12/03 - Model Validation

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

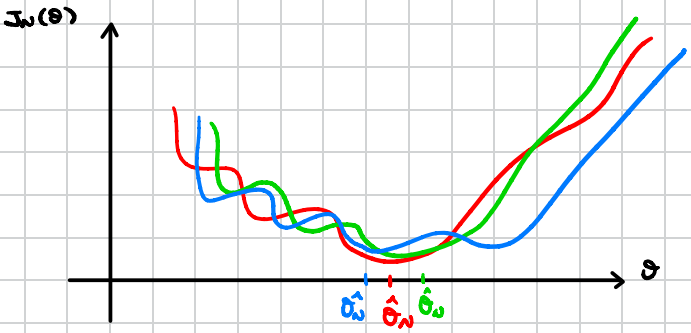
Model class: $M_\Theta = \{M(\vartheta),\ \vartheta \in \Theta \subseteq \mathbb{R}^m\}$

M. param: $\hat{\vartheta}_N = \arg\min_\vartheta J_N(\vartheta)$

Avg squared error: $J_N(\vartheta) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon(t|t-1,\vartheta)^2$

Prediction error: $\varepsilon(t|t-1,\vartheta) = y(t) - \hat{y}(t|t-1,\vartheta)$

Is $M_{\hat{\vartheta}_N}$ a good model for the system? Consider a system with $\dim[\vartheta] = 1$ and a fixed number of samples N



Each realization of the process $y(t)$ are slightly different because $y(t)$ is stochastic. So also $\hat{\vartheta}_N$ changes with the dataset

For $N \to +\infty$, it can be proven that they all converge to a single cost $\bar{\vartheta}$

THEOREM: Under current assumption, as $N \to +\infty$, the predicted cost $J_N(\vartheta, s) \to \bar{J}(\vartheta) = E_s[\varepsilon(t|t-1,\vartheta,s)^2]$

Moreover, by belief $\Delta = \{\vartheta^*,\ \bar{J}(\vartheta^*) \le \bar{J}(\vartheta),\ \forall\vartheta\}$ be the set of global minimum points of $\bar{J}(\vartheta)$

we have

$$\hat{\vartheta}_N(s) \to \Delta \qquad \text{for } N \to +\infty$$

with probability 1.

Corollary: if $\Delta = \{\vartheta^*\}$, i.e. $\bar{J}(\vartheta)$ has one unique minimum point, $\hat{\vartheta}_N(s) \to \vartheta^*$ w.p. 1 for $N \to +\infty$

Are we happy with $\hat{\vartheta}_N \to \vartheta^*$? What we are really interested in is to guarantee that if $S \in M_\Theta$ ($\exists\,\vartheta^\circ : S = M(\vartheta^\circ)$), then $\hat{\vartheta}_N \to \vartheta^*$

Theorem: $\vartheta^\circ \in \Delta$ (so, if $\Delta = \{\vartheta^*\}$ singleton, so $\vartheta^\circ = \vartheta^*$)

Proof:

Prediction error: $\varepsilon(t|t-1,\vartheta) = y(t) - \hat{y}(t|t-1,\vartheta)$

$= \underbrace{y(t) - \hat{y}(t|t-1,\vartheta^\circ)}_{e(t)} + \hat{y}(t|t-1,\vartheta^\circ) - \hat{y}(t|t-1,\vartheta)$

$e(t)$: optimal predictor

$y(t) = a\,y(t-1) + e(t) \Rightarrow \hat{y}(t|t-1) = a\,y(t-1) \Rightarrow \varepsilon = y - \hat{y} = e(t) \Rightarrow \mathrm{VAR}[\varepsilon(t|t-1)] = \mathrm{VAR}[e(t)]$

for the 1-step ahead predictor

In this way the identification criteria $\bar{J}(\vartheta)$, so the variance of the prediction error $\varepsilon(t|t-1,\vartheta)$ will be

$\bar{J}(\vartheta) = \mathrm{VAR}[\varepsilon(t|t-1,\vartheta)] = E[\varepsilon^2(t|t-1,\vartheta)]$

$= E[(e(t) + \hat{y}(t|t-1,\vartheta^\circ) - \hat{y}(t|t-1,\vartheta))^2]$

$= E[e(t)^2] + E[(\hat{y}(t|t-1,\vartheta^\circ) - \hat{y}(t|t-1,\vartheta))^2] +$

