KTH Royal Institute of Technology

Embedded Intelligence
IL2233 VT24

Lab 1 - Time-series data visualization and feature extraction

Humblet Raphaël
Yao Tianze

April 2024

Academic year 2023-2024

Contents

1 Task 1. Exploratory Data Analysis **2**

1.1 Task 1.1. White noise series 2

1.2 Task 1.2. Random-walk series 3

1.3 Task 1.3. Global land temperature anomalies series 5

2 Task 2. Feature Extraction **9**

2.1 Task 2.1. Frequency components of a synthetic time-series signal . . . 9

2.2 Task 2.2. Statistical features and discovery of event-related potential 10

2.3 Task 2.3. Features of observed rhythms in EEG 10

1 Task 1. Exploratory Data Analysis

1.1 Task 1.1. White noise series

A white noise series with $N = 1000$ data points has been generated with a Gaussian function $\mathcal{N}(0, 1)$ (the histogram and density plots obtained illustrates this).

The mean that has been calulated was $mean = -0.071204$ and the standard deviation $std = 0.992283$. This is close to the "predicted" mean of 0 and std of 1.

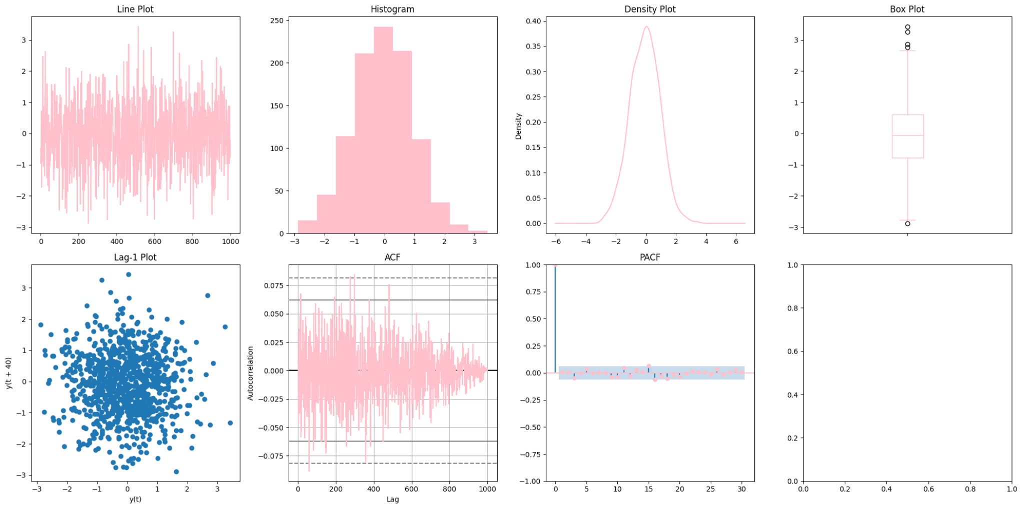

Figure 1: Plots for the white noise series

As we can see on the PACF of figure 1 that there is no correlation except with the value that itself.

Then 100 random series of 1000 data points have been implemented, and a mean series has been produced from it (see figure 2). The expected mean should be of 0. The actual value obtained was $mean = 0.003204$, which is pretty close to the predicted one.

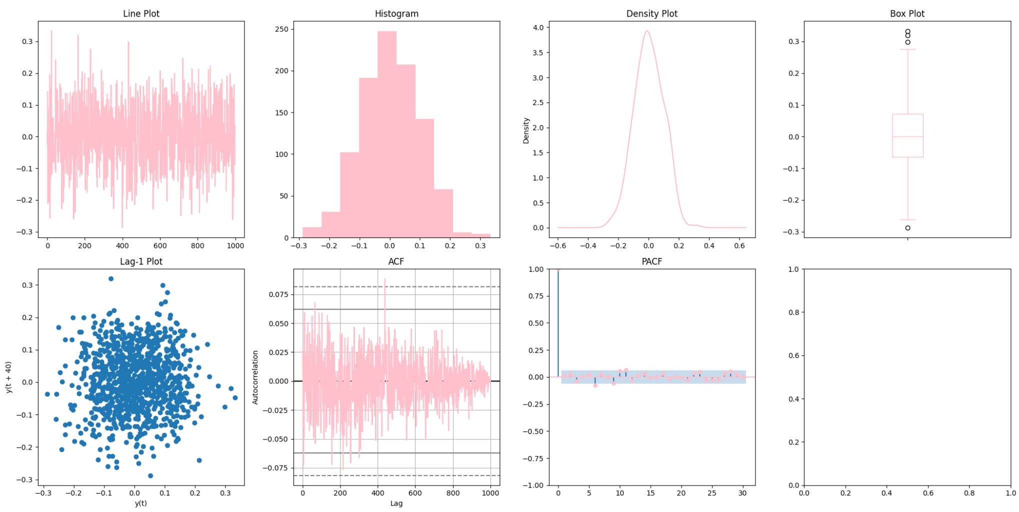

Figure 2: Plots for the average white noise series

The standard deviation $std = 0.097248$.

The Ljung-Box test has been executed for the white noise series to see if the series is random or not. As we can see in table 1, the p-values are bigger than 0.05 so the series is **random**.

	lb_stat	lb_pvalue
1	0.209612	0.647072
2	0.913240	0.633421
3	2.511843	0.473155
4	2.534945	0.638389
5	2.997271	0.700407
6	8.317121	0.215779
7	8.891058	0.260573
8	8.913012	0.349687
9	11.346760	0.252687
10	13.932070	0.176113

Table 1: Randomness test result for the white noise series

The Augmented Dickey-Fuller (ADF) test has been performed to analyse the stationarity of the series. The p-value obtained is $2.7058925365876226 \cdot 10^{-15} <<< 0.05$. Hence, we reject the null hypothesis and judge that the series is **stationary**.

1.2 Task 1.2. Random-walk series

The random-walk series generated has a mean of $mean = 3.510000$ and a standard deviation $std = 8.063059$.

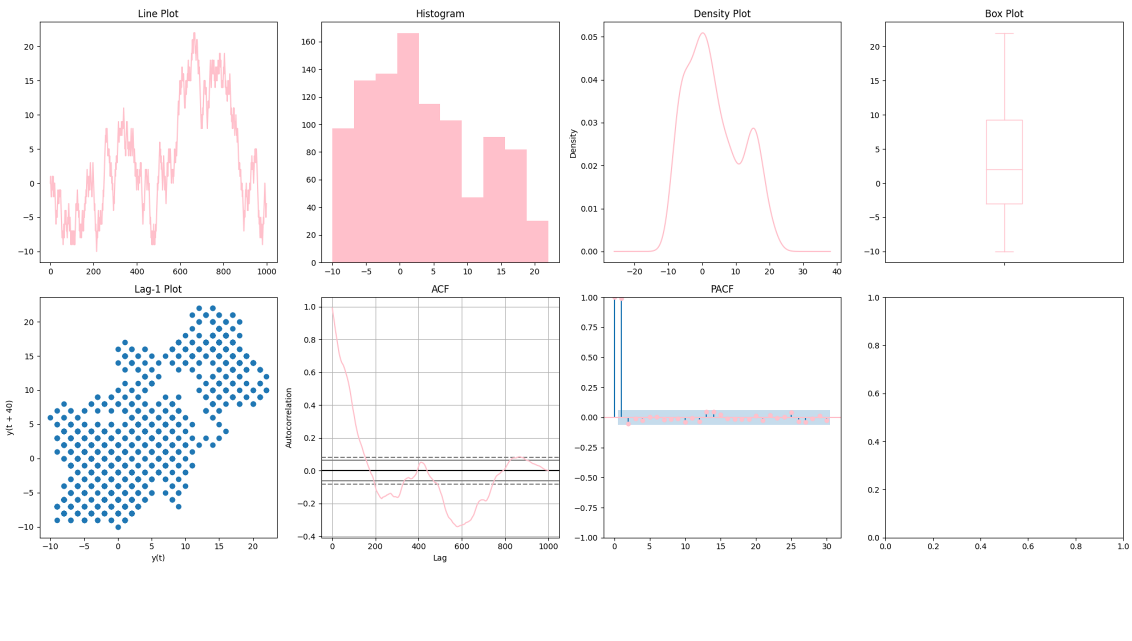

Figure 3: Plots for random-walk series

We can observe on the PCAF that there is a huge correlation between each value and its previous lag. This will be helpful to make it stationary afterwards.

When performing the Ljung-Box test, we obtain p-values that are smaller than 0 (see table 2. Hence, the series is random.

	lb_stat	lb_pvalue
1	986.796507	1.331216e-216
2	1957.001386	0.000000e+00
3	2910.467447	0.000000e+00
4	3846.756314	0.000000e+00
5	4766.185350	0.000000e+00
6	5669.244598	0.000000e+00
7	6555.657519	0.000000e+00
8	7425.217100	0.000000e+00
9	8277.954405	0.000000e+00
10	9113.026291	0.000000e+00

Table 2: Randomness test results for the random-walk series

The p-value obtained while doing the ADF test is $p{-}value = 0.2912501414668911 > 0.05$. Thus, it is not stationary. The series being described by the following equation:

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

It can be turned into a stationary series by first-order differencing.

$$y'_t = y_t - y_{t-1} = \epsilon'_t$$

with ϵ'_t being randomly chose between either -1 or 1. Hence, **stationary** as it is time independent.

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

Use the Ljung-Box test and check the p-value, it is random if p-value < 0.05. Or we can check the Lag plot and see if we recognize some sort of pattern.

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

Use the Augmented Dickey-Fuller (ADF) test, it is stationary if p-value < 0.05. On the line plot, if the points have a tendency to stay around a "mean" value, it is stationary. Otherwise, they have a tendency to go higher or lower than they are not stationary.

Why is white noise important for time-series prediction?

White noise is important for time-series prediction it serves as a reference point for evaluating the performance of time-series prediction models, diagnosing model adequacy.

What is the difference between a white noise series and a random walk series?

A white noise series is stationary, while a random walk series is not. Meaning that random walk series has time dependency while white noise is time independent.

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.

Yes it is possible with first order differencing. This is because of the definition of the random walk. This allows to eliminate the temporal dependencies.

Differencing is allowed and widely utilized in time-series analysis because it aligns with statistical principles, theoretical foundations, and practical considerations.

1.3 Task 1.3. Global land temperature anomalies series

After extracting the data, a series has been created with it and plotted in figure 4. As we can see there is a clear trend on the line plot. And from the PACF we can see that there is a dependency with the two previous lags.

