VT2023: IL2233 Lab 1

Time Series Visualization and Feature Extraction

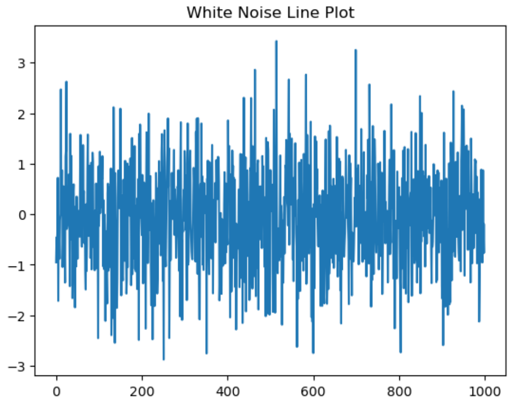
Conglei Xiang   Yuqi Sun

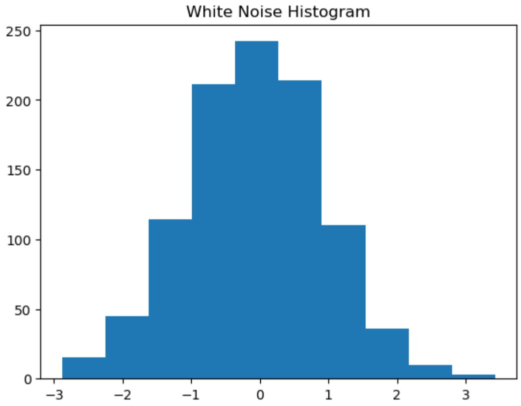
April 6, 2023

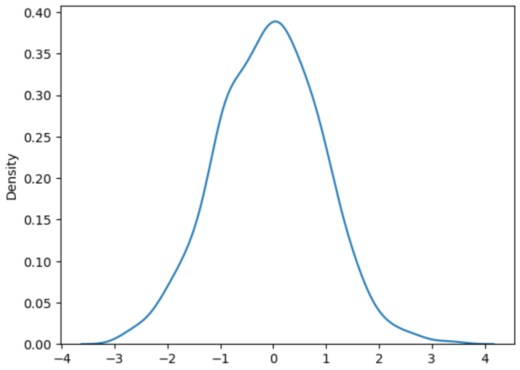
Task 1

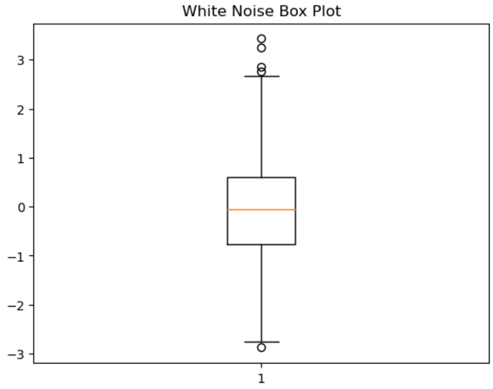
1.1 White noise series

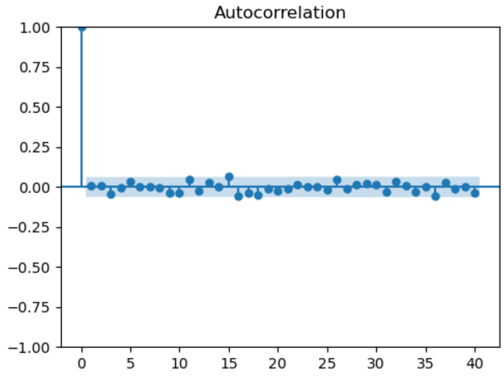
1. Generate a white noise series with N data points (e.g. N can be 100, 1000, 5000, or 10000). Then find its actual mean, standard deviation, and draw its line plot, histogram, density plot, box plot, lag-1 plot, ACF and PACF graphs (lags up to 40).

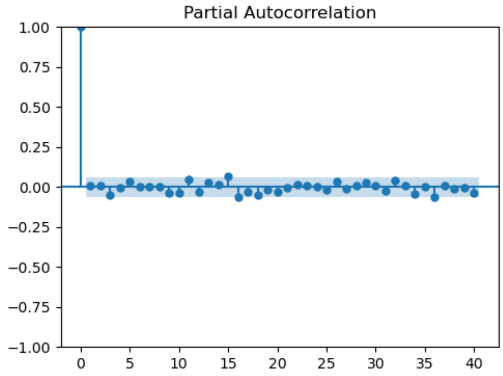


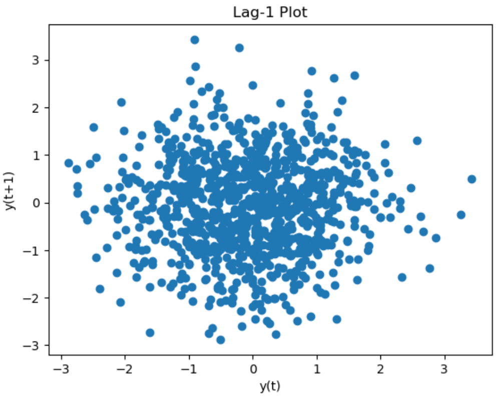












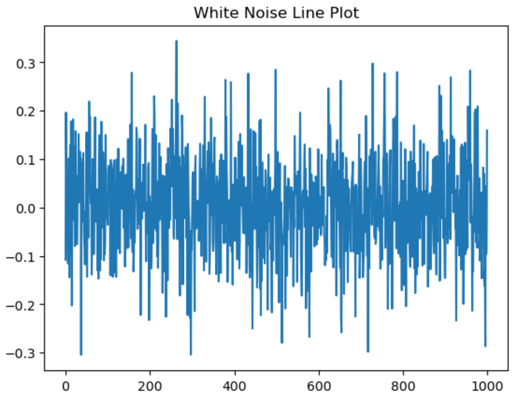
In this case, we generated the white noise series with N=1000.

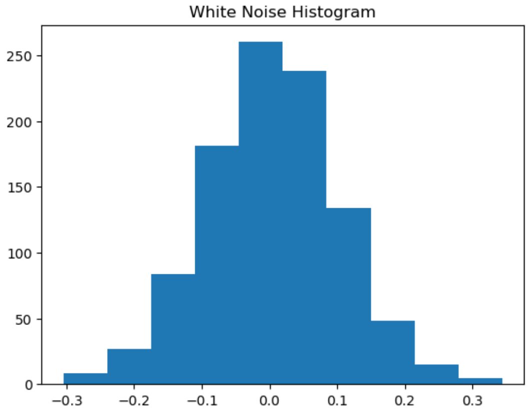
Actual Mean: -0.07120429237582042

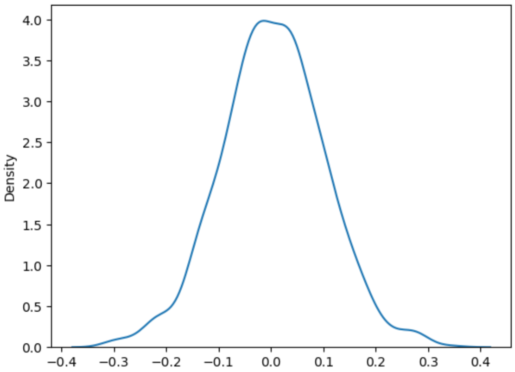
Actual Standard Deviation: 0.9922827480232803

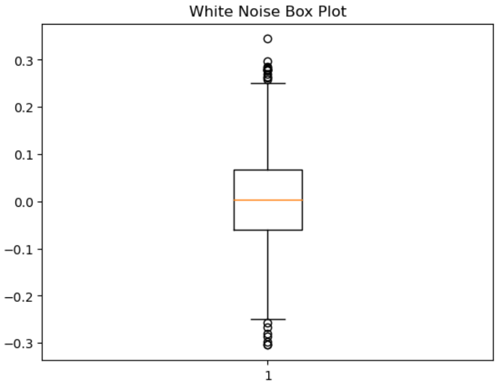
It can be observed that the mean value is close to 0, and the standard deviation is close to 1.

