KTH Royal Institute of Technology

Embedded Intelligence
IL2233 VT24

Lab 2 - ARIMA Model and Prediction

Humblet Raphaël
Yao Tianze

May 2024

Academic year 2023-2024

Contents

1 Task 1: Stationarity of AR models

The models that will be studied in task 1 are the following:

1. $AR(1) : x_t = 0.8x_{t-1} + \epsilon_t$
2. $AR(1) : x_t = -1.1x_{t-1} + \epsilon_t$
3. $AR(2) : x_t = x_{t-1} - 0.5x_{t-2} + \epsilon_t$
4. $AR(2) : x_t = x_{t-1} + 0.5x_{t-2} + \epsilon_t$

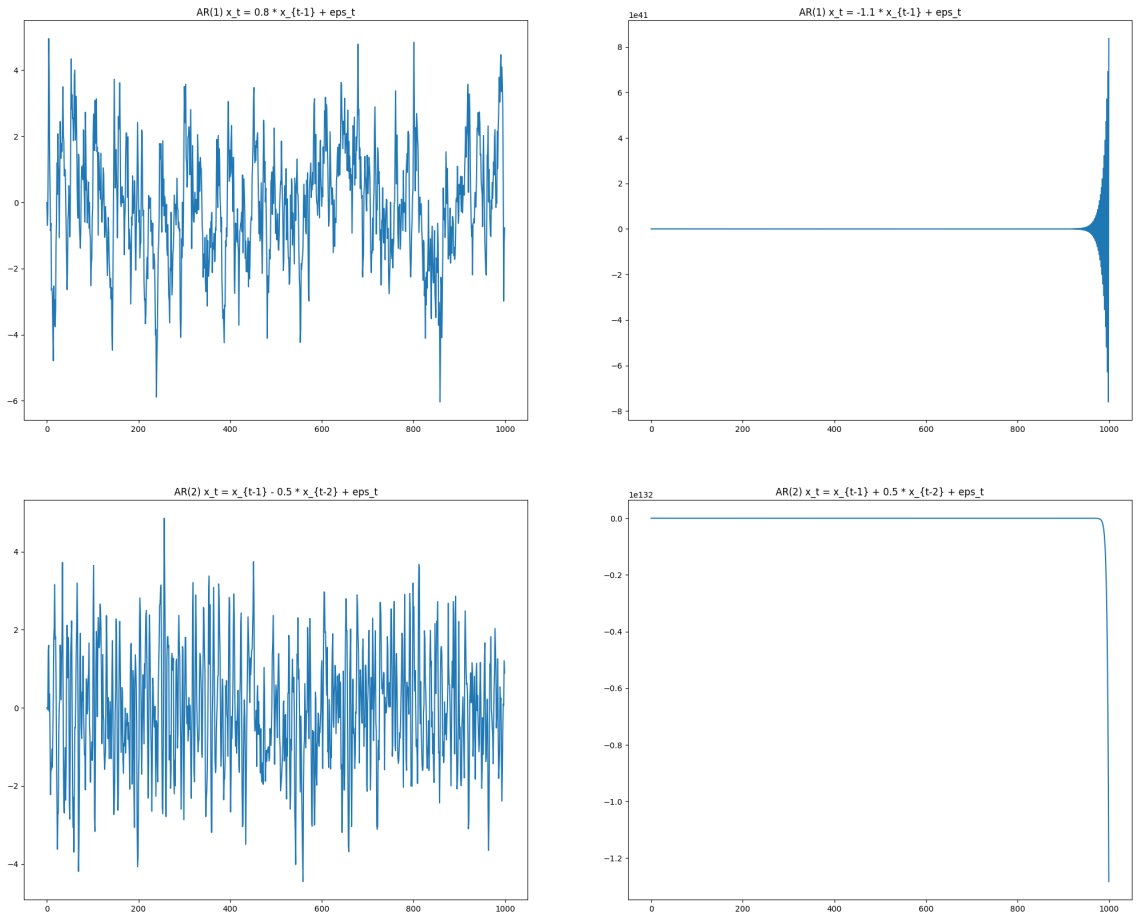

Figure 1: Task 1 series line plots

By visual inspection on figure 1, we can see that series 1 and 3 are stationary while series 2 and 4 are not, even though series 2 is almost stationary (except for the end).

With the auto-regressive coefficients, the series 1 and 3 are stationary but series 2 and 4 is not, it has been tested with the code.

But the unit-root test (ADF test) said that series 2 was also stationary.

```
Test: 1 p-value: 1.4811673763172147e-20
Test: 2 p-value: 0.0
Test: 3 p-value: 0.0
Test: 4 p-value: 1.0
```

Figure 2: Result of ADF test for task 1 series

With visual inspection, how do you identify if a time series is stationary or not?
By looking at the line-plot. If depending on the time period we analyse, the trend is different, then it is time-dependent and thus not stationary.

How do you judge the stationarity of time series using the unit-root method? Does it always give correct results?
In the Augmented Dickey-Fuller (ADF) test, we examine the p-value. If the p-value is less than 0.05, we reject the null hypothesis that the series has a unit root, and we may consider the series to be stationary.

However, this is not always accurate. For instance, a series might appear to be stationary over a certain period but later exhibit non-stationary behavior. This could lead to a Type II error, where we fail to reject the null hypothesis when it is false. In other words, we incorrectly classify a non-stationary series as stationary. This is particularly likely when the series is close to being stationary, but has subtle non-stationary characteristics, such as a slowly changing mean or variance.

What is the role of component ϵ_t in the model? Why is it important?
This is the error term/residuals, it is important for multiple reasons:

- **Model Adequacy:** used to check the adequacy of the model. If the model fits well, the residuals should be white noise.
- **Predictive power:** represent the unpredictability inherent in the process being modelled. A smaller error term indicates a model that can explain a larger portion of the variation in the data, and thus potentially has better predictive power.
- **Model improvement:** By analyzing the error term, we can identify ways to improve the model.

To have an AR(p) model be stationary, is there any requirement on the auto-regressive coefficients? List the constraints for AR(1) and AR(2) models
In order for the AR model to be stationary we have the following constraints.

For **AR(1)** models of the form

$$y_t = c + \phi_1 y_{t-1} + \epsilon_t$$

We must have $-1 < \phi_1 < 1$.

For **AR(2)** models:

$$y_t = c + \phi_2 y_{t-2} + y_{t-1} + \epsilon_t$$

We must have $-1 < \phi_2 < 1$ and $\phi_2 + \phi_1 < 1$ and $\phi_2 - \phi_1 < 1$.

2 Task 2: ACF, PACF of AR models

The series that will be used for task 2 are the following:

1. AR(1) $x_t = 0.8 * x_{t-1} + \varepsilon_t$
2. AR(1) $x_t = -0.8 * x_{t-1} + \varepsilon_t$
3. AR(2) $x_t = x_{t-1} - 0.5 * x_{t-2} + \varepsilon_t$
4. AR(2) $x_t = -x_{t-1} - 0.5 * x_{t-2} + \varepsilon_t$

As we can see on figure 3, the line plots show that all series are stationary. This can also be seen by analysing the auto-regressive coefficients.