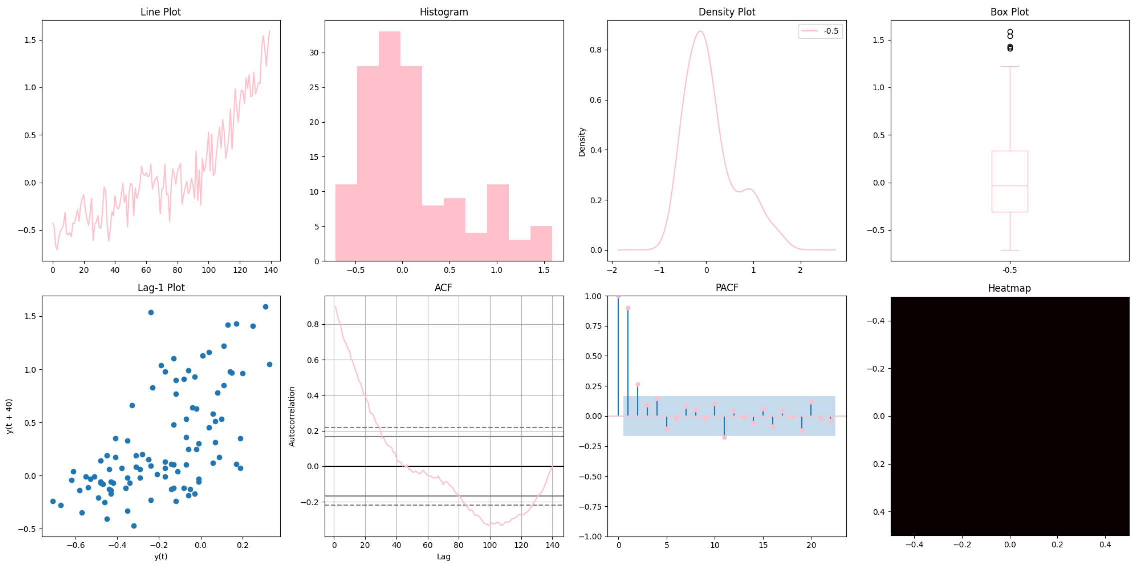

Figure 4: Plots for Global land temperature anomalies series data

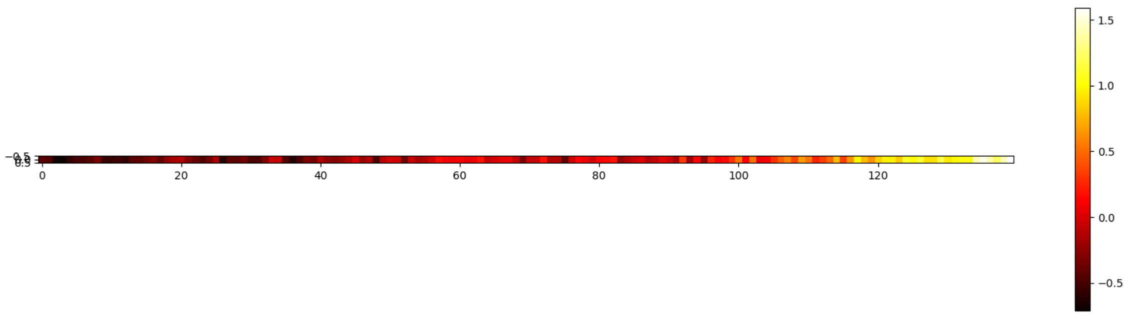

Figure 5: Heatmap for Global land temperature anomalies

In order to make this series stationary, we can do the first order differencing (as seen on previous section). We then obtain figure 6.

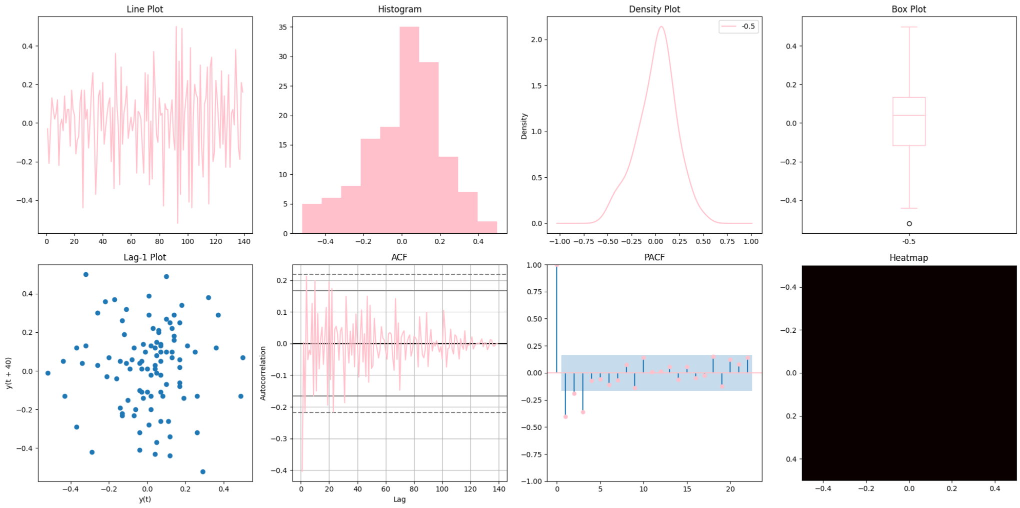

Figure 6: Plots after applying first order difference to temperature anomaly dataset series

