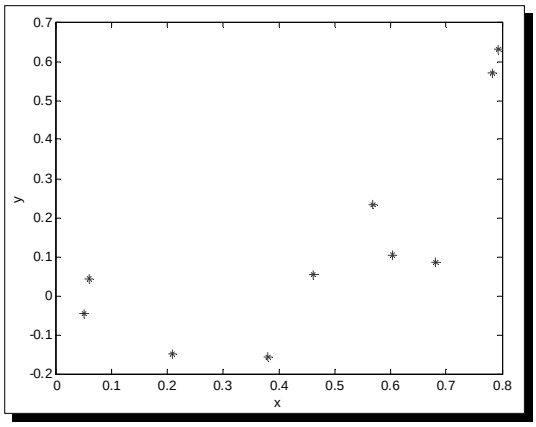
Model structure selection

Michel Verleysen, Amaury Lendasse
Université catholique de Louvain (Louvain-la-Neuve, Belgium)
Electricity and Applied Mathematics departments

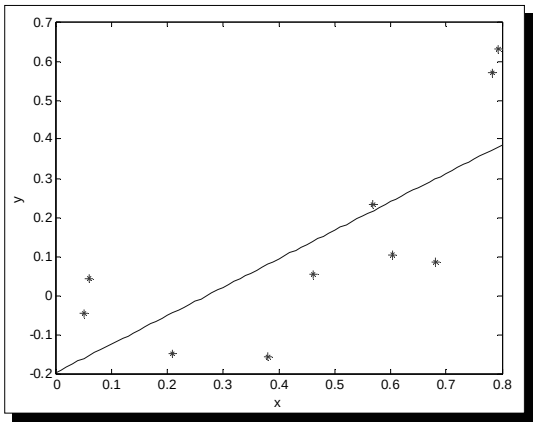
October 2002

Motivation: « good » model?



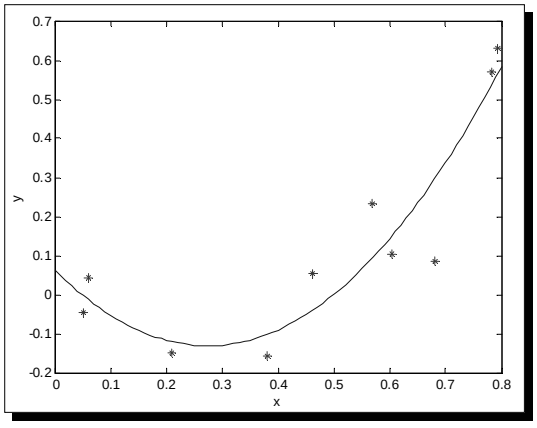
Motivation: « good » model?

$$y_t = a + bx_t$$



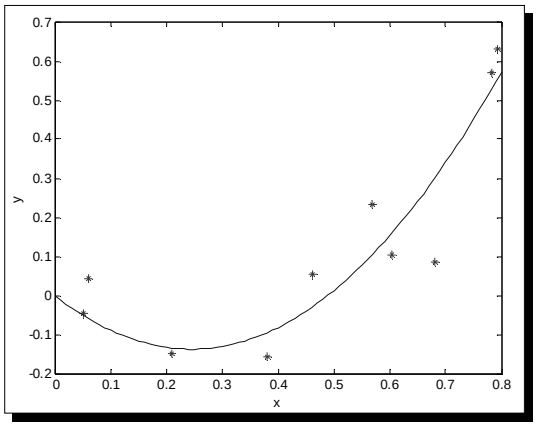
Motivation: « good » model?

$$y_t = a + bx_t + cx_t^2$$



Motivation: « good » model?

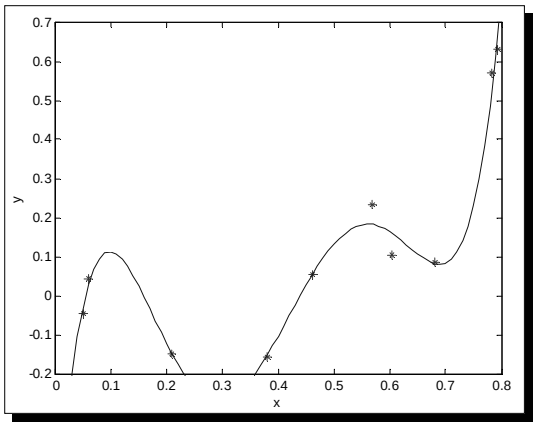
$$y_t = bx_t + cx_t^2$$



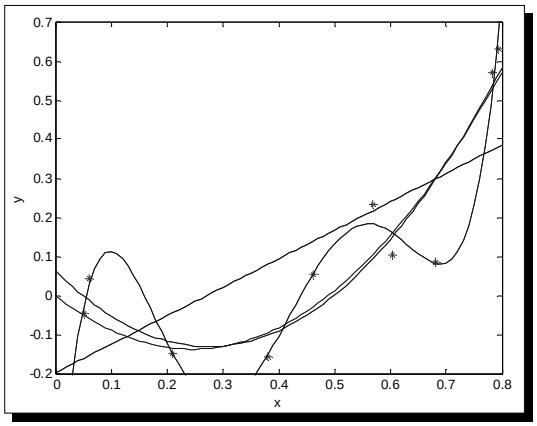
- The one used to generated the data (with noise)…

