What is a Time Series

- a sequence of values indexed by time
  - ↳ order is important
  - ↳ time interval is important.

index → t	x	
1	2	$=1^2+1$
2	5	$=2^2+1$
3	10	$=3^2+1$
4	17	$=4^2+1$
5	26	⋮
6	37	
⋮		
100	10001	

- differences with Regression / Classification
  - order cannot be changed
  - future observations depend on past history
  - rows are not independent of each other

Structure of Time Series

t	x_1	x_2	y
0	x_{10}	x_{20}	y_0
1	x_{11}	x_{21}	y_1
2	x_{21}	x_{22}	y_2
			y_3

index

$\underbrace{x_1 \quad x_2}_{\text{exogenous variables}}$

→ univariate TS

→ target

↳ if present, then multivariate time series

$$y_t = f(\underbrace{y_{t-k}}_{\text{depends on historical target variables}}, \underbrace{x_{1,t-k}, x_{2,t-k}}_{\text{depends on historical exog var.}})$$

→ to find this f. ← Time series problem

$$y_t = f(x_{1t}, x_{2t})$$ → Regression / classification problem.

Else, univariate time series (t, y_t)

Components of Time Series

residuals

$$y_t \equiv T_t + S_t + L. + X_t + \epsilon_t$$

trend (long term)

DES

seasonality (short term)

econometrics

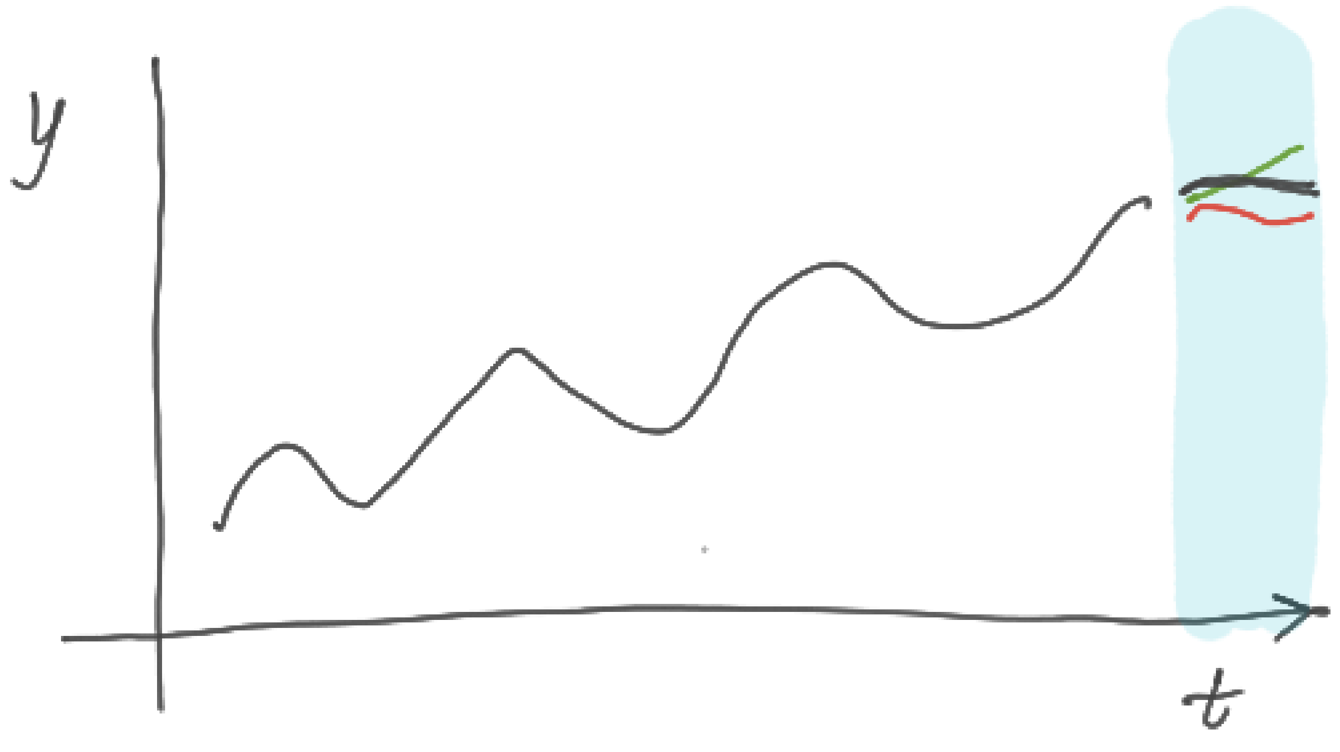
(constant)