$\underline{2\,E[e(t)(\hat{y}(t|t-1,\vartheta^\circ) - \hat{y}(t|t-1,\vartheta))] = 0}$

predictors computed based on data until $t-1$

So the asymptotic cost

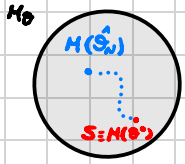
$$\bar{J}(\vartheta) = \lambda^2 + E[(\hat{y}(t|t-1,\vartheta^\circ) - \hat{y}(t|t-1,\vartheta))^2]$$

where $E[(\hat{y}(t|t-1,\vartheta^\circ) - \hat{y}(t|t-1,\vartheta))^2]$ is function of $\vartheta \ge 0$, so the minimum will be found for $\vartheta = \vartheta^\circ$

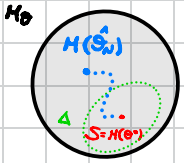
So $\bar{J}(\vartheta^\circ) \le \bar{J}(\vartheta)\ \forall\vartheta$ which means that $\vartheta^\circ \in \Delta$

4 Possible scenarios:

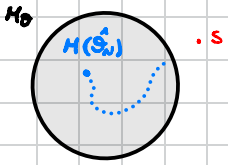
We can encounter 4 possible scenarios:



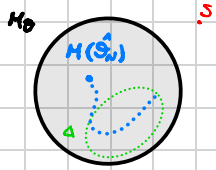
a) $S \in M_\theta$ and Δ is singleton ($S \equiv M(\theta^\circ)$, $\Delta = \{\theta^\circ\}$)



b) $S \in M_\theta$ but Δ is not singleton (Δ contains more than 1 global optimum). Not possible if everything is well posed, but it might happen



c) $S \notin M_\theta$ and Δ is singleton



d) $S \notin M_\theta$ and Δ is not singleton

Selecting M_θ is the most critical part of the process

Model order selection

Selecting $M_\theta = \{ M(\vartheta) : \vartheta \in \Theta \subseteq \mathbb{R}^{n_\theta} \}$. We cannot guarantee that $S \in M_\theta$ because S is unknown and we don't know how to enlarge M_θ in case $S \notin M_\theta$. For simplicity , we will consider $n = m = p$ (only 1 degree of freedom and simplify the process). How to select n properly (Scalar optimization problem)?

Let us take the following procedure

set $n = 1$

repeat until $n = n_{max}$

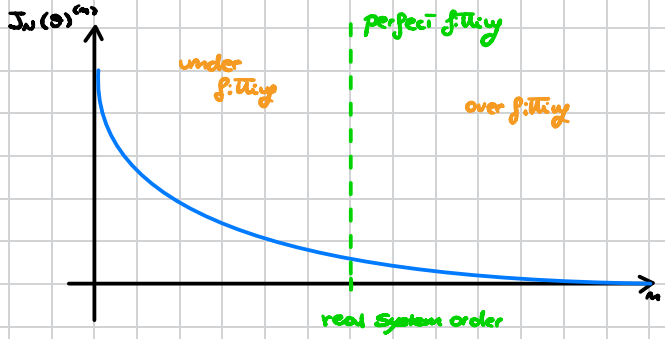
$$M_\theta^{(n)} = \{ M(\vartheta) : \vartheta \in \Theta^{(n)} \subseteq \mathbb{R}^{n_\theta} \}$$

$$\hat{\vartheta}_N^{(n)} = \arg\min_\vartheta J_N^{(n)}(\vartheta) \qquad J_N^{(n)}(\vartheta) = \frac{1}{N} \sum (y(t) - \hat{y}(t|t-1, \vartheta^{(n)}))^2$$

$$n := n+1$$

Consider by starting with: $A(z) = 1 - a_1 z^{-1}$ $B(z) = b_0 + b_1 z^{-1}$ $C(z) = 1 + c_1 z^{-1}$ and then add terms to each component.

This is not suitable for the selection of n , since this point cannot be detected just by looking at the behaviour of $J_N(\hat{\vartheta}_N^{(n)})$ as function of n



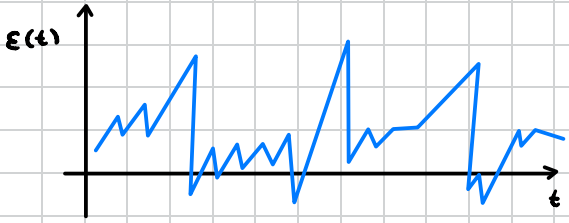
3 different criteria for model order selection

a) whiteness test on residual

b) cross validation

c) identification with model order penalties

Whiteness Test on Residuals (Anderson's Test)



