t(data) %*% one
xbar
mean <- matrix(data = 1, nrow = n) %*% cbind(xbar[[1]], xbar[[2]], xbar[[3]])
meanC <- data - mean
meanC
aMat <- c(1,1,-1)
a <- matrix(aMat, nrow = 1, ncol = 3, byrow = TRUE)
zero <- a %*% t(meanC)
a
t(meanC)
zero
covar <- 1/(n-1) * t(meanC) %*% meanC
covar
a %*% t(covar)
data
xMat <- c(1, 4, 3,
          6, 2, 6,
          8, 3, 3)
X <- matrix(xMat, nrow = 3, ncol = 3, byrow = TRUE)
bMat <- c(1,1,1)
b <- matrix(bMat, nrow = 1, ncol = 3, byrow = TRUE)
bX <- b %*% X
cMat <- c(1,2,-3)
c <- matrix(cMat, nrow = 1, ncol = 3, byrow = TRUE)
cX <- c %*% X
x1 <- t(b) * X[1,1:3]
bx1 <- b %*% x1
x2 <- t(b) * X[2,1:3]
bx2 <- b %*% x2
x3 <- t(b) * X[3,1:3]
bx3 <- b %*% x3
# cat
cat("The observed value of bX1 is:", bx1, "\n")
cat("The observed value of bX2 is:", bx2, "\n")
cat("The observed value of bX3 is:", bx3)
m1 <- (bx1 + bx2 + bx3)/3
cat("The sample mean of bX is:", m1)
v1 <- ((bx1 - m1)^2 + (bx2 - m1)^2 + (bx3 - m1)^2)/2
cat("The sample variance of bX is:", v1)
x1 <- t(c) * X[1,1:3]
cx1 <- c %*% x1
x2 <- t(c) * X[2,1:3]
cx2 <- c %*% x2
x3 <- t(c) * X[3,1:3]
cx3 <- c %*% x3
```

```r
# mean/var
m2 <- (cx1 + cx2 + cx3)/3
v2 <- ((cx1 - m2)^2 + (cx2 - m2)^2 + (cx3 - m2)^2)/2
# cat
cat("The observed value of cX1 is:", cx1, "\n")
cat("The observed value of cX2 is:", cx2, "\n")
cat("The observed value of cX3 is:", cx3, "\n")
cat("The sample mean of cX is:", m2, "\n")
cat("The sample variance of cX is:", v2)
num1 <- (bx1 - m1)*(cx1 - m2)
num2 <- (bx2 - m1)*(cx2 - m2)
num3 <- (bx3 - m1)*(cx3 - m2)
covar1 <- (num1 + num2 + num3)/2
cat("The sample covariance of b'X and c'X is:", covar1)
# xbar
one <- as.matrix(rep(1, dim(X)[1]))
n <- dim(X)[1]
xbar <- 1/n*t(X) %*% one
# S
mean <- matrix(data = 1, nrow = n) %*% cbind(xbar[[1]], xbar[[2]], xbar[[3]])
meanX <- X - mean
covar <- 1/(n-1) * t(meanX) %*% meanX
xbar
covar
sm1 <- b %*% xbar
sm2 <- c %*% xbar
sV1 <- b %*% covar %*% t(b)
sV2 <- c %*% covar %*% t(c)
covar2 <- b %*% covar %*% t(c)
# cat
cat("The sample mean of b'X is", sm1, "\n")
cat("The sample mean of c'X is", sm2, "\n")
cat("The sample variance of b'X is", sV1, "\n")
cat("The sample variance of c'X is", sV2, "\n")
cat("The sample covariance of b'X and c'X is", covar2)
cat("The sample mean of bX is:", m1, "\n")
cat("The sample mean of cX is:", m2, "\n")
cat("The sample variance of bX is:", v1, "\n")
cat("The sample variance of cX is:", v2, "\n")
cat("The sample covariance of b'X and c'X is:", covar1)
xbarMat <- c(0.766, 0.508, 0.438, 0.161)
xbar <- matrix(xbarMat, nrow = 4, ncol = 1, byrow = TRUE)
sMat <- c(0.856, 0.635, 0.173, 0.096,
          0.635, 0.568, 0.128, 0.067,
          0.173, 0.127, 0.171, 0.039,
          0.096, 0.067, 0.039, 0.043)
S <- matrix(sMat, nrow = 4, ncol = 4, byrow = TRUE)
oneMat <- c(1,1,1,1)
one <- matrix(oneMat, nrow = 1, ncol = 4, byrow = TRUE)
one %*% xbar
one %*% S %*% t(one)
minMat <- c(1,-1,0,0)
min <- matrix(minMat, nrow = 1, ncol = 4, byrow = TRUE)
min %*% xbar
min %*% S %*% t(min)
```