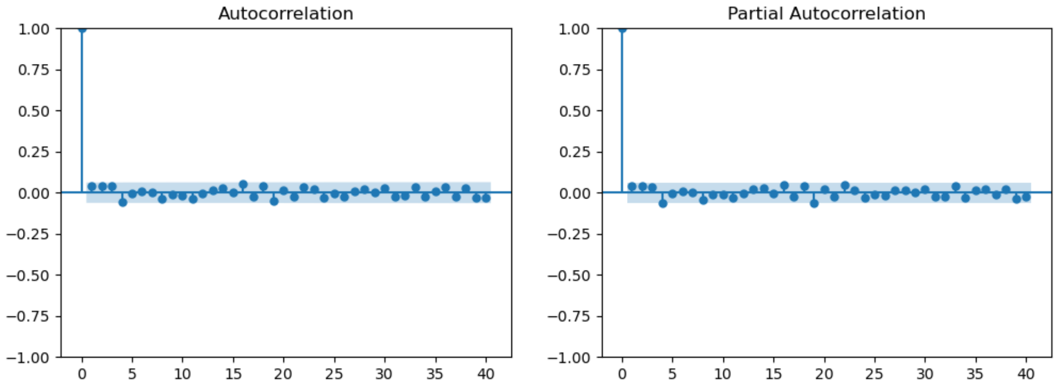
2. Generate 100 random series with length 1000 data points, then use the average values at each time to produce an average value series. Then repeat the same process above.

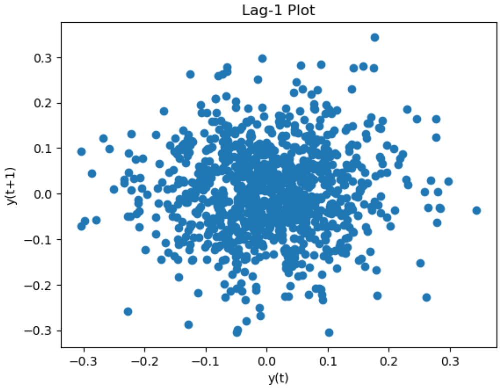












Actual Mean: 0.004170518169858113

Actual Standard Deviation: 0.009877098562222227

It can be observed that the mean value is close to 0, and the standard deviation is close to 0.01.

3. Perform randomness test on the white noise series using the Ljung-Box test.

lb_stat=5.167016

lb_pvalue=0.879745

From the result, it can be observed that lb_pvalue > 0.05 (default threshold), then accept the Null hypothesis that the series is independent, meaning that the series is random.

4. Perform stationarity test on the white noise series using the Augmented Dickey-Fuller (ADF) test.

ADF Statistic: -9.942463

p-value: 0.000000

Critical Values:

1%: -3.498

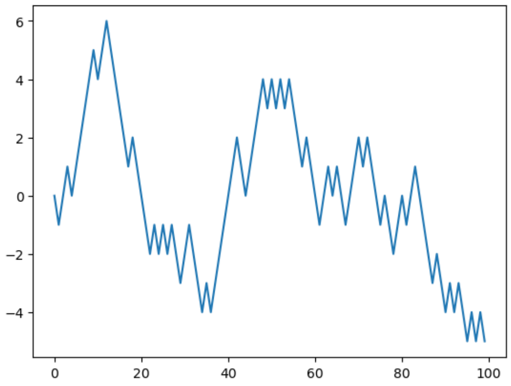
5%: -2.891

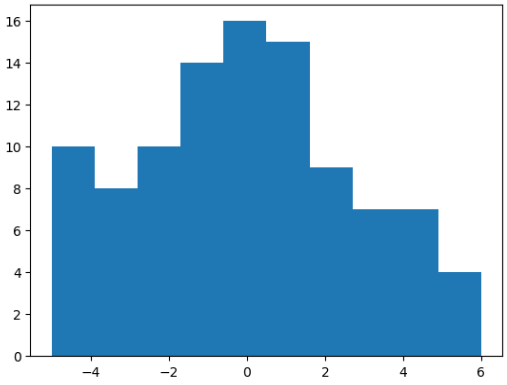
10%: -2.583

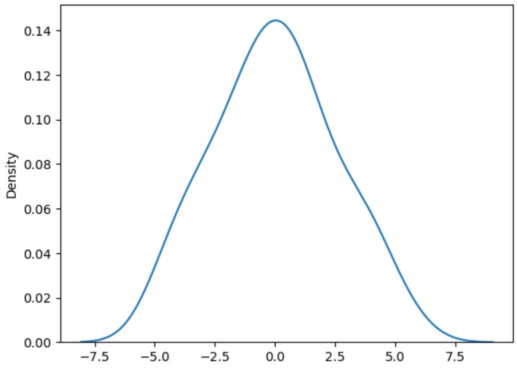
The p-value is much less than the significance level of 0.05 and hence we can reject the null hypothesis and judge that the series is stationary.

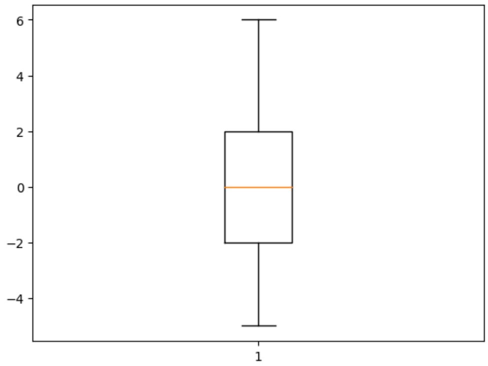
1.2 Random-walk series

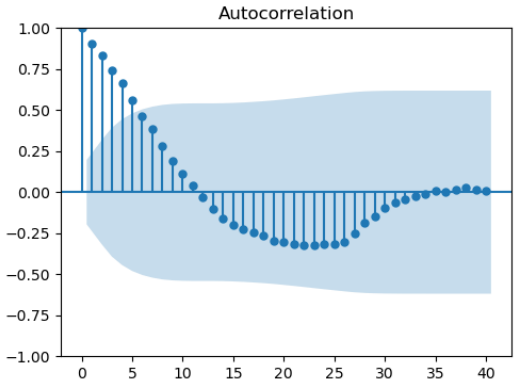
1, 2. Realize a simple random walk series with N data points (e.g. N can be 100, 1000, 5000, or 10000) starting from an initial value of 0 ($y_0 = 0$), and $x_t = +1, -1$. Then find its actual mean, standard deviation, and draw its line plot, histogram, density plot, box plot, lag-1 plot, ACF and PACF graphs (lags up to 40).

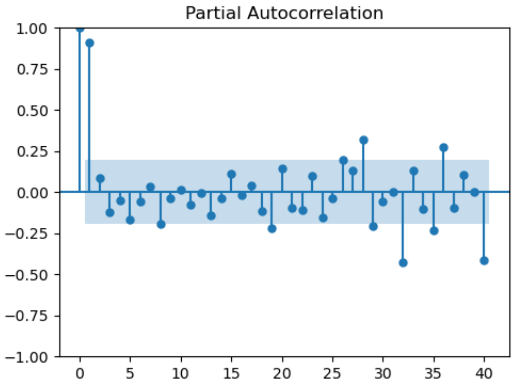


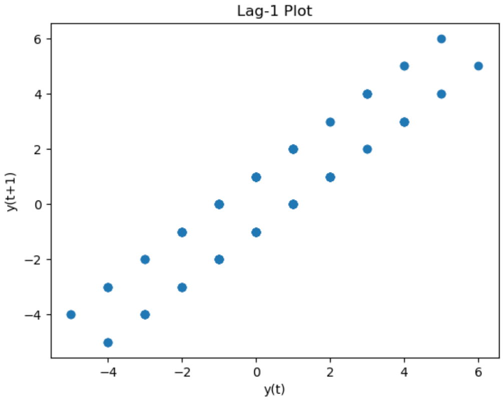












In this case, we generated the series with N=100.

