332:525 – Optimum Signal Processing
Computer Experiment 3 – Due February 17, 2011

The Box-Jenkins airline data set has served as a benchmark in testing seasonal ARIMA models. In particular, it has led to the popular "airline model", which, for monthly data with yearly periodicity, is defined by the following innovations signal model:

$$\boxed{(1-Z^{-1})(1-Z^{-12})y_n = (1-bZ^{-1})(1-BZ^{-12})\varepsilon_n} \tag{1}$$

where Z^{-1} denotes the delay operator and b, B are constants. In this experiment, we briefly consider this model, but then replace it with the following ARIMA model, mainly because we have developed enough material in the course so far to handle it:

$$\boxed{(1-Z^{-12})y_n = \frac{1}{A(Z)}\,\varepsilon_n}, \quad A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_p z^{-p} \tag{2}$$

The airline data set can be loaded with the MATLAB commands:

```
Y = loadfile('airline.dat');
Y = Y'; Y = Y(:);                    % concatenate rows
y = log(Y);                          % log data
```

The data represent monthly airline passengers for the period Jan. 1949 to Dec. 1960. There are $N = 144$ data points. In this experiment, we will work with a subset of the first $n_0 = 108$ points, that is, y_n, $0 \le n \le n_0 - 1$, and attempt to predict the future 36 months of data (108 + 36 = 144.)

a. Plot Y_n and the log-data $y_n = \ln Y_n$ versus n and note the yearly periodicity. Note how the log-transformation tends to equalize the apparent increasing amplitude of the original data.

b. Compute and plot the normalized sample ACF, $\rho_k = R(k)/R(0)$, of the zero-mean log-data for lags $0 \le k \le 40$ and note the peaks at multiples of 12 months.

c. Let $x_n = (1-Z^{-1})(1-Z^{-12})y_n$ in the model of Eq. (1). The signal x_n follows an MA model with spectral density:

$$S_{xx}(z) = \sigma_\varepsilon^2 (1-bz^{-1})(1-bz)(1-Bz^{-12})(1-Bz^{12})$$

Multiply the factors out and perform an inverse z-transform to determine the autocorrelation lags $R_{xx}(k)$. Show in particular, that

$$\frac{R_{xx}(1)}{R_{xx}(0)} = -\frac{b}{1+b^2}, \qquad \frac{R_{xx}(12)}{R_{xx}(0)} = -\frac{B}{1+B^2} \tag{3}$$

Filter the subblock $y_n, 0 \le n \le n_0 - 1$ through the filter $(1-z^{-1})(1-z^{-12})$ to determine x_n. You may discard the first 13 transient outputs of x_n. Use the rest of x_n to calculate its sample ACF and apply Eq. (3) to solve for the model parameters b, B. This simple method gives values that are fairly close to the Box/Jenkins values determined by maximum likelihood methods, namely, $b = 0.4$ and $B = 0.6$ (see Reference in the `airline.dat` file).

d. Consider, next, the model of Eq. (2) with a second-order AR model i.e., $A(z)$ filter order of $p = 2$. Define $x_n = (1-Z^{-12})y_n = y_n - y_{n-12}$ and denote its autocorrelation function by R_k. Since x_n follows an AR(2) model, its model parameters $a_1, a_2, \sigma_\varepsilon^2$ can be computed from:

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = -\begin{bmatrix} R_0 & R_1 \\ R_1 & R_0 \end{bmatrix}^{-1} \begin{bmatrix} R_1 \\ R_2 \end{bmatrix}, \quad \sigma_\varepsilon^2 = R_0 + a_1 R_1 + a_2 R_2 \tag{4}$$

Using the data subset $y_n, 0 \le n \le n_0 - 1$, calculate the signal x_n and discard the first 12 transient samples. Using the rest of x_n, compute its sample ACF, $\hat{R}_k$ for $0 \le k \le M$ with $M = 40$, and use the first 3 computed lags $\hat{R}_0, \hat{R}_1, \hat{R}_2$ in Eqs. (4) to estimate $a_1, a_2, \sigma_\varepsilon^2$ (in computing the ACF, the signal x_n need not be replaced by its zero-mean version).

Because of the assumed autoregressive model, we will see in Ch.5 that it is possible to calculate all the autocorrelation lags R_k for $k \ge p+1$ from the first $p+1$ lags. This can accomplished by the MATLAB "autocorrelation sequence extension" function:

```
M=40; Rext = acext(R(1:p+1), zeros(1,M-p));
```

On the same graph, plot the sample and extended ACFs, $\hat{R}(k)$ and $R_{\text{ext}}(k)$ versus $0 \le k \le M$ normalized by their lag-0 values.

e. Let $\hat{x}_n$ denote the estimate/prediction of x_n based on the data subset $\{x_m, m \le n_0 - 1\}$. Clearly, $\hat{x}_n = x_n$, if $n \le n_0 - 1$. Writing Eq. (2) recursively, we obtain for the predicted values into the future beyond n_0:

$$\begin{aligned} x_n &= -\big(a_1 x_{n-1} + a_2 x_{n-2}\big) + \varepsilon_n \\ \hat{x}_n &= -\big(a_1 \hat{x}_{n-1} + a_2 \hat{x}_{n-2}\big), \quad \text{for } n \ge n_0 \end{aligned} \tag{5}$$

where we set $\hat{\varepsilon}_n = 0$ because all the observations are in the strict past of ε_n when $n \ge n_0$.

