STAT 720 sp 2019 hw 1

due on Wednesday, Feb 6th, 2019

1. Do problems 1.1, and 1.15 from B&D Intro.

2. Let $Y_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2}$, for $t \in \mathbb{Z}$, where $\{Z_t, t \in \mathbb{Z}\}$ is $\mathrm{WN}(0, \sigma^2)$.

   (a) Find the autocorrelation function $\rho(\cdot)$ of $\{Y_t, t \in \mathbb{Z}\}$.

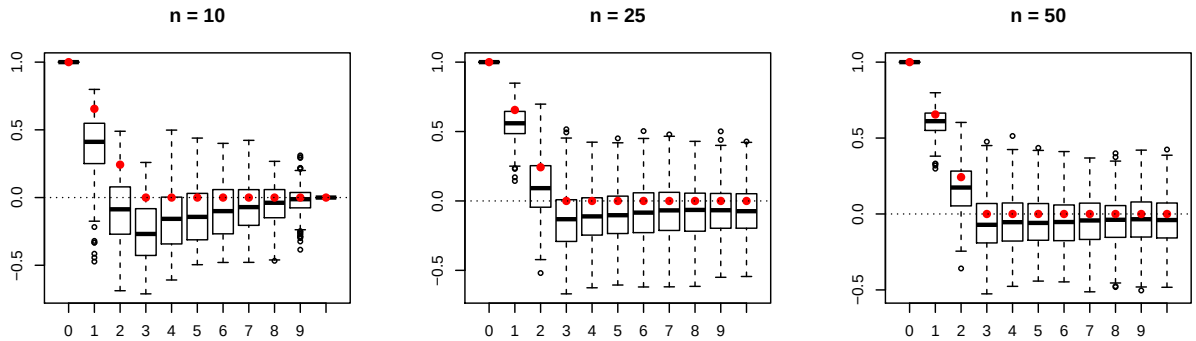
   We begin by finding the autocovariance function $\gamma(\cdot)$. We have

$$\begin{aligned}
\gamma(h) &= \mathrm{Cov}(Y_{t+h}, Y_t) \\
&= \mathbb{E} Y_{t+h} Y_t \\
&= \mathbb{E}[(Z_{t+h} + \theta_1 Z_{t+h-1} + \theta_2 Z_{t+h-2})(Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2})] \\
&= \mathbb{E}[Z_{t+h} Z_t + \theta_1 Z_{t+h} Z_{t-1} + \theta_2 Z_{t+h} Z_{t-2} \\
&\quad + \theta_1 Z_{t+h-1} Z_t + \theta_1^2 Z_{t+h-1} Z_{t-1} + \theta_1 \theta_2 Z_{t+h-1} Z_{t-2} \\
&\quad + \theta_2 Z_{t+h-2} Z_t + \theta_1 \theta_2 Z_{t+h-2} Z_{t-1} + \theta^2 Z_{t+h-2} Z_{t-2}] \\
&= \begin{cases} (1 + \theta_1^2 + \theta_2^2)\sigma^2, & h = 0 \\ (\theta_1 + \theta_1 \theta_2)\sigma^2, & |h| = 1 \\ \theta_2 \sigma^2, & |h| = 2 \\ 0, & |h| \geq 3. \end{cases}
\end{aligned}$$

   This gives

$$\rho(h) = \begin{cases} 1, & h = 0 \\ \dfrac{\theta_1 + \theta_1 \theta_2}{1 + \theta_1^2 + \theta_2^2}, & |h| = 1 \\ \dfrac{\theta_2}{1 + \theta_1^2 + \theta_2^2}, & |h| = 2 \\ 0, & |h| \geq 3. \end{cases}$$

   (b) Set $\theta_1 = 0.9$ and $\theta_2 = 0.5$ and let $\{Z_t, t \in \mathbb{Z}\}$ be independent $\mathrm{Normal}(0, 1)$ random variables. Then, for the sample sizes $n = 10$, $n = 25$, and $n = 50$, generate 500 realizations of $Y_1, \ldots, Y_n$. On each realization, compute the sample autocorrelation function at lags $h = 0, 1, \ldots, 10$. Then make boxplots of the estimated values of the autocorrelations from the 500 simulated time series with the true values of the autocorrelations overlaid. The resulting plots should look like this:



Turn in your plots and your R code. In addition, comment on the performance of the sample autocorrelation function as an estimator of the true autocorrelation function as the sample size increases.

```
## Function to compute sample autocorrelations:

my.acf <- function(x,max.lag=12)
{

        n <- length(x)
        x.bar <- mean(x)

        gamma.hat <- numeric(max.lag+1)

        for(h in 0:min(max.lag,n-1))
        {

                gamma.hat[h+1] <- 0

                for(t in 1:(n-h))
                {

                        gamma.hat[h+1] <- gamma.hat[h+1]+(x[t]-x.bar)*(x[t+h]-x.bar)

                }

        }

        gamma.hat <- gamma.hat / n
        rho.hat <- gamma.hat / gamma.hat[1]


        output <- list( gamma.hat = gamma.hat,
                                        rho.hat = rho.hat,
```

```
                                        lags = 0:max.lag)

        return(output)

}

######

theta1 <- .9
theta2 <- .5
sigma <- 1

rho <- c(   1,
            (theta1 + theta1*theta2)/(1+theta1^2+theta2^2),
            theta2/(1 + theta1^2 + theta2^2),
            rep(0,8) )

theta <- c(1,.9,.5)
q <- length(theta) - 1
S <- 500
max.lag <- 10

nn <- c(10,25,50)
acf.est <- array(NA,dim=c(S,max.lag+1,length(nn)))

for(j in 1:length(nn))
{
        n <- nn[j]

        for(s in 1:S)
        {

                Z <- rnorm(n+q,0,1)
                X <- numeric(n)

                for( t in 1:n)
                {
                        ind <- q + t:(t-q)
                        X[t] <- sum( theta * Z[ind] )
                }

                acf.est[s,,j] <- my.acf(X,max.lag)$rho.hat

        }

}
```

```
par(mfrow=c(1,3))
for(j in 1:length(nn))
{
        boxplot(acf.est[,,j],main=paste("n=",nn[j]),names=c(0:10))
        points(rho,pch=19,col="red")
        abline(h=0,lty=3)
}
```

3. Let $Y_t = m_t + \varepsilon_t$ for $t \in \mathbb{Z}$, where $m_t = c_0 + c_1 t + c_2 t^2$ and $\{\varepsilon_t, t \in \mathbb{Z}\}$ is a stationary time series with mean zero.

   (a) Show that the series $\nabla^2 Y_t, t \in \mathbb{Z}$ is stationary.

We have

$$\begin{aligned}
\nabla^2 Y_t &= \nabla(Y_t - Y_{t-1}) \\
&= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \\
&= c_0 + c_1 t + c_2 t^2 + \varepsilon_t - 2[c_0 + c_1(t-1) + c_2(t-1)^2 + \varepsilon_{t-1}] \\
&\quad + [c_0 + c_1(t-2) + c_2(t-2)^2 + \varepsilon_{t-2}] \\
&= c_2 t^2 - 2c_2(t^2 - 2t + 1) + c_2(t^2 - 4t + 4) + \varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2} \\
&= 2c_2 + \varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2}.
\end{aligned}$$

Since this is a constant plus a linear combination of realizations of a stationary process, it is stationary.

   (b) Find choices of $a_2 = a_{-2}$, $a_1 = a_{-1}$, and a_0 such that the filter

$$\hat{m}_t = \sum_{j=-2}^{2} a_j Y_{t-j} \tag{1}$$

is an unbiased estimator of m_t for all t.