Actual Mean: 0.02

Actual Standard Deviation: 2.5622748730869063

3. Perform randomness test on the random walk series using the Ljung-Box test.

lb_stat = 84.282798

lb_pvalue = 4.288301e-20

From the result, it can be observed that lb_pvalue << 0.05 (default threshold), then we reject the null hypothesis and accept the series is not random.

4. Perform stationarity test on the random walk series using the Augmented Dickey-Fuller (ADF) test.

ADF Statistic: -1.469495

p-value: 0.548544

Critical Values:

1%: -3.498

5%: -2.891

10%: -2.583

From the result, it can be observed that p-value > 0.05 , which means the p-value is larger than the significance level, then we accept the null hypothesis, and consider that the series is not stationary.

The series will become stationary by generating a one order difference.

Questions and Answers

- What methods can be used to check if a series is random? Describe both visualization and statistic test methods.

  A: As for visualization methods, we can use Line plot, Autocorrelation plot, Spectral density plot, and so on. And for statistic test methods, we can use ACF test to identify the presence of any autocorrelation in the series. Ljung-Box test also can be used to test the null hypothesis that the series is a random sample from a normal distribution with zero mean and constant variance.
- What methods can be used to check if a series is stationary? Describe both visualization and statistic test methods.

  A: To check if a series is stationary, we can use visualization methods like line plots, as well as statistical test methods such as ADF tests.
- Why is white noise important for time-series prediction?

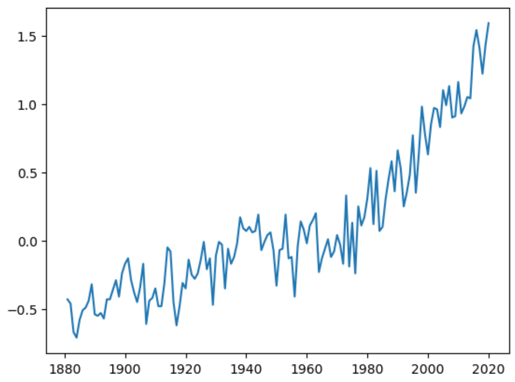
  A: White noise is a type of time series data that has random, independent, and identically distributed values with a mean of zero and constant variance. It is important for time series prediction because it serves as a baseline model, helps diagnose model performance, detects signals in the data, and measures irreducible error in prediction.
- What is the difference between a white noise series and a random walk series?

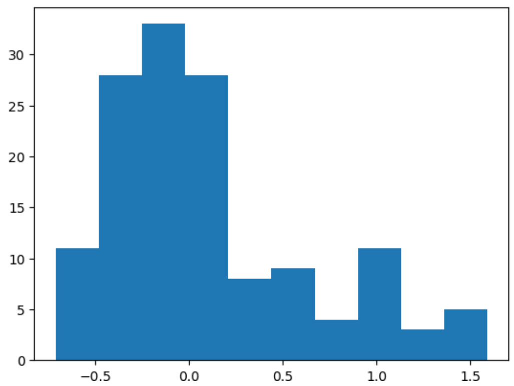
  A: The main difference between a white noise series and a random walk series is that the former has no underlying trend or pattern, while the latter has a strong trend that is dependent on previous values. White noise is often used as a baseline model for time series analysis, while random walk series require special attention and techniques such as differencing to make them stationary.
- Is it possible to change a random walk series into a series without correlation across its values? If so, how? Explain also why it can.

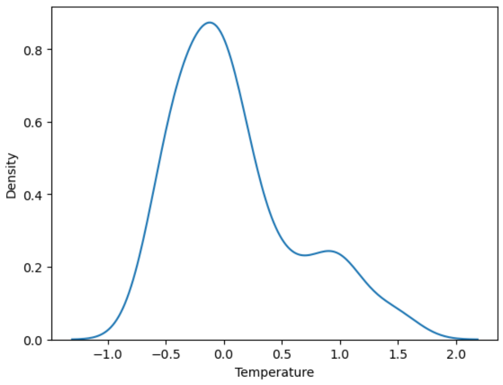
  A: Yes, it is possible to change a random walk series into a series without correlation across its values by differencing. Specifically, differencing a random walk series involves subtracting each observation from the previous observation, resulting in a new series where the trend has been removed.

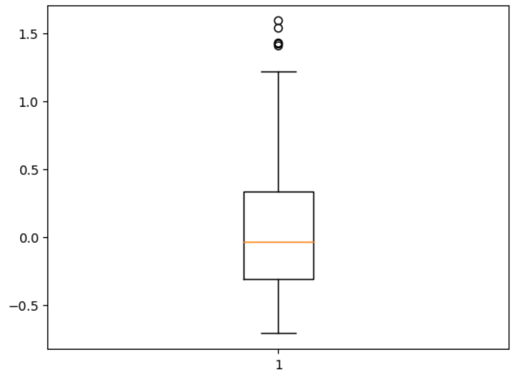
1.3 Global land temperature anomalies series

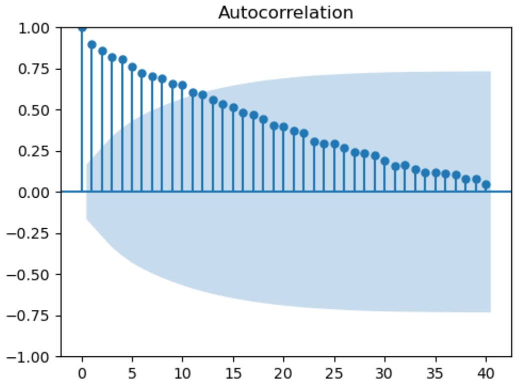
1. Use the global land temperature anomalies data to draw line plot, histogram, density plot, box-plot, heatmap, lag-1 plot, auto-correlation function (acf) and partial acf (pacf) graphs (lags up to 40).

