STAT GU4221/GR5221 Homework 3 [100 pts]
Due: Friday, April 14th at 11:59pm (ET)

Problem 1

The goal of problem 1 is to explore the intricate details of the **sample partial autocorrelation function** ($PACF$). Further, students should understand how the sample $PACF$ can be used to identify the model order ($p > 0$) of a pure autoregressive process $AR(p)$. Assume that the observed data `HW3_AR_Data1.csv` and `HW3_AR_Data2.csv` come from pure autoregressive models with respective orders $p_1 > 0$ and $p_2 > 0$.

1.i Build your own sample $PACF$ using `R` or `Python` and test the function on datasets `HW3_AR_Data1.csv` and `HW3_AR_Data2.csv`. Report the sample $PACF$ values for lags $h = 0, 1, \ldots, 20$. **Note:** You can extract the lag-correlations $\hat{\rho}(h)$ and/or the lag-covariances $\hat{\gamma}(h)$ directly from the `R` function `acf()`. Also, there is no need to include the 95% confidence bands.

1.ii Use the `R` function `pacf()` to compute the sample partial autocorrelations of datasets `HW3_AR_Data1.csv` and `HW3_AR_Data2.csv` and compare these results with part 1.i. The output should be the same!

1.iii Based on the `R` function `pacf()`, identify the autoregressive orders ($p_1 > 0$ and $p_2 > 0$) for datasets `HW3_AR_Data1.csv` and `HW3_AR_Data2.csv`. Briefly explain how you chose the AR model orders from the $PACF$ plots. **Note:** For this problem part, you will use the 95% confidence bands provided by the `pacf()` output.

Problem 2

The goal of problem 2 is to estimate autoregressive parameters ϕ_j and noise variance σ^2 based on the sample Yule-Walker equations (or method of moments). You can assume a causal $AR(p)$ model with zero mean.

2.i Build your own sample Yule-Walker estimator/function using `R` or `Python`. Test your estimator on datasets `HW3_AR_Data1.csv` and `HW3_AR_Data2.csv`. The model orders ($p_1 > 0$ and $p_2 > 0$) should be consistent with problem 1.iii.

2.ii Run your Yule-Walker estimates on the dataset `HW3_AR_Data2.csv` using $AR(p)$ orders $p = 1, 2, 3, 4, 5$. Display all estimated parameters in a tabular format and interpret the final result.

2.iii Construct a 95% confidence interval for ϕ_1 based on the dataset `HW3_AR_Data1.csv`. Using the 95% interval, test the null/alternative pair

$$H_0 : \phi_1 = -1 \quad \text{versus} \quad H_A : \phi_1 \neq -1.$$

Interpret your testing procedure and its results in a few sentences.

Problem 3

Consider the sample $X_1, X_2, \ldots, X_n$, where each $\{X_t\}$ is generated by the $AR(1)$ process

$$X_t = \phi X_{t-1} + Z_t, \quad \text{where} \quad Z_t \overset{iid}{\sim} N(0, \sigma^2),$$

and σ^2 is known. Assuming that the realized time series is centered ($x_t := x_t - \bar{x}$), define the sample autocovariance function as

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{i=1}^{n-|h|} x_{t+|h|} x_t, \quad -n < h < n.$$

3.i Show that the sample Yule-Walker estimator of a causal $AR(1)$ is

$$\hat{\phi} = \frac{\sum_{i=1}^{n-1} X_{i+1} X_i}{\sum_{i=1}^{n} X_i^2}.$$

Note: This is an easy problem. You can simply reference the sample Yule-Walker equations and plug in $\hat{\gamma}(h)$.

3.ii Now consider the sample $X_0, X_1, \ldots, X_n$, where each $\{X_t\}$ is generated by the $AR(1)$ process and we allow for the non-stationary solution $\phi = 1$. Also assume that σ^2 is unknown. One method of testing the null/alternative pair:

$$H_0 : \phi = 1 \quad \text{versus} \quad H_A : \phi \neq 1,$$

is based on the test statistic

$$T_n = \frac{\hat{\phi}_n - 1}{\sqrt{\frac{S_n^2}{\sum_{i=1}^{n} X_{i-1}^2}}},$$

where

$$\hat{\phi}_n = \frac{\sum_{i=1}^{n} X_{i-1} X_i}{\sum_{i=1}^{n} X_{i-1}^2} \quad \text{and} \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\phi}_n X_{i-1})^2.$$

Rejecting the null hypothesis indicates that the series exhibits stationary behavior against the random walk null. It can be shown, under $H_0 : \phi = 1$, that the limiting distribution of T_n is

$$T_n \xrightarrow{d} \frac{[W(1)]^2 - 1}{2\sqrt{\int_0^1 [W(r)]^2 dr}},$$

where $W(r)$ is standard Brownian Motion.