The equation

$$\sum_{j=-2}^{2} a_j[c_0 + c_1(t-j) + c_2(t-j)^2] = c_0 + c_1 t + c_2 t^2$$

can be written as

$$(a_2 + a_1 + a_0 + a_1 + a_2)(c_0 + c_1 t + c_2 t^2) + c_2(2a_1 + 8a_2) = c_0 + c_1 t + c_2 t^2,$$

and this is satisfied if

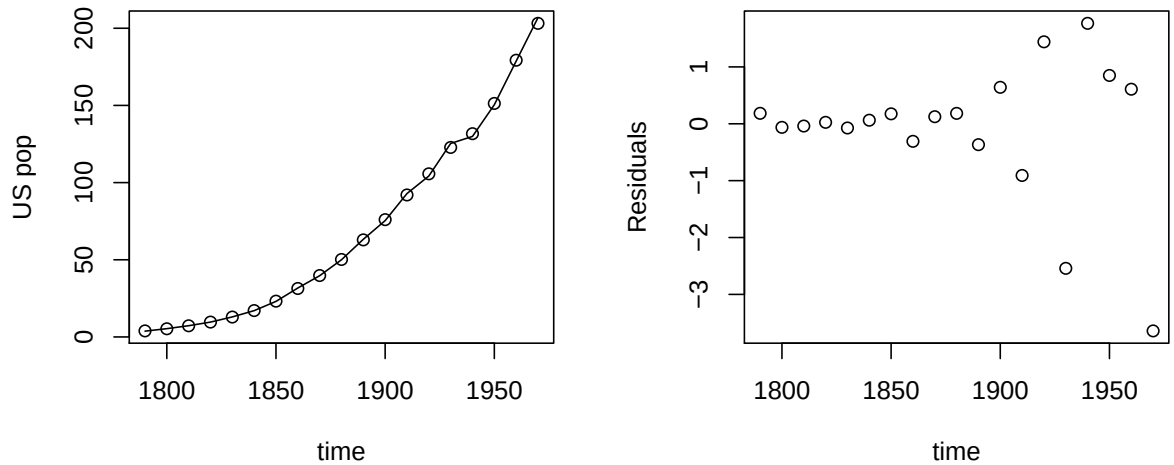
$$\begin{aligned}
a_2 + a_1 + a_0 + a_1 + a_2 &= 1 \\
2a_1 + 8a_2 &= 0 \quad (\iff a_1 = -4a_2).
\end{aligned}$$

So, for example, the choices $a_0 = 18/12$, $a_1 = -4/12$, and $a_2 = 1/12$ will work.

(c) Following part (b), set $a_0 = 18/12$ and a_1 and a_2 accordingly. Then apply the smoothing in (1) to the R data set `uspop`, using $Y_{-1} = Y_0 = Y_1$ and $Y_{n+1} = Y_{n+2} = Y_n$ to take care of estimation at the start and end of the series. Plot the time series with the estimated $\hat{m}_1, \ldots, \hat{m}_n$ overlaid and make a plot of the residuals $Y_t - \hat{m}_t$ versus t for all $t = 1, \ldots, n$. Supply the R code and the plots.

The following R code applies the smoother and produces the plots:

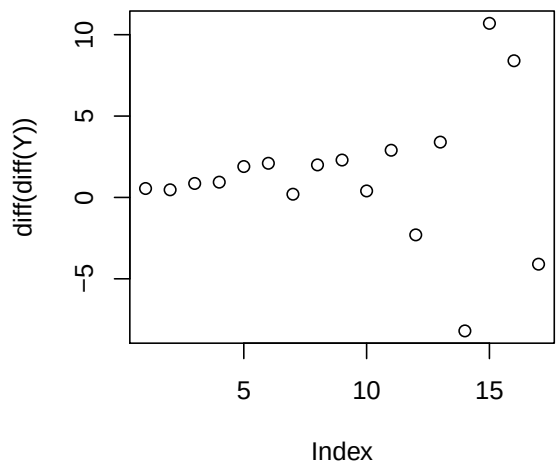
```
data(uspop)

Y <- as.numeric(uspop)
tt <- seq(1790,1970,by=10)
n <- length(Y)
a <- c(1/12,-4/12,18/12,-4/12,1/12)
q <- (length(a)-1)/2

Ymod <- c(rep(Y[1],q),Y,rep(Y[n],q))
m.hat <- numeric(n)

for(t in 1:n)
{
    m.hat[t] <- sum( a * Ymod[t:(t+2*q)])
}

par(mfrow=c(1,2))

plot(Y~tt,xlab="time",ylab="US pop")
lines(m.hat~tt)

plot(Y-m.hat~tt,xlab="time",ylab="Residuals")
```



(d) Since the trend in the `uspop` data appears to be quadratic, and since quadratic trends can be eliminated by differencing the series with the operator ∇^2, apply the differencing ∇^2 to the `uspop` data and plot the resulting series. Comment on whether you think the trend has been eliminated by the differencing.