Motivation: « good » model?

$$y_t = a + bx_t + cx_t^2 + dx_t^3 + ex_t^4 + fx_t^5$$



Motivation: « good » model?



Best model

- Notations $x_t \in R^d, y_t \in R$

  $\hat{y}_t = g(x_t, \theta)$

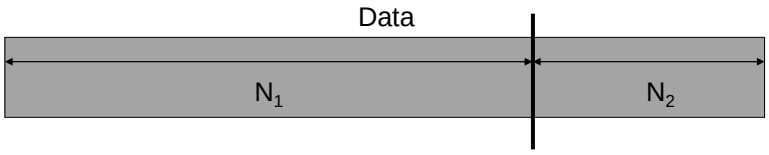
- Generalization error

$$E_{gen}(\theta) = \lim_{T \to \infty} \sum_{t=1}^{T} \frac{(g(x_t, \theta) - y_t)^2}{T} \longrightarrow \boxed{\hat{E}_{gen}(\theta)}$$

Content

- Methods
  - validation
  - cross-validation (Monte-Carlo + k-fold + leave-one-out)
  - bootstrap
  - AIC + BIC
- Some theoretical insights
- Examples

Validation



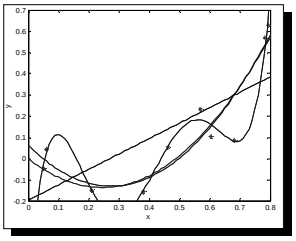
Learning set — Validation set

A model is built → Error

$$\hat{E}_{gen} = \frac{\sum_{t \in VS} (\hat{y}_t - y_t)^2}{N_2}$$

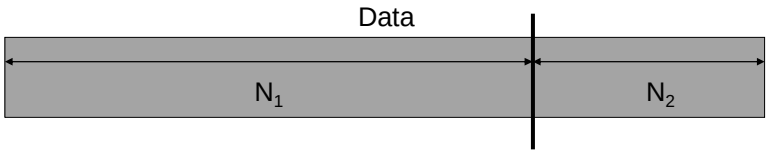
Validation

$$\hat{E}_{gen} = \frac{\sum_{t \in VS} (\hat{y}_t - y_t)^2}{N_2}$$



k	linear model	quadratic model	quadratic model without independent term	5th-order model
1	0.0370	0.0181	0.0118	0.0038

Cross-validation



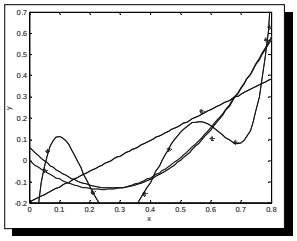
Learning set → A model is built → Error (Validation set)

$$\hat{E}_{gen} = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{t \in VS} \left(\hat{y}_t^k - y_t \right)^2}{N_2}$$

Experience repeated *K* times

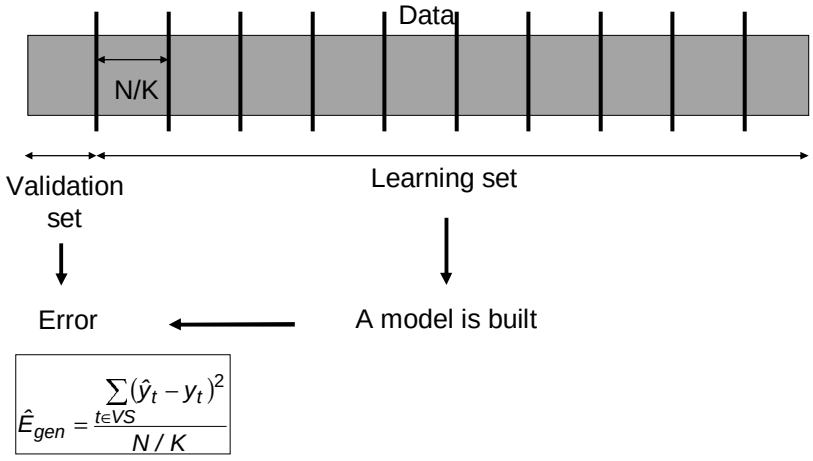
Cross-validation

$$\hat{E}_{gen} = \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{t\in VS}\left(\hat{y}_t^k - y_t\right)^2}{N_2}$$



