Homework 4

1 Data analysis problems

This week we will be analyzing a number of data sets. We are going to build ARIMA models using the steps outlined in class. It is also a good idea to read section 3.8 from our textbook. Make sure to outline the steps used in analyzing the data. If there are two (or more) competing models, make sure you discuss each of these. Make a decision about which model you think is best and support this decision with plots and and other information. I suggest using the `sarima()` function to do these fits. Here are short summaries about the data sets to analyze.

1. The data in `cow.dat` are the daily morning temperature readings for a cow.

   [Sol] It is better to take logarithm of the data and preform the analysis. But it is also ok if students do not take the logarithm. In either case, students need to take the difference of the data.

   [If students take logarithm of the data] ACF plot cuts off $q = 1$ and PACF tails off, it imply MA(1). It is also possible that both ACF and PACF tail off exponentially, so $(p, q) = (1, 1)$. So, we have two possible models ARIMA(0,1,1) and ARIMA(1,1,1). After fitting these three models, we find that the first model can be rejected by looking at the p-values plot of the diagnostic output. The second model ARIMA(1,1,1) seems to fit the data fairly well, it passes all diagnostic tests, and the smaller AIC value = -2.88, but the coefficient of AR is not significant. However, the coefficient of AR is fairly significant and the diagnostics of ARIMA(1,1,1) is better and AIC is smaller than one of ARIM(0,1,1). Thus, we conclude that ARIMA(1,1,1) is the best model. Here is the output:

```
Call:
arima(x = data, order = c(p, d, q), seasonal = list(order = c(P, D, Q),
period = S), xreg = constant, include.mean = F)

Coefficients:
         ar1     ma1  constant
      0.1815  -1.000   -0.0043
s.e.  0.1161   0.076    0.0009

sigma^2 estimated as 0.01916:  log likelihood = 39.36,  aic = -70.71

$AIC
[1] -2.875046

$AICc
[1] -2.840761

$BIC
[1] -3.782347
```

[If students do not take the logarithm] Obviously ACF cuts off after the first lag and PACF tails off. We can try other models but I find ARIMA(0,1,1) is the best model. Here is the output:

```
Call:
arima(x = data, order = c(p, d, q), seasonal = list(order = c(P, D, Q),
period = S), xreg = constant, include.mean = F)

Coefficients:
          ma1  constant
      -0.8724   -0.2380
s.e.   0.0983    0.1339

sigma^2 estimated as 65.06:  log likelihood = -260.21,  aic = 526.41

$AIC
[1] 5.22869

$AICc
```

```
[1] 5.259864

$BIC
[1] 4.29049
```

2. The data in `sheep.dat` are the sheep population (in millions) for England and Wales from 1867-1939.

   [Sol] **The time series plot indicates non-stationarity. Hence, differencing is needed. The patterns in ACF and PACF plot are not clear. ACF and PACF plots indicates $q = 4$ and $p = 3$, respectively. It is also possible that both ACF/PACF tail off exponentially, and hence $(p, q) = (1, 1)$. So, there are three possible models ARIMA(3,1,0), ARIMA(0,1,4) and ARIMA(1,1,1). After fitting these models, ARIMA(3,1,0) and ARIMA(0,1,4) seem to be reasonable because both models passes all the diagnostic tests provided by `sarima` function. If one is concerned with the normality assumption, ARIMA(3,1,0) appears to be better. But ARIMA(0,1,4) has the smallest AIC value. Here is the output:**

```
## ARIMA(3,1,0)
Call:
arima(x = data, order = c(p, d, q), seasonal = list(order = c(P, D, Q),
period = S), xreg = constant, include.mean = F)

Coefficients:
         ar1      ar2      ar3  constant
      0.4134  -0.2044  -0.3115   -5.8649
s.e.  0.1192   0.1357   0.1241    7.4555

sigma^2 estimated as 4742:  log likelihood = -407.26,  aic = 824.51

$AIC
[1] 9.573821

$AICc
[1] 9.613486

$BIC
```

```
[1] 8.699326

## ARIMA(0,1,4)
Call:
arima(x = data, order = c(p, d, q), seasonal = list(order = c(P, D, Q),
period = S), xreg = constant, include.mean = F)

Coefficients:
         ma1      ma2      ma3      ma4
      0.3109  -0.2357  -0.5094  -0.5657
s.e.  0.1190   0.1085   0.1229   0.1169
      constant
       -7.5410
s.e.    1.3763

sigma^2 estimated as 4348:  log likelihood = -405.72,  aic = 823.45

$AIC
[1] 9.514487

$AICc
[1] 9.559319

$BIC
[1] 8.671367
```

3. The data in `bicoal.dat` are the annual bituminous coal production levels in the US from 1930-1968 in millions of net tons per year.