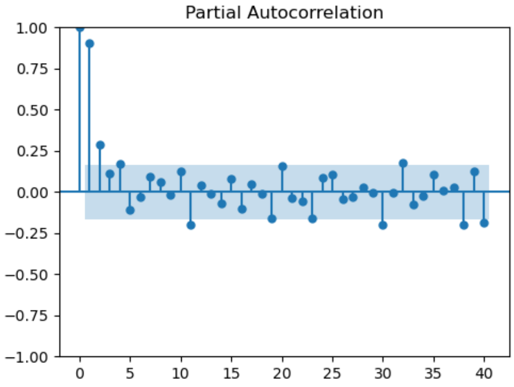


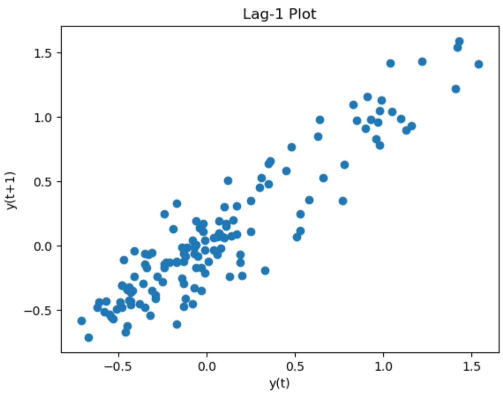


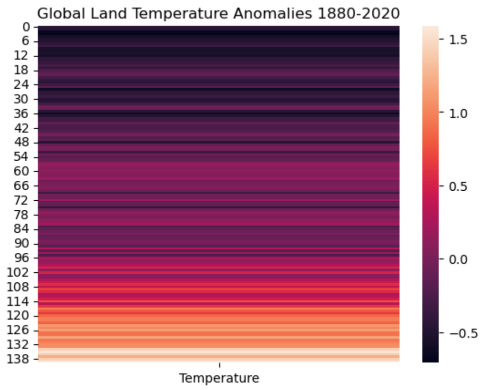




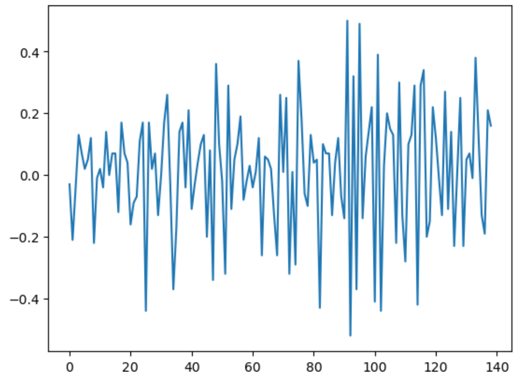


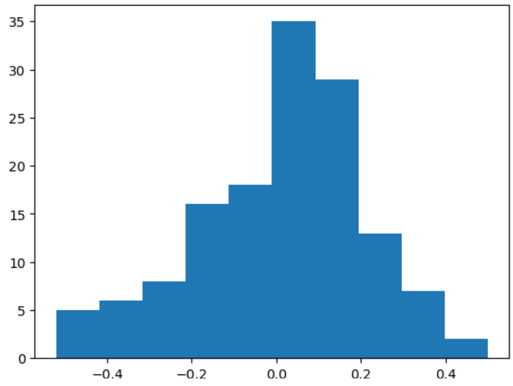


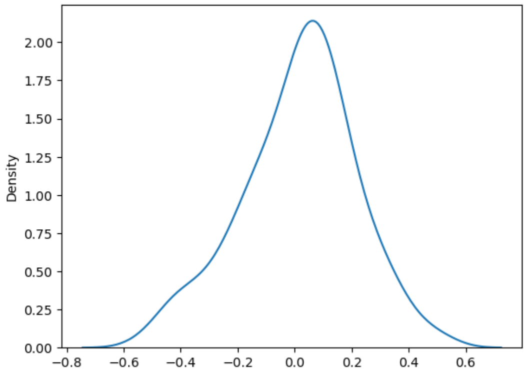


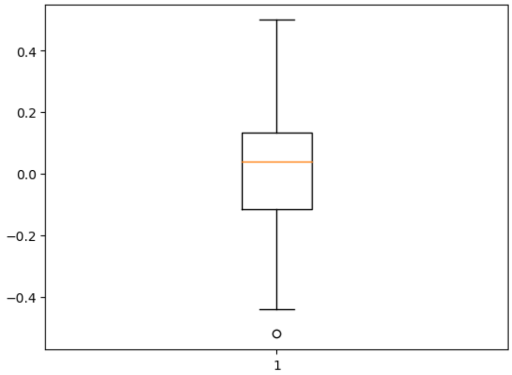


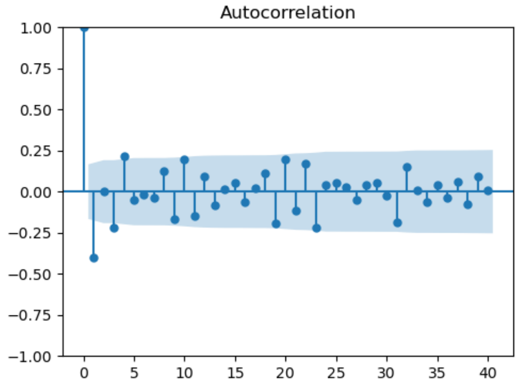
2. Take the first order difference of the temperature anomaly dataset. Draw line plot, histogram, density plot, box-plot, heatmap, lag-1 plot, acf and pacf graphs (lags up to 40).

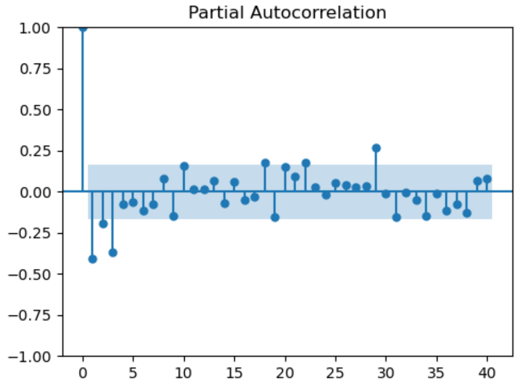


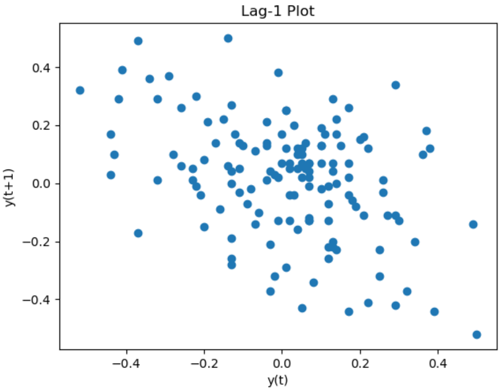












3. Test if the original and the differenced temperature anomaly series are random or not.

- Original:

  lb_stat = 116.259661

  lb_pvalue = 4.169678e-27

  lb_pvalue << 0.05 (default threshold), then reject the Null hypothesis and accept the alternative hypothesis that the series is dependent, meaning that the series is not random.

- First-order differencing:

  lb_stat = 23.320097

  lb_pvalue = 0.000001

  lb_pvalue << 0.05 (default threshold), then reject the Null hypothesis and accept the alternative hypothesis that the series is dependent, meaning that the series is not random.

4. Test if the original and the differenced temperature anomaly series are stationary or not.

- Original:

  ADF Statistic: 0.938293

  p-value: 0.993570

  Critical Values:

  1%: -3.479

  5%: -2.883

  10%: -2.578

  p-value > 0.05, so accept the null hypothesis, and consider that the series is not stationary.

- First-order differencing:

  ADF Statistic: -12.165503

  p-value: 0.000000

  Critical Values:

  1%: -3.479

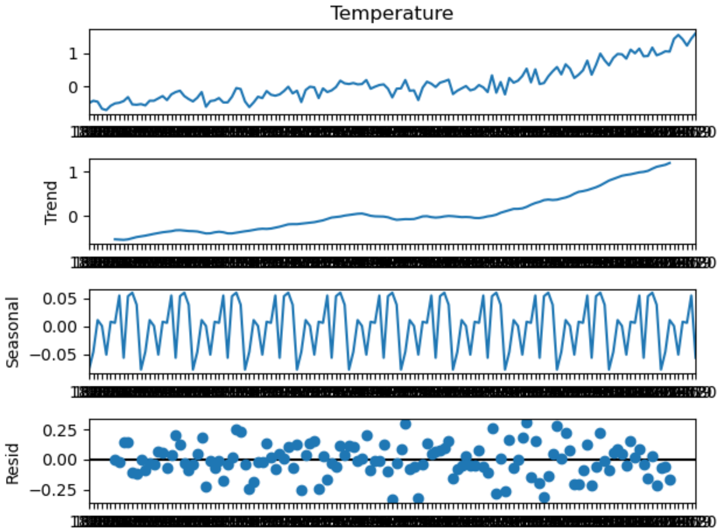
  5%: -2.883

  10%: -2.578

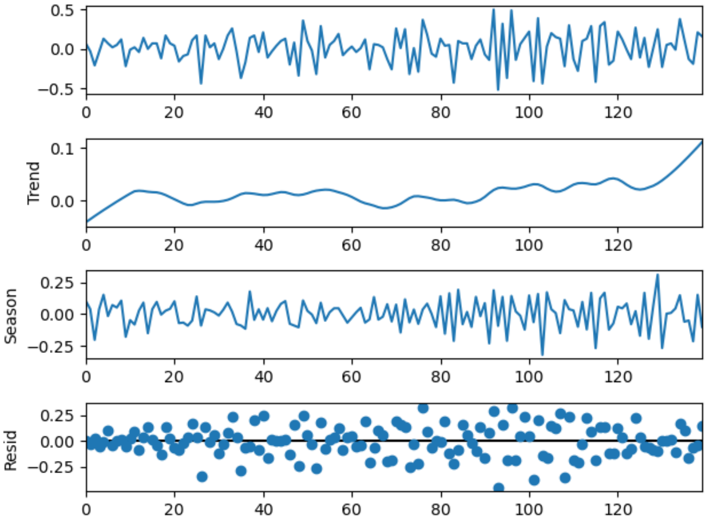
  p-value < 0.05, then reject the null hypothesis and consider that the series is stationary

5. Perform the classical decomposition and STL decomposition on the dataset.

- Original:



- First-order differencing:



Questions and answers:

- What is a stationary time series?

  A stationary time series is one whose properties do not depend on the time at which the series is observed. Stationary time series have constant statistical properties over time, such as constant mean and variance, and the covariance between the series and its lags does not depend on time.
- If a series is not stationary, is it possible to transform it into a stationary one? If so, give one technique to do it?

  Yes. One potential technique for transforming a non-stationary series into a stationary one is differencing. Differencing involves taking the difference between consecutive observations in the series. This can be done once or multiple times until the resulting series is stationary. Also it could be decomposed, and the seasonal component could be stationary. But this technique doesn't adapt to white noise series and for some data we need higher ordering to transform it into a stationary one.
- Is the global land temperature anomaly series stationary? Why or why not?

No, because p value is larger than 0.05, then we can accept the null hypothesis and judge that the series is not stationary

- Is the data set after the first-order difference stationary?

  Yes, because p value is less than 0.05, then we can reject the null hypothesis and judge that the series is stationary.

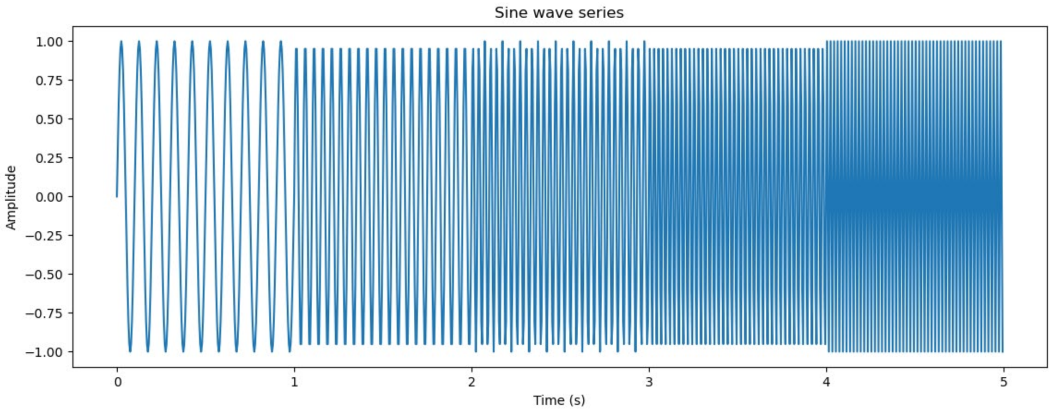
- Why is it useful to decompose a time series into a few components? What are the typical components in a time-series decomposition?

  Decomposition provides a useful abstract model for thinking about time series generally and for better understanding problems during time series analysis and forecasting. Trend-cycle component, seasonal component and remainder component are the typical components in a time-series decomposition.

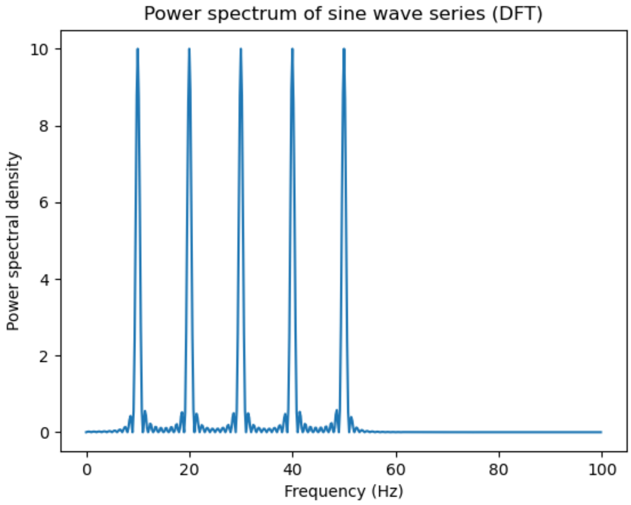
Task 2

2.1 Frequency components of a synthetic time-series signal

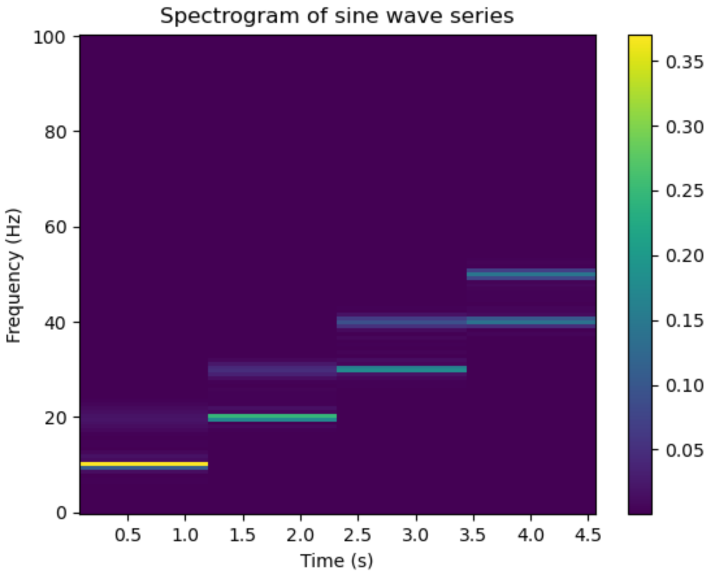
1. Draw a line plot of the series.



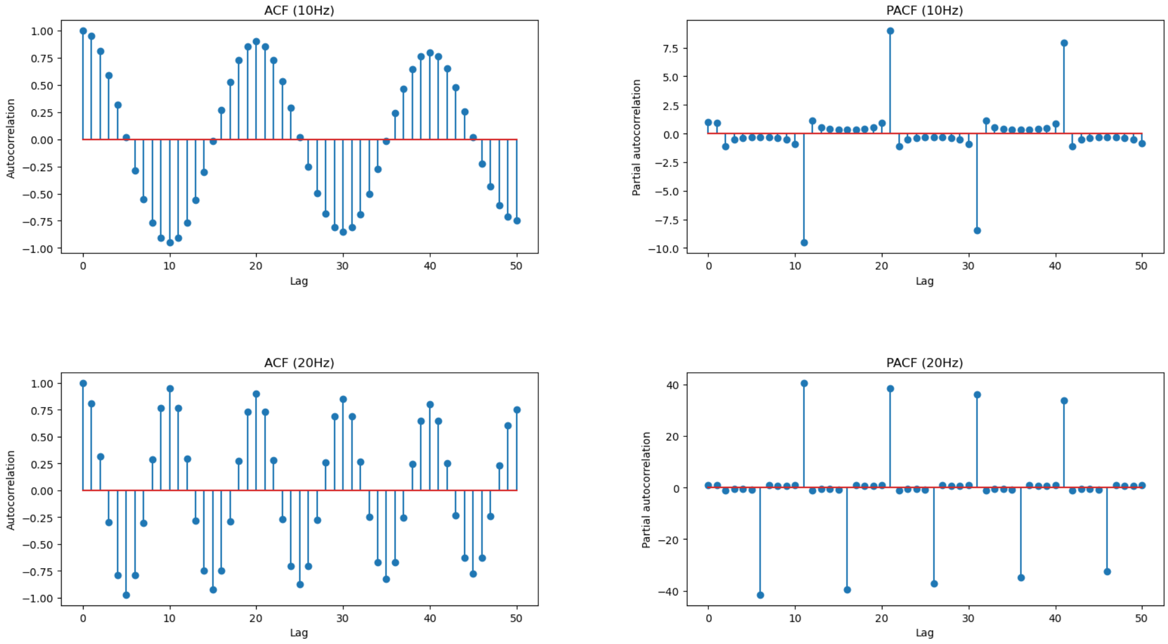
2. Draw power spectrum (power density graph) of the series.