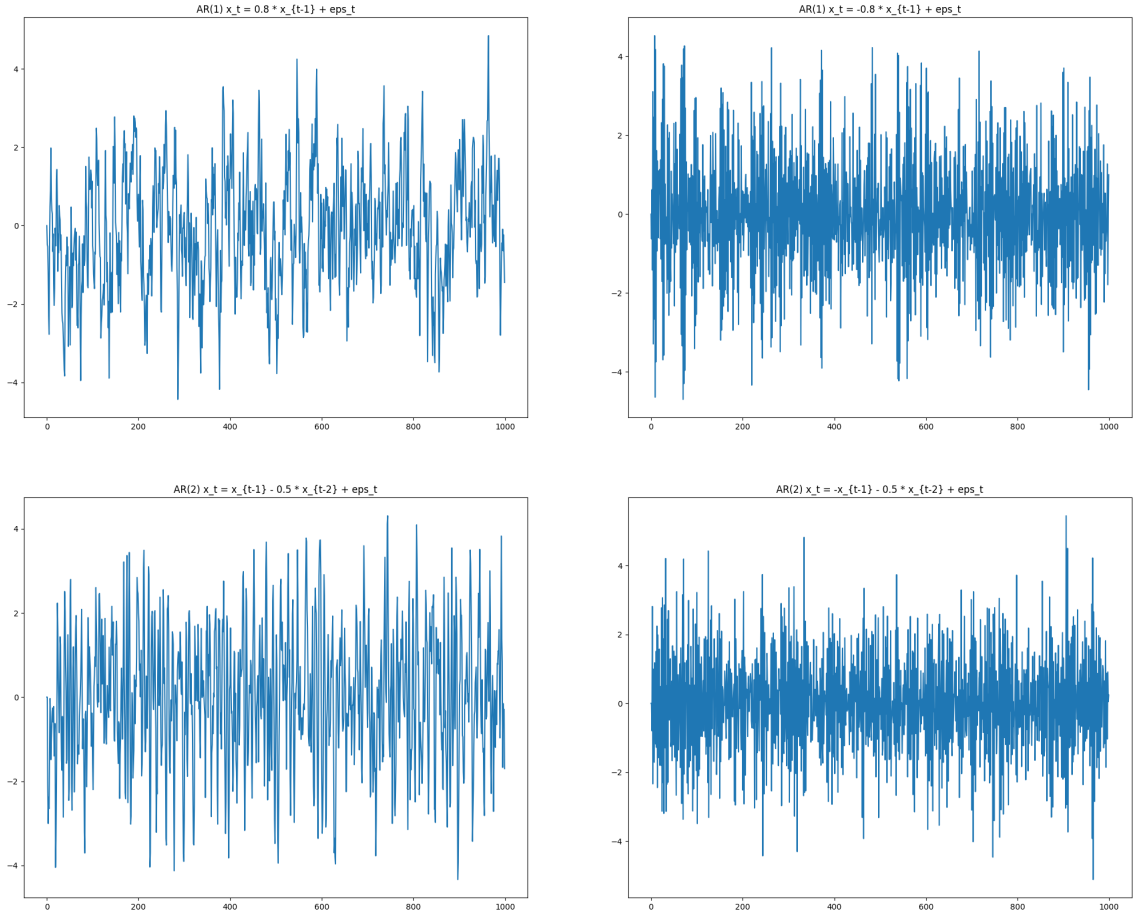

Figure 3: Line plots of the series

As shown by figure 4, there are some outliers in each series. This can be seen in the box-plots, with the points at the extremities of the graph.

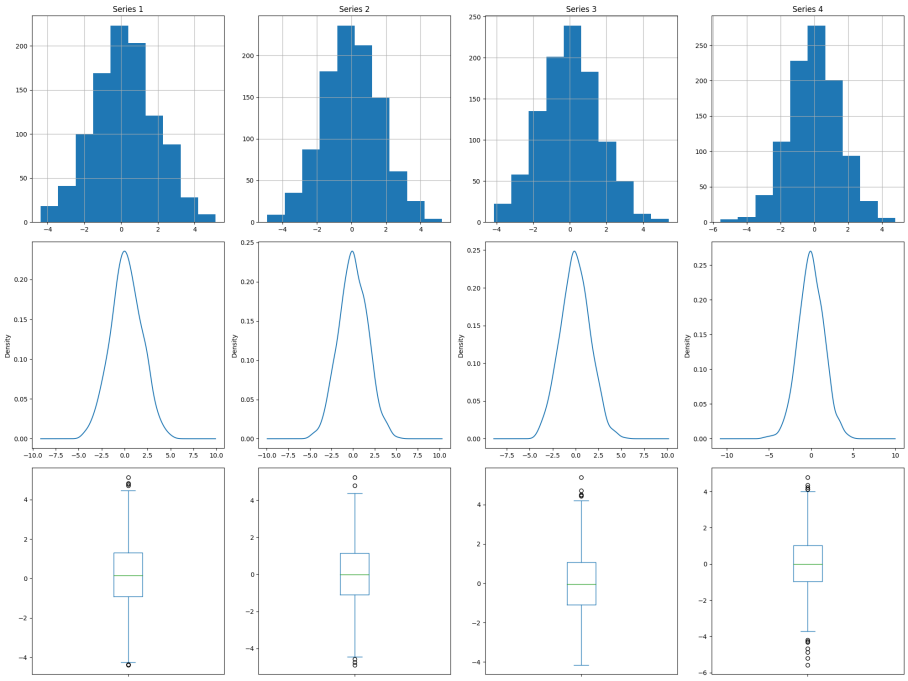

Figure 4: Density plot of task 2's series

There is a sort of linear correlation in the lag1 graphs of figure 5.

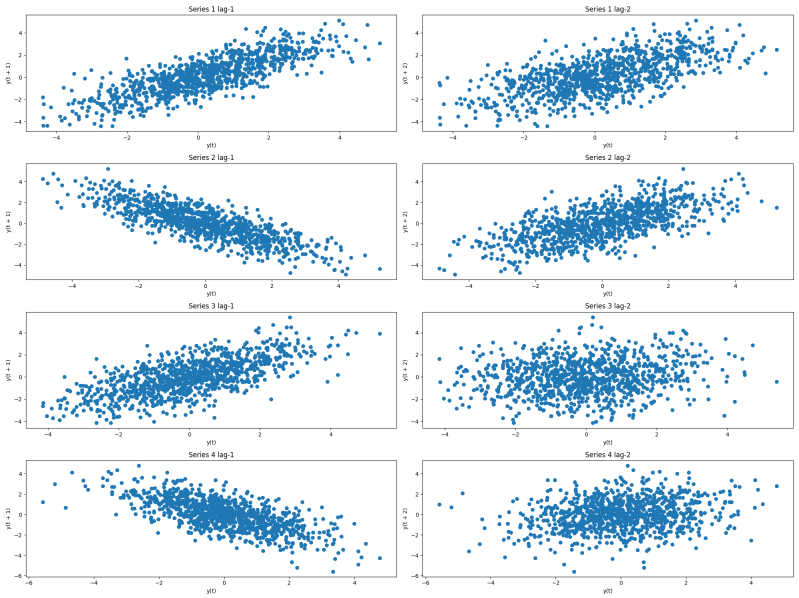

Figure 5: Lag plots for task 2's series

What characteristics can you observe from the ACF graphs of the AR(p) models?
As seen on the left of figure 6, we can not observe any specific characteristics on the ACF graph of AR(p) models except that it tails off gradually.

What characteristics can you observe from the PACF graphs of the AR(p) models?
As seen on the right of figure 6, for the two first series, which are of order $p = 1$,

the value goes to zero after the first lag.