Moving Avg

SES

(exogenous factors)

TS Prediction & Evaluation



How to find which forecast is better and by how much?

$$y_{actual} = [15, 17, 14, 13]$$

$$y_{p1} = [13, 15, 12, 13]$$

$$y_{p2} = [14, 15, 16, 17]$$

- Mean Abs Error

$$\frac{1}{4}\left[|15-13| + |17-15| + \cdots \cdots\right]$$

- Mean Abs % Error

$$100 \times \frac{1}{4}\left[\frac{|15-13|}{15} + \frac{|17-15|}{17} + \cdots\right]$$

TS Prediction : Smoothing

t	y
0	y_0
1	y_1
2	y_2
⋮	
154	y_{154}
155	?

1. Forecast for tomorrow is the same as today's observation

$$y_{t+1} = y_t$$

2. Forecast for tomorrow depends on today, yesterday and day before yesterday .... values

$$y_{t+1} = f(\underbrace{y_t, y_{t-1}, y_{t-2} \cdots y_{t-k}}_{k \text{ numbers}})$$

moving avg $\leftarrow$ $y_{t+1} = \frac{1}{k}(y_t + y_{t-1} + y_{t-2} + \cdots y_{t-k})$

Weighted Moving Average

$$y_{t+1} = w_0 y_t + w_1 y_{t-1} + w_2 y_{t-2} + w_3 y_{t-3}$$

$$\begin{cases} w_0 > w_1 > w_2 > w_3 \\ w_0 + w_1 + w_2 + w_3 = 1 \end{cases}$$

Exponential smoothing.

$$y_{t+1} = \alpha y_t + \alpha(1-\alpha) y_{t-1} + \cdots$$

↪ weights in GP ⇒ Simple Exp. Smoothing.

$\alpha \uparrow$ ⇒ higher responsiveness of forecasts $\alpha \in [0, 1]$

Double Exp Smoothing

$\alpha \uparrow$ ⇒ higher responsiveness of level

$\beta \uparrow$ ⇒ " " " trend

$$\Delta y_t = y_t - \underbrace{y_{t-1}}_{\text{lag } 1}$$

$$\Delta_s y_t = y_t - \underbrace{y_{t-24}}_{\text{lag } 24}$$

help make TS stationary

auto-correlation fn

ACF : plot k vs $\text{corr}(y_t, y_{t-k})$

PACF : plot k vs α_k

partial auto-correlation fn

$$y_t = \alpha_1 y_{t-1} + \epsilon_t$$

$$y_t = \beta_1 y_{t-1} + \alpha_2 y_{t-2} + \epsilon_t$$

t	y	
0	y_0	
1	y_1	y_0
2	y_2	y_1
$\vdots$		$\vdots$
k	y_k	y_{k-1}
$k+1$	y_{k+1}	y_k
$\vdots$	$\vdots$	$\vdots$

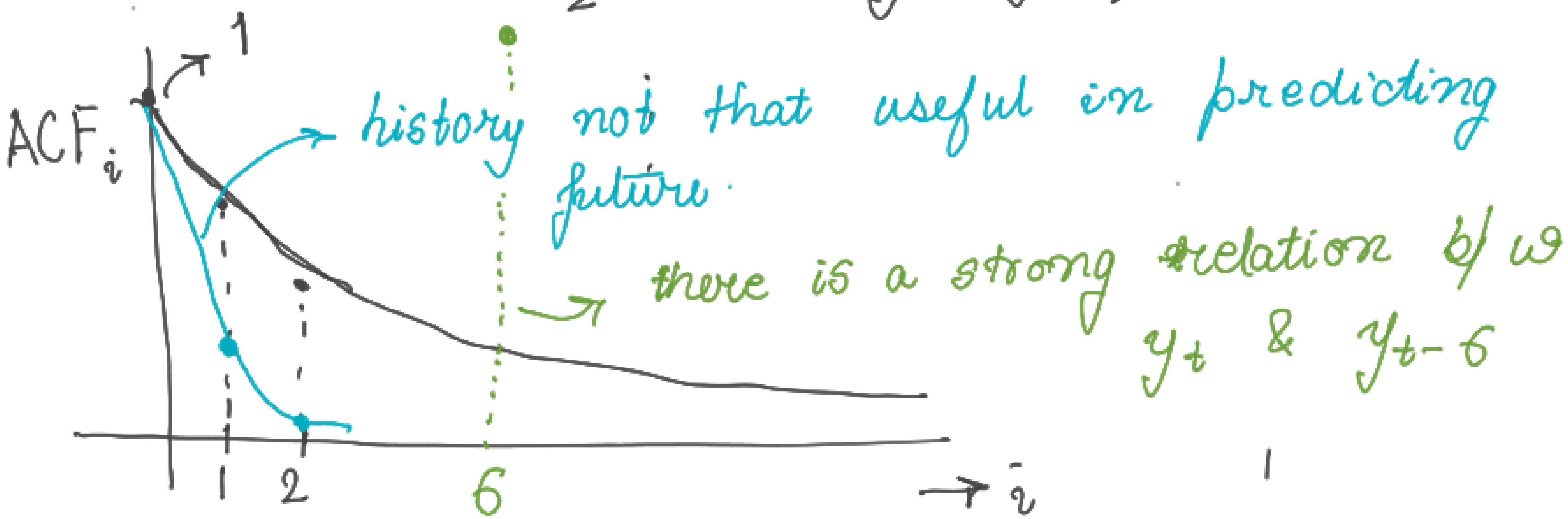
y → univariate time series

shifted by 1

t : $y_0 \quad y_1 \quad y_2 \; \cdots \; y_k \quad y_{k+1} \; \cdots \; y_n$

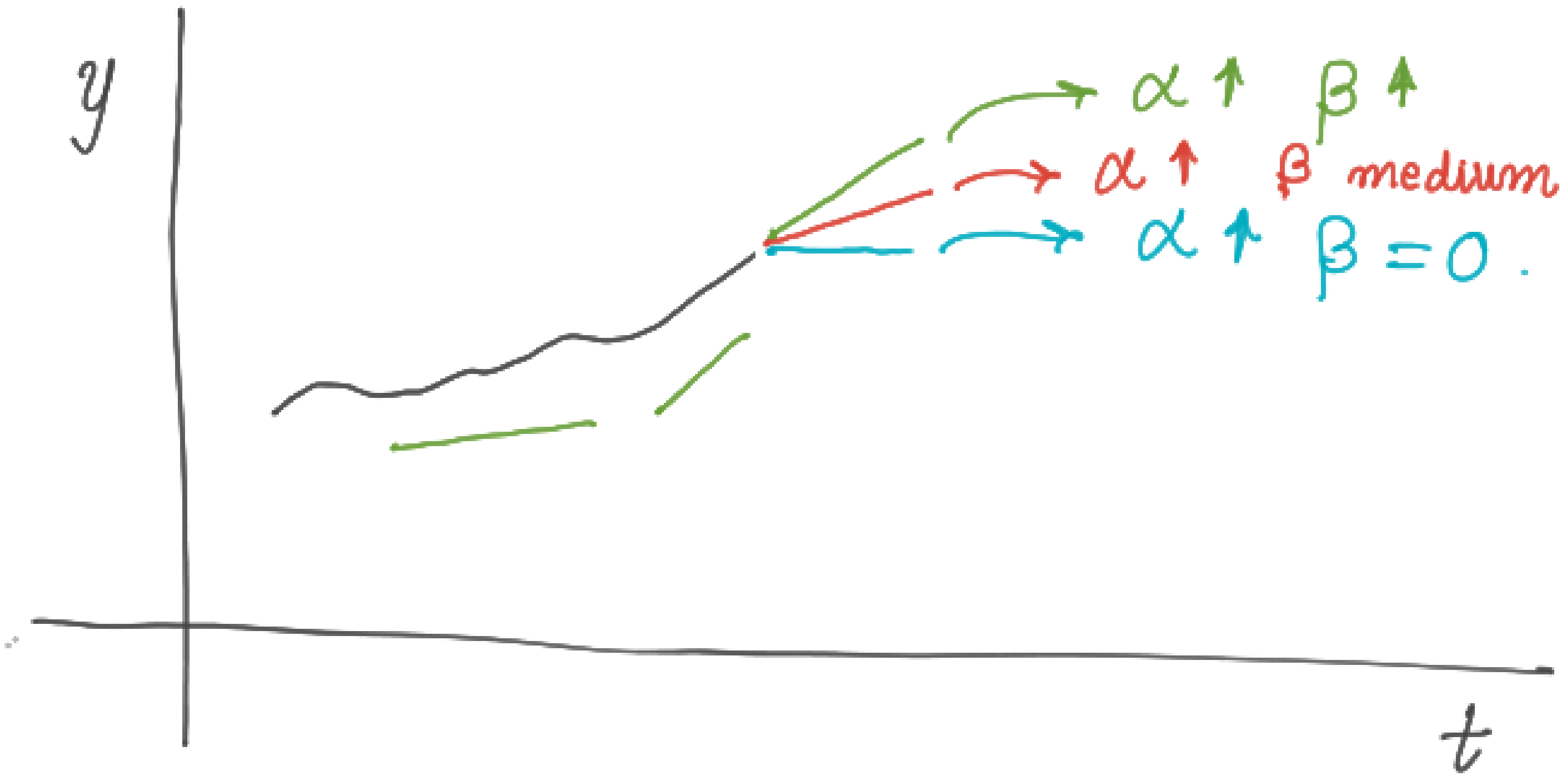
$L1$: $\cdot \quad \underbrace{y_0 \quad y_1 \; \cdots \; y_{k-1} \quad y_k \; \cdots \; y_{n-1}}_{ACF_1 \,=\, corr(y_t,\, y_{t-1})} \quad y_n$

$L2$: $\cdot \quad \cdot \quad \underbrace{y_0 \quad y_1 \; \cdots \; y_{k-2} \quad y_{k-1} \; \cdots \; y_{n-2}}_{ACF_2 \,=\, corr(y_t,\, y_{t-2})} \quad y_{n-1} \quad y_n$



$t,\ y_t$

$$y_t = f\Big(\underbrace{\overset{\alpha}{\overset{\downarrow}{y_{t-k}}}}_{\text{hist. obsv.}},\ \underbrace{\overset{\beta}{\overset{\downarrow}{\Delta y_{t-k}}}}_{\text{historical rate of change}}\Big).$$



TS → Regression

lagged features

t	y	y_{L1}	y_{L2}
0	y_0	–	–
1	y_1	y_0	–
2	y_2	y_1	y_0
3	y_3	y_2	y_1
⋮	⋮	⋮	⋮
k	y_k	y_{k-1}	y_{k-2}
$k+1$	y_{k+1}	y_k	y_{k-1}
⋮	⋮	⋮	⋮
n	y_n	y_{n-1}	y_{n-2}

dataset for regression

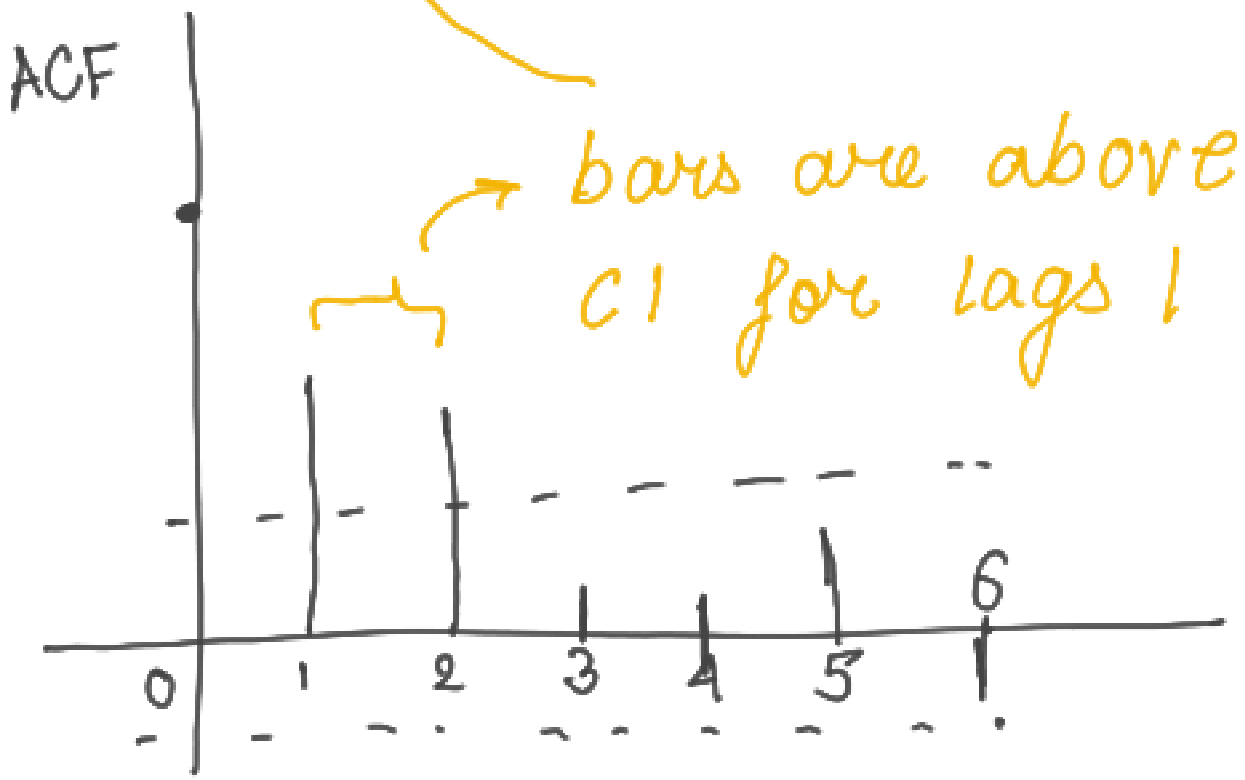
We assume that the future is ONLY affected by the past TWO days.