The following R code produces a plot of the twice-difference series:

```
pdf(height=5,width=5,"hw01_plot02.pdf")

par(mfrow=c(1,1))
plot(diff(diff(Y)))

dev.off()
```



4. Show that for any random variables $X_1, \ldots, X_m$ and real numbers $a_1, \ldots, a_m$, we have

$$\mathrm{Var}(\sum_{i=1}^{m} a_i X_i) = \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j \,\mathrm{Cov}(X_i, X_j).$$

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^{m} a_i X_i\right) &= \mathbb{E}\left(\sum_{i=1}^{m} a_i X_i\right)^2 - \left[\mathbb{E}\left(\sum_{i=1}^{m} a_i X_i\right)\right]^2 \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j \mathbb{E} X_i X_j - \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j \mathbb{E} X_i \mathbb{E} X_j \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j \,\mathrm{Cov}(X_i, X_j).
\end{aligned}$$

5. Let $\{X_t, t \in \mathbb{Z}\}$ be a stationary time series with autocovariance function $\gamma(\cdot)$ and show that

$$\mathrm{Var}(\sqrt{n}\bar{X}_n) = \sum_{h=-(n-1)}^{n-1} \left(1 - \frac{|h|}{n}\right) \gamma(h),$$

where $\bar{X}_n = (X_1 + \cdots + X_n)/n$.

We have

$$\begin{aligned}
\operatorname{Var}(\sqrt{n}\bar{X}_n) &= n \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Cov}(X_i, X_j) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \gamma(i-j) \\
&= \frac{1}{n}\sum_{h=-(n-1)}^{n-1} (n-|h|)\gamma(h) \\
&= \sum_{h=-(n-1)}^{n-1} (1-|h|/n)\gamma(h).
\end{aligned}$$

6. Consider the MA(q) process defined by

$$X_t = \theta_0 Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \qquad \text{for} \qquad t \in \mathbb{Z},$$

where $\{Z_t, t \in \mathbb{Z}\}$ is $\mathrm{WN}(0, \sigma^2)$.

(a) Show that

$$\operatorname{Cov}(X_t, X_{t+h}) = \begin{cases} \sigma^2 \displaystyle\sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|} & \text{for } |h| \le q \\ 0 & \text{for } |h| > q \end{cases}$$

for all $t \in \mathbb{Z}$.

We have $\mathbb{E}X_t = 0$ for all $t \in \mathbb{Z}$. Moreover

$$\begin{aligned}
\operatorname{Cov}(X_t, X_{t+h}) &= \operatorname{Cov}\left(\sum_{j=0}^{q} \theta_j Z_{t-j}, \sum_{k=0}^{q} \theta_k Z_{t+h-k}\right) \\
&= \sum_{j=0}^{q}\sum_{k=0}^{q} \theta_j\theta_k \operatorname{Cov}(Z_{t-j}, Z_{t+h-k}) \\
&= \sum_{j=0}^{q}\sum_{k=0}^{q} \theta_j\theta_k \gamma_Z(t-j-(t+h-k)) \\
&= \sum_{j=0}^{q}\sum_{k=0}^{q} \theta_j\theta_k \gamma_Z(t-j-(t+h-k)) \\
&= \sum_{j=0}^{q}\sum_{k=0}^{q} \theta_j\theta_k \mathbb{1}(k=j+h)\gamma_Z(0) \\
&= \begin{cases} \sigma^2 \displaystyle\sum_{j=0}^{q-|h|} \theta_j\theta_{j+|h|} & \text{for } |h| \le q \\ 0 & \text{for } |h| > q. \end{cases}
\end{aligned}$$

(b) Show that

$$\lim_{n\to\infty} \operatorname{Var}(\sqrt{n}\bar{X}_n) = \sigma^2 \left(\sum_{j=0}^{q} \theta_j\right)^2.$$

Assuming $n > q$, we have

$$\begin{aligned}
\operatorname{Var}(\sqrt{n}\bar{X}_n) &= \sum_{h=-(n-1)}^{n-1} (1-|h|/n)\gamma(h) \\
&= \sum_{h=-(n-1)}^{n-1} (1-|h|/n)\sigma^2 \sum_{j=0}^{q-|h|} \theta_j\theta_{j+|h|} \\
&= \sum_{h=-q}^{q} (1-|h|/n)\sigma^2 \sum_{j=0}^{q-|h|} \theta_j\theta_{j+|h|}
\end{aligned}$$

Now we have

$$\lim_{n\to\infty} \operatorname{Var}(\sqrt{n}\bar{X}_n) = \sigma^2 \sum_{h=-q}^{q}\sum_{j=0}^{q-|h|} \theta_j\theta_{j+|h|} = \sigma^2 \left(\sum_{j=0}^{q} \theta_j\right)^2.$$

To see the last equality, consider summing all the entries of the matrix

$$\begin{pmatrix} \theta_0\theta_0 & \dots & \theta_0\theta_q \\ \vdots & \ddots & \vdots \\ \theta_q\theta_0 & \dots & \theta_q\theta_q \end{pmatrix}.$$

(c) Set $(\theta_0, \theta_1, \theta_2, \theta_3) = (1, 0.8, 0.5, -0.2)$ and let $\{Z_t, t \in \mathbb{Z}\}$ be independent $\text{Normal}(0, 1)$ random variables.

i. Give the asymptotic variance of the quantity $\sqrt{n}\bar{X}_n$ as $n \to \infty$ based on this time series.

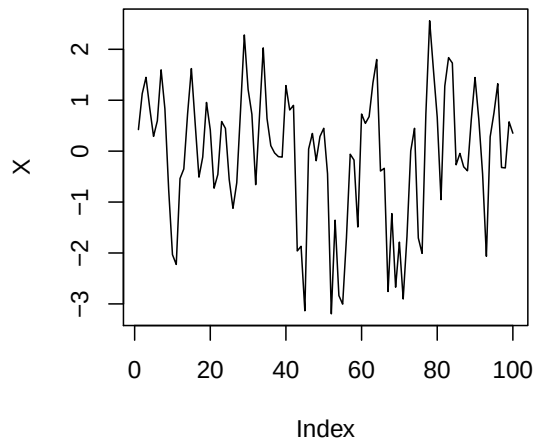
The asymptotic variance is

`sum(c(1,.8,.5,-.2))**2` $= 4.41$.

ii. Use R to generate a realization of length $n = 100$ of the time series; produce a plot of your time series using `plot(X,type="l")`. *Hint: You will need to generate* $n + q = 103$ *values from the* $\text{Normal}(0, 1)$ *distribution and then use these to construct the realizations* $X_1, \dots, X_n$.

The following R code does this:

```
n <- 100
theta <- c(1,.8,.5,-.2)
q <- length(theta) - 1

Z <- rnorm(n+q,0,1)
X <- numeric(n)

for(t in 1:n)
{
        ind <- q + t:(t-q)
        X[t] <- sum( theta * Z[ind] )
}

plot(X,type="l")
```



iii. Under the same settings, generate 1000 realizations of length $n = 100$ of the times series and compute the sample means of each of the 1000 realizations. Then compute the variance of $\sqrt{n}\bar{X}_n$ over the 1000 realizations. Compare this value to the theoretical asymptotic variance of $\sqrt{n}\bar{X}_n$.

The following R code runs the simulation:

```
n <- 100
theta <- c(1,.8,.5,-.2)
q <- length(theta) - 1
S <- 1000

X.bar <- numeric(S)

for(s in 1:S)
{

        Z <- rnorm(n+q,0,1)
        X <- numeric(n)

        for( t in 1:n)
        {
                ind <- q + t:(t-q)
                X[t] <- sum( theta * Z[ind] )
        }

        X.bar[s] <- mean(X)

}
```

```
var(sqrt(n) * X.bar)
sum(theta)^2
```

The empirical variance of $\sqrt{n}\bar{X}_n$ should be close to 4.41.