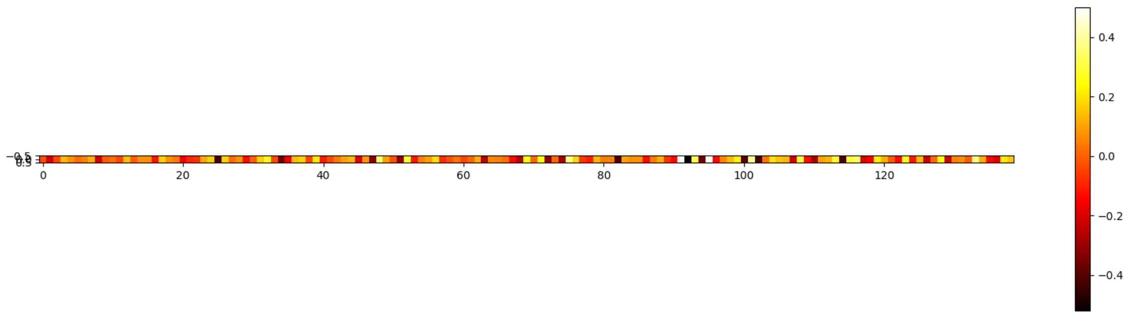

Figure 7: Heatmap of the first order differenced series.

We can see there in the PACF that the series is more stationary than before.

To check whether the series are random, we use the Ljung-Box test. As we can see in table 3, the p-values are lower than 0.05 for both series, and thus **both series are random**.

	Global lb_stat	Global lb_pvalue		Diff lb_stat	Diff lb_pvalue
1	115.515630	6.067891e-27		23.132636	1.512024e-06
2	221.656946	7.375742e-49		23.135046	9.468661e-06
3	319.673865	5.487167e-69		29.890551	1.455185e-06
4	414.868360	1.703920e-88		36.480896	2.304036e-07
5	499.604656	9.718316e-106		36.866671	6.369614e-07
6	576.934531	2.200049e-121		36.929100	1.817822e-06
7	650.628528	3.022260e-136		37.112014	4.466959e-06
8	721.693488	1.526452e-150		39.528658	3.920768e-06
9	787.344054	1.126176e-163		43.657261	1.632204e-06
10	851.935748	1.399223e-176		49.636784	3.112541e-07

Table 3: Randomness test on Global and diff series

By performing the ADF test, we obtain the following results:

- Global is not stationary $p - value = 0.9922499670941117$
- Differenced is stationary $p - value = 1.1335188437723516 \cdot 10^{-22}$

This can also be seen graphically.

Classical and STL decomposition have been made on the dataset as seen in figure 8 and 9.