ACF/PACF plot after de-seasonalizing

$$y_t = f(y_{t-1}, y_{t-2})$$?

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \epsilon_t$$

→ Linear Regression



bars are above CI for lags 1 & 2

$$\left.\begin{aligned} y_2 &= \alpha_1 y_1 + \alpha_2 y_0 \\ y_3 &= \alpha_1 y_2 + \alpha_2 y_1 \\ y_4 &= \alpha_1 y_3 + \alpha_2 y_2 \\ &\vdots \end{aligned}\right\} \Rightarrow \alpha_1, \alpha_2$$

TS → Regression (Contd...)

lagged features

statistic features

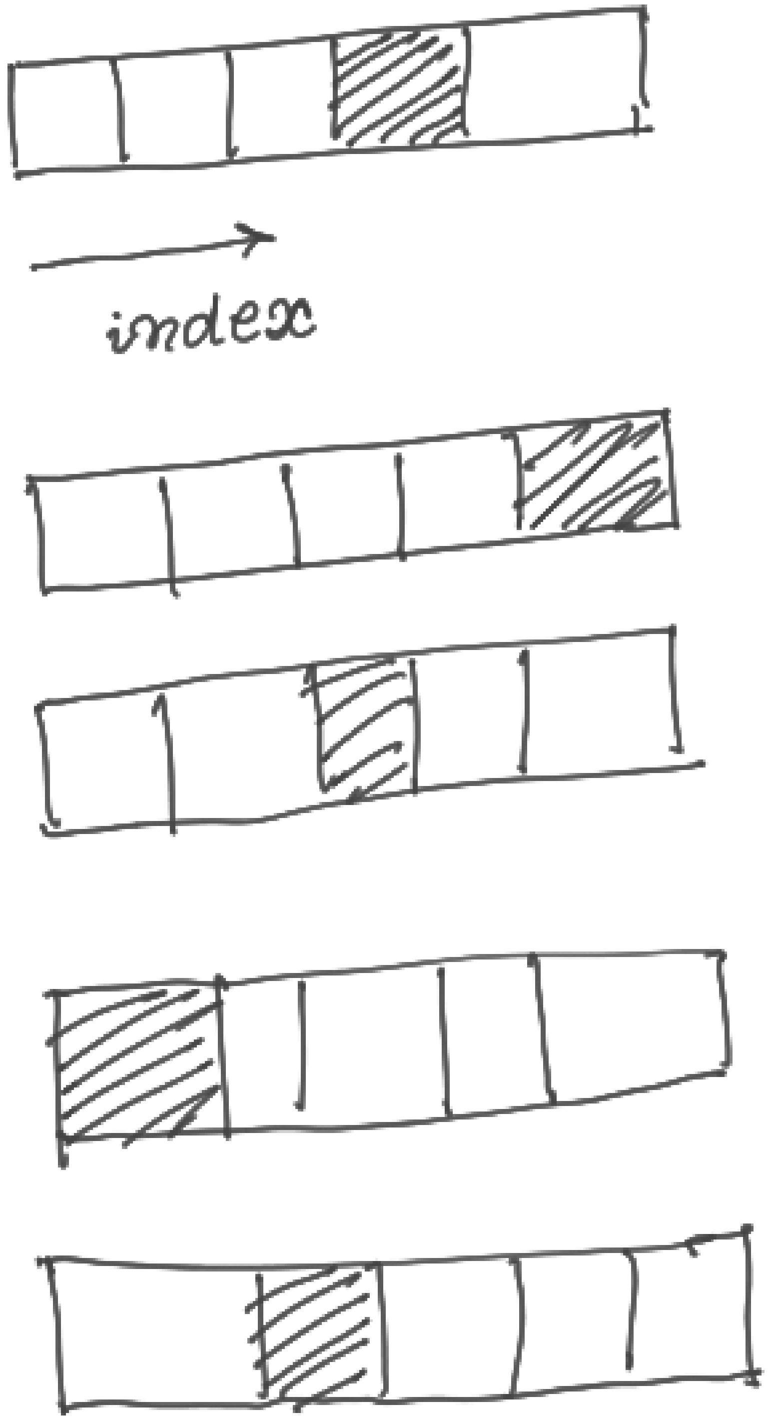
time feature

t	y	y_{L1}	y_{L2}	y_{L3}	y_{max}	y_{min}	y_{mean}	y_{std}	$t_{feature}$
0	y_0				0				
1	y_1	y_0			y_0				