For the 3rd and 4th series, the value of the PACF goes to zero after the second lag, which also proves that the order of these series is $p = 2$.

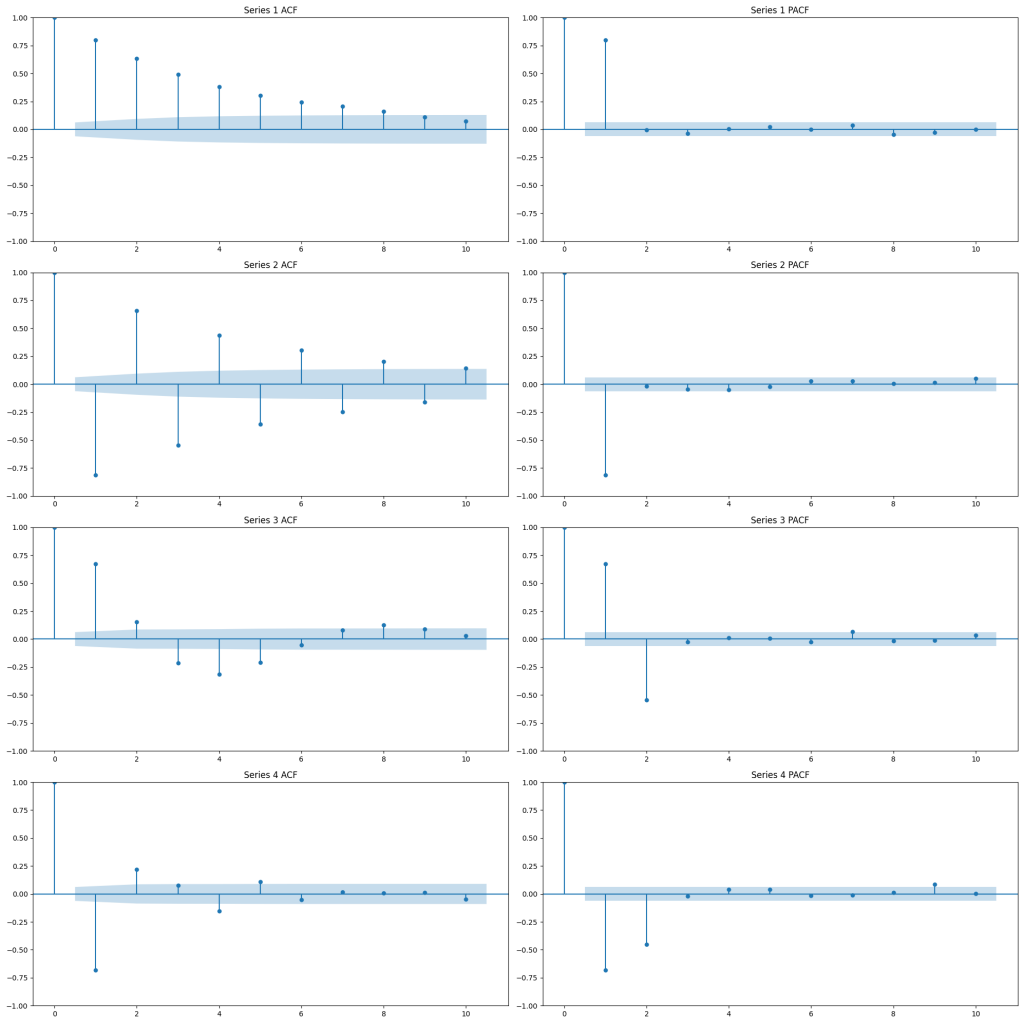

Figure 6: ACF and PACF for task 2's series

3 Task 3: Invertibility, ACF, PACF of MA models

The series that will be studied in this task are based on the following models:

1. MA(1) $x_t = \varepsilon_t - 2 * \varepsilon_{t-1}$
2. MA(2) $x_t = \varepsilon_t - 0.5 * \varepsilon_{t-1}$
3. MA(3) $x_t = \varepsilon_t - \frac{4}{5} * \varepsilon_{t-1} + \frac{16}{25} * \varepsilon_{t-2}$
4. MA(4) $x_t = \varepsilon_t - \frac{5}{4} * \varepsilon_{t-1} + \frac{25}{16} * \varepsilon_{t-2}$

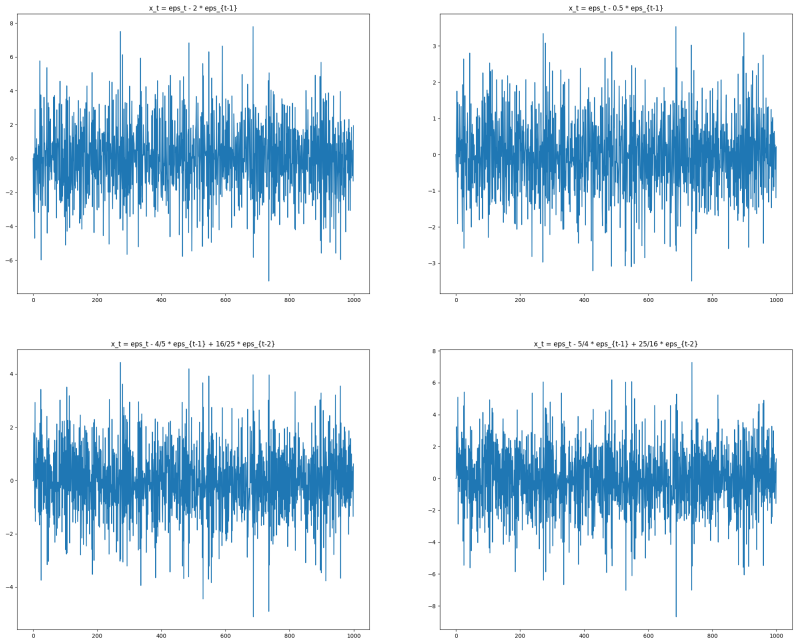

Figure 7: Line plots of task 3 series

It is not easy to look at invertibility just from the line plots. Therefore it is better to look at it by looking at the model auto-regressives parameters.

This is done by using the premade function `isinvertible` fromt `statsmodel` library. It says that only model 2 and 3 are invertible.

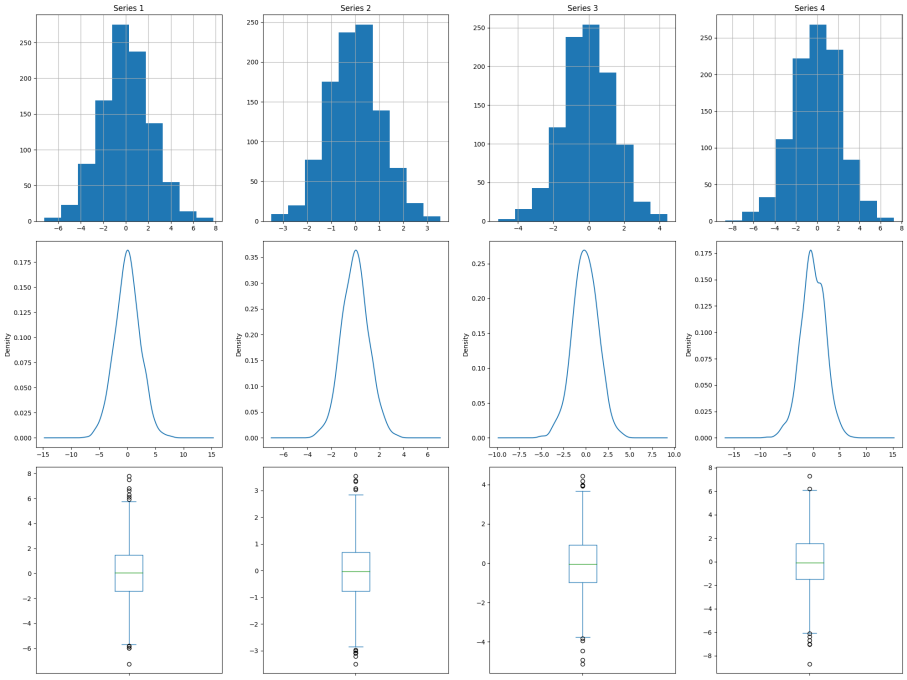

Figure 8: Histogram, density plot and box plot of task 3's series

As we can see on 8, there are some outliers. We can spot them on the box plot. This is because of the normal distribution, sometimes there can be outliers.

