

## Homework Policy Update

- Starting with problem set 4, all code must be properly indented to receive full credit. This means that any time you have an if/else, for, and while block, you must indent the code inside the block. See me if you have questions about this.
- I cannot guarantee timely responses to emails sent the night before the homework is due.
- If you want me to look at your code to help fix errors, bring it in to office hours. Restrict email questions to specific parts of the code, with your questions clearly stated.
- Please start homeworks earlier so that you could take advantage of my office hours on Thursday, for which I have 2 hours to help clarify material and help you with the homework.

## An example of Indented code:

```
M = array(0, c(5,5,5) )
for( i in 1:5 )
{
  for( j in 1:5 )
  {
    for( k in 1:5)
    {
      M[i,j,k]=i+j+k
    }
  }
}
```

## An example of code NOT indented

```
M = array(0, c(5,5,5) )
for( i in 1:5 )
{
for( j in 1:5 )
{
for( k in 1:5)
{
M[i,j,k]=i+j+k
}
}
}
```

## Properties of Covariance

For random variables  $X, Y, U, V$  and constants  $a, b, c, d$ :

$$\text{cov}(X, Y) = E[XY] - E[X]E[Y] \quad (1)$$

$$\text{cov}(Y, X) = \text{cov}(X, Y) \quad (2)$$

$$\text{cov}(X, X) = E[XX] - E[X]E[X] = \text{var}(X) \quad (3)$$

$$\text{cov}(aX + c, bY + d) = ab \text{cov}(X, Y) \quad (4)$$

$$\text{cov}(X + Y, U + V) = \text{cov}(X, U) + \text{cov}(X, V) + \text{cov}(Y, U) + \text{cov}(Y, V) \quad (5)$$

For a sequence of random variables  $X_1, \dots, X_n$ :

$$\text{var}\left(\sum_{i=1}^n X_i\right) = \text{cov}\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i\right) = \sum_{i=1}^n \sum_{j=1}^n \text{cov}(X_i, X_j) \quad (6)$$

## The Covariance Matrix

Consider the sequence of random variables  $X_1, \dots, X_p$  defined as:

$$X_i = \theta Z_i + Z_{i-1}$$

where  $Z_0, \dots, Z_p$  are iid  $N(0,1)$ , and  $|\theta| < 1$ . This model is an example of a *moving average model of order 1* denoted MA(1).

1. Calculate the covariance matrix of  $X_1, \dots, X_p$  analytically.

### Solution:

We wish to find the matrix  $\Sigma$  where  $\Sigma_{ij} = \text{cov}(X_i, X_j)$  for  $1 \leq i, j \leq p$ . Apply properties (4) and (5):

$$\begin{aligned} \text{cov}(X_i, X_j) &= \text{cov}(\theta Z_i + Z_{i-1}, \theta Z_j + Z_{j-1}) \\ &= \text{cov}(\theta Z_i, \theta Z_j) + \text{cov}(\theta Z_i, Z_{j-1}) + \text{cov}(Z_{i-1}, \theta Z_j) + \text{cov}(Z_{i-1}, Z_{j-1}) \\ &= \theta^2 \text{cov}(Z_i, Z_j) + \theta \text{cov}(Z_i, Z_{j-1}) + \theta \text{cov}(Z_{i-1}, Z_j) + \text{cov}(Z_{i-1}, Z_{j-1}) \end{aligned}$$

For  $i = j$  apply property (3):

$$\text{cov}(X_i, X_i) = \text{var}(X_i) = \theta^2 \text{var}(Z_i) + \text{var}(Z_{i-1}) = \theta^2 + 1$$

For  $j = i - 1$  we have:

$$\begin{aligned} \text{cov}(X_i, X_{i-1}) &= \theta^2 \text{cov}(Z_i, Z_{i-1}) + \theta \text{cov}(Z_i, Z_{i-2}) + \theta \text{cov}(Z_{i-1}, Z_{i-1}) + \text{cov}(Z_{i-1}, Z_{i-2}) \\ &= \theta \text{cov}(Z_{i-1}, Z_{i-1}) = \theta \text{var}(Z_{i-1}) \\ &= \theta \end{aligned}$$