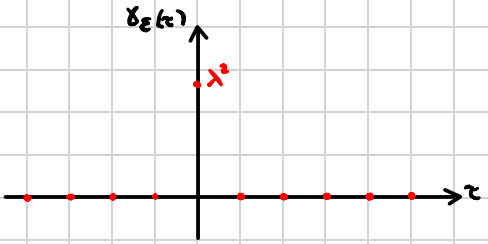
We know that, if $S \in M_\theta$, for a certain model order n , $\hat{\vartheta}_N^{(n)} \to \vartheta^\circ$, and

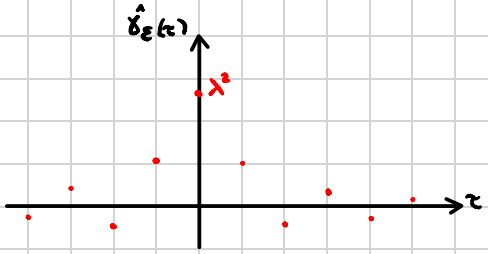
$$\varepsilon(t|t-1, \hat{\vartheta}_N^{(n)}) \to e(t) \sim WN(0, \lambda^2)$$

Then, for large N:

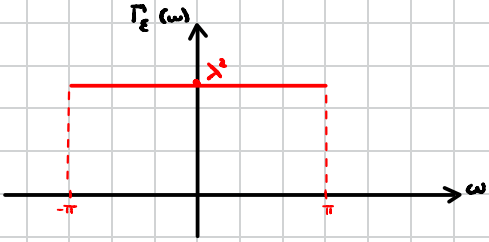
$$\varepsilon(t|t-1, \hat{\vartheta}_N^{(n)}) = y(t) - \hat{y}(t|t-1, \hat{\vartheta}_N^{(n)})$$

should be a realization of a WN

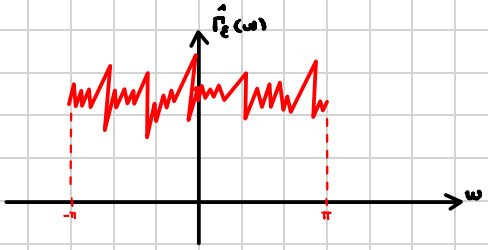




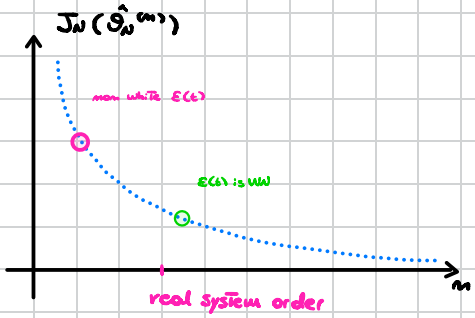
If $\varepsilon(t)$ is WN, we expect the real values. However, we obtain some fluctuation that make difficult to recognize a WN.



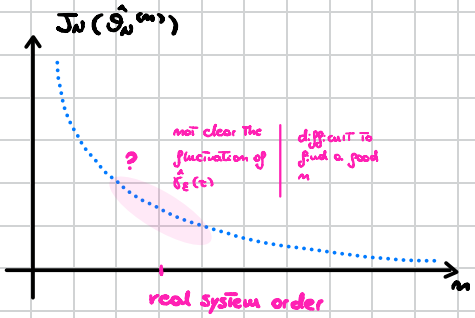
$\Rightarrow$



IDEAL



REALITY



Cross Validation

Consider the dataset $D_N = \{(u(1),y(1)),\dots,(u(m),y(m))\}$ and the cost function we want to minimize $J_N(\vartheta) = \frac{1}{N}\sum(y(t)-\hat{y}(t|t-1,\vartheta))^2$. Just looking at the dataset we can subdivide it into training set and validation set. We can follow the procedure:

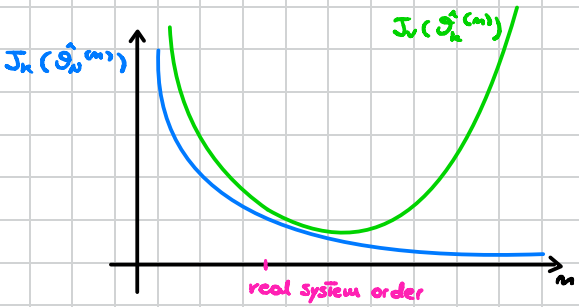
For $n=1$ To $n=n_{max}$

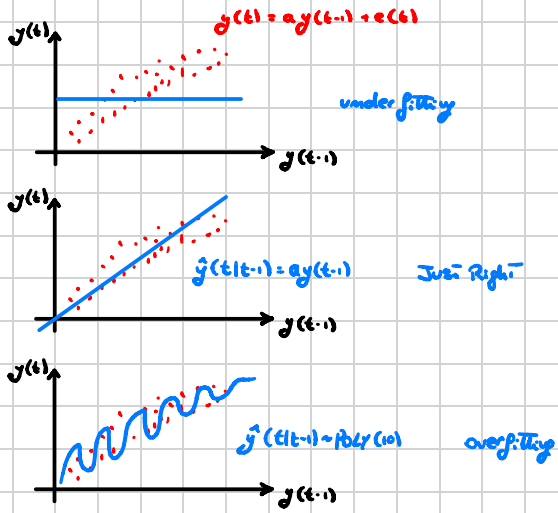
find $\hat{\vartheta}_k^{(m)} = \arg\min_\vartheta J_k^{(m)}(\vartheta)$

evaluate $J_V(\hat{\vartheta}_k^{(m)}) = \frac{1}{N-k}\sum_{k+1}^{N}(y(t)-\hat{y}(t|t-1,\hat{\vartheta}_k^{(m)}))^2$

Choose n (optimal model order) as the one minimizing J_V

$\hat{n} = \arg\min_n J_V(\hat{\vartheta}_N^{(m)})$





In ML, k-fold cross validation is often used: divide the dataset in k block, using $1/k$ of it as validation and the other as training. This is not doable here since the system is dynamical and we cannot attach different time windows (this would lead to "fake" temporal relations). Some problems for validation

Identification costs with model order penalties

Using Cross Validation we have split the dataset and obtained the same cost for Training and validation

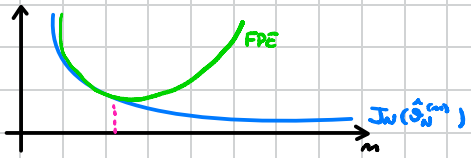
Now: different costs for model order selection and parameter selection ⇒ Single DS

Instead of minimizing $J_N(\theta) = \frac{1}{N}\sum (y(t) - \hat{y}(t|t-1,\theta))^2$ we can minimize 3 alternatives

1) Final prediction error (FPE)
2) Akaike information criterion (AIC)
3) Minimum description length (MDL)

FPE: our ideal objective is $\min_\theta E[J_N(\theta)]$ but is not affordable. So, FPE is a numerical approximation of this

$$FPE(m) = \frac{N+m}{N-m} J_N(\hat{\theta}_N^{(m)})$$



AIC: $$AIC(m) = \underbrace{\log\left(J_N(\hat{\theta}_N^{(m)})\right)}_{\text{decrease with } m} + 2\underbrace{\frac{m}{N}}_{\text{penalty}}$$

MDL: $$MDL(m) = \underbrace{\log\left(J_N(\hat{\theta}_N^{(m)})\right)}_{} + \underbrace{\log(N)\frac{m}{N}}_{}$$

FPE vs AIC ?

$$\log(FPE) = \log\left(\frac{N+m}{N-m} J_N(\hat{\theta}_N^{(m)})\right) = \log\left(\frac{1+m/N}{1-m/N} J_N(\hat{\theta}_N^{(m)})\right)$$

m/N is small ⇒ $\frac{\log(1+x)}{x} \to 1$ as $x \to 0$

$$= \log(1+m/N) - \log(1-m/N) + \log(J_N(\hat{\theta}_N^{(m)}))$$
$$\approx \frac{m}{N} - (-\frac{m}{N}) + \log(J_N(\hat{\theta}_N^{(m)}))$$
$$= \log(J_N(\hat{\theta}_N^{(m)})) + 2\frac{m}{N} = AIC$$

⇒ So AIC and FPE return the same optimal $\hat{m}$

AIC vs MDL? Being

$$AIC = \log\left(J_N(\hat{\theta}_N^{(m)})\right) + 2\frac{m}{N} \qquad MDL = \log\left(J_N(\hat{\theta}_N^{(m)})\right) + \log(N)\frac{m}{N}$$

if $\log(N) \geq 2$ ⇒ the model order is penalized more by MDL ⇒ $N > 8$ (basically always since $N \sim > 10000$)

If $S \in M_O$ and M_O is the set of ARX models, MDL is right. In the general case, $S \in M_O$ we usually prefer to slightly overfitting and we use AIC