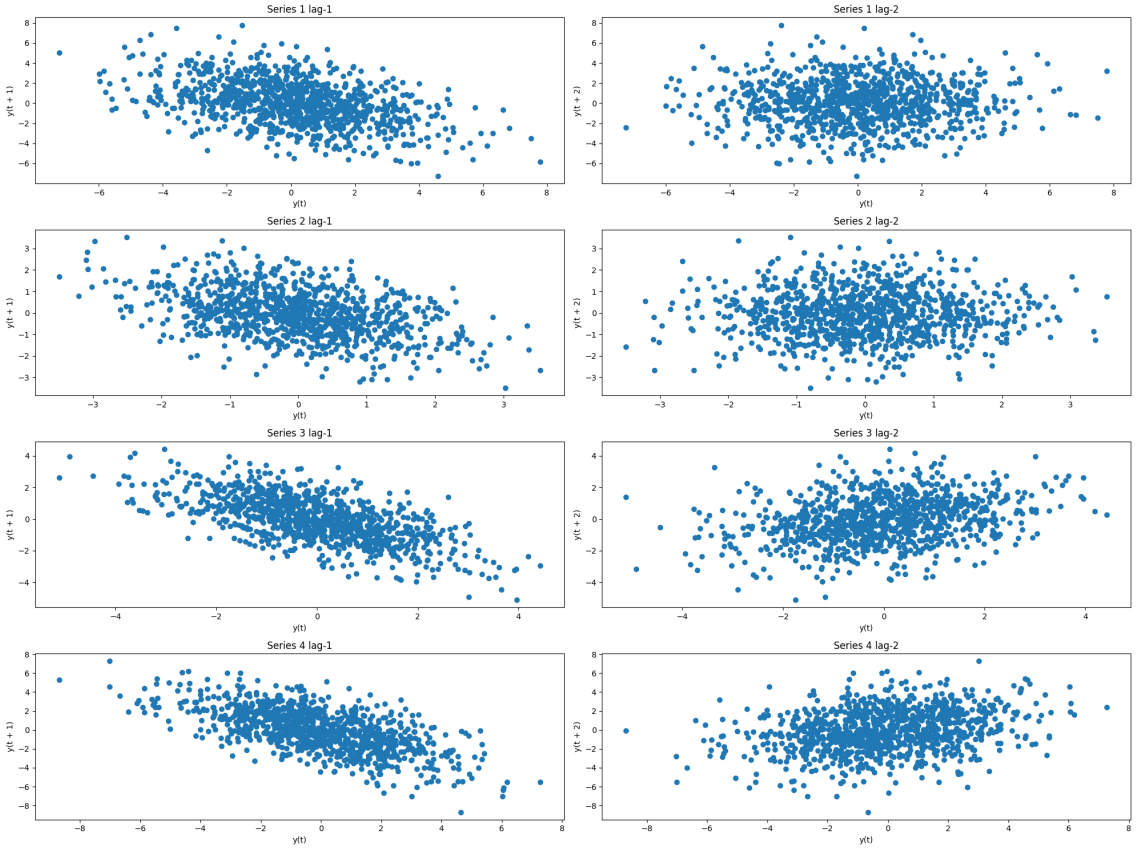

Figure 9: Lag-1 plot (left) and Lag-2 plots (right) of task 3 series

From the lag-1 and lag-2 plot on figure 9, there is no clear pattern of auto-correlation.

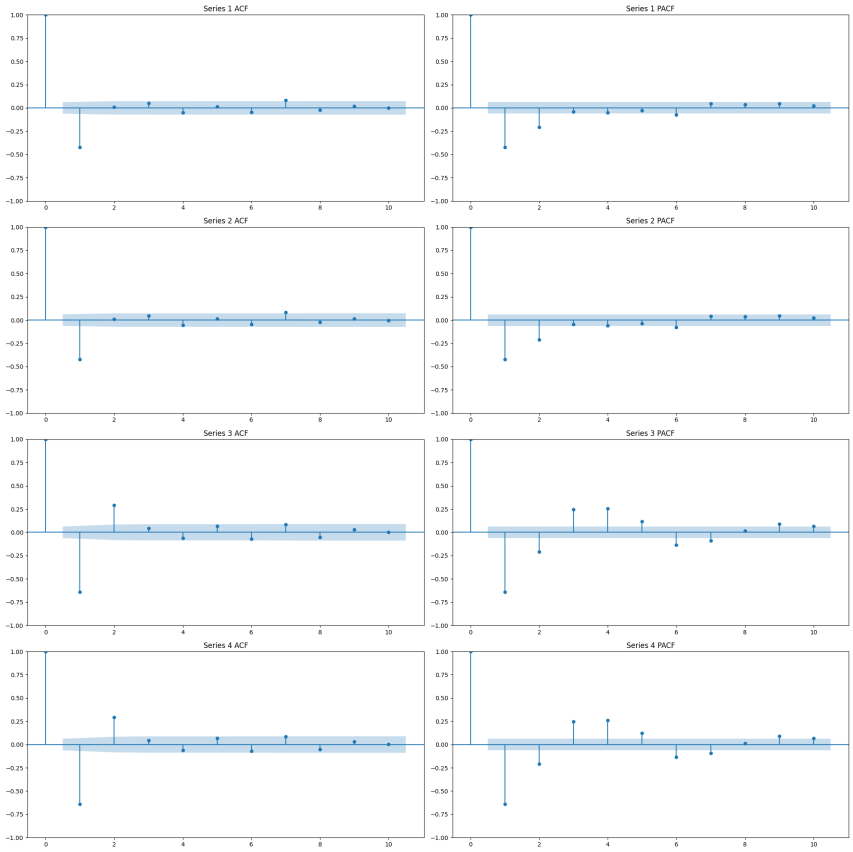

Figure 10: Task 3 ACF (left) and PACF (right)

Are all the MA models invertible? If not, which ones are invertible and which ones are not invertible?
No, some MA models are not invertible. A MA model is invertible if it ensures at one-to-one mapping to an ACF graph.

What characteristics can you observe from the ACF graphs of the MA(q) models?
We can observe the order q of the MA model. It can be seen for series 1 and 2 (resp. 3 and 4) that after $q = 1$ (resp. $q = 2$), all values of the ACF graph tends to zero.

What characteristics can you observe from the PACF graphs of the MA(q) models?
That it tails off gradually.