For  $j = i + 1$  we have:

$$\begin{aligned} \text{cov}(X_i, X_{i+1}) &= \theta^2 \text{cov}(Z_i, Z_{i+1}) + \theta \text{cov}(Z_i, Z_i) + \theta \text{cov}(Z_{i-1}, Z_{i+1}) + \text{cov}(Z_{i-1}, Z_i) \\ &= \theta \text{cov}(Z_i, Z_i) = \theta \text{var}(Z_i) \\ &= \theta \end{aligned}$$

For  $|i - j| \geq 2$  we have  $\text{cov}(X_i, X_j) = 0$ . In summary we found that:

$$\text{cov}(X_i, X_j) = \begin{cases} \theta^2 + 1 & \text{if } i = j \\ \theta & \text{if } |i - j| \leq 1 \\ 0 & \text{if } |i - j| \geq 2 \end{cases}$$

2. Use simulation to estimate the covariance matrix of  $X_1, \dots, X_p$  in R. You can use the `cov(X)` function, which computes the sample covariance of a matrix `X` for which the columns are reps realizations of random variable  $X_i, 1 \leq i \leq p$ . Let  $p = 8$  and  $\theta = 0.5$ .

```
reps = 1e5
theta = 0.5
p=8

## Generate reps realizations of Z_0, ..., Z_p
Z = array(rnorm(reps*(p+1)),c(reps,p+1))

## Find X in terms of Z
X = array(0, c(reps,p))
for (i in 1:p)
{
  ## X_i = theta * Z_i + Z_{i-1} (where the Z is one column off)
  X[,i] = theta*Z[,i+1] + Z[,i]
}

## Compute the sample covariance matrix
SigmaHat = cov(X)
round(SigmaHat,3)
```

3. Give an exact expression for  $\text{var}(\bar{X})$

$$\begin{aligned} \text{var}(\bar{X}) &= \text{var}\left(\frac{1}{p} \sum_{i=1}^p X_i\right) = \frac{1}{p^2} \sum_{i=1}^p \sum_{j=1}^p \text{cov}(X_i, X_j) \\ &= \frac{1}{p^2} \left( \sum_{i=1}^p \text{var}(X_i) + 2 \sum_{i=2}^p \text{cov}(X_i, X_{i-1}) \right) \\ &= \frac{1}{p^2} (p(\theta^2 + 1) + 2(p-1)\theta) \\ &= \frac{1}{p} (\theta^2 + 2\theta + 1) - 2\theta/p^2 \end{aligned}$$

4. Verify the formula for the last problem by estimating the  $\text{var}(\bar{X})$  with simulated data:

```

reps = 1e5
theta = 0.5
p=8

## Generate reps realizations of Z_0, ..., Z_p
Z = array(rnorm(reps*(p+1)),c(reps,p+1))

## Find X in terms of Z
X = array(0, c(reps,p))
for (i in 1:p)
{
  ## X_i = theta * Z_i + Z_{i-1} (where the Z is one column off)
  X[,i] = theta*Z[,i+1] + Z[,i]
}

## Compute the sample covariance matrix
SigmaHat = cov(X)

## Find the simulated var(Xbar)
Xbar = apply(X, 1, mean)
var(Xbar)

## Find the exact var(Xbar)
exact = (1/p)*(theta^2 + 2*theta + 1) - 2*theta*p^(-2)
exact

```

5. If the  $X_i$  were independent, with the same variances as the  $X_i$  defined here, what would the variance of  $\bar{X}$  be?

**Solution:**

$$\text{var}(\bar{X}) = \text{var}\left(\frac{1}{p} \sum_{i=1}^p X_i\right) = \frac{1}{p^2} \sum_{i=1}^p \text{var}(X_i) = (1 + \theta^2)/p$$

## Approximate Confidence Intervals

Let data  $X_1, \dots, X_n$  be iid Uniform(0,  $\theta$ ). An unbiased estimator of  $\theta$  is

$$\hat{\theta} = \frac{n+1}{n} \max(X_1, \dots, X_n)$$

Where unbiased means that  $E[\hat{\theta}] = \theta$ . This happens to be a scaled version of the biased maximum likelihood estimator  $\max(X_1, \dots, X_n)$ . We know that:

$$\text{var}(\hat{\theta}) = \theta^2 \left( \frac{(n+1)^2}{n(n+2)} - 1 \right)$$

