Lecture 2. Basics on Time-Series Analysis

Exploratory data analysis and White noise series

Prof. Zhonghai Lu

KTH Royal Institute of Technology

March 20, 2024

Outline

1. Exploratory Data Analysis (EDA)
2. Data visualization
3. Application of EDA
4. White noise and Randomness
5. IPython demo

Exploratory Data Analysis (EDA)

- In statistics, EDA is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.
- Most EDA techniques are graphical in nature with a few quantitative techniques.
- By its very nature, the main role of EDA is to open-mindedly explore, and reveal its structural secrets, and to gain some new, often unsuspected, insight into the data.

Exploratory data analysis chapter: Engineering statistics handbook. Chapter 1. Explore. (`https://www.itl.nist.gov/div898/handbook/` The link was working. It seems down now.).

Common types of plots for time series

In the statistical and time-domain analysis, seven common types of plots are generally interesting for time-series analysis.

1. Line plot.
2. Histogram and density plot.
3. Box & Whisker plot (Box plot).
4. Heat map.
5. Lag scatter plot.
6. Auto-Correlation function (ACF) plot.
7. Partial Auto-Correlation Function (PACF) plot.

The focus is on univariate time series, but the techniques are just as applicable to multivariate time series, when you have more than one observation at each time step.

Why useful?

- Explore the *temporal structure* of time series with line plots, lag plots, and autocorrelation plots.
- Understand the *distribution of observations* using histograms and density plots.
- Capture the *change in distribution over intervals* using box and whisker plots and heat map plots.

Plots of the raw sample data can provide valuable diagnostics to identify temporal structures like trends, cycles, and seasonality that can influence the choice of model.

A data set

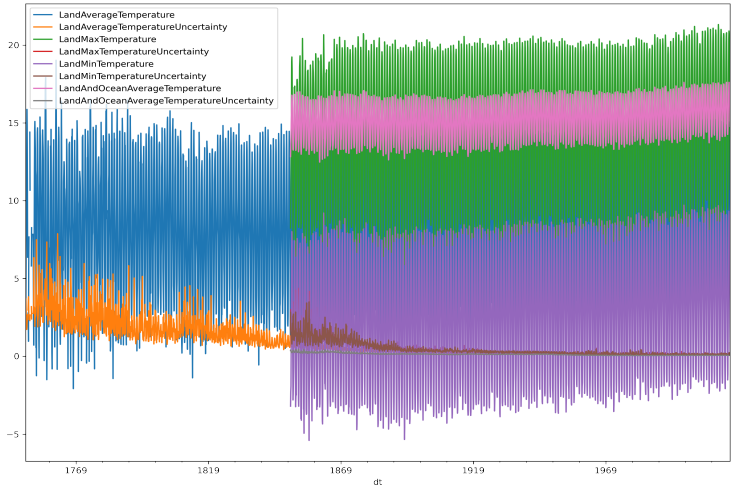
- The Global Land Temperatures data set records temperatures from 1750 to 2015, spanning 266 years.
- Each year has per-month data, i.e., 12 numbers. In total, 266x12=3192.
- The data set is in .csv (comma separated values) format, which has 3192 + 1 rows, 8 + 1 columns. One column for one type of data and one for date-time.
- There are many missing data points (incomplete record).

The data set is from Kaggle.

Line plot

- The first and perhaps most popular visualization for time series is the line plot, also known as run sequence plot.
- In this plot, time is shown on the x-axis with observation values along the y-axis.

Line plot


LandAverageTemperature
LandAverageTemperatureUncertainty
LandMaxTemperature
LandMaxTemperatureUncertainty
LandMinTemperature
LandMinTemperatureUncertainty
LandAndOceanAverageTemperature
LandAndOceanAverageTemperatureUncertainty
20
15
10
5
0
−5
1769
1819
1869
1919
1969
dt

Line plot

Let's focus on the LandAverageTemperature data.

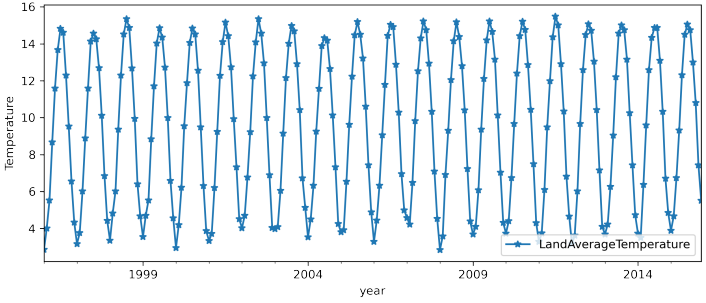


Figure: The land average temperature in last 20 years

Data distribution

- Another important visualization is about the distribution of observations themselves, a plot of the values without the temporal ordering.
- Some linear time series forecasting methods assume a well-behaved distribution of observations (i.e. a bell curve or normal distribution). This can be explicitly checked using tools like statistical hypothesis tests.
- But plots can provide a useful first check of the distribution of observations both on raw observations and after data transformation.

Histogram

- A histogram groups values into bins (value intervals), and the frequency or count of observations in each bin can provide insight into the underlying distribution of the observations.
- The example creates a histogram plot of the observations in the temperatures dataset.

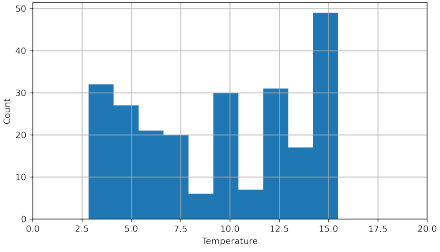


Figure: A histogram

Density plot

- A density plot is a representation of the distribution of a numeric variable. It uses *kernel density estimation* (KDE) to show the *probability density function* of the variable.
- A density plot is a smoothed version of histogram. It provides a clearer summary of the distribution of observations.
- In statistics, KDE is a non-parametric way to estimate the probability density function of a random variable. It is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample.

Density plot example

- The density plot shows two peaks, a little asymmetrical.
- Statistical hypothesis tests may be used to formally check if the distribution is Gaussian.
- Data preprocessing techniques may be used to reshape the distribution, like the BoxCox transform.

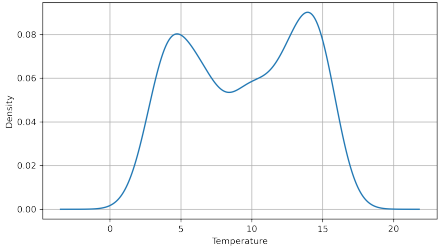


Figure: A density plot

Box plot

Histograms and density plots provide insight into the distribution of all observations, but we may be interested in the distribution of values in certain intervals, and identifying if there is any outlier in the data.
For this, we can draw the box and whisker plot, which is a standardized way of displaying the distribution of data based on the **five-number summary**:

- Minimum (Q0 or 0th percentile): the lowest data point excluding any outliers
- First quartile (Q1 or 25th percentile): also known as the lower quartile q_n(0.25), it is the median of the lower half of the dataset.
- Median (Q2 or 50th percentile): the middle value.
- Third quartile (Q3 or 75th percentile): also known as the upper quartile q_n(0.75), it is the median of the upper half of the dataset.
- Maximum (Q4 or 100th percentile): the highest data point excluding any outliers.

Box plot

- This plot draws a box around the 25th and 75th percentiles of the data that captures the middle 50% of observations.
- A line is drawn at the 50th percentile (the median) and whiskers are drawn above and below the box to summarize the general extents of the observations.
- Dots are drawn for outliers outside the whiskers or extents of the data.

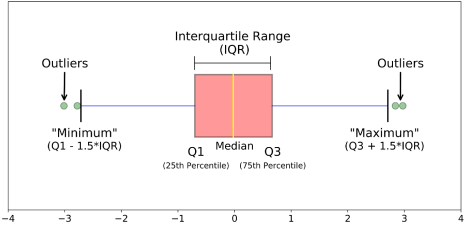


Figure: Explanation of box plot with outliers

Source online, e.g. `https://builtin.com/data-science/boxplot`

Box plot

A box plot can show you:

- if the data is symmetrical or skewed
- how tightly the data are grouped
- the values of outliers

An example of box plot for the temperature data set.

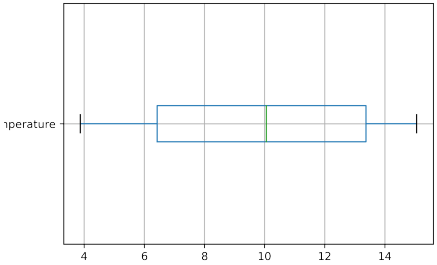


Figure: Box plot for year 2015 data

Box plot

- Box and whisker plots can be created and compared for each interval in a time series, such as years, months, or days.
- The example below creates 20 box plots, one for each of the last 20 years.

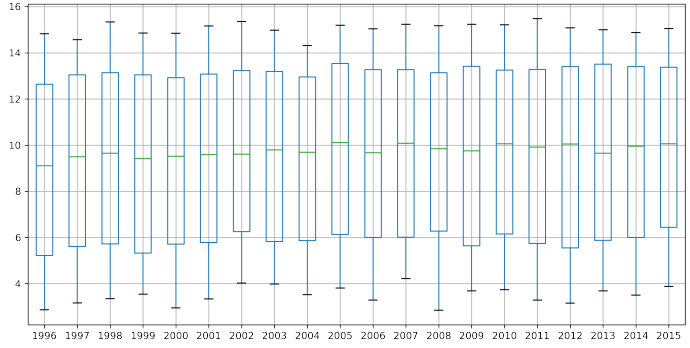
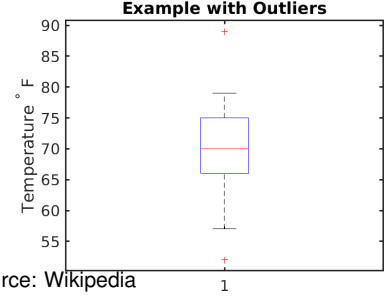
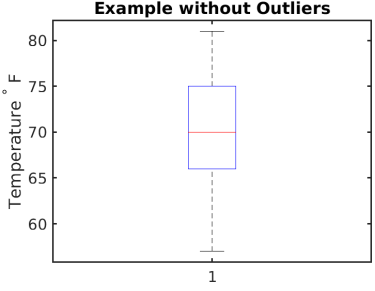


Figure: Box plot for 20 years of data (1996 to 2015)

Box plot with/without outliers

- Note that the whiskers can stand for other things, for example, the minimum and the maximum value of the data set. In this case, we assume no outliers in the data set.
- Example: A series of hourly temperatures were measured throughout the day in degrees Fahrenheit.
  The recorded values are listed in order as follows (°F): 57, 57, 57, 58, 63, 66, 66, 67, 67, 68, 69, 70, 70, 70, 70, 72, 73, 75, 75, 76, 76, 78, 79, 89.





Source: Wikipedia

Heatmap

- A matrix of numbers can be plotted as a surface, where the values in each cell of the matrix are assigned a unique color.
- Such a graph is called a heatmap, since larger values can be drawn with warmer colors (yellows and reds) and smaller values can be drawn with cooler colors (blues and greens).
- Like the box plots, we can compare observations between intervals using a heat map.

Example of heatmap

- For the Land Average Temperatures, the observations can be arranged into a matrix of year columns and month-rows, with temperature in the cell for each month.
- A heat map of this matrix can then be plotted. Each column represents one year and each row one month.

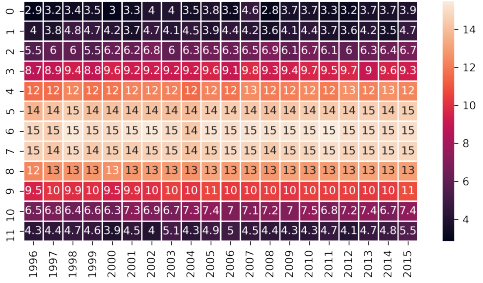


Figure: Heatmap for last 20 years of the temperature data

Lag plot

- A "lag" is a fixed amount of passing time. The kth lag is the time period that happened "k" time points **before** time t.
  For example, given a series $Y_1, Y_2..., Y_T$. $Lag_1(Y_2) = Y_1$ and $Lag_4(Y_9) = Y_5$.
  Its Lag-1 series is $Y_1, Y_2..., Y_{T-1}$.
- Lag plots can be generated for any arbitrary lag, though the commonly used lag is 1.
- A plot of lag 1 is a scatter plot of the values of Y_t versus Y_{t-1}:
  Vertical axis: Y_t for all t;
  Horizontal axis: Y_{t-1} for all t
- Lag plots can provide answers to the following questions:
  - Are the data random?
  - Is there serial correlation in the data?

Lag plot example

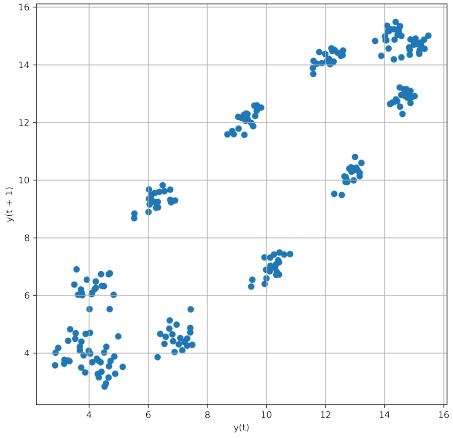


Figure: Lag-1 plot of the land temperature dataset

Autocorrelation

- Covariance and correlation measure the extent of linear relationship between two stochastic variables (Y and X)
- Auto-covariance and autocorrelation measure the linear relationship between lagged values of a time series y.
- We measure the relationship between y_t and y_{t-1}, y_t and y_{t-2}, y_t and y_{t-3}, etc.

Auto-correlation and correlogram

Given a time series y: $y_1, y_2..., y_T$. $\bar{y}$ is its mean value.
Denote the sample auto-covariance at lag k by c_k.
Denote the sample auto-correlation at lag k by γ_k.

$$c_k = \frac{1}{T} \sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})$$

$$\gamma_k = c_k / c_0$$

- γ_1 indicates how successive values of y relate to each other.
- γ_k is a normalized. $\gamma_0 \equiv 1$.
- The autocorrelations at lags $k = 1, 2, ...$ make up the autocorrelation fucntion (ACF) of the series. Such a plot is known as a correlogram.

ACF example

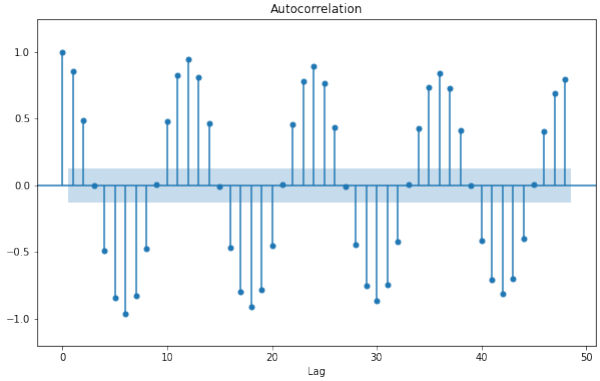


Figure: ACF of the land temperature dataset in the last 20 years

Autocorrelations measure correlation (linear relationship) between y_t and y_{t-k}.

PACF

- Auto-Correlation γ_k between two variables y_t and y_{t-k} can result from a mutual linear dependence on intermediate variables.
- Partial autocorrelation α_k is the autocorrelation between y_t and y_{t-k} after the removal of any linear dependence on the intermediate variables y_{t-1}, y_{t-2}, ..., $y_{t-(k+1)}$.
- Partial auto-correlations α_k at lags $k = 1, 2, ...$ make up the partial autocorrelation function (PACF) of the series y.

PACF example

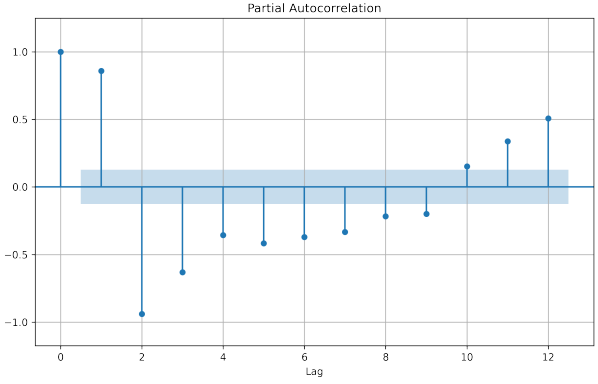


Figure: PACF of the land temperature dataset in the last 20 years

Partial autocorrelations measure correlation (linear relationship) between y_t and y_{t-k} when the inferences (effects) of other time lags, $1, 2, 3, \cdots, k-1$ are removed.

Application of EDA

EDA allows to visualize data and extract relevant information about the data. This helps to check for

- Outliers: Identify data points with extremely high or low values.
- Randomness: Is there a pattern in the data? Is the data white noise? Lag plot
- Modeling suitability: Is the data ready for modeling? E.g check for stationarity.
- Structure of the data. E.g. is there a long-term trend in the data? If so, decompose the series.
- Seasonality: Periodic fluctuations that happen at regular periods.
  Frequency-domain analysis will further expose the spectral features of periodic data.
- Find serial/auto-correlation: Identify a proper model for the data, e.g. ARIMA modeling methodology.

An example: Boxplot for outlier detection

Let x = $\{8, 10, 4, 5, 5, 4, 7, 16, 8, 9, 9, 7, 10, 1, 9, 6\}$.

- First, sort the 16 data in ascending order.
  x = [1, 4, 4, 5, 5, 6, 7, 7, 8, 8, 9, 9, 9, 10, 10, 16]
  MinimumThreshold=Q1-k·(Q3-Q1)
  MaximumThreshold=Q3+k·(Q3-Q1)
- The first quartile Q1 (the 4th value ?) is 5, and the third quartile Q3 (the 12nd value ?) is 9.
  By k=1.5, the maximum threshold is 15 (9+1.5(9-5)=15), and the minimum threshold is -1 (5-1.5(9-5)=-1).
- Is 16 an outlier?
  Yes, since 16 is larger than the maximum threshold 15, so it is an outlier.

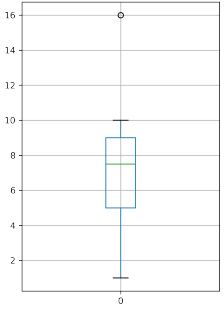


Figure: Box plot for a synthetic data set

How to calculate the quartile values?

Assume T is the length of a series.
Q(n) is a quartile, where $n = 0.25, 0.5, 0.75$, indicating the first, second, third quartile, i.e., Q1, Q2, Q3, respectively.

How to get Q(n)?

1. Calculate $z = T \cdot n$
2. If z is an integer, Q(n) is the average of the values at positions z and $z + 1$.
   If z is not an integer, Q(n) is the value at position $\lceil z \rceil$.
   $\lceil z \rceil$ is the ceiling function, e.g. $\lceil 2.3 \rceil = 3$

How to calculate the quartile values?

- Let x = $\{8, 10, 4, 5, 5, 4, 7, 16, 8, 9, 9, 7, 10, 1, 9, 6\}$.
- Ordered x = [1, 4, 4, 5, 5, 6, 7, 7, 8, 8, 9, 9, 9, 10, 10, 16]
- Q1, Q2, Q3 = ?
  T=16
  Q1: $z = 16 \cdot 0.25 = 4$. Q1 = (5+5)/2=5 (Average of the 4th and 5th values)
  Q2: $z = 16 \cdot 0.5 = 8$. Q2 = (7+8)/2=7.5 (Average of the 8th and 9th values)
  Q3: $z = 16 \cdot 0.75 = 12$. Q3 = (9+9)/2=9 (Average of the 12th and 13th values)

Box plot

The box plot is often used to identify outliers, where are *abnormal* data points.

- MinimumThreshold=Q1-k·(Q3-Q1)
  MaximumThreshold=Q3+k·(Q3-Q1)
- Smaller k means a more tight threshold and data are more likely to be considered as outliers.
- Usually, k equals to 1.5. Why?

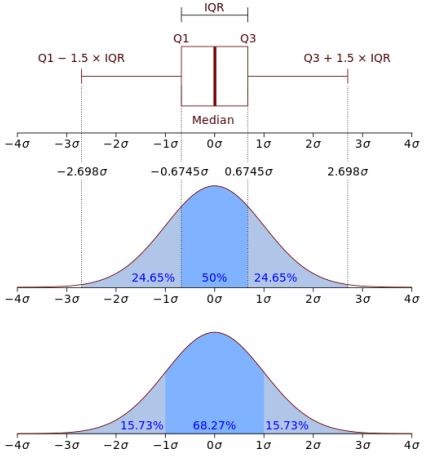


Figure: Box plot outliers in normal distribution data

Standard score

The z-score, also known as the standard score, is a common indicator which measures the deviation of a data point from the whole series data.

$$z = (x - \mu)/\sigma$$

- $z = 3$. The reason is due to the 3σ rule.
- Consider a normal distribution, about 68% of values are within one standard deviation (σ) away from the mean; about 95% of the values lie within 2σ; and about 99.7% are within 3σ.
- This fact is known as the empirical 68-95-99.7 rule, or the 3σ rule.

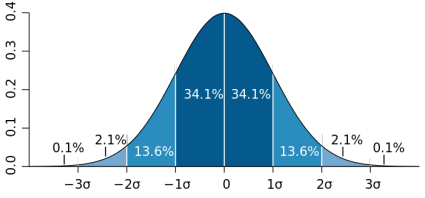


Figure: Outliers in normal distribution data

White noise and randomness

- White noise plays an important role in the time-series analysis.
- Check a series for randomness: How?

White noise

A white noise (WN) series is a stochastic random process, with each stochastic variable is iid (independently and identically) distributed.
The white noise series plays a special role in time-series analysis.

- Predictability: A white noise series has no dependent structure in data, and thus nothing is predictable.
- Model validity: The residual series of a time-series model should be white noise.

The starting point for time-series analysis is that the time series is **not** a white noise series.

A white noise series

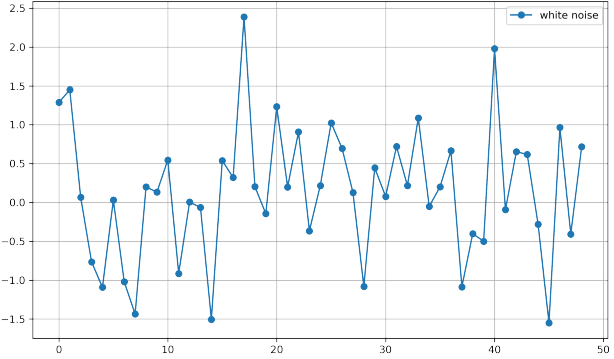


Figure: A white noise series

Histogram of a WN series

We can first draw histogram and density plot of the WN series.

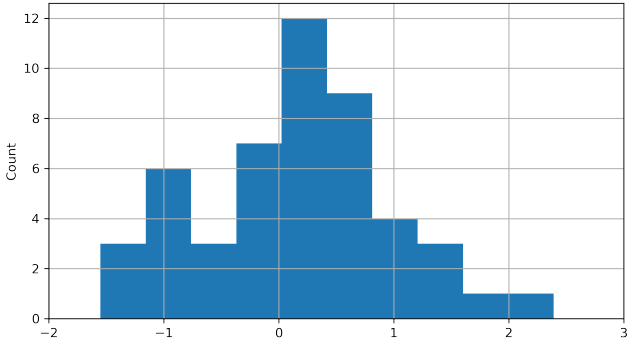


Figure: Histogram of a white noise series

Histogram and density of a WN series

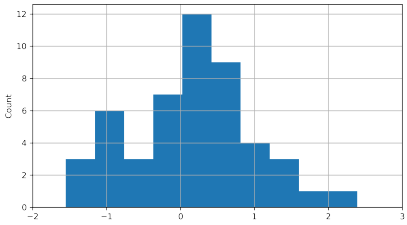


Figure: Histogram

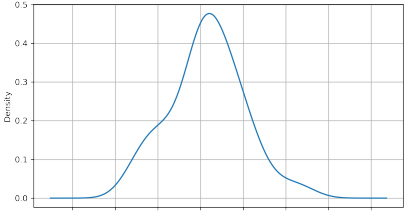


Figure: Density plot

ACF of a white noise series

- Sample autocorrelations for a white noise series are close to zero.
- Sampling distribution of γ_k for white noise series is asymptotically standard Gaussian (normal distribution) $N(0, 1/T)$, where T is number of observations.
- 95% confidence interval: 95% of all γ_k must lie within $\pm 1.96/\sqrt{T}$.
- It is common to plot lines at $\pm 1.96/\sqrt{T}$ when plotting ACF, as they are critical values.
- If this is not the case, the series is probably not white noise.

ACF of White noise series

All autocorrelation coefficients lie within these limits, confirming that the data are white noise. (More precisely, the data cannot be distinguished from white noise.)
Example: T = 49. Critical values at $\pm 1.96/\sqrt{49} = \pm 0.28$.

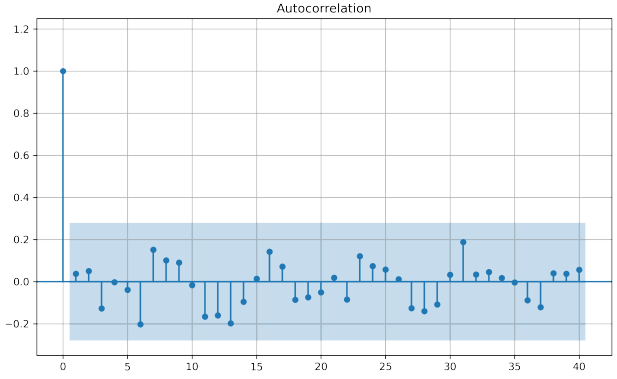


Figure: ACF of a white noise series

PACF of White noise series

All partial autocorrelation coefficients lie within these limits, confirming that the data are white noise. (More precisely, the data cannot be distinguished from white noise.)
Example: T = 49. Critical values at $\pm 1.96/\sqrt{49} = \pm 0.28$.

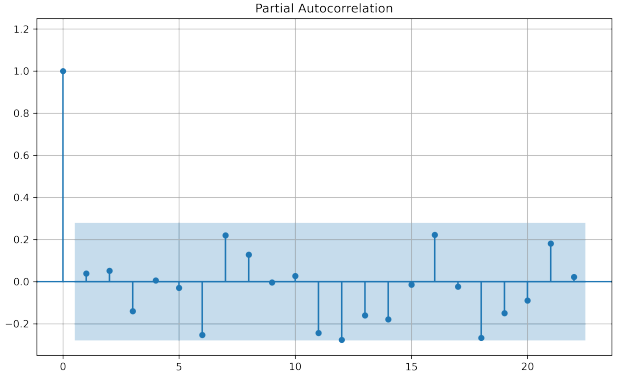


Figure: PACF of a white noise series

ADF test and Lag plot of white noise

- Results of Augmented Dickey-Fuller Test:

Test Statistic	p-value
-6.630003e+00	5.746791e-09

Since p-value < 0.05, we reject the Null hypothesis. Thus the series is stationary.

- To show better the result, the lag plot is generated with 500 random points.

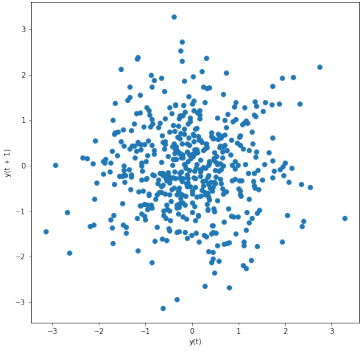


Figure: Lag plot of a white noise

Check for Randomness

Can we check randomness without visualizing graphs?

Yes, the Ljung-Box test is a statistical hypothesis test that checks if auto-correlations are significantly away from zero in a time series. It is often used to test if the residual series after prediction is random noise.

$$Q = T(T+2)\sum_{k=1}^{h}(T-k)^{-1}\gamma_k^2$$

where h is the max lag considered, and T number of observations.

It has the following two hypotheses:

- Null hypothesis (H0): The (residual) series is independently distributed, i.e. random.
- Alternative hypothesis (Ha): The (residual) series is not independently distributed but exhibits serial correlation, thus not random.

Ljung-Box test

To perform the Ljung-Box test, we can use the acorr_ljungbox() function from statsmodels:

```
acorr_ljungbox(x, lags=None)
```

where x is the series and lags the number of lags to test.

https://www.statsmodels.org/dev/generated/statsmodels.stats.diagnostic.acorr_ljungbox.html

Ljung-Box test example

```
import statsmodels.api as sm
data = sm.datasets.sunspots.load_pandas().data
res = sm.tsa.ARMA(data["SUNACTIVITY"], (1,1)).fit(disp=-1)
sm.stats.acorr_ljungbox(res.resid, lags=[10], return_df=True)
       lb_stat     lb_pvalue
10  214.106992  1.827374e-40
```

The lb_stat is the value of the test statistic, and lb_pvalue is the p value of the test.

- If lb_pvalue > 0.05 (default threshold), then accept the Null hypothesis that the series is independent, meaning that the series is random.
- If 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.

In this example, lb_pvalue $<< 0.05$, thus we reject the null hypothesis and accept the alternative hypothesis. The residuals are not random.

Ljung-Box test on the random number series

```
from random import gauss
from random import seed
import pandas as pd
import statsmodels.api as sm

seed(1)
series = [gauss(0, 1) for i in range(1000)]
series = pd.Series(series)
sm.stats.acorr_ljungbox(series, lags=[10], return_df=True)
       lb_stat     lb_pvalue
10     5.723556    0.877927
```

The lb_stat is the value of the test statistic, and lb_pvalue is the p value of the test. Since lb_pvalue > 0.05, we accept the null hypothesis that the series is random.

IPython demo

This lecture uses the following iPython demos under directory IL2233VT22 Lecture_2:

- Global land temperatures. Data_visualization_global_land_temperatures.ipynb
- White noise. white_noise.ipynb

References

- Box, G. E. P.; Jenkins, G. M.; Reinsel, G. C. (2008). Time Series Analysis, Forecasting and Control (4th ed.). Hoboken, NJ: Wiley. ISBN 9780470272848.
- Brockwell, Peter; Davis, Richard (2009). Time Series: Theory and Methods (2nd ed.). New York: Springer. ISBN 9781441903198.