Task: Assuming that the data generating process is a random walk, simulate the limiting distribution of the test statistic T_n using $n = 10, 30, 100, 1000$. Your final answer should be presented as four histograms, one for each sample size n.

Note: For this problem, you will simulate the limiting distribution by generating $R = 10000$ random walk processes and then computing the resulting estimators $\hat{\phi}_n^{(1)}, \ldots, \hat{\phi}_n^{(10000)}$ and test statistics $T_n^{(1)}, \ldots, T_n^{(10000)}$. Then construct histograms of the T_n's, one for each sample size $n = 10, 30, 100, 1000$.

3.iii If $H_0 : \phi = 1$ is true, argue the below limit:

$$T_n \xrightarrow{d} \frac{[W(1)]^2 - 1}{2\sqrt{\int_0^1 [W(r)]^2 dr}},$$

where $W(r)$ is standard Brownian Motion.

Note: You can directly use the below limits, i.e., under the random walk process:

$$\frac{1}{n}\sum_{i=1}^{n} X_{i-1}Z_i \xrightarrow{d} \frac{1}{2}\sigma^2([W(1)]^2 - 1) \quad \text{and} \quad \frac{1}{n^2}\sum_{i=1}^{n} X_{i-1}^2 \xrightarrow{d} \sigma^2 \int_0^1 [W(r)]^2 dr,$$

Problem 4

Consider the time series process $\{X_i\}$ with mean $E[X_i] = 0$ and covariance function $\gamma(i,j) = E[X_i X_j]$. Define the n^{th} innovation as $U_n = X_n - \hat{X}_n$, where

$$\hat{X}_n = \begin{cases} 0 & \text{if} \quad n = 1, \\ P(X_n | X_{n-1}, \ldots, X_1) & \text{if} \quad n = 2, 3, \ldots. \end{cases}$$

4.i In the innovations algorithm, show that for each $n \geq 2$, the innovation $U_n = X_n - \hat{X}_n$ is uncorrelated with $X_1, \ldots, X_{n-1}$. Conclude that $U_n = X_n - \hat{X}_n$ is uncorrelated with the innovations $X_1 - \hat{X}_1, \ldots, X_{n-1} - \hat{X}_{n-1}$.

4.ii Derive the update step for θ in the innovations algorithm, i.e., derive the expression

$$\theta_{n,n-k} = v_k^{-1}\left(\gamma(n+1, k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}v_j\right), \quad 0 \leq k < n.$$

Note: You can assume that the update step for v is:

$$v_n = \gamma(n+1, n+1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 v_j.$$

Hint: Start with the expression

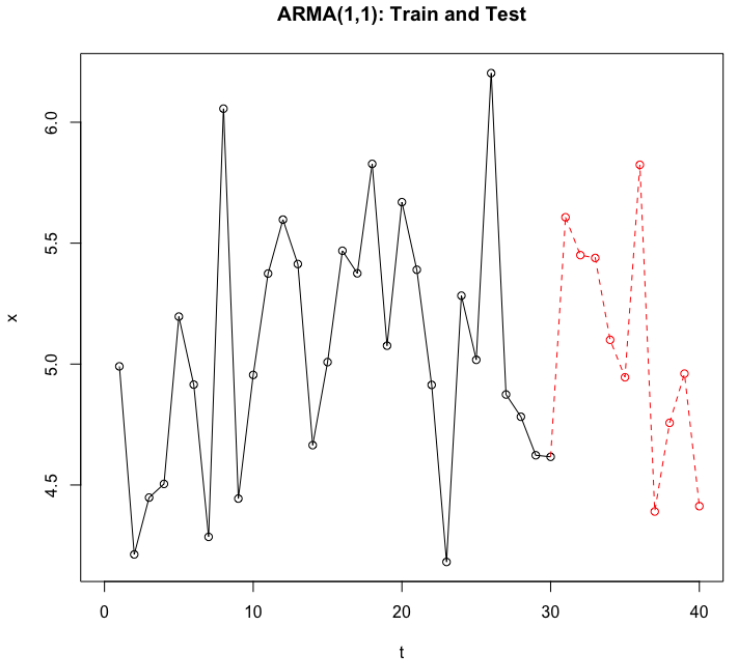
$$X_{n+1} = \sum_{j=0}^{n} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j})$$

and multiply each side by innovation $U_{k+1} = (X_{k+1} - \hat{X}_{k+1})$.