1. Verify by simulation that this is the correct formula for  $\text{var}(\hat{\theta})$

```
n=20
reps = 1e5
theta = 5
X = array(runif(n*reps)*theta,c(n,reps))
theta.hat = (n+1)/n * apply(X, 2, max)
sim.var = var(theta.hat)
sim.var

formula.var = theta^2 * ((n+1)^2 / (n*(n+2)) - 1)
formula.var
```

2. Derive an approximate confidence interval for  $\theta$ , centered at  $\hat{\theta}$ , based on an iid sample of size  $n$ . You should begin by standardizing  $\hat{\theta}$ , then treat this standardized value as a standard normal value.

$$\begin{aligned} 0.95 &\approx P\left(-1.96 \leq \frac{\hat{\theta} - \theta}{\sqrt{\hat{\theta}^2 \left(\frac{(n+1)^2}{n(n+2)} - 1\right)}} \leq 1.96\right) \\ &= P\left(\hat{\theta} - 1.96\sqrt{\hat{\theta}^2 \left(\frac{(n+1)^2}{n(n+2)} - 1\right)} \leq \theta \leq \hat{\theta} + 1.96\sqrt{\hat{\theta}^2 \left(\frac{(n+1)^2}{n(n+2)} - 1\right)}\right) \end{aligned}$$

Giving the confidence interval:

$$\hat{\theta} \pm 1.96\sqrt{\hat{\theta}^2 \left(\frac{(n+1)^2}{n(n+2)} - 1\right)}$$

3. Use simulation to assess the coverage probabilities of your interval when the data  $X_1, \dots, X_n$  are iid Uniform(0,  $\theta$ ) for  $n = 20$  and  $\theta = 2, 5, 10$ .

```
nrep = 1e4
n = 20
theta = c(2,5,10)

CP = NULL
for (k in 1:3) {
  X = runif(nrep*n, min=0, max=theta[k])
  X = array(X, c(nrep,n))
  M = apply(X, 1, max)
  theta.hat = (n+1)*M/n
```

```

LB = theta.hat - 1.96*sqrt(theta.hat^2 * ( (n+1)^2/(n*(n+2)) - 1 ) )
UB = theta.hat + 1.96*sqrt(theta.hat^2 * ( (n+1)^2/(n*(n+2)) - 1 ) )

CP[k] = mean( (LB<theta[k]) & (theta[k]<UB) )
}

```

4. For those interested in verifying mathematically that  $E[\hat{\theta}] = \theta$  and  $\text{var}[\hat{\theta}] = \theta^2 \left( \frac{(n+1)^2}{n(n+2)} - 1 \right)$ .

$$P(\max(X_1, \dots, X_n) \leq t) = P(X_1 \leq t) \cdots P(X_n \leq t) = (t/\theta)^n$$

$$f_{\max(X_1, \dots, X_n)}(t) = nt^{n-1}/\theta^n$$

Now we find  $E[\max(X_1, \dots, X_n)]$  and  $E[\max(X_1, \dots, X_n)^2]$

$$E[\max(X_1, \dots, X_n)] = \int_0^\theta \frac{nt^n}{\theta^n} dt = n\theta/(n+1)$$

$$E[\max(X_1, \dots, X_n)^2] = \int_0^\theta \frac{nt^{n+1}}{\theta^n} dt = n\theta^2/(n+2)$$

Since our estimator is  $\hat{\theta} = \frac{n+1}{n} \max(X_1, \dots, X_n)$ , we just scale  $E[\max(X_1, \dots, X_n)]$  by  $\frac{n+1}{n}$ , yielding:

$$E[\hat{\theta}] = E\left[\frac{n+1}{n} \max(X_1, \dots, X_n)\right] = \frac{n+1}{n} E[\max(X_1, \dots, X_n)] = \frac{n+1}{n} n\theta/(n+1) = \theta$$

To find the variance of  $\hat{\theta}$  we just use the formula  $\text{var}[\hat{\theta}] = E[\hat{\theta}^2] - E[\hat{\theta}]^2$ :

$$E[\hat{\theta}^2] = E\left[\left(\frac{n+1}{n}\right)^2 \max(X_1, \dots, X_n)^2\right] = \left(\frac{n+1}{n}\right)^2 n\theta^2/(n+2)$$

$$\text{var}[\hat{\theta}] = \left(\frac{n+1}{n}\right)^2 n\theta^2/(n+2) - \theta^2$$

$$\text{var}(\hat{\theta}) = \theta^2 \left( \frac{(n+1)^2}{n(n+2)} - 1 \right)$$