3. Draw the spectrogram of the series.



4. Draw and compare the ACF and PACF graphs of the first one-second (frequency 10Hz) and the second one-second series (frequency 20Hz), with lags up to 50.



In the ACF plot for the 10 Hz signal, we see that the autocorrelation values drop off quickly and become close to zero after a lag of around 5, which suggests that there is not much correlation between samples beyond this point.

In the PACF plot for the 10 Hz signal, we see that the partial autocorrelation values are mostly small, except for the first two lags, which suggests that the signal may have some autoregressive structure.

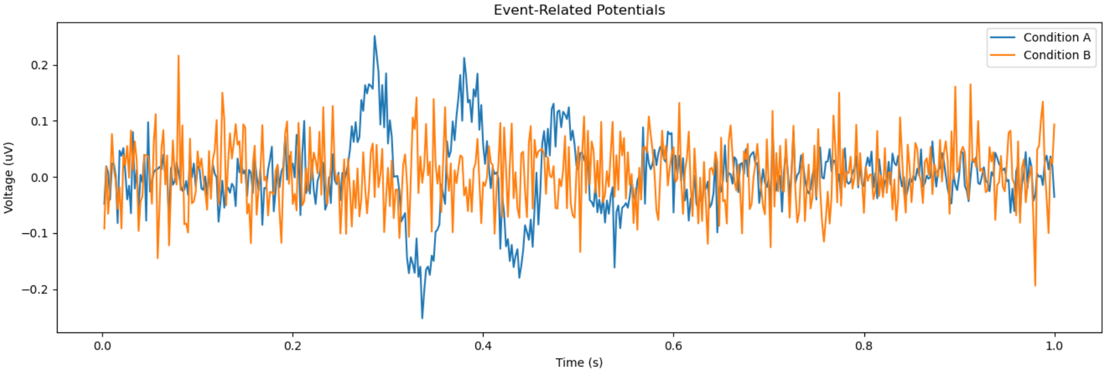
In contrast, in the ACF plot for the 20 Hz signal, we see that the autocorrelation values decay more slowly and remain relatively high for longer lags, which suggests that there may be more correlation between samples at larger lags.

In the PACF plot for the 20 Hz signal, we see that the partial autocorrelation values are mostly small, except for the first few lags, which again suggests that the signal may have some autoregressive structure, although not as pronounced as in the 10 Hz signal.

Overall, we can see that the 10 Hz signal exhibits more pronounced autocorrelation and partial autocorrelation structure compared to the 20 Hz signal, which may indicate that it has a more complex and structured underlying process.

2.2. Statistical features and discovery of event-related potential

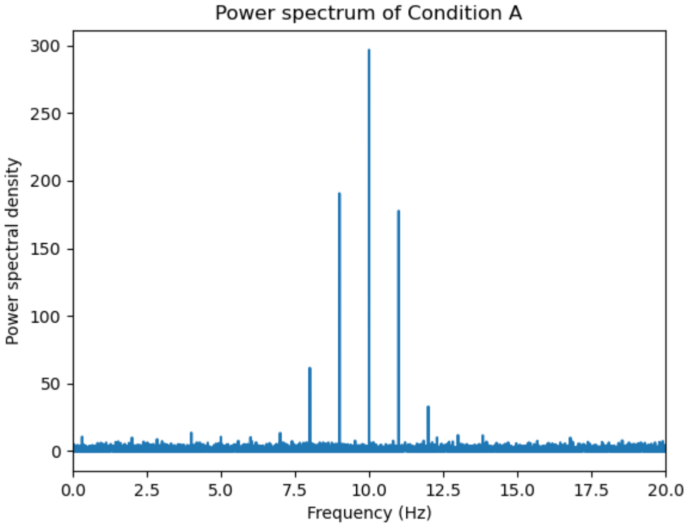
1. Visualize the response, i.e., ERP of the EEG, in the two conditions, A and B.



2. Find the brain activity frequency in the data of condition A (see below for condition A and B).

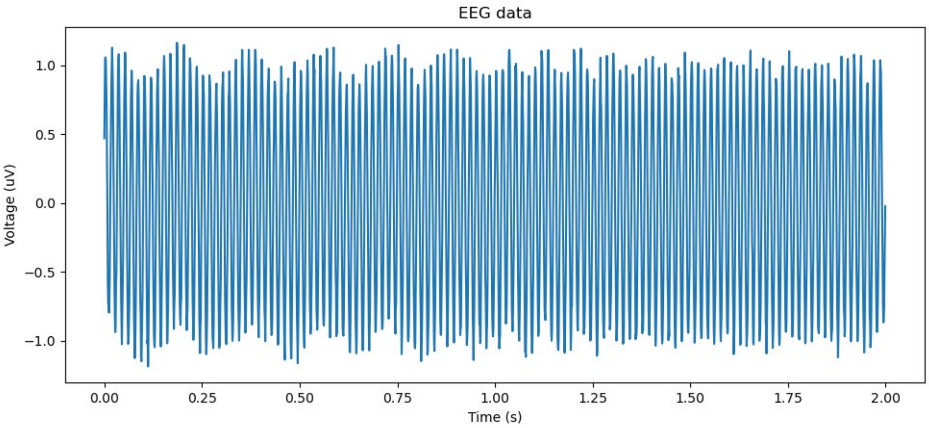
To find the brain activity frequency, we decide to use the power spectrum.

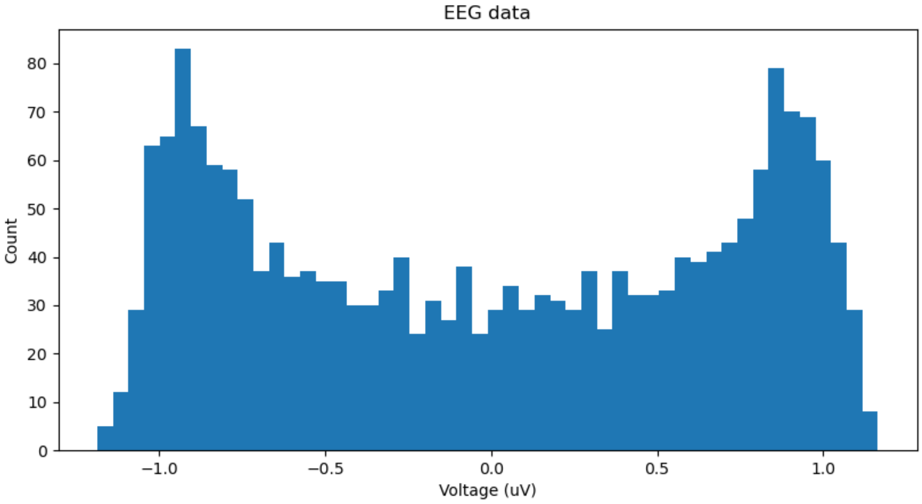
As the following picture, the most active frequency of brain is 10 Hz.