4.iii Consider the observed data `HW3_MA_Data.csv`. For this exercise, use the `R` function `ia()` from the `itsmr` library to estimate the pure $MA(q)$ model. You will also have to choose the MA order $q > 0$ based tools discussed in class. Report how you chose the model order and your estimated θ's, i.e., report $\hat{\theta}$'s.

Problem 5

The following figure shows the training data `ARMA_11_train.csv` and testing data `ARMA_11_test.csv` used to solve Problem 5.



Consider the training sample $X_1, X_2, \ldots, X_{30}$, where each $\{X_t\}$ is generated by the causal and invertible $ARMA(1,1)$ process

$$X_t = \mu + \phi(X_{t-1} - \mu) + Z_t + \theta Z_{t-1}, \quad \text{where} \quad Z_t \overset{iid}{\sim} N(0, \sigma^2).$$

Based on the Gaussian error structure (Z_t), the $ARMA(1,1)$ process can be described by a multivariate normal distribution:

$$f(x_1, \ldots, x_n | \mu, \phi, \theta, \sigma) = (2\pi)^{-n/2} |\mathbf{\Sigma}^{-1}|^{1/2} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{\Sigma}^{-1})(\mathbf{x} - \boldsymbol{\mu}) \right\}$$

$$-\infty < x_1, \ldots, x_n < \infty.$$

Hence the likelihood can be expressed as:

$$\mathcal{L}(\mu, \phi, \theta, \sigma | x_1, \ldots, x_n) = (2\pi)^{-n/2} |\mathbf{\Sigma}^{-1}|^{1/2} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{\Sigma}^{-1})(\mathbf{x} - \boldsymbol{\mu}) \right\}$$

Solve problems (5.i) - (5.v)

5.i Write down the exact likelihood for the $ARMA(1,1)$ in terms of parameters $\mu, \theta, \phi, \sigma$. Note that you must express the mean vector $(\boldsymbol{\mu})$ and covariance matrix $(\mathbf{\Sigma})$ in terms of $\mu, \theta, \phi, \sigma$.

5.ii Consider the dataset `ARMA_11_train.csv`. Write the negative log-likelihood as a `R` or `Python` function. Test your function at the point $(\mu = 0, \theta = .9, \phi = .9, \sigma = 1)$.

5.iii Use a built in-minimizer to optimize the negative log-likelihood. In `R`, I recommend using `nlm()`. If you are using `nlm()`, try the starting point

```
p=c(mean(train),0,0,sd(train))
```

If you fail to solve problems (5.ii) and/or (5.iii), you can run the following code for partial credit:

```
arima(train, order=c(1,0,1), method="ML", include.mean = T)
```

5.iv Using the data `ARMA_11_test.csv`, forecast $h = 1, 2, 3, \ldots, 10$ steps ahead, i.e, compute

$$\hat{X}_{n+1}, \hat{X}_{n+2}, \hat{X}_{n+3}, \ldots, \hat{X}_{n+10}.$$

Also compute the corresponding mean square prediction errors. Solve this problem in two ways:

5.iv.a Assuming the $ARMA(1,1)$ structure, replace θ, ϕ, σ with the respective MLE's $\hat{\theta}, \hat{\phi}, \hat{\sigma}$ and solve the prediction equation, i.e., find the coefficients $\hat{\mathbf{a}}$ that satisfy $\hat{\mathbf{\Gamma}}\hat{\mathbf{a}} = \hat{\boldsymbol{\gamma}}$. Think about how to incorporate μ for the MLE case. You can check you answer with the following code:

```
arma11_R = arima(train, order=c(1,0,1), method="ML",include.mean = T)
predict(arima11_R, n.ahead = 10)
```

5.iv.b Don't assume an $ARMA(1,1)$ and set up $\hat{\mathbf{\Gamma}}, \hat{\boldsymbol{\gamma}}$ directly using the sample acvf values, i.e., `acf()`. Then find the coefficients $\hat{\mathbf{a}}$ that satisfy $\hat{\mathbf{\Gamma}}\hat{\mathbf{a}} = \hat{\boldsymbol{\gamma}}$. Think about how to incorporate μ for this case.

5.v Using the data `ARMA_11_test.csv`, compute the test error for each method described above. See class notes for more details.

Note: The datasets `ARMA_11_train.csv` and `ARMA_11_test.csv` are posted on Canvas.