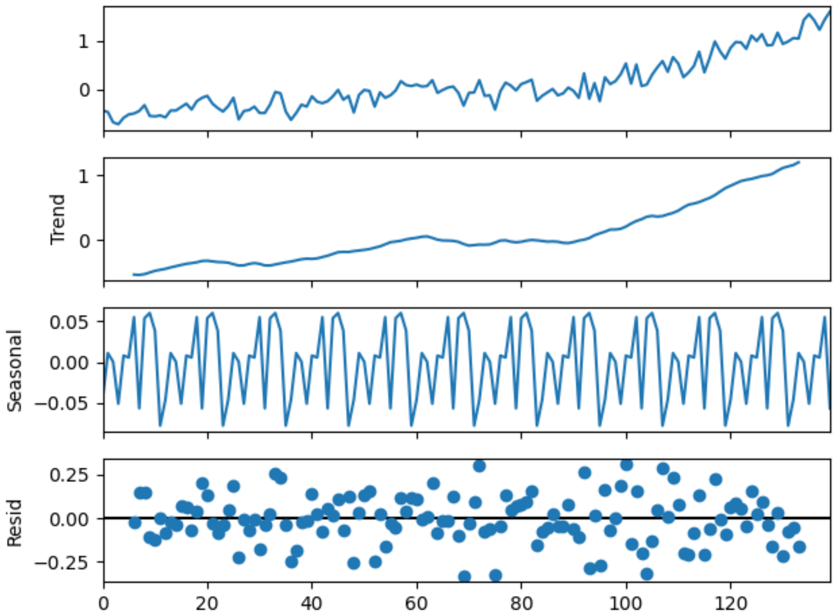

Figure 8: Additive decomposition

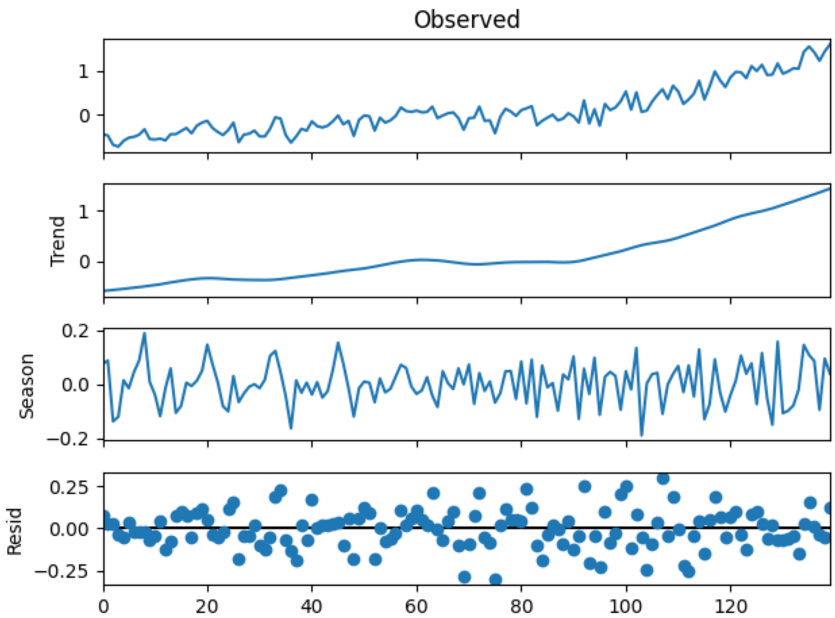

Figure 9: STL decomposition

What is a stationary time series?
A stationary time series is a series that has no time dependency. It doesn't matter which time period is taken to analyse the trend.

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 it is, by differencing for example. Other techniques exist but have not been seen here.

Is the global land temperature anomaly series stationary? Why or why not?
No it is not stationary because as we can see the trend is going higher, also the p-value is > 0.05.

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

Why is it useful to decompose a time series into a few components? What are the typical components in a time-series decomposition?
The typical components are trend, seasonality, residual.

It is useful to have some insight into underlaying patterns, modelling and forecasting.

2 Task 2. Feature Extraction

2.1 Task 2.1. Frequency components of a synthetic time-series signal

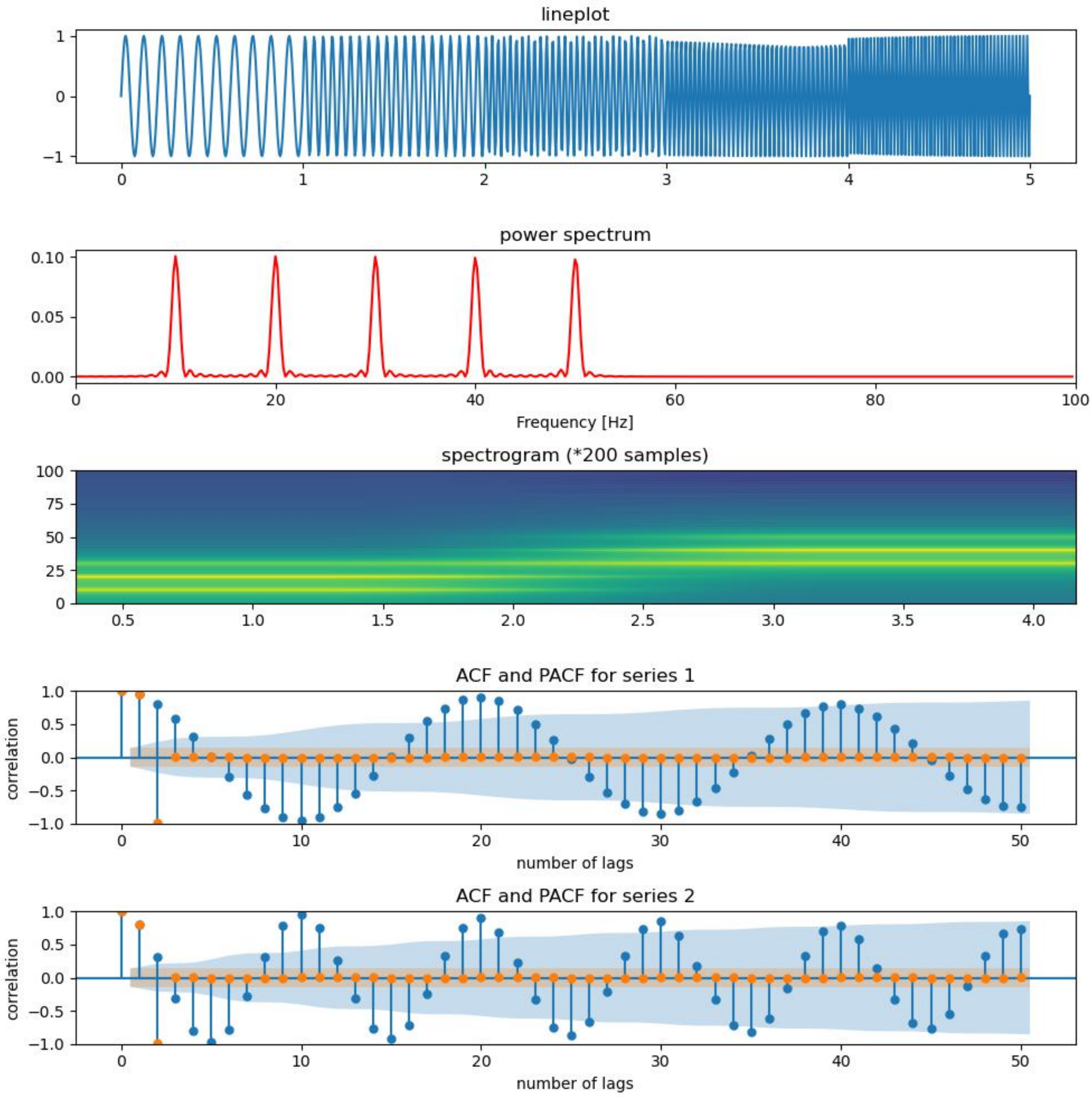

Figure 10: Plot, power spectrum, spectrogram, ACF and PACF

1. Line plot for the series is shown in the first subfigure of Figure.10
2. Powerspectrum for the series is shown in the second subfigure of Figure.10
3. Spectrogram for the series is shown in the third subfigure of Figure.10
4. ACF and PACF for the first 1s series is shown in the 4^{th} subfigure of Figure.10, where the blue curve represents ACF and the orange curve represents PACF.
5. ACF and PACF for the first 1s series is shown in the 5^{th} subfigure of Figure.10, where the blue curve represents ACF and the orange curve represents PACF.

2.2 Task 2.2. Statistical features and discovery of event-related potential

1. The visualized ERPs of EEG in two conditions are shown in Figure.11

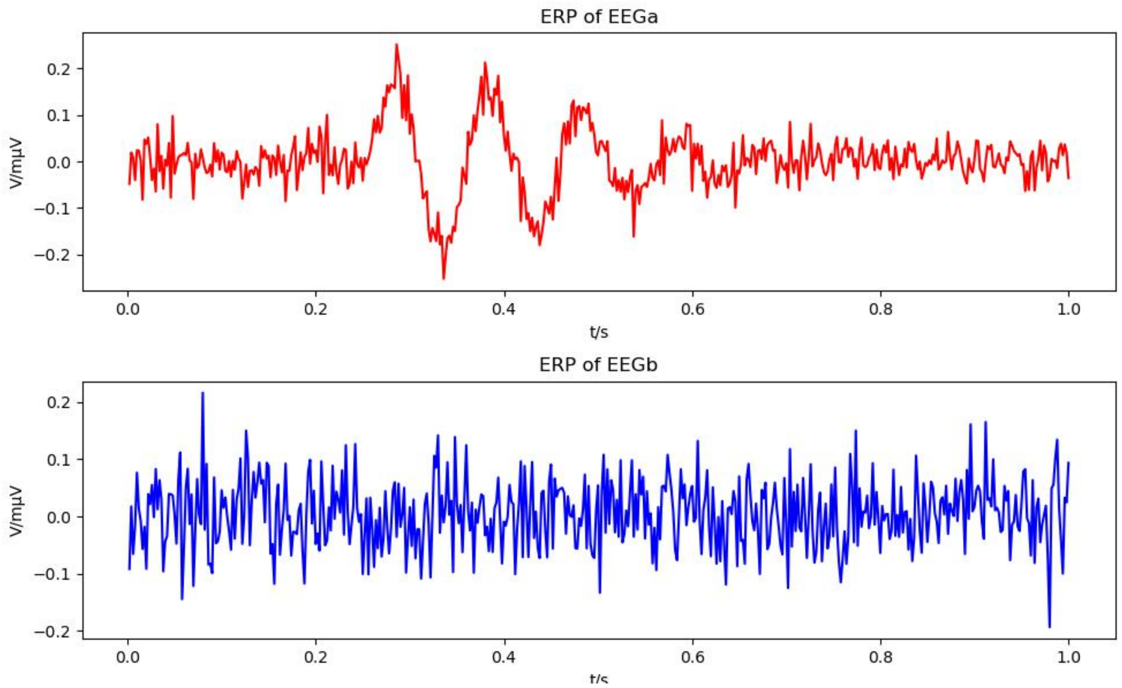

Figure 11: ERP of EEGs

2. Focus on the main wave of ERP in condition A, measure that from $t_0 = 0.25s$ to $t_1 = 0.57s$, there are 3 complete cycles. So, the brain activity in response to condition A

$$f = \frac{3}{t_1 - t_0} = 9.375Hz$$

2.3 Task 2.3. Features of observed rhythms in EEG

1. The statistic figures regarding the EEG data are shown in Figure. 12

2. With the help of functions $np.mean(), np.var(), np.std()$, calculate that

$$mean = 2.731 * 10^{-17}, \quad variance = 0.505, \quad standard\ deviation = 0.710$$

3. The auto-covariance plot is shown in the 7^{th} subfigure of Figure.12 and the matrix for auto-covariance is shown in Figure. 13, the main auto-covariance is 0.47229842. From the power spectrum, we can derive that EEG data has only one peak at 60 Hz, which means a oscillation cycle of accurate $1/60 = 0.0167s$, so the auto-covariance should also have a cycle of 0.0167s.

4. The power-spectrum plot of the EEG data is shown in the 8^{th} and 9^{th} subfigures of Figure.12.

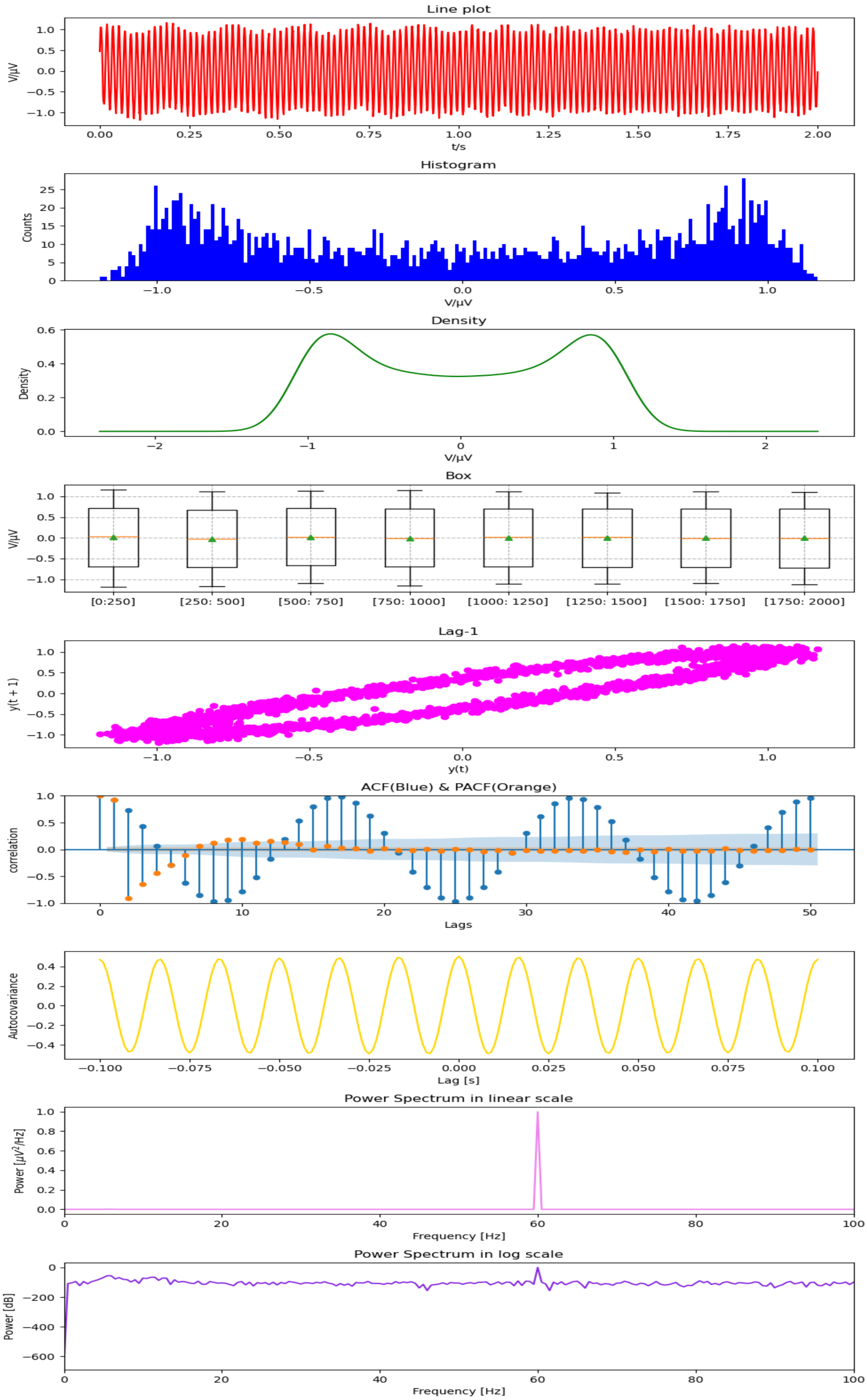

Figure 12: Plots for Task 3

```
[ 0.47229842  0.43936817  0.34439294  0.20087657  0.0287394  -0.14793023
 -0.30412595 -0.41784278 -0.4732382  -0.4623532  -0.3866724  -0.25671984
 -0.09083437  0.08798794  0.25434194  0.38492611  0.46134043  0.47294446
  0.41786849  0.30372059  0.14682436 -0.03131054 -0.20546305 -0.35086075
 -0.44751636 -0.4814903  -0.44801858 -0.35172519 -0.20617956 -0.03166733
  0.14729362  0.30550066  0.4207532   0.47683742  0.46571907  0.38901121
  0.25716677  0.08907744 -0.0920967  -0.26078416 -0.39304554 -0.47052432
 -0.48225329 -0.42630462 -0.31055885 -0.15129641  0.02911834  0.20567814
  0.35312791  0.45104831  0.48541652  0.4514943   0.35388821  0.20632646
  0.02947187 -0.1518187  -0.31211829 -0.42895535 -0.48558539 -0.47417804
 -0.39613497 -0.26262438 -0.09193173  0.09164858  0.26260838  0.39678673
  0.47523141  0.48702814  0.43041353  0.31329782  0.15210301 -0.03071293
 -0.20907849 -0.35833183 -0.45718525 -0.49184249 -0.45724036 -0.35804382
 -0.20837326 -0.02898616  0.15491067  0.31731468  0.43574648  0.49321951
  0.48174448  0.40278713  0.26780289  0.09511703 -0.090701   -0.26360871
 -0.39904793 -0.47833561 -0.49007471 -0.43243309 -0.3136667  -0.15016277
  0.03509614  0.2159301   0.36714711  0.46739139  0.50471724  0.46739139
  0.36714711  0.2159301   0.03509614 -0.15016277 -0.3136667  -0.43243309
 -0.49007471 -0.47833561 -0.39904793 -0.26360871 -0.090701    0.09511703
  0.26780289  0.40278713  0.48174448  0.49321951  0.43574648  0.31731468
  0.15491067 -0.02898616 -0.20837326 -0.35804382 -0.45724036 -0.49184249
 -0.45718525 -0.35833183 -0.20907849 -0.03071293  0.15210301  0.31329782
  0.43041353  0.48702814  0.47523141  0.39678673  0.26260838  0.09164858
 -0.09193173 -0.26262438 -0.39613497 -0.47417804 -0.48558539 -0.42895535
 -0.31211829 -0.1518187   0.02947187  0.20632646  0.35388821  0.4514943
  0.48541652  0.45104831  0.35312791  0.20567814  0.02911834 -0.15129641
 -0.31055885 -0.42630462 -0.48225329 -0.47052432 -0.39304554 -0.26078416
 -0.0920967   0.08907744  0.25716677  0.38901121  0.46571907  0.47683742
  0.4207532   0.30550066  0.14729362 -0.03166733 -0.20617956 -0.35172519
 -0.44801858 -0.4814903  -0.44751636 -0.35086075 -0.20546305 -0.03131054
  0.14682436  0.30372059  0.41786849  0.47294446  0.46134043  0.38492611
  0.25434194  0.08798794 -0.09083437 -0.25671984 -0.3866724  -0.4623532
 -0.4732382  -0.41784278 -0.30412595 -0.14793023  0.0287394   0.20087657
  0.34439294  0.43936817  0.47229842]
```

Figure 13: Auto-covariance matrix

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

First is the mean value and box plot, they mainly point out the average scale of a time series, which is good to evaluate the values of series in general.

Second is the auto-covariance and auto-correlation, both of them measure the linear relationship between different time (lagged value) of a time series, where we can derive the change of series in timescale and see if it is periodic or random.

Third is the line plot, this can directly uncover the original time series for analysis.

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

Time series specifically has the time character. Each value in time series all matches with a time point, so time series is changing all time when time goes. However, non-time series doesn't have such characters and may keep constant with time. Both of them have statistic features such as mean, standard deviation and variance, but only time series has auto-correlation and auto-covariance.

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

Auto-covariance: show the dependent structure in data.

$$r_{xx}[L] = \frac{1}{N-L} \sum_{n=1}^{N-L} (x_{n+L} - \overline{x})(x_n - \overline{x})$$

Auto-correlation: give the relationship between lagged values of time series.

$$\gamma_k = \frac{r_{xx}[k]}{r_{xx}[0]}$$

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.
The manually calculations are shown in the Table.4

Lags	Auto-covariance	Auto-correlation
0	4.73	1
1	3.55	0.75
2	2.07	0.44
3	0.50	0.11
4	-0.97	-0.21
5	-2.14	-0.45
6	-2.81	-0.59
7	-2.77	-0.58
8	-1.83	-0.39
9	-0.93	-0.20
10	-0.13	-0.03
11	0.50	0.10
12	0.89	0.19
13	0.99	0.21
14	0.71	0.15

Table 4: Auto-covariance and auto-correlation values computed manually

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

The validation program's output results are shown in Figure.14, which proves the manual calculations.

```
[ 0.71140741  0.98725926  0.89422222  0.49896296 -0.13185185 -0.93155556
 -1.83348148 -2.77096296 -2.80622222 -2.13925926 -0.97007407  0.50133333
  2.07496296  3.55081481  4.72888889  3.55081481  2.07496296  0.50133333
 -0.97007407 -2.13925926 -2.80622222 -2.77096296 -1.83348148 -0.93155556
 -0.13185185  0.49896296  0.89422222  0.98725926  0.71140741]
[ 1.          0.75087719  0.43878446  0.10601504 -0.20513784 -0.45238095
 -0.59342105 -0.58596491 -0.3877193  -0.19699248 -0.02788221  0.10551378
  0.18909774  0.20877193  0.1504386 ]
```

Figure 14: Python computed values of auto-covariance and auto-correlation

(3) Draw the ACF graph for the time series
Figure.15 shows the ACF graph.

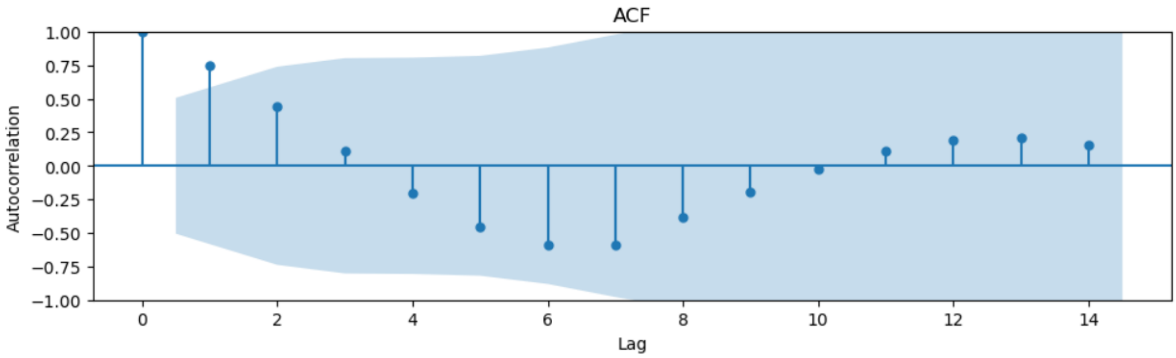

Figure 15: ACF for the time series