2	y_2	y_1	y_0						
3	y_3	y_2	y_1	y_0					
4	y_4	y_3	y_2	y_1					
⋮	⋮	⋮	⋮	⋮					
$k-1$	y_{k-1}	y_{k-2}	y_{k-3}	y_{k-4}					
k	y_k	y_{k-1}	y_{k-2}	y_{k-3}					

y X

$$y = Xw + \epsilon \rightarrow \text{Linear Regression}$$

TS Cross Validation



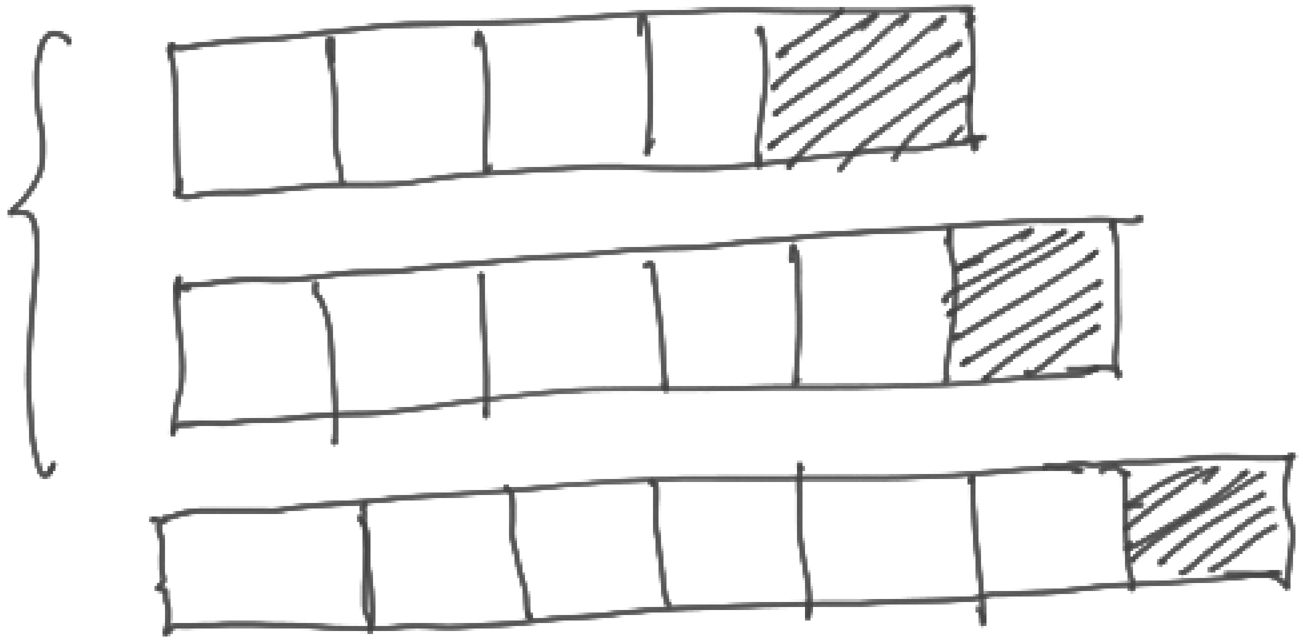
Regression

using future info to predict past. → inaccurate CV strategy

For Time Series

Need a time aware CV strategy.

Walk forward CV



Regularization

$$\frac{1}{n}\sum |y_i - X_i w|^2 \quad + \underbrace{\underbrace{\lambda_1 \sum_i |w_j|}_{L1} \quad + \underbrace{\lambda_2 \sum_i w_j^2}_{L2}}_{\text{Elastic Net}}$$