Feature Engineering

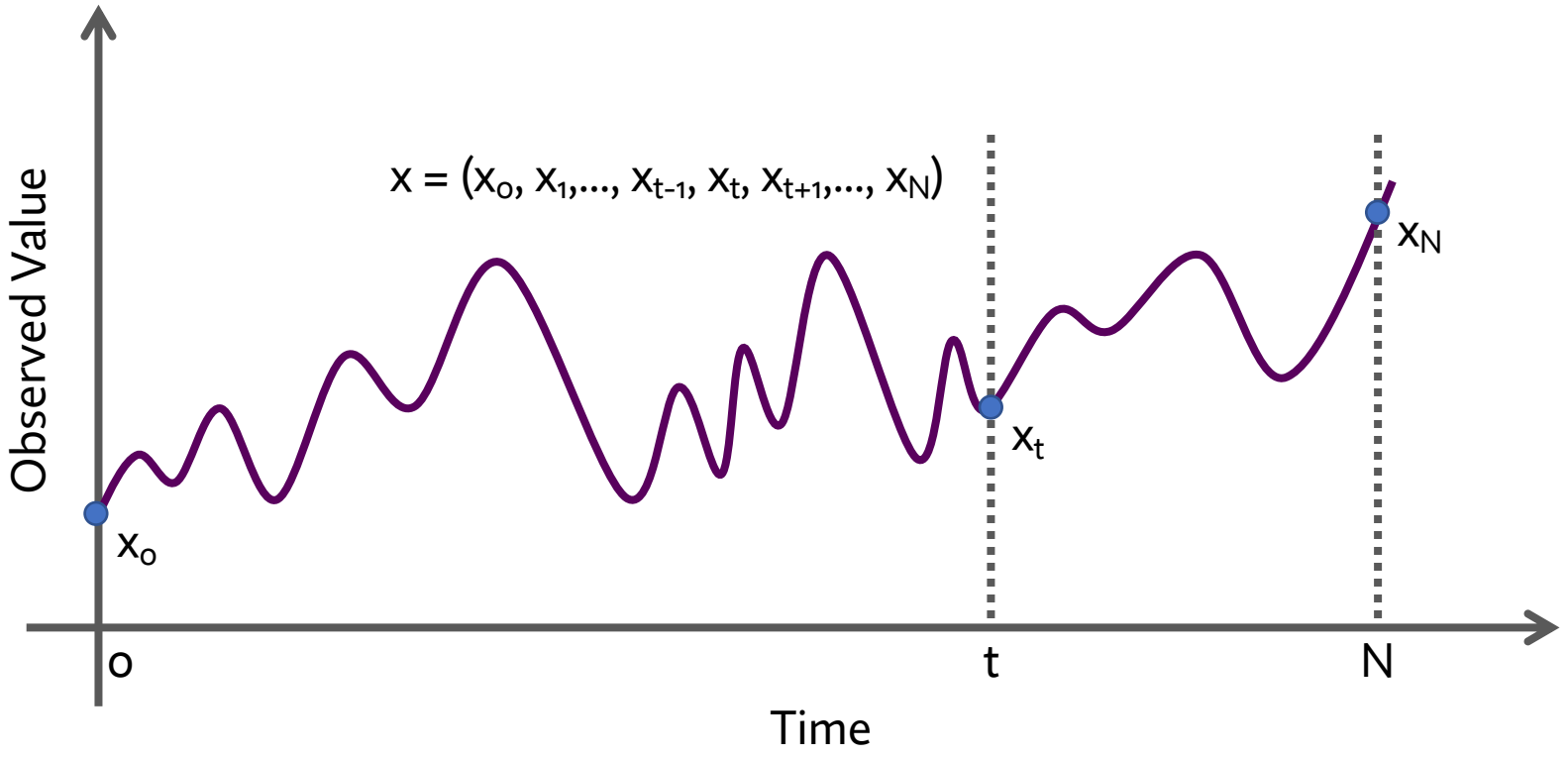
Papangkorn Inkeaw, Ph.D.

Feature Extraction

Chapter 5 (Part III) - Feature Extraction for Time Series Data

Time Series – Basic Knowledge

Time series is a sequence of data points collected over an interval of time.



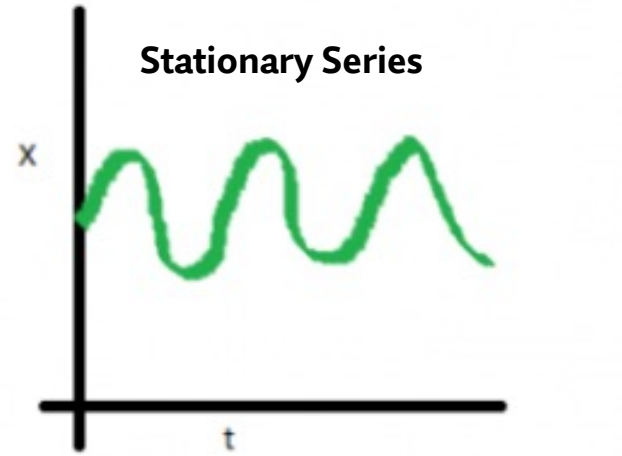
Time Series – Basic Knowledge

Characteristics of Time Series Data

Stationary

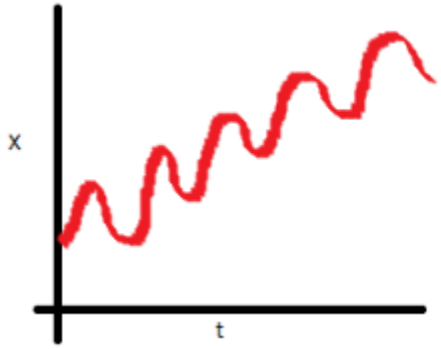
Statistical properties do not change over time.

- Mean
- Variance
- Covariance

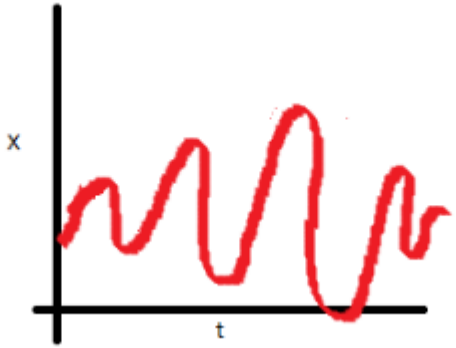


Source: https://medium.com/greyatom/time-series-b6ef79c27d31

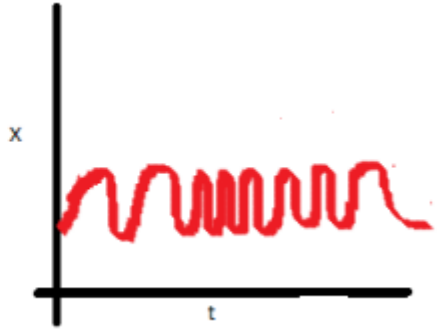
Non-stationary Series



Mean increases with time.



Variance of the series is a function of time.



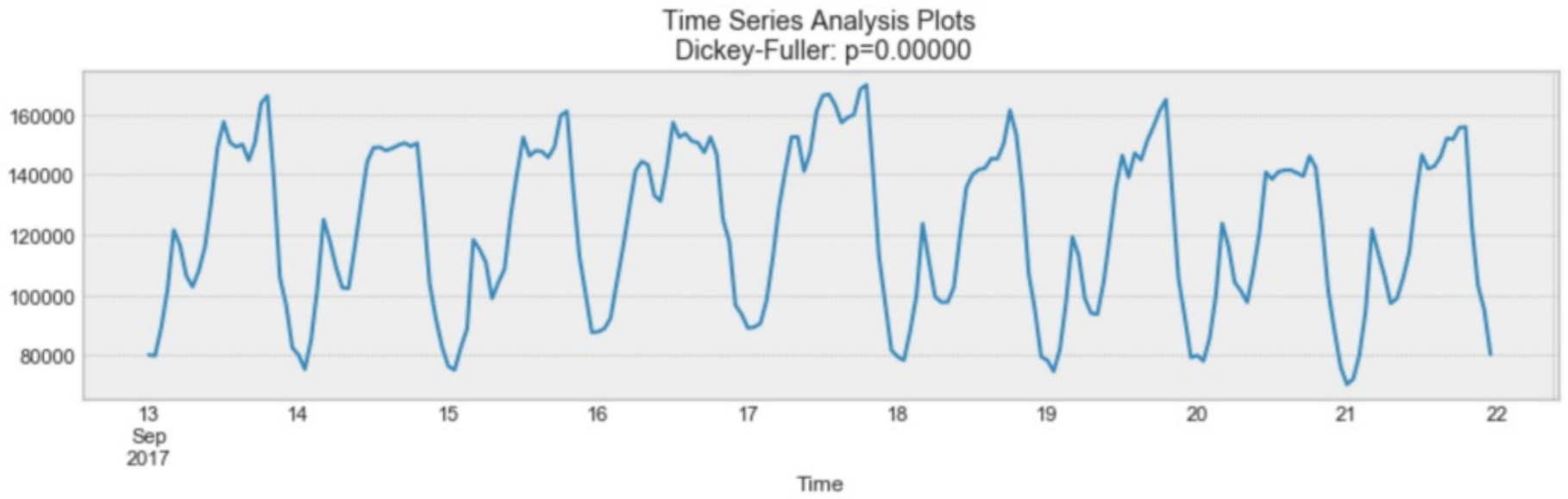
The spread becomes closer as the time increases.

Time Series – Basic Knowledge

Characteristics of Time Series Data

Seasonality

Periodic fluctuations - pattern that recurs or repeats over regular intervals.



Example of seasonality

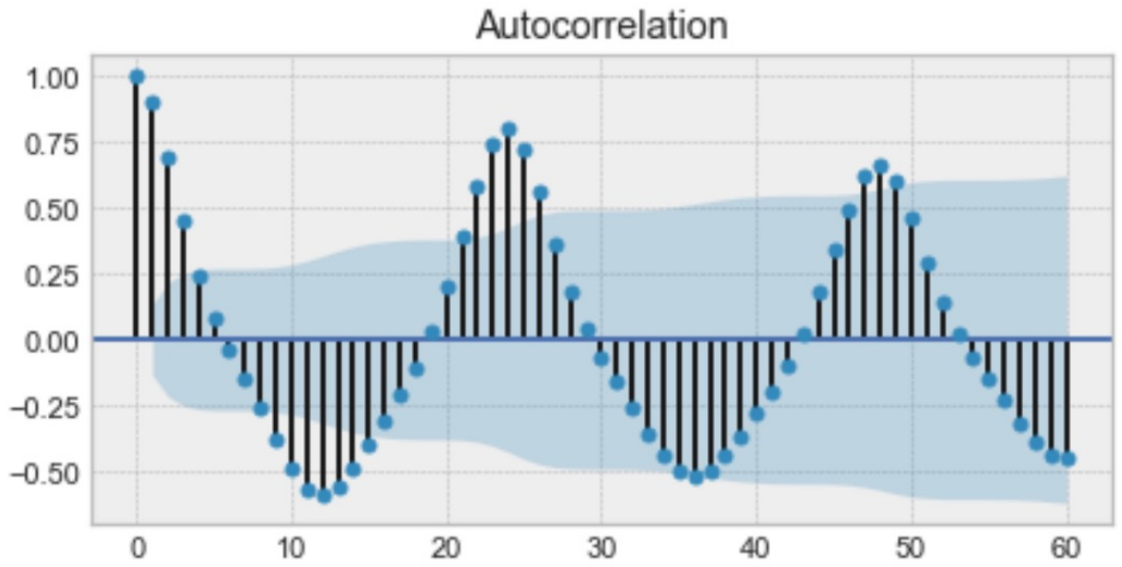
Source: https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775

Time Series – Basic Knowledge

Characteristics of Time Series Data

Autocorrelation

- Internal correlation in a time series.
- The similarity between observations as a function of the time lag between them.



Example of an autocorrelation plot - we will find a very similar value at every 24 unit of time.
Source: https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775

Global Features

Sample Mane and Variance

- Given $x = (x_o, x_1, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_N)$ is a time series of length N.
- The sample mean is:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- The (unbiased) sample variance can be calculated by:

$$s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^2$$

- The mean and variance is independent of the ordering of values in x.

Global Features

Stationary

- Capture how temporal dependences vary over time.
- The mean stationarity can be measured by:

$$StatAv = \frac{\mathrm{std}\left(\left\{\overline{x_{1:w}}, \overline{x_{w+1:2w}}, \ldots, \overline{x_{(m-1)w:mw}}\right\}\right)}{\mathrm{std}(x)}$$

- It divides x into non-overlapping windows of length w.
- The standard deviation is taken across the set of means computed in each window.

Global Features

Autocorrelation

- The correlation between time-series values separated by a given time lag τ.
- It can be estimated by:

$$C(\tau) = \langle x_t, x_{t+\tau} \rangle = \frac{1}{s_x^2 (N - \tau)} \sum_{t=1}^{N-\tau} (x_t - \bar{x})(x_{t+\tau} - \bar{x})$$

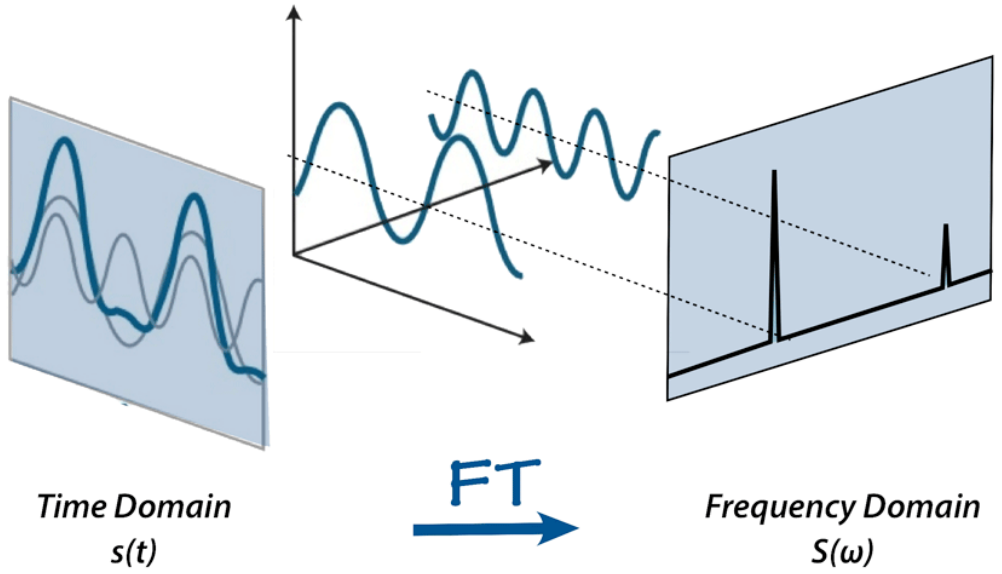
Global Features

Fourier Transform

- Represent the time series as a linear combination of frequency components.
- Each frequency component can be computed by

$$\tilde{x}_k = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} x_n e^{2\pi i k n / N}$$

- $\tilde{x}_k$ composes of the real and complex parts that encode the amplitude and phase of that component.



Source: https://mriquestions.com/fourier-transform-ft.html

Global Features

Entropy

- Quantify predictability in a time series.
- Approximate Entropy is defined as the logarithmic likelihood that the sequential patterns of the data of length *m* that are closed to each other within a threshold *r*:

$$ApEn(m,r) = \Phi^{m}(r) - \Phi^{m+1}(r)$$

where

$$\Phi^{m}(r) = \frac{1}{N-m+1}\sum_{i=1}^{N-m+1} \log C_i^m(r)$$

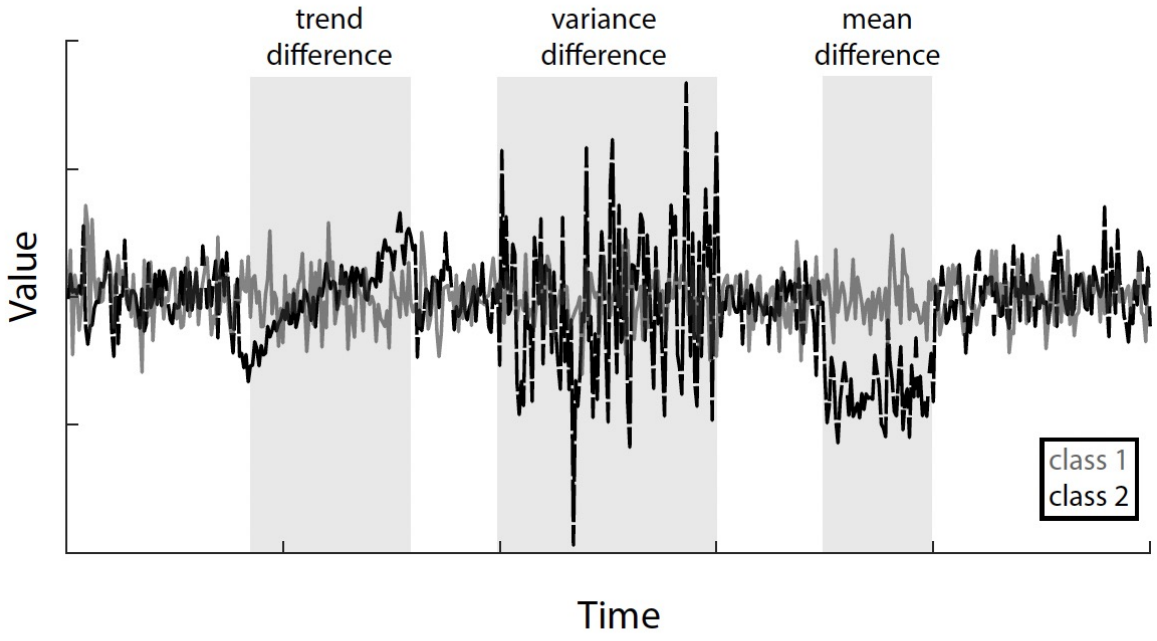
and

$$C_i^m(r) = \frac{number\ of\ u(j)\ such\ that\ d\big(u(i),u(j)\big) \le r}{N-m+1}$$

Subsequence Features

Interval Feature

- Some time-series classification problems may involve class differences in time-series properties that are restricted to specific discriminative time intervals.
- Given $s = (x_k, x_{k+1},..., x_{k+l-1})$ is a subsequence of length l taken from a time series x.
- Simple features, such as mean and standard deviation, can be used to represent the signal on an interval.
- Dealing with whole time-series data:
  - Divide a time series into overlapping/non-overlapping windows.
  - Random sampling of time intervals and accumulate classifiers.



Source: Guozhu Dong and Huan Liu. (2020). *Feature Engineering for Machine Learning and Data Analytics*. CRC Press.

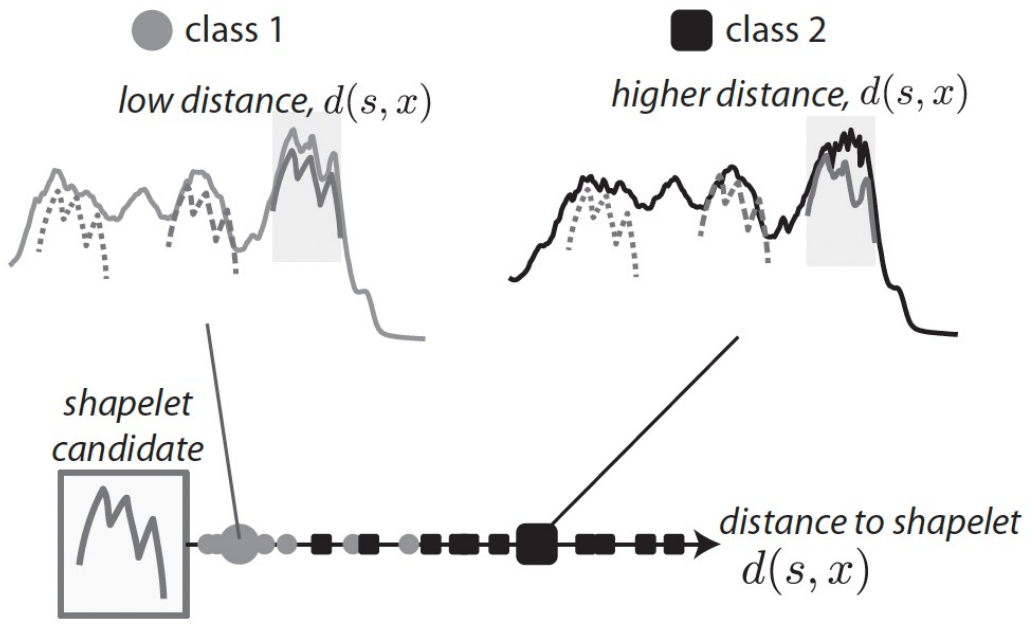
Subsequence Features

Shapelets

- A **shapelet** is defined as a contiguous subsequence of a time series.
- The distance between a shapelet and a time series is defined as:

$$d(s, x) = \min_k d(s, x_{k:k+l})$$

- The distance can be thought of as a "feature" extracted from the time series.
- How to determine shapelets:
  - Searching all possible candidate subsequences or random sampling subsequences
  - Select most discriminative shapelets given a criterion (mutual information or F-scores)



Subsequence Features

Pattern Dictionaries

- Similar to the bag-of-words representation of text.
- Given a set of subsequence patterns.
- A time series is represented by a histogram that counts the number of matches to the given set of subsequence patterns.
- The time series is firstly transformed into sequences of symbols using:
  - Bag of words transformation.
  - Symbolic-Fourier-Approximation Symbols.
  - Word ExtrAction for time SEries cLassification.

References & Study Resources

- Guozhu Dong and Huan Liu. (2020). *Feature Engineering for Machine Learning and Data Analytics*. CRC Press.
- https://pyts.readthedocs.io/en/stable/modules/transformation.html
- https://mriquestions.com/fourier-transform-ft.html
- https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775
- https://medium.com/greyatom/time-series-b6ef79c27d31