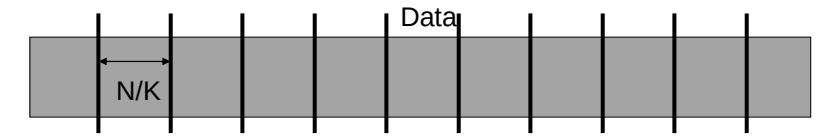
k	linear model	quadratic model	quadratic model without independent term	5th-order model
10	0.0184	0.0229	0.0201	0.0097
100	0.0652	0.0275	0.0251	13.8062
1000	0.0743	0.0276	0.0257	154.8485
10000	0.0798	0.0267	0.0250	208.5566

K-fold cross-validation



$$\hat{E}_{gen} = \frac{\sum_{t\in VS}\left(\hat{y}_t - y_t\right)^2}{N/K}$$

K-fold cross-validation



Validation set

Learning set

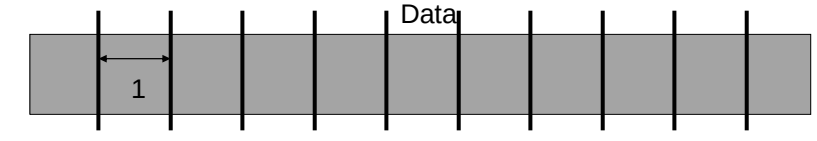
Error

A model is built

$$\hat{E}_{gen} = \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{t\in VS}\left(\hat{y}_t^k - y_t\right)^2}{N/K}$$

Experience repeated *K* times

Leave-one-out



Validation set

Learning set

Error

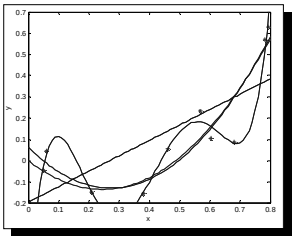
A model is built

$$\hat{E}_{gen} = \frac{1}{N}\sum_{k=1}^{N}\left(\hat{y}_t^k - y_t\right)^2$$

Experience repeated *K* times

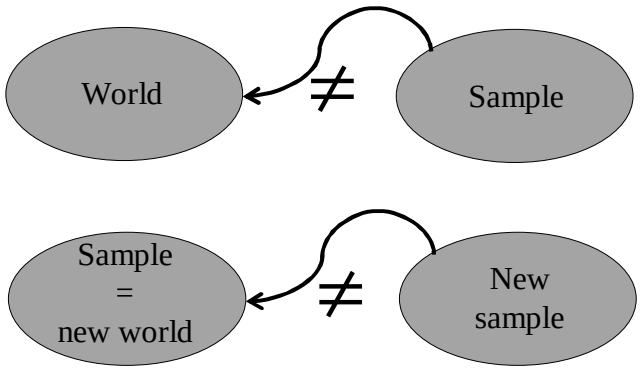
Leave-one-out

$$\hat{E}_{gen} = \frac{1}{N}\sum_{k=1}^{N}\left(\hat{y}_t^k - y_t\right)^2$$



linear model	quadratic model	quadratic model without independent term	5th-order model
0.0488	0.0153	0.0146	0.0045

Bootstrap: plug-in principle



Bootstrap : plug-in principle

1
2
3
4
…
10

2
10
6
8
…
10

Sample = new world

New sample

Bootstrap : plug-in principle

$E_{gen} = E_{sample} + \Delta E$

World

Sample

Learning $\longrightarrow E_{sample}$

$\Delta E = E_{new\,world} - E_{new\,sample}$

Sample = new world

New sample

$E_{new\,world}$

Learning $\longrightarrow E_{new\,sample}$

Bootstrap

$$E_{gen}(\theta) = E_{gen}(\theta) - E_{sample}(\theta) + E_{sample}(\theta)$$

Definition: $D(\theta) \triangleq E_{gen}(\theta) - E_{sample}(\theta)$

$$\Rightarrow E_{gen}(\theta) = D(\theta) + E_{sample}(\theta)$$

Estimate: $\hat{D}(\theta) = E_{new\,world}(\theta) - E_{new\,sample}(\theta)$

Bootstrap

$$E_{gen}(\theta) = E_{gen}(\theta) - E_{sample}(\theta) + E_{sample}(\theta)$$

Definition: $D(\theta) \triangleq E_{gen}(\theta) - E_{sample}(\theta)$

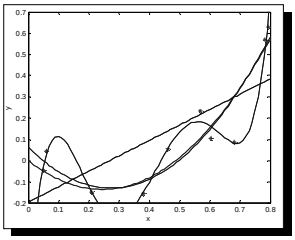
$$\Rightarrow E_{gen}(\theta) = D(\theta) + E_{sample}(\theta)$$

Estimