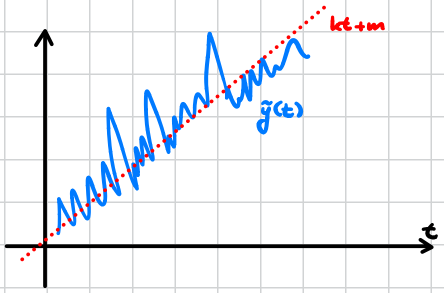
25/03 - Data preprocessing

What if $y(t)$ is not stationary? For example: Trend, seasonality are two source of non-stationarity

To work with non-SSP we first have need to estimate possible trends as seasonalities, then we have to remove them from the original signal and work with the remaining SSP.

One possible application could be forecasting using trend removal

Trend removal



If $y(t)$ is not an SSP we will rewrite it as linear combination of two part: a trend $(kt+m)$ and a SSP $\tilde{y}(t)$

$$y(t) = (kt+m) + \tilde{y}(t)$$

The only available information is the dataset

$$D_N = \{ y(1), \dots, y(N) \}$$

What if I need to compute $\hat{y}(t+1|t)$?

> Algorithm: to forecasting $\hat{y}(t+1|t)$
> - estimate m and k characterizing the trend $\rightarrow (\hat{k}, \hat{m})$ ← This is a critical part
> - $\hat{\tilde{y}}(t) = y(t) - (\hat{k}t + \hat{m})$
> - $\hat{\tilde{y}}(t+1|t)$ from the formulas
> - $\hat{y}(t+1|t) = \hat{\tilde{y}}(t+1|t) + (\hat{k}(t+1) + \hat{m})$

To estimate $\hat{m}$ and $\hat{k}$ we notice that $E[y(t) - (kt+m)] = E[\tilde{y}(t)] = 0$, so inspired by this observation, we can get

$$(\hat{m}, \hat{k}) = \min_{k,m} \frac{1}{N} \sum_{t}^{N} (y(t) - kt - m)^2$$

This is a least-squares problem (remember: in ARX identification $\frac{1}{N} \sum (y(t) - \vartheta^T \varphi(t))^2$. Now $\vartheta = [k \ m]^T$ and $\varphi(t) = [t \ 1]^T$)

So you obtain the following system

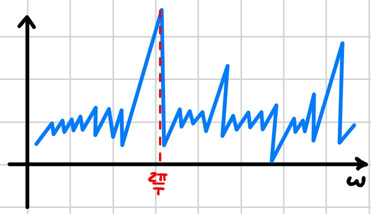
$$\begin{bmatrix} \hat{k} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} \sum t^2 & \sum t \\ \sum t & N \end{bmatrix}^{-1} \begin{bmatrix} \sum t\, y(t) \\ \sum y(t) \end{bmatrix}$$

This work with different slope of the trends (written in linear term)

$$y(t) = \tilde{y}(t) + \underline{s(t)} \quad \text{not linear}$$

Seasonality removal

Now consider the output process as $y(t) = \tilde{y}(t) + s(t)$ with $s(t)$ being non linear. Assume that $s(t) = s(t + Tk)$ with $k \in \mathbb{Z}$ and T the seasonality period. Usually T is given; otherwise we would look at the spectrum of $y(t)$



In order to forecast $\hat{y}(t+1|t)$ we first need to estimate $\hat{s}(t)$, then we will remove it from $y(t)$ and then we could work with the estimated SSP $\hat{\tilde{y}}(t)$. How to estimate $\hat{s}(t)$? We assume that in the dataset we have M periods, so the predicted $\hat{s}(t)$ is

$$\hat{s}(t) = \frac{1}{M} \sum_{h=0}^{M-1} y(t + hT) \qquad \forall t = 1, \dots, T$$

where we should have $MT \leq N$. The underlying idea is:

$$\hat{s}(t) = \frac{1}{M} \sum y(t+hT) = \frac{1}{M} \sum (\tilde{y}(t+hT) + s(t+hT))$$
$$= \frac{1}{M} \sum \tilde{y}(t+hT) + \frac{1}{M} \sum s(t+hT)$$

where we have that

- if M is large: $1/M \sum \tilde{y}(t+hT) \approx E[\tilde{y}(t)] = 0$
- being $s(t) = s(t+hT)$ $\forall h \in \mathbb{Z}$ we can write
  $1/M \sum s(t+hT) = 1/M \sum s(t) = \frac{1}{M} M s(t) = s(t)$

So we can say that, if $M \to +\infty$, $\hat{s}(t) \to s(t)$