   [Sol] **The time series plot indicates non-stationarity. Hence, differencing is needed. The patterns in ACF and PACF plot are not clear and it seem to be that both cut off. ACF and PACF plots indicates $q = 2$ and $p = 2$, respectively. It is also possible that both ACF/PACF tail off exponentially, and hence $(p, q) = (1, 1)$. So, there are three possible models ARIMA(2,1,0), ARIMA(0,1,2), ARIMA(1,1,1) and ARIMA(2,1,2). After fitting these models, ARIMA(0,1,2) and ARIMA(2,1,1) seem to be reasonable: both have smaller AIC values than other models and passes all the diagnostic tests. Note that ARIMA(2,1,1) is obtained from ARIMA(2,1,2)**

**because the second coefficient of MA part was not significant.
Here is the output:**

```
## ARIMA(0,1,2)
Call:
arima(x = data, order = c(p, d, q), seasonal = list(order = c(P, D, Q),
period = S), xreg = constant, include.mean = F)

Coefficients:
         ma1      ma2  constant
      0.0985  -0.4666    0.0641
s.e.  0.1312   0.1355    5.1830

sigma^2 estimated as 3051:  log likelihood = -260.93,  aic = 529.86

$AIC
[1] 9.145703

$AICc
[1] 9.205072

$BIC
[1] 8.261529

## ARIMA(2,1,1)
Call:
arima(x = data, order = c(p, d, q), seasonal = list(order = c(P, D, Q),
period = S), xreg = constant, include.mean = F)

Coefficients:
          ar1      ar2     ma1  constant
      -0.6653  -0.4345  0.6713    0.5941
s.e.   0.1946   0.1442  0.1729    6.1946

sigma^2 estimated as 2859:  log likelihood = -259.45,  aic = 528.9

$AIC
[1] 9.121623

$AICc
```

```
[1] 9.190916

$BIC
[1] 8.276058
```

Other answers are ok if reasonable explanation is given.

2 Theoretical problems

1. Identify the following as a specific ARIMA model. That is, what are p, $d(\geq 0)$ and q, and what are the values of the parameters?

$$x_t = x_{t-1} - 0.25x_{t-2} + w_t - 0.1w_{t-1}$$

where $w_t \sim wn(0, \sigma_w^2)$. We need to check the causal and invertible conditions (for causality condition, use the results of Example 3.8 in the textbook).

[Sol] This looks like an ARMA(2,1) model with $\phi_1 = 1$, $\phi_2 = -0.25$ and $\theta_1 = -0.1$. We need to check the stationarity conditions : $\phi_1 + \phi_2 = 0.75 < 1$, $\phi_2 - \phi_1 = -1.25 < 1$ and $\mid \phi_2 \mid = 0.25 < 1$. So, the process is a stationary and invertible ARMA(2,1).

2. Compute the values for $E(\nabla x_t)$ and $Var(\nabla x_t)$ in the following ARIMA model

$$x_t = 10 + 1.25x_{t-1} - 0.25x_{t-2} + w_t - 0.1w_{t-1}$$

where $w_t \sim wn(0, \sigma_w^2)$ (Hint : for $Var(\nabla x_t)$, you use the formula in Example 3.13 of the textbook).

[Sol] $\nabla x_t = x_t - x_{t-1} = 10 + 0.25(x_{t-1} - x_{t-2}) + w_t - 0.1w_{t-1}$. So the model is a causal and invertible ARIMA(1,1,1) with $\phi_1 = 0.25$, $\theta_1 = -0.1$. Hence $E(\nabla x_t) = \frac{10}{1-0.25} = \frac{40}{3}$ and $Var(\nabla x_t) = \frac{1+2\phi_1\theta_1+\theta_1^2}{1-\phi_1^2}\sigma_w^2 = 1.024\sigma_2^2$.

3. Suppose that x_t is generated according to $x_t = w_t + cw_{t-1} + cw_{t-2} + cw_{t-3} + \cdots + +cw_0$ for $t > 0$ where $w_t \sim wn(0, \sigma_w^2)$.

1) Find the mean and autocovariance functions for x_t, $E(x_t)$ and $cov(x_t, x_s)$ for $t < s$. Is $\{x_t\}$ stationary?

[Sol] $E(x_t) = 0$ **and** $Var(x_t) = (1 + tc^2)\sigma_w^2$ **which, in general, varies with** t**. Assume that** $t < s$**. Then** $cov(x_t, x_s) = cov(w_t + cw_{t-1} + cw_{t-2} + cw_{t-3} + \cdots + +cw_0, w_s + cw_{s-1} + cw_{s-2} + cw_{s-3} + \cdots + +cw_0) = (c + c^2 t)\sigma_w^2 = c(1 + ct)\sigma_w^2$**.**

2) Is ∇x_t stationary? Why or why not?

[Sol] $\nabla x_t = (w_t + cw_{t-1} + cw_{t-2} + cw_{t-3} + \cdots + +cw_0) - (w_{t-1} + cw_{t-2} + cw_{t-3} + cw_{t-4} + \cdots + +cw_0) = w_t - (1 - c)w_{t-1}$**. So this is stationary for any value of** c**.**

3) Identify x_t as a specific ARIMA process.

[Sol] The process $\{\nabla x_t\}$ **is an MA(1) process so that** $\{x_t\}$ **is ARIMA(0,1,1) process with** $\theta_1 = c - 1$**. The** $\{\nabla x_t\}$ **process is invertible if** $\mid c \mid< 1$**.**