Online Machine Learning

Time Series Analysis and Forecasting

Antonio Carta

antonio.carta@unipi.it

Outline

change of scenario: we have the entire time series at once and we want to detect, model, or remove the non-stationarity.

- definition of stationarity
- modeling and removing trends and seasonality
- TS forecasting

TS Analysis and Forecasting

Today, we talk only about **time series**

- Offline setting: we have already collected the entire time series
- Example: financial data, weather forecasting, …
- how do we model nonstationarity in the offline setting?

Problems:

- Time Series **Analysis**: explaining the past, understanding seasonality, finding patterns, …
- Time Series **Forecasting**: predicting future values

Stochastic Processes and Stationarity

Stochastic Process

- **definition**: a stochastic process is a collection of random variables indexed by some set (time in TSA)
- notation: $\{X_t\}_{t \in T}$

We need some structure/property to be able to model time series.

Strong Stationarity

given

- $\{X_t\}_{t \in T}$ stochastic process
- $F_X\left(x_{t_1+\tau}, \dots, x_{t_n+\tau}\right)$ its cumulative distribution function

Def. Strong stationarity:

$F_X\left(x_{t_1+\tau}, \dots, x_{t_n+\tau}\right) = F_X\left(x_{t_1}, \dots, x_{t_n}\right) \quad$ for all $\tau, t_1, \dots, t_n \in \mathbb{R}$ and for all $n \in \mathbb{N}_{>0}$

intuitively: the probability distribution does not depend on time.

Weak Stationarity

- $m(t)$ = expected value at time t
- $K(t_1, t_2)$ = covariance between values at t_1 and t_2

Properties:

- Mean is constant (time-independent)
- Covariance only depend on the lag t1 – t2
- Finite second moment

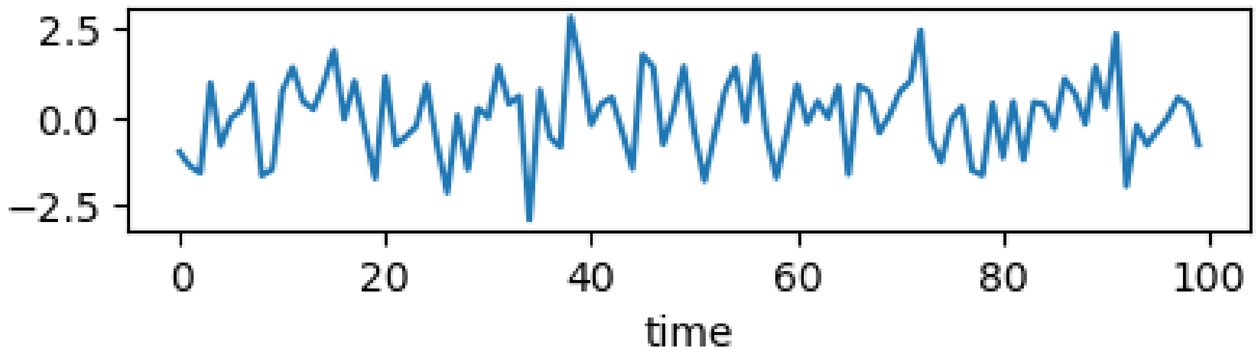
$$
\begin{aligned}
&m_X(t) = m_X(t+\tau) && \text{for all } \tau, t \in \mathbb{R} \\
&K_{XX}(t_1, t_2) = K_{XX}(t_1 - t_2, 0) && \text{for all } t_1, t_2 \in \mathbb{R} \\
&\mathrm{E}[|X_t|^2] < \infty && \text{for all } t \in \mathbb{R}
\end{aligned}
$$

Why do we care?

- Many forecasting models assume stationarity
  - We need our TS to be stationary
  - If they are not, we want to make them stationary via preprocessing
- stationarity $\rightarrow$ predictable

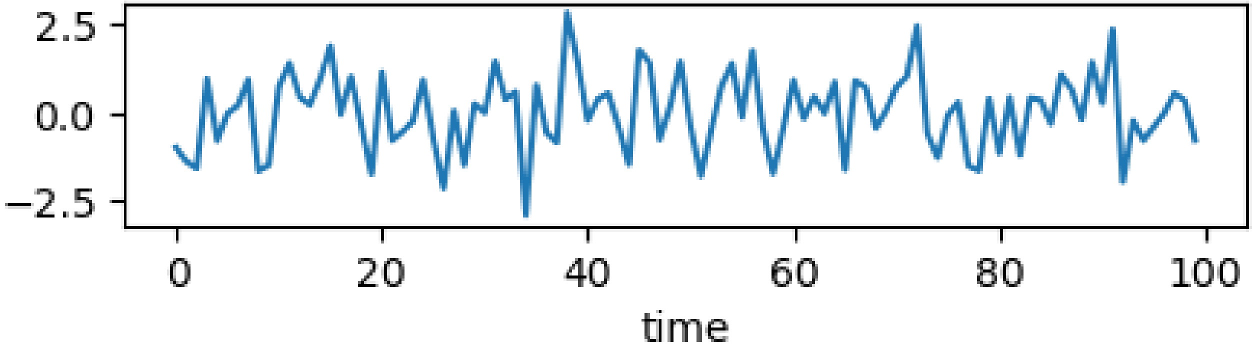
Example - White Noise

- a time series with constant mean and variance and no seasonality is stationary
- example: white noise

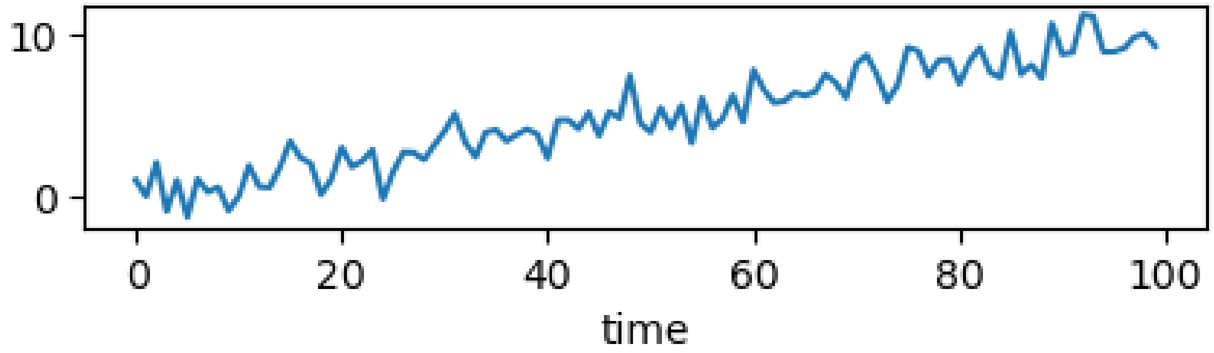


Example - Linear Trend

Stationary:

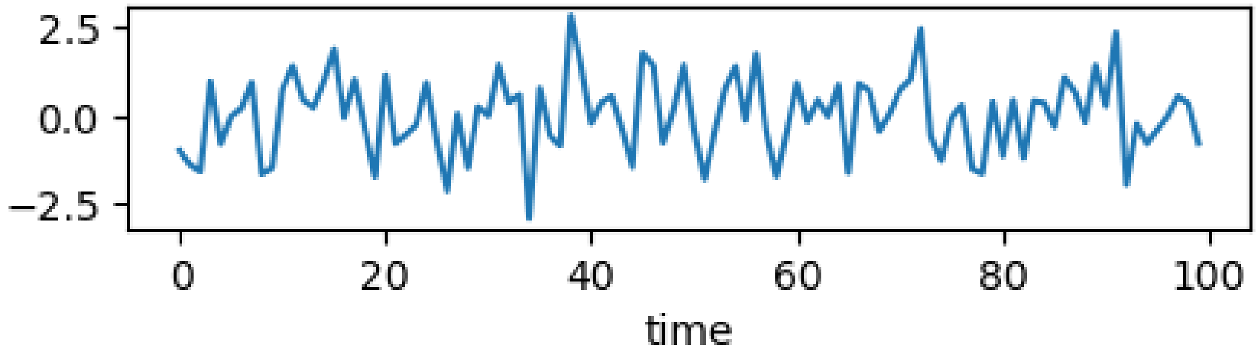


Non-constant mean (trend):

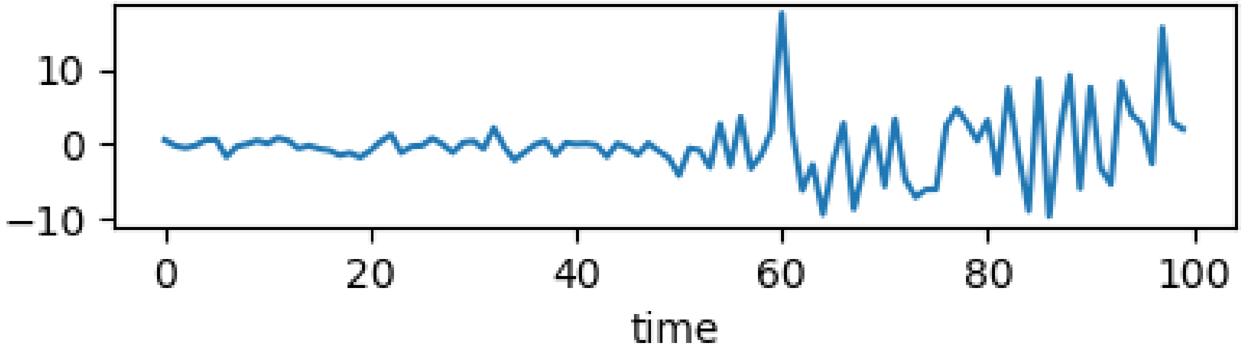


Example - Non-constant Variance

Stationary:

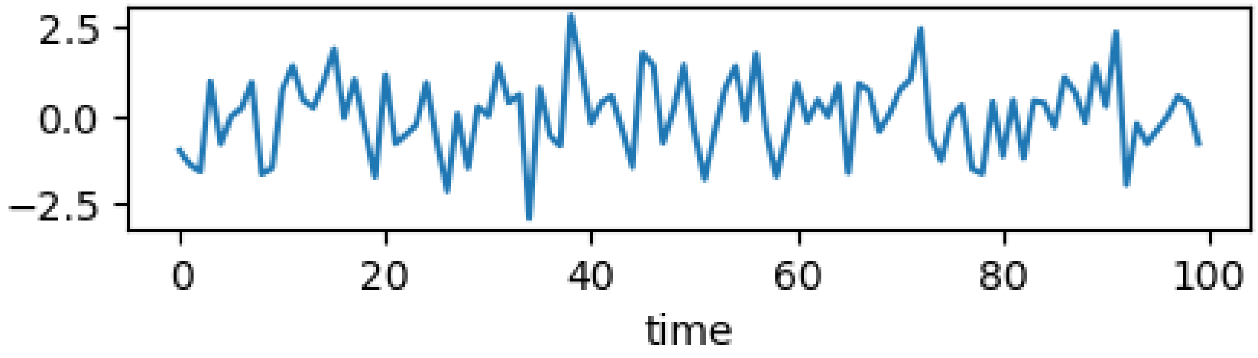


Non-constant Variance:

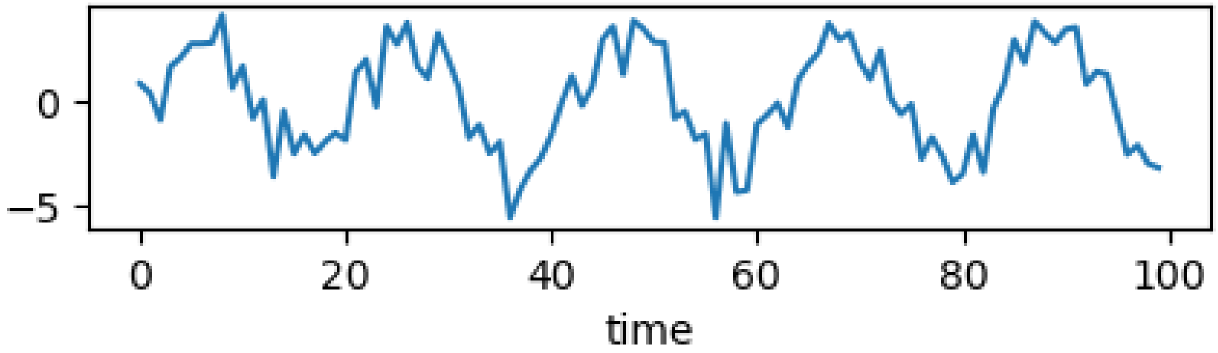


Example - Seasonality

Stationary:



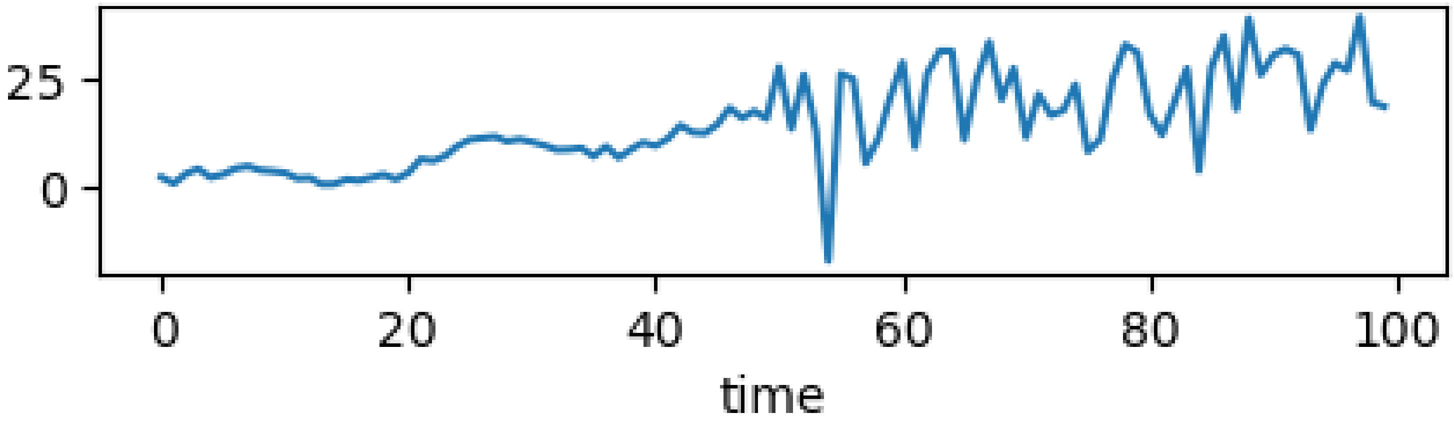
With Seasonality



Nonstationarity - Putting it together

often, nonstationary time series have all of these components mixed together:

- non-constant mean
- non-constant variance
- seasonality



Testing for Non-stationarity

Nonstationarity Tests - Ideas

Sometimes, nonstationarity is obvious

- plot the time series: do you see any trends/seasonality?
- compare mean and variance for different chunks of the time series

Nonstationarity - Statistical Tests

More principled approach: statistical tests

- examples on the notebook
- Unit Root Tests
  - ADF Test
  - KPSS Test

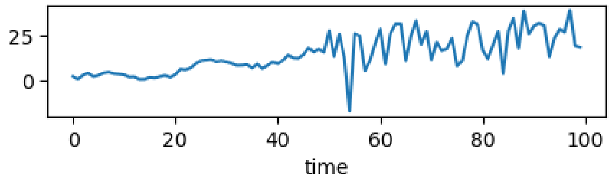
Decomposition and Detrending

TS Decomposition

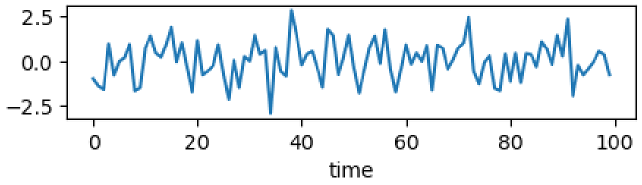
IDEA: we want a stationary TS. Can we remove the nonstationarity?

- we can model the different forms of nonstationarity (mean, variance, seasonality)
- remove nonstationarity from the raw TS via preprocessing

Raw TS



Residual TS after decomposition



TS Components

TS components: trend x seasonal x residual

- **trend**: persistent and long-term change
- **seasonal**: periodic fluctuations
- **residual**: stationary component

RAW TS = $T \odot S \odot R$ ← what operator can we use to combine the different components?