Calculate the predicted values 36 steps into the future, $\hat{x}_n$ for $n_0 \le n \le N-1$, using the fact that $\hat{x}_n = x_n$, if $n \le n_0 - 1$. Once you have the predicted x_n's, you can calculate the predicted y_n's by the recursive equation:

$$\hat{y}_n = \hat{y}_{n-12} + \hat{x}_n \tag{6}$$

where you must use $\hat{y}_n = y_n$, if $n \le n_0 - 1$. Compute $\hat{y}_n$ for $n_0 \le n \le N-1$, and plot it on the same graph with the original data y_n, $0 \le n \le N-1$. Indicate the start of the prediction horizon with a vertical line at $n = n_0$ (see example graph at end.)

f. Repeat parts (d,e) using an AR(4) model (i.e., $p = 4$), with signal model equations:

$$x_n = y_n - y_{n-12}, \quad x_n = -\big(a_1 x_{n-1} + a_2 x_{n-2} + a_3 x_{n-3} + a_4 x_{n-4}\big) + \varepsilon_n$$

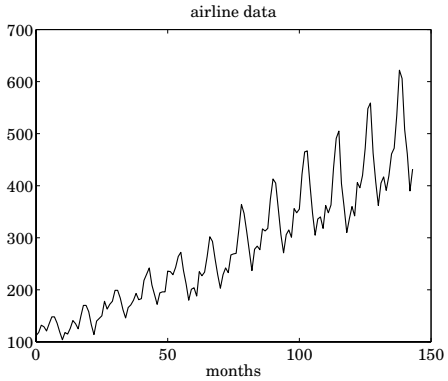
and normal equations:

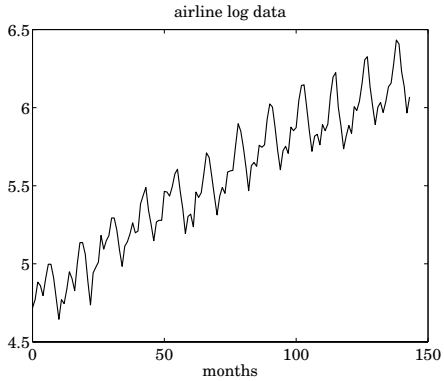
$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = -\begin{bmatrix} R_0 & R_1 & R_2 & R_3 \\ R_1 & R_0 & R_1 & R_2 \\ R_2 & R_1 & R_0 & R_1 \\ R_3 & R_2 & R_1 & R_0 \end{bmatrix}^{-1} \begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \end{bmatrix}, \quad \sigma_\varepsilon^2 = R_0 + a_1 R_1 + a_2 R_2 + a_3 R_3 + a_4 R_4$$

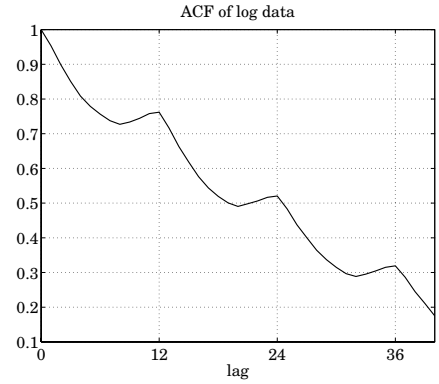
with predictions:

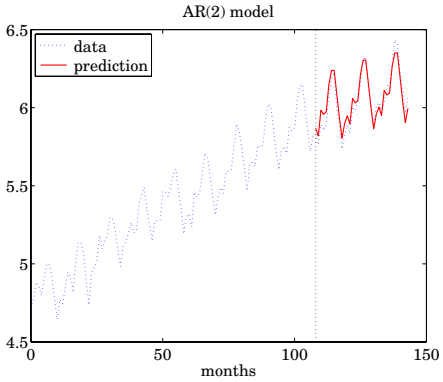
$$\begin{aligned} \hat{x}_n &= -\big(a_1 \hat{x}_{n-1} + a_2 \hat{x}_{n-2} + a_3 \hat{x}_{n-3} + a_4 \hat{x}_{n-4}\big) \\ \hat{y}_n &= \hat{y}_{n-12} + \hat{x}_n \end{aligned}$$

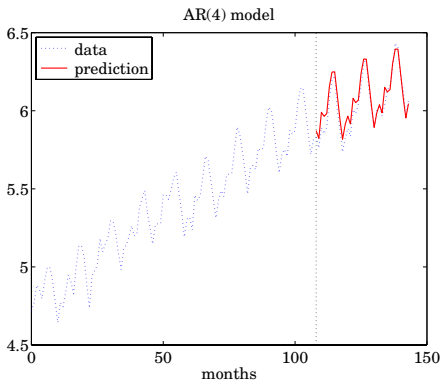
taking into account the properties that $\hat{x}_n = x_n$ and $\hat{y}_n = y_n$, if $n \le n_0 - 1$.

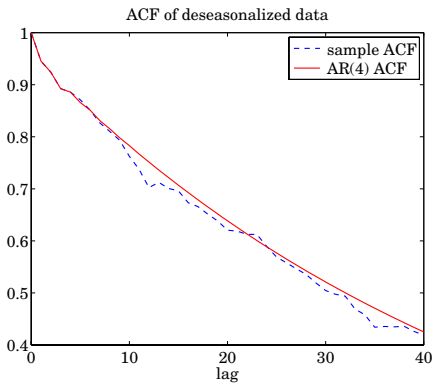

airline data
700
600
500
400
300
200
100
0
50
100
150
months


airline log data
6.5
6
5.5
5
4.5
0
50
100
150
months


ACF of log data
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
12
24
36
lag


AR(2) model
data
prediction
6.5
6
5.5
5
4.5
0
50
100
150
months


AR(4) model
data
prediction
6.5
6
5.5
5
4.5
0
50
100
150
months


ACF of deseasonalized data
sample ACF
AR(4) ACF
1
0.9
0.8
0.7
0.6
0.5
0.4
0
10
20
30
40
lag