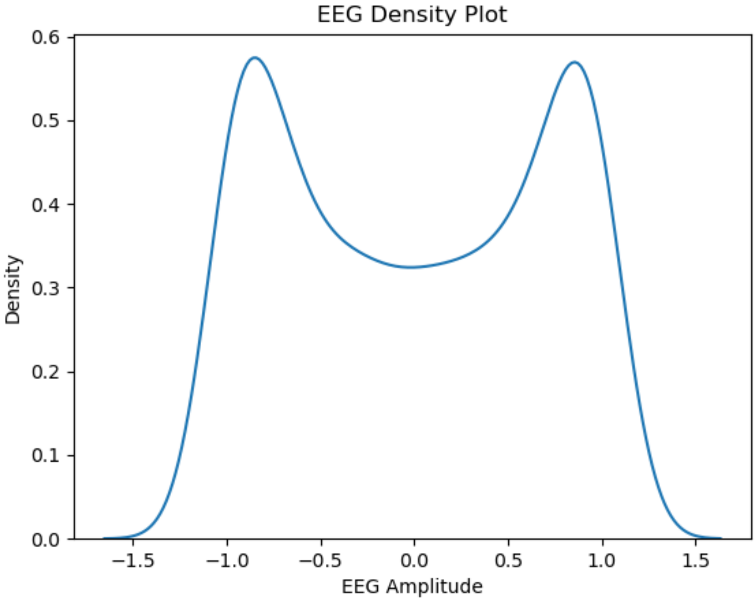


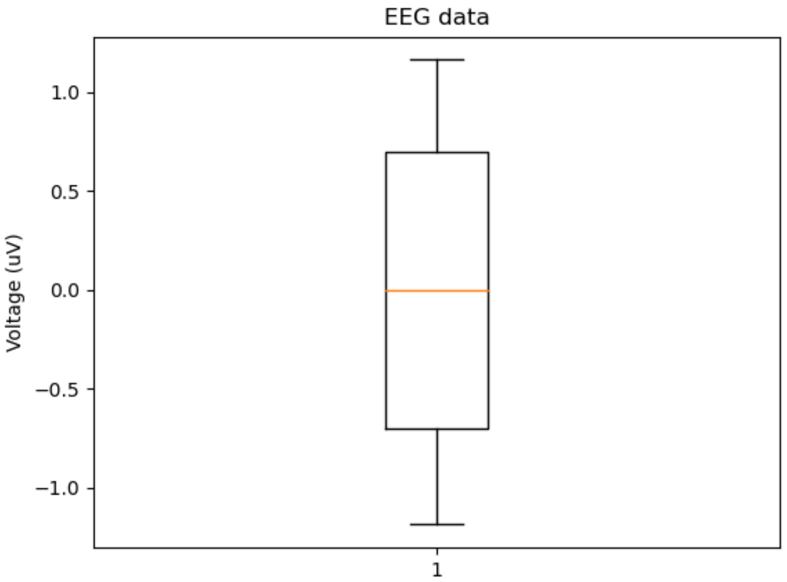
2.3 Features of observed rhythms in EEG

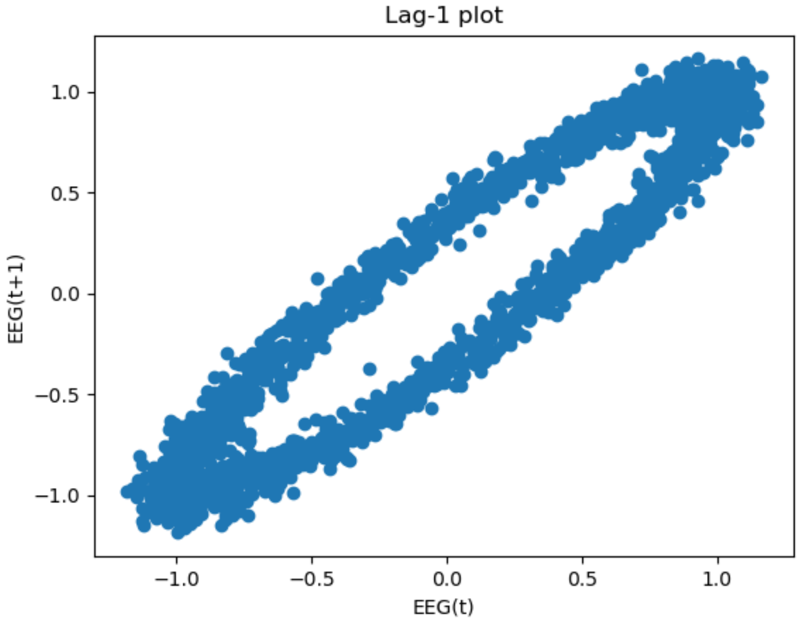
1. For the EEG data set, draw the line plot, histogram, density plot, box plot, lag-1 plot, ACF and PACF graphs (lags up to 50).

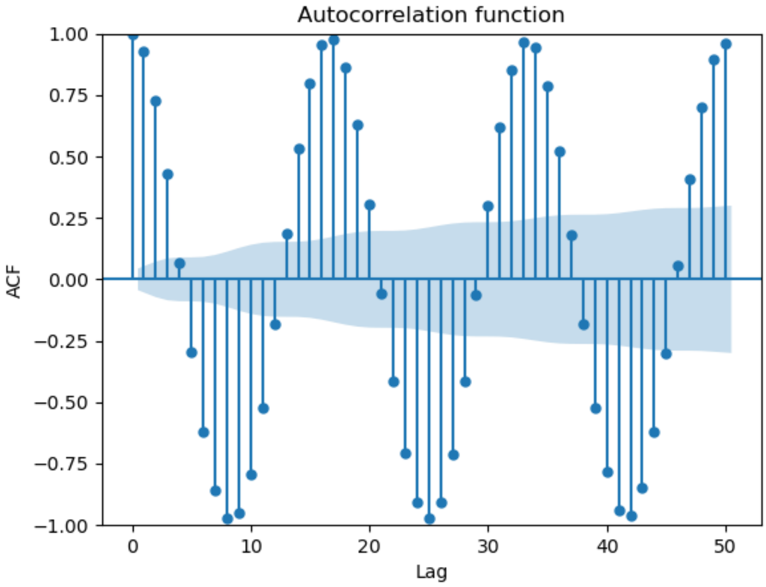

EEG data
Voltage (uV)
1.0
0.5
0.0
−0.5
−1.0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Time (s)

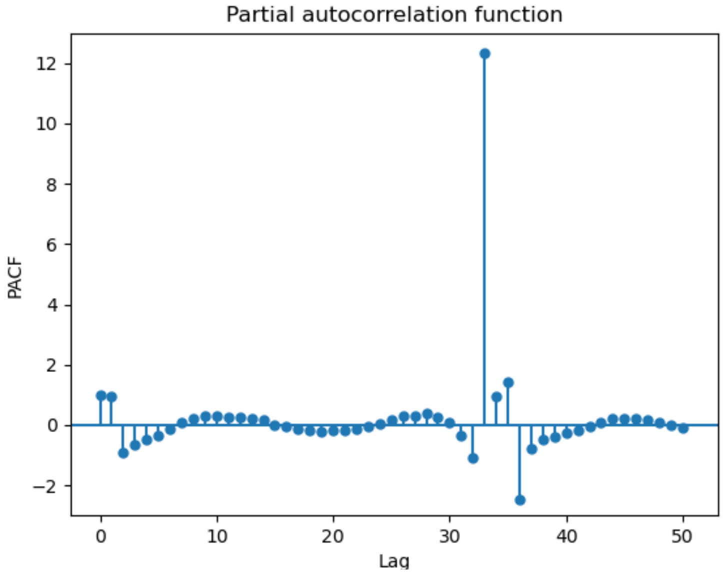

EEG data
Count
80
70
60
50
40
30
20
10
0
−1.0
−0.5
0.0
0.5
1.0
Voltage (uV)


EEG Density Plot
Density
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
EEG Amplitude


EEG data
Voltage (uV)
1.0
0.5
0.0
−0.5
−1.0
1


Lag-1 plot
EEG(t+1)
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
EEG(t)



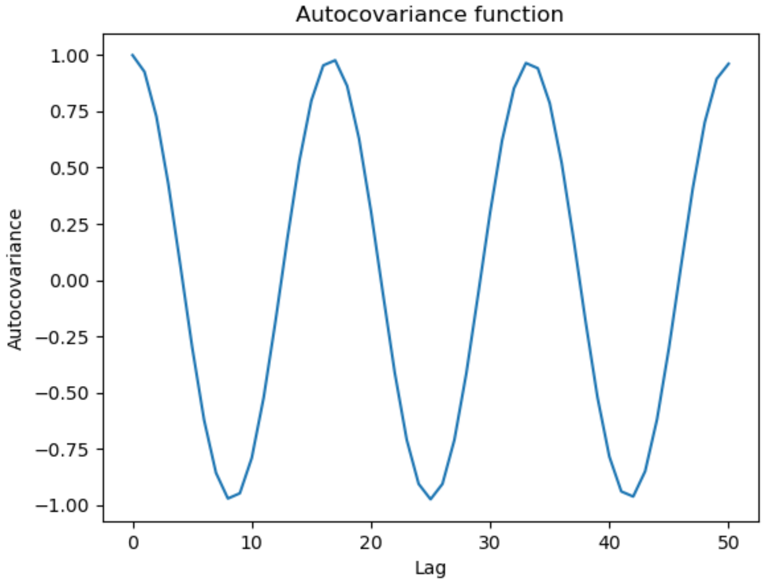


2. Show the statistical characteristics of the EEG data, such as mean, variance, standard deviation.

Mean of EEG data: 0.0000
Variance of EEG data: 0.5047
Standard deviation of EEG data: 0.7104

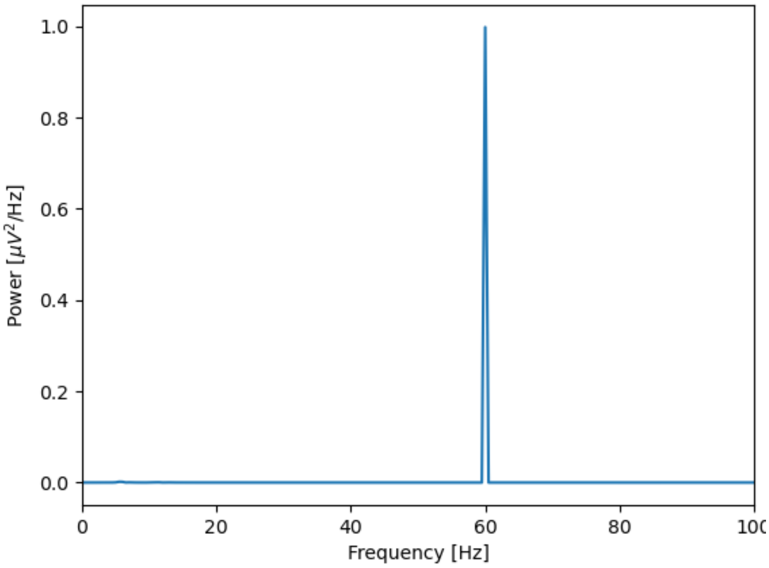
3. Compute the auto-covariance of the EEG data. Draw and save the auto-covariance graph. Examine the auto-covariance plot, why does the auto-covariance exhibit repeated peaks and troughs approximately every 0.0166 s?

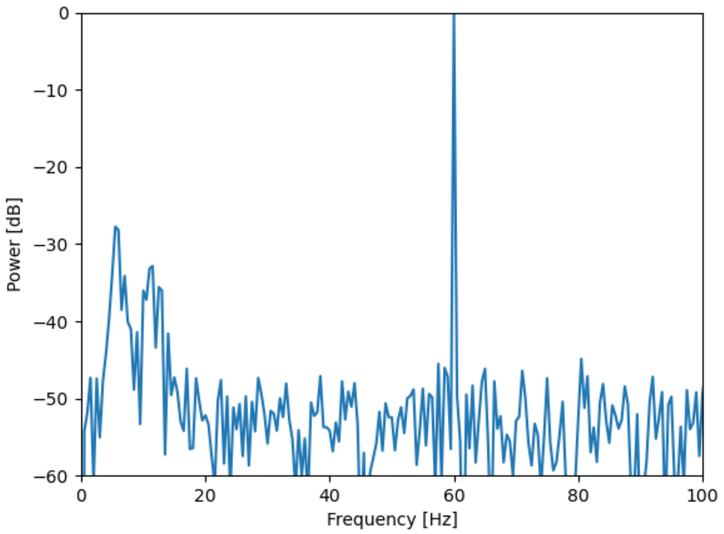
Through the power spectrum, we can find that it peaks at 60Hz, which can be turned into 0.0166s.



4. Compute and plot the power-spectrum of the EEG data. Show it in both linear scale and log (dB) scale. To emphasize low-amplitude rhythms hidden by large-amplitude oscillations is to change the scale of the spectrum to decibels.

Here we use log scale to amplify the data which don't change dramatically. As the following two pictures showing, we can see the change of the data clearly in log scale.





Questions

- What features do you typically consider useful for analyzing and modeling time-series data?

Trend: Identifying the overall behavior or tendency of the time series, e.g., increasing or decreasing over time.

Seasonality: Identifying repeating patterns or cycles in the time series, e.g., daily, weekly, or yearly cycles.

Autocorrelation: Identifying the correlation between the values of the time series at different points in time.

Stationarity: Identifying whether the statistical properties of the time series, such as mean and variance, are consistent over time.

- What features are specific for time-series, and what are general for both time-series and non-time-series data?

Specific to time-series data: Trend, Seasonality, Autocorrelation, Stationarity

General for both time-series and non-time-series data: Mean, Variance, Standard Deviation,

- How are auto-covariance and auto-correlation are defined for a time series? Give mathematical formulas for the definitions.

Autocovariance: Measures the linear dependence between two points in a time series separated by a certain time lag.

$$Cov(y_t, y_{t-k}) = E[(y_t - \mu_t)(y_{t-k} - \mu_{t-k})]$$

E means expectations, y_t and y_{t-k} mean the observed value at time t and t-k, respectively. and $\mu_t$and μ_{t-k} mean average value at time t and t-k, respectively.

Autocorrelation: Measures the linear relationship between two points in a time series separated by a certain time lag, after accounting for the overall variability in the time series.

$$Corr(y_t, y_{t-k}) = \frac{Cov(y_t, y_{t-k})}{\sqrt{Var(y_t)Var(y_{t-k})}}$$

$Var(y_t)$ and $Var(y_{t-k})$ mean the variance at time t and t-k, respectively.

- Assume a short time-series {1, 2, 3, 4, 5, 6, 7, 8, 7, 6, 5, 4, 3, 2, 1}.

(1) Calculate the auto-covariance and auto-correlations for all valid lags. Do the calculations manually.

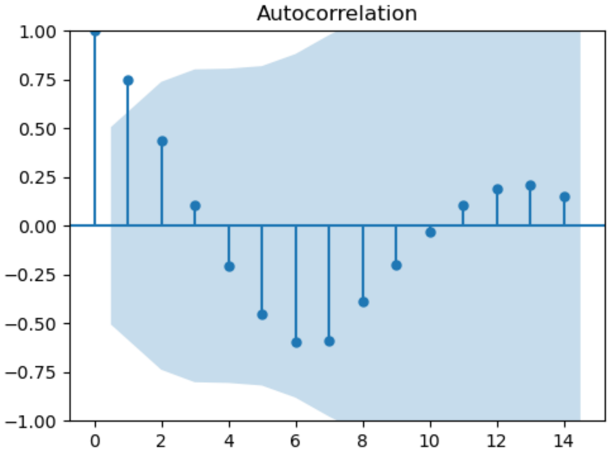
Here, we use the formula with biased.

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1+2+\cdots+2+1}{15} = 4.2667$$

lags	Auto-covariance	Auto-correlation
0	4.7289	1
1	3.5508	0.7569
2	2.0750	0.4388
3	0.5013	0.1060
4	-0.9701	-0.2051
5	-2.1393	-0.4524
6	-2.800	-0.5934
7	-2.7710	-0.5860
8	-1.833	-0.3877
9	-0.9316	-0.1970
10	-0.1319	-0.0279
11	0.4990	0.1053
12	0.8942	0.1891
13	0.9873	0.2288
14	0.7114	0.1504

(2) Write a Python program to validate your calculations.

(3) Draw the ACF graph for the time series.


Autocorrelation