Additive and Multiplicative Models

- given:
  - m_t trend component
  - s_t seasonal component
  - Y_t residual component
- **additive** model: $X_t = m_t + s_t + Y_t$
- **multiplicative** model: $X_t = m_t s_t Y_t$
- additive model assumes the seasonal is approximately constant over time
- multiplicative model is better when the seasonal component changes over time according to the general trend

Example

- example on notebook using statsmodels
- comparison between additive and multiplicative model

Non-seasonal Decomposition

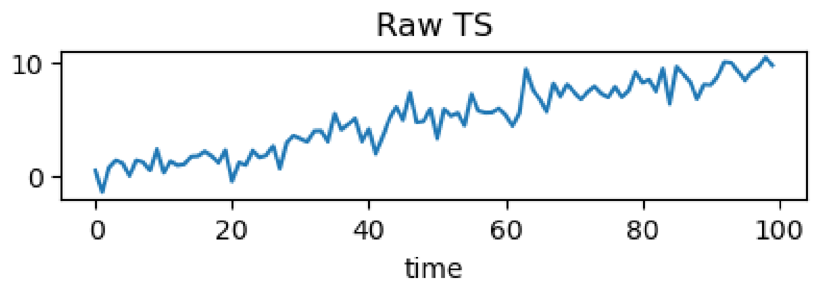
IDEA: let's ignore the seasonal component

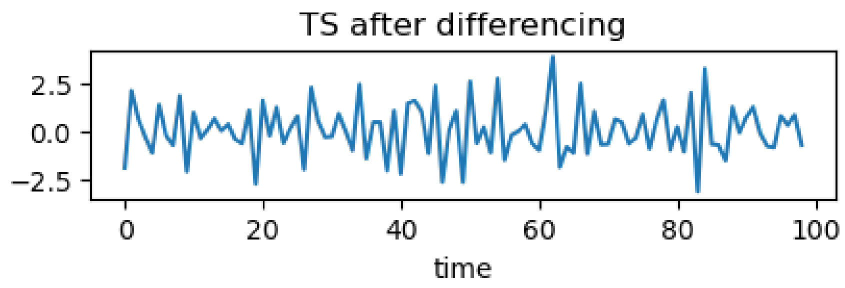
We want to model the trend only

- Trend elimination via differencing
- Trend estimation

Trend Elimination - Differencing

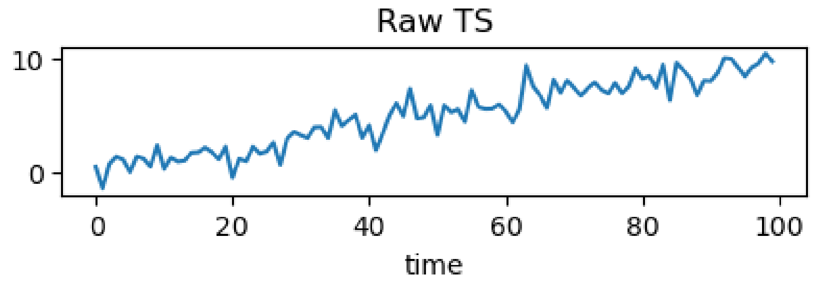
- simple method to remove trends
- the value of the new TS are the difference between consecutive methods
- new TS: $Y_t = x_t - x_{t-1}$
- **question:** what kind of trends are we removing with this method?

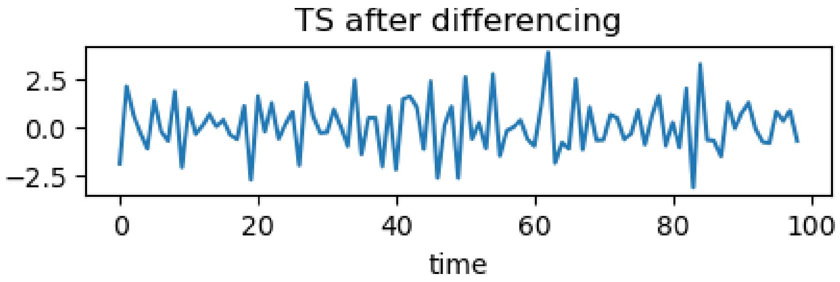




Difference Removes Linear Trend

- differencing removes linear trend
- consequence: if you have a polinomial trends (degree n) you need to difference n times





Trend Estimation

Trend Estimation

- you have a trend model (e.g. linear or polinomial trend)
- fit the model to your data
- remove the trend

Example with linear trend:

- x_t original TS, Y_t detrended TS, $\overline{x}_t$ trend component
- trend model: $\overline{x}_t = at + b$
  - a, b parameters
  - $x_t = Y_t + \overline{x}_t$
- fit a, b
- remove the trend: $Y_t = x_t - \overline{x}_t$

Seasonality Removal -Seasonal Differencing

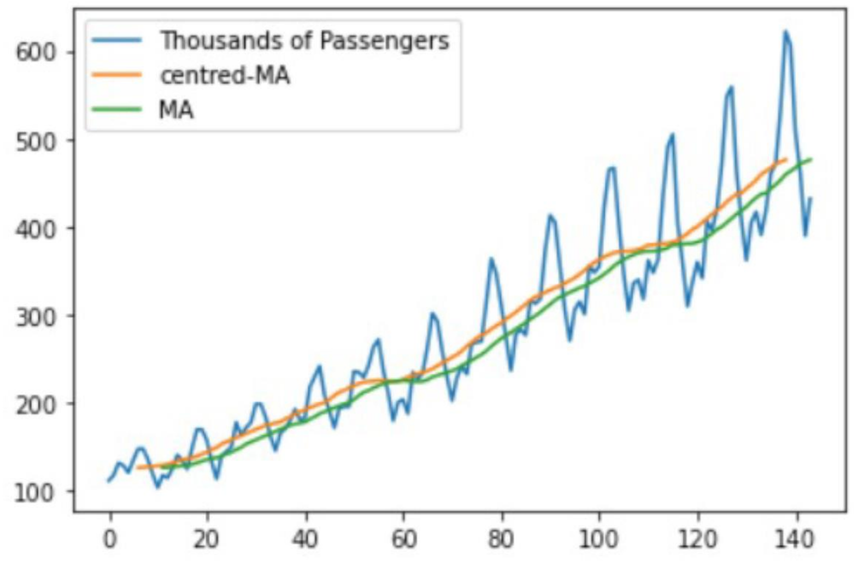
- we assume the seasonal component has a fixed and known period d
- **example**: the periodic behavior due to the calendar (weekly/daily…) has a known period

Seasonal Differencing:

- remove the trend
- apply differencing with period d: $Y_t = x_t - x_{t-d}$

Trend with (Centered) Moving Average

- **Moving average**: estimate average at time t using the last q elements
- **Centered MA**: estimate average at time t using a windows centered around t



Centered moving average:

Even:

$$\widehat{m_t} = \left(0.5x_{t-q} + x_{t-q+1} + \cdots + x_{t+q-1} + 0.5x_{t+q}\right)/d$$

Odd:

$$\widehat{m_t} = \left(x_{t-q} + x_{t-q+1} + \cdots + x_{t+q-1} + x_{t+q}\right)/d$$

First and last q elements are ignored

Seasonality with Moving Averages

For a known period d

- Detrend the TS
- Divide the TS in windows of length d
- Compute $w_k, k = 1, \ldots, d$ as the average of all the windows in position k
- $\widehat{s_k} = w_k - 1/d \sum_{j=1}^{d} w_j \, , k = 1, \cdots, d$ <- subtract the average

Forecasting

Forecasting

Problem:
- given a time series $x_1, \ldots, x_t$, predict x_{t+k}
- one-step $k = t + 1$, but it can be multi-step.

what do you want to predict?
- trend
- seasonal component
- residual: short-term variations

two approaches:
- first, preprocess to make the TS stationary, then train the forecasting model
- train the forecasting model directly

Evaluation

- mean absolute error $MAE = \sum_i \frac{|Y_i - \widehat{Y}_i|}{n}$
- mean absolute percent error $\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \frac{|Y_i - \widehat{Y}_i|}{Y_i}$
- mean squared error $\text{MSE} = \sum_{i=1}^{n} \frac{\left(\text{Y}_i - \widehat{\text{Y}}_i\right)^2}{n}$
- root MSE $\text{RMSE} = \sqrt{\text{MSE}}$

Simple Baselines

- **average**: $\widehat{y}_{t+1} = \frac{1}{N}\sum_i y_i$
- **window-based**: $\widehat{y}_{t+1} = \frac{1}{k}\sum_{i=0}^{k-1} y_{t-i}$.
- **last value**: $\widehat{y}_{t+1} = y_t$

Exponential Smoothing

Exponential Smoothing

exponential smoothing methods are weighted averages of past observations, with the weights decaying exponentially as the observations get older

- generates reliable forecasts quickly
- we will see methods that don't need a preliminary TS decomposition

simple exponential smoothing (SES)

forecasting data with no clear trend or seasonal pattern

SES Equation

$\widehat{y}_{T+1|T} = \alpha y_T + (1-\alpha)\,\widehat{y}_{T|T-1}$

Component form

- Forecast equation $\widehat{y}_{t+h|t} = \ell_t$
- Smoothing equation $\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}$

parameters: α and l_0

With Trends - Holt's Linear Trend Method

Forecast equation $\hat{y}_{t+h|t} = \ell_t + hb_t$

Level equation $\ell_t = \alpha y_t + (1 - \alpha)(\ell_{t-1} + b_{t-1})$

Trend equation $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1}$

level, b_t trend, α level smoothing, β^* trend smoothing

Damped trends

- Holt's method assumes a trend constant goes on indefinitely
  - tends to overestimate the real trend
  - solution: trend dampening

$$
\begin{aligned}
\hat{y}_{t+h|t} &= \ell_t + \left(\phi + \phi^2 + \cdots + \phi^h\right) b_t \\
\ell_t &= \alpha y_t + (1-\alpha)(\ell_{t-1} + \phi b_{t-1}) \\
b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}
\end{aligned}
$$

- parameter $\phi \in [0,1]$ controls dampening factor
- $\phi = 1 \rightarrow$ Holt's method

Damped Trends

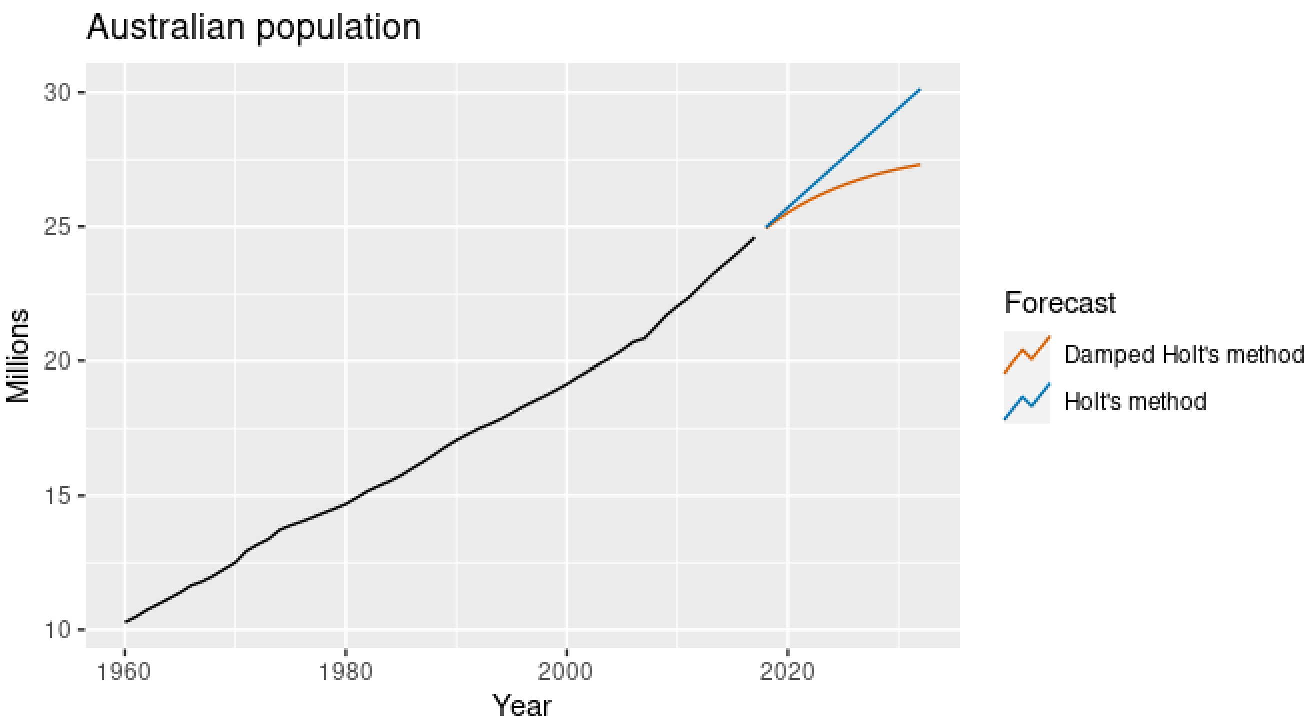


figure from https://otexts.com/fpp3/holt.html

With Seasonality - Holt-Winters Method

- add seasonal component s_t with smoothing factor γ
- $m =$ seasonality period
- two methods:
  - additive: assuming roughly constant seasonal variations
  - multiplicative: seasonal variations are proportional to the level

Holt-Winters Additive model

$$
\begin{aligned}
\hat{y}_{t+h|t} &= \ell_t + hb_t + s_{t+h-m(k+1)} \\
\ell_t &= \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \\
s_t &= \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m},
\end{aligned}
$$

where $k = floor(\frac{h-1}{m})$

Holt-Winters - Multiplicative model

$$
\begin{aligned}
\hat{y}_{t+h|t} &= (\ell_t + hb_t)s_{t+h-m(k+1)} \\
\ell_t &= \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \\
s_t &= \gamma \frac{y_t}{(\ell_{t-1} + b_{t-1})} + (1-\gamma)s_{t-m}.
\end{aligned}
$$

Dampened Holt-Winters

with a damped trend and multiplicative seasonality (additive is also possible):

$$
\begin{aligned}
\hat{y}_{t+h|t} &= \left[\ell_t + \left(\phi + \phi^2 + \cdots + \phi^h\right) b_t\right] s_{t+h-m(k+1)} \\
\ell_t &= \alpha(y_t / s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1}) \\
b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1} \\
s_t &= \gamma \frac{y_t}{(\ell_{t-1} + \phi b_{t-1})} + (1-\gamma) s_{t-m}.
\end{aligned}
$$

Taxonomy

Table 8.5: A two-way classification of exponential smoothing methods.

Trend Component	Seasonal Component		
	N (None)	A (Additive)	M (Multiplicative)
N (None)	(N,N)	(N,A)	(N,M)
A (Additive)	(A,N)	(A,A)	(A,M)
A_d (Additive damped)	(A_d,N)	(A_d,A)	(A_d,M)

Some of these methods we have already seen using other names:

Short hand	Method
(N,N)	Simple exponential smoothing
(A,N)	Holt's linear method
(A_d,N)	Additive damped trend method
(A,A)	Additive Holt-Winters' method
(A,M)	Multiplicative Holt-Winters' method
(A_d,M)	Holt-Winters' damped method

Figure from https://otexts.com/fpp3/taxonomy.

full equations for all the 9 methods in the reference

Conclusion

Take-home Messages

- even in the offline setting you still need to be aware of nonstationarity
- for TS **analysis**: trends and seasonality are fundamental properties of TS
- in **forecasting** probems: you can model it and remove it to improve the performance of your model

References

References:

- notebooks and slides from the TSA module of "Streaming Data Analytics" course https://github.com/emanueledellavalle/streaming-data-analytics by Emanuele Della Valle and Giacomo Ziffer
  - Slower pace than this lecture and notebooks that you can use to play with the models
- **Online Book**: Forecasting: Principles and Practice (3rd ed) Rob J Hyndman and George Athanasopoulos free html version
  - Chapter 3: TS Decomposition
- Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics) 4th ed. 2017 Edition

Next lecture

We start the **Knowledge Transfer and Adaptation** module

- Deep neural networks
- Trained on multiple tasks
- Main question: how to **generalize between different tasks** effectively (and efficiently)