To have an MA(q) model be invertible, is there any requirement on the auto-regressive coefficients? List the constraints for MA(1) and MA(2) models.
In order for the MA model to be invertible we have the following constraints.

For **MA(1)** models of the form

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1}$$

We must have $-1 < \theta_1 < 1$.

For **MA(2)** models:

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}$$

We must have $-1 < \theta_2 < 1$ and $\theta_2 + \theta_1 > -1$ and $\theta_1 - \theta_2 < 1$.

4 Task 4 : Stationarity, ACF and PACF of ARMA models

The models studied in this task are the following:

1. AR(1) $x_t = 0.8 * x_{t-1} + \varepsilon_t$
2. MA(1) $y_t = \varepsilon_t + 0.7 * \varepsilon_{t-1}$
3. ARMA(1, 1) $z_t = 0.8 * z_{t-1} + \varepsilon_t + 0.7 * \varepsilon_{t-1}$
4. ARMA(1, 1) $z_t = -0.8 * z_{t-1} + \varepsilon_t - 0.7 * \varepsilon_{t-1}$

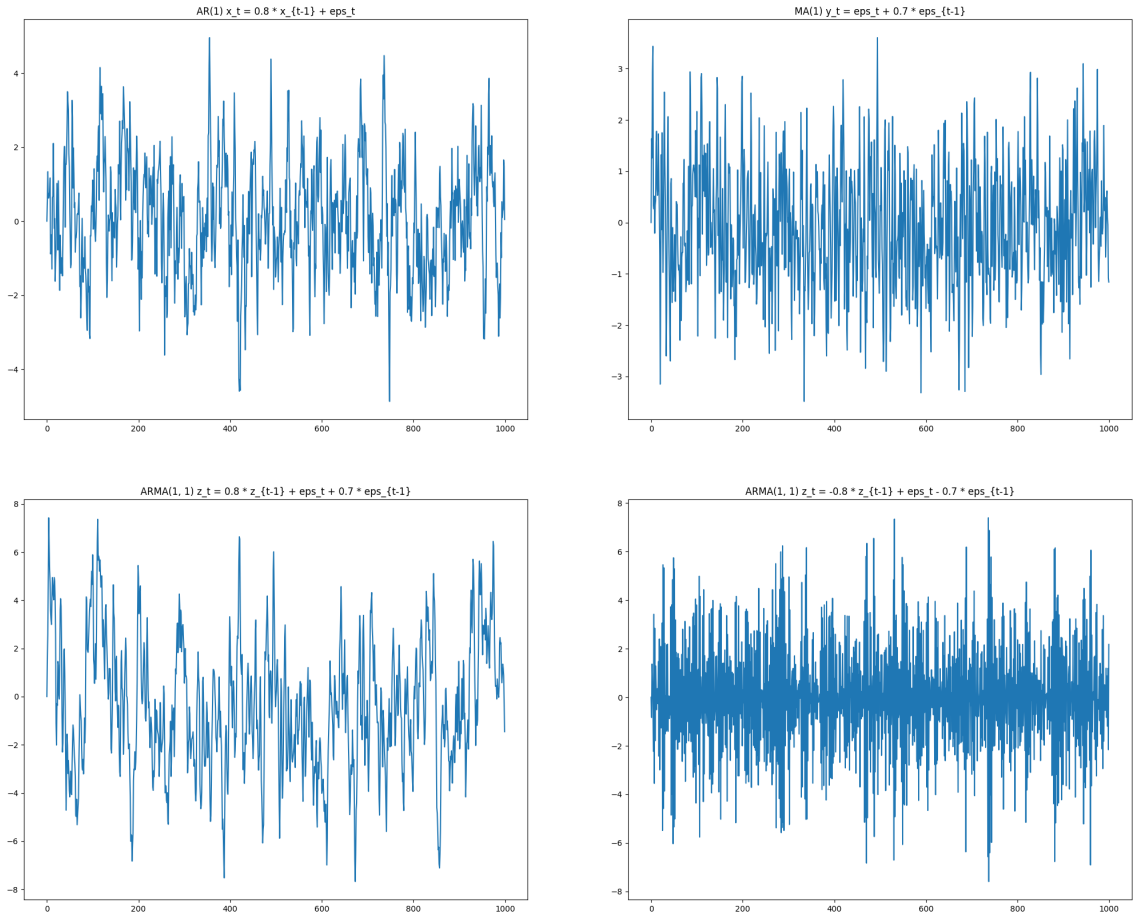

Figure 11: Task 4 series line plots

Visually, all series look stationary, except maybe the 3rd one (bottom left of figure 11).

As seen in figure 12, all series are stationary according to the ADF test since all p-values are smaller than 0.05.

```
ADF Test: AR(1) p-value: 1.509720236924587e-20
ADF Test: MA(1) p-value: 2.235695728185697e-15
ADF Test: ARMA(1, 1) p-value: 2.1361323655308316e-07
ADF Test: ARMA(1, 1) p-value: 0.0
```

Figure 12: Result of the ADF test of task 4 series

By calling the function `isstationary` and `isinvertible` of statsmodel, we obtain that all series are both stationary and invertible.

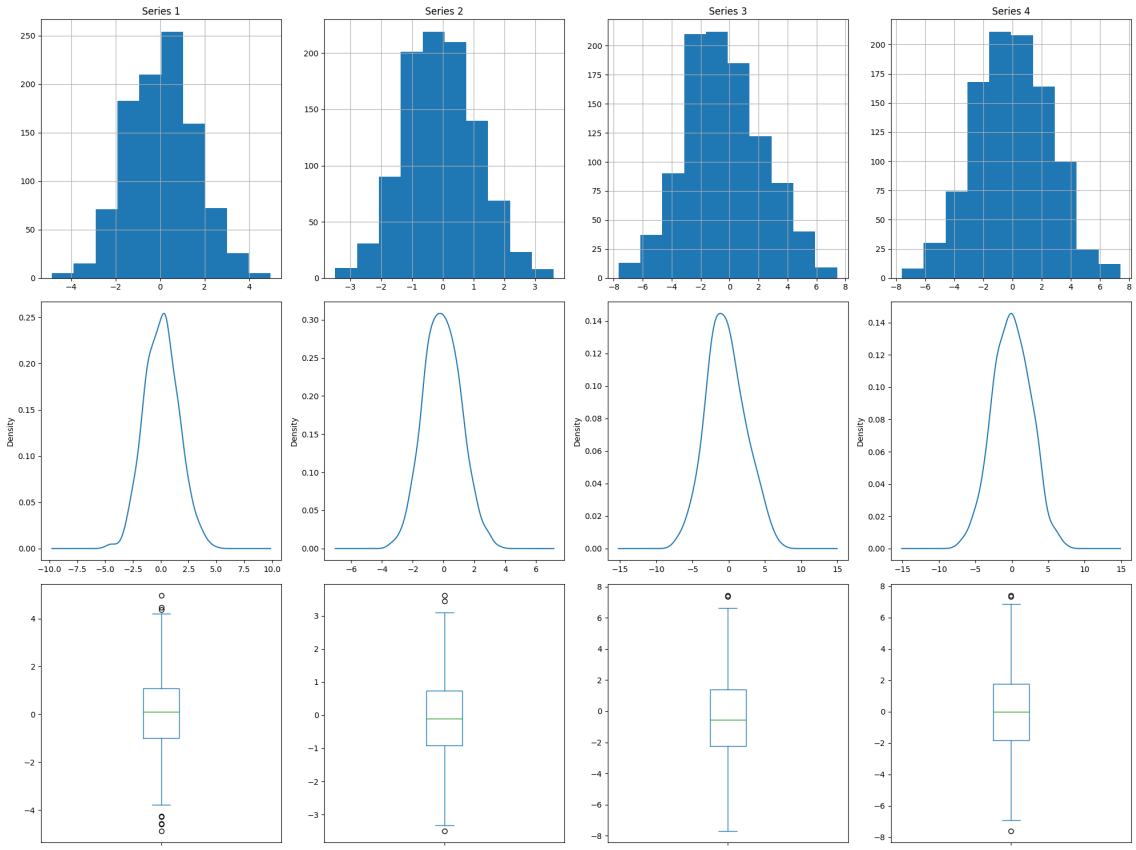

Figure 13: Task 4 histograms, density plots and Box-plots

As seen on figure 13, there are some outliers but less in the ARMA model.

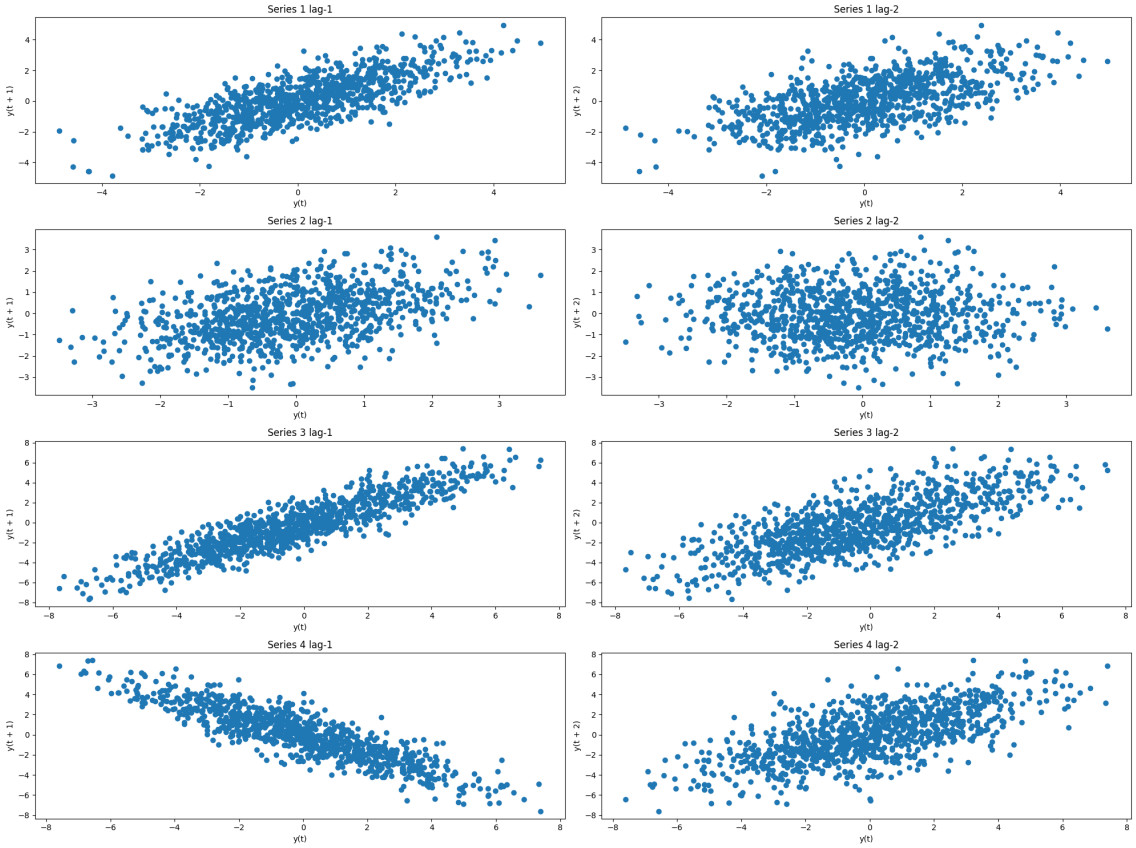

Figure 14: Lag-1 (left) and Lag-2 (right) plot of task 4 series

For the 3rd and 4th series (ARMA(1,1)) models, we can see a correlation (decreasing line) but covered with a lot of noise on Lag-1 plot on figure 14.

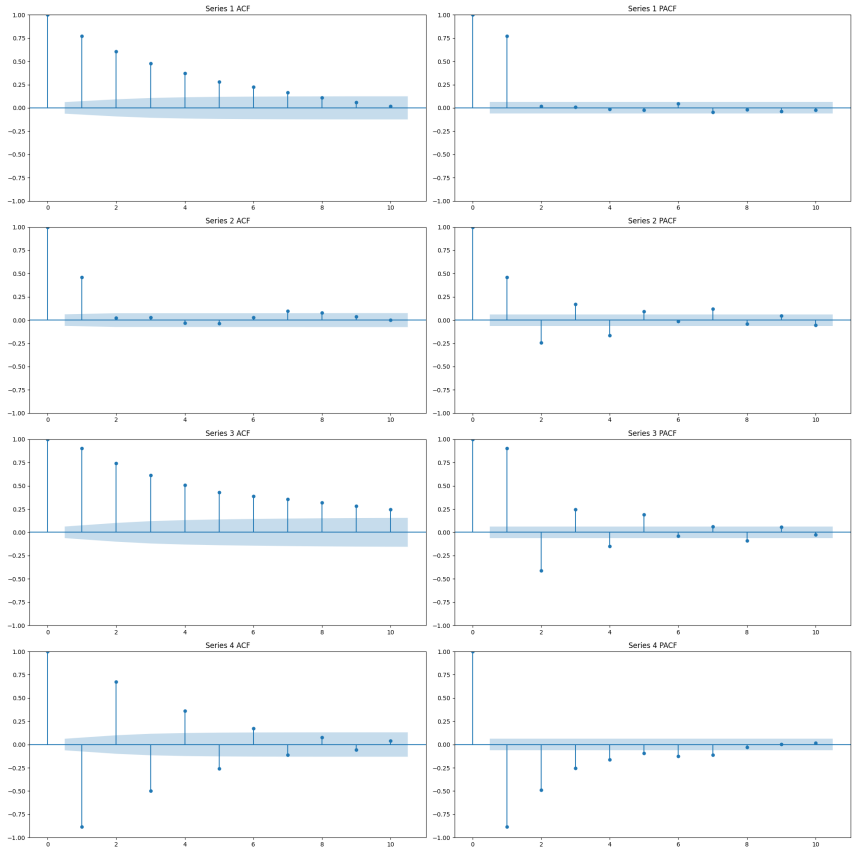

Figure 15: Task 4's series ACF (left) and PACF (right)

What characteristics can you observe from the ACF, PACF graphs of the AR(p) model?
For the AR model (1st line in figure 15), the ACF does not give any particular information except that it tails off gradually but the PACF gives the order of the AR model p. Here it is 1 as, after the second point the values tend to 0.

What characteristics can you observe from the ACF, PACF graphs of the MA(q) model?
For the MA model (1st line in figure 15), the PACF does not give any particular information except that it tails off gradually but the ACF gives the order of the AR model q. Here it is 1 as, after the second point the values tend to 0.

What characteristics can you observe from the ACF, PACF graphs of the ARMA(p, q) model? Both ACF and PACF tails off gradually.

Model	ACF	PACF
AR(p)	Tails off gradually	Cuts off after p lags
MA(q)	Cuts off after q lags	Tails off gradually
ARMA(p, q)	Tails off gradually	Tails off gradually

Table 1: The ACF and PACF characteristics of ARMA models

5 Task 5: ARIMA modeling and prediction

The plots needed in task5 are shown in Figure.16

- **Draw a line plot for the time series data. Do you observe any trend, season in the data?**
  Both trend and season characters are found in the line plot. It's obvious that it has an increased trend and there is a periodic-like curve in the plot.

- **Draw histogram, density plot, heat map, and box plot for the time series data. Are there any outliers? Why?**
  Yes. Because the data contains white noise which interrupts the whole line plot and the data's value count.

- **Draw lag-1 and lag-2 plots for the time series data. Do you observe any auto-correlation from the lag plots?**
  Yes. From two lag plots (Blue: lag-1, Red: lag-2), it's found that they have many overlapping points which indicates that auto-correlation exists in the data, the data has some linear relationships between it and lags.

- **Is the series random? How do you check it? Do the three methods (line plot, lag-1 plot, and Ljung-Box test) give the same results?**
  No. The three method (line plot, lag-1 plot, and Ljung-Box test) all gives this conclusion. For line plot, there is an obvious trend and season exist. For lag-1 plot, the points form an ellipse across the block, which shows that the data isn't a white noise, neither a random data. For Ljung-Box test, the result is shown in Table.2.The result shows all p-value are less than 0.05 which shows the data shouldn't be random.

lb_stat	lb_pvalue
65.914271	4.709690e-16
114.149261	1.632311e-25
151.955524	9.974838e-33
189.507131	6.764012e-40
220.874422	9.652744e-46
253.060924	9.093751e-52
276.676346	5.742962e-56
301.607312	1.873081e-60
324.274857	1.832207e-64
350.976775	2.472290e-69

Table 2:

- **Is the series stationary? Try with visual inspection and ADF test. Do they give the same results?**
  No. Both inspection and ADF test give the same result that the data is non-stationary. The ADF test gives the p-value that is $0.32 > 0.05$, which proves the non-stationarity.

- **If the series is not stationary, how do you make it stationary? If you use the differencing operation, how do you decide a proper order of**

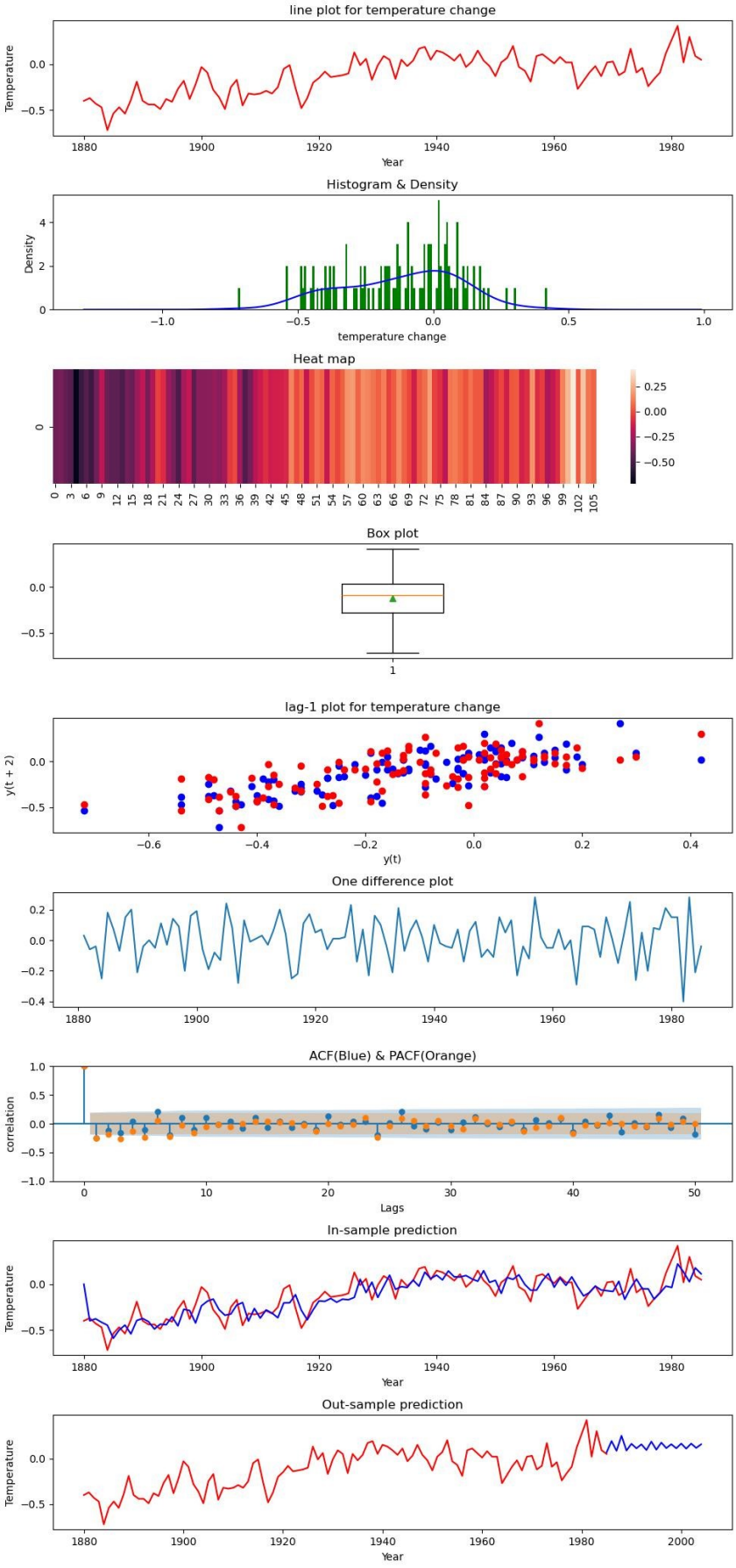

Figure 16: Figures for task 5

differencing without under-differencing/over-differencing?
We can use the differencing operation. The differenced series should first be evaluated for its trend. If the trend is more like a linear trend, then probably only one operation is enough, i.e., one order. The order of differencing operation depends on whether the differenced series is stationary.

- **What (p, d, q) values do you use? How do you determine them?**
  We use (7, 1, 1) as the test values. For d, it can be determined by the times of differencing operations it takes until the time series is stationary. For p and q, they can be derived by inspections on ACF and PACF plot or AIC/BIC tests.

- **After model fitting, is the reminder series (in-sample prediction) considered to be white noise?**
  Yes. By doing Ljung-Box test on the residual series, it can be found that the residual series' p-value is larger than 0.05, thus the residual series is white noise.

- **What is the MSE of the fitted model for the data?**
  MSE means Mean Square Error, which symbolizes the difference scale between in-sample prediction and original series. The formula is

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  In this case, we get a MSE of 0.016.

- **For out-of-sample prediction, do the predicted values (10 steps) reflect the trend and fluctuation of the series?**
  The prediction result is shown in Figure.16. Yes, because we use the original time series for training, so the forcasting series should repeat the trend and fluctuation.

6 Task6: Series transformation

- **What are the common transformation techniques applicable to turn a non-stationary series into a stationary series?**
  The commonest way is differencing operation. Besides we can also use Box-Cox transformation or seasonal differencing operation.

- **What is the Box-Cox transform? Give its definition and explain its generality.**
  Box-Cox transformation is used to transform a non-normal distributed series to a normal distributed series. The formula is given below

$$y(\lambda) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ ln(y), & \lambda = 0 \end{cases}$$

  where λ is the coefficient depends which kind of Box-Cox transformation the series will take.

- **Can a differencing operation remove a linear trend? Give an example by generating a synthetic series, and draws the series before differencing and after differencing.**
  Yes. As shown in the Figure.17, the linear trend can be wiped out by differencing because the common differences between time series points are same.

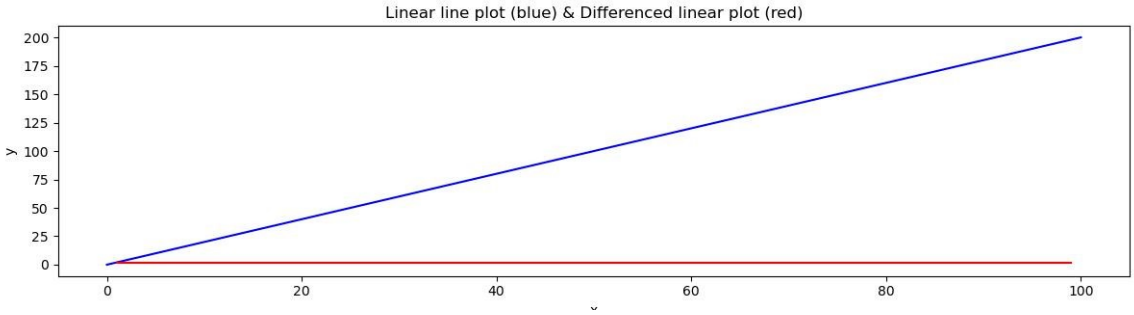

Figure 17:

- **Can a differencing operation remove an exponential trend? If not, which additional transformation needs to be taken? Give an example by generating a synthetic series, and plots the series before transformation and after transformation, before differencing and after differencing.**
  No. In this case the differences become larger when the time grows, so differencing operation is not capable. To make it works again, a log operation should be applied to the series first to turn the series into linear one, and then carry on differencing operation, as shown in Figure.18

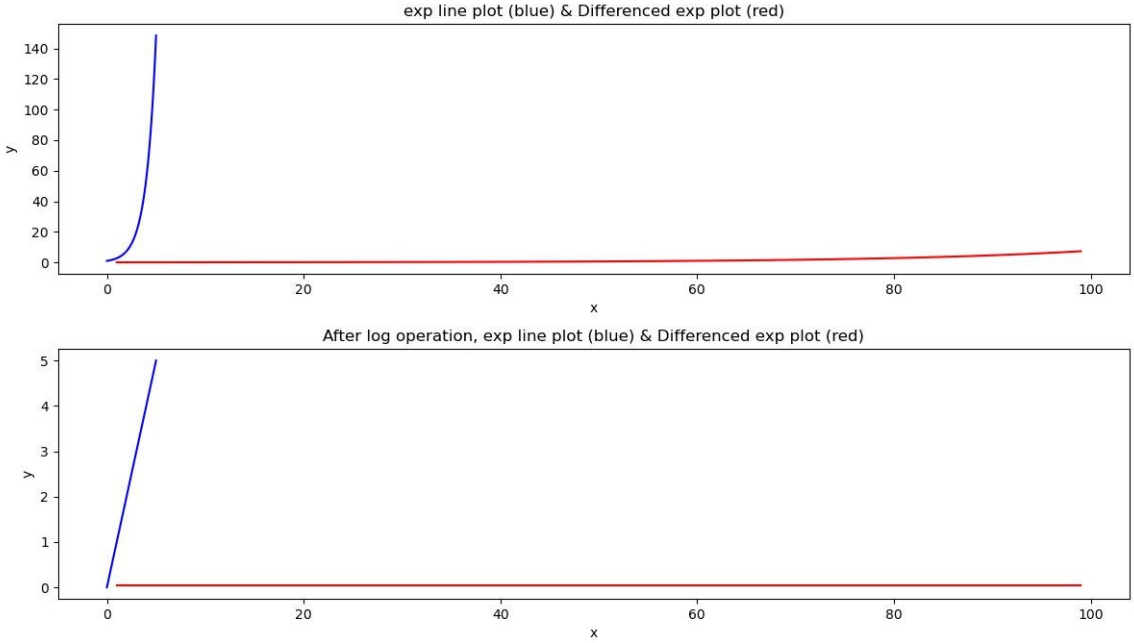

Figure 18:

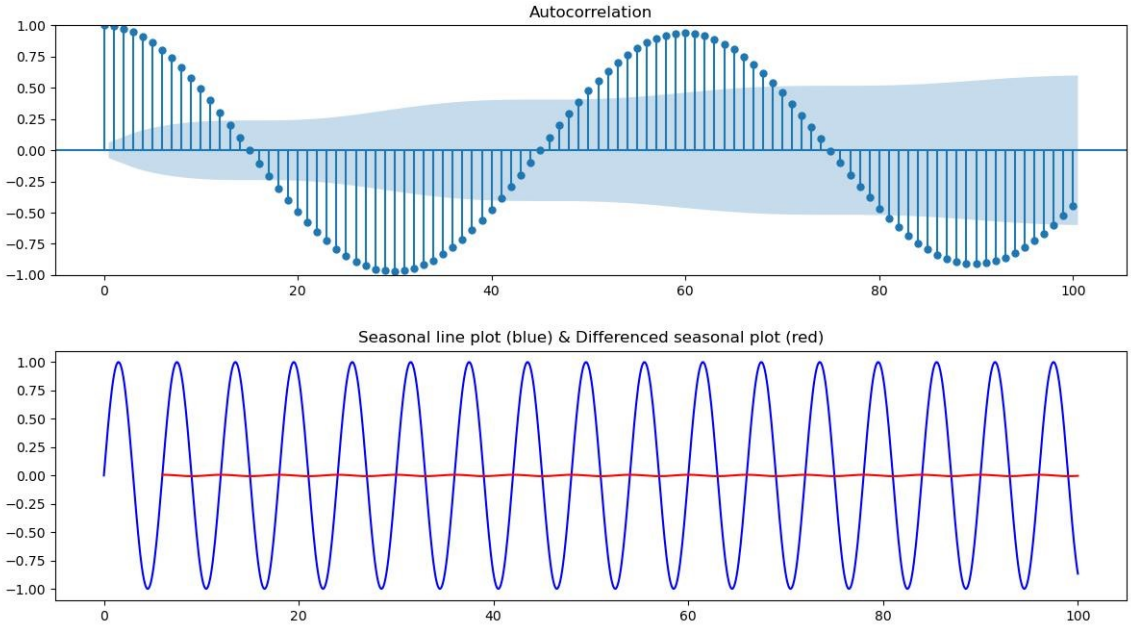

Figure 19:

- **Can a differencing operation remove a seasonal (periodic) trend? If yes, under what condition? Give an example by generating a synthetic series, draw its ACF, and draws the series before differencing and after differencing with different step length.**
  Yes. For seasonal trend, the times of differencing operation should equal to the cycle of seasonal series. Like Figure.19 , we know from ACF graph that the cycle is 60, so a 60 times differencing operation is applied to the series, and it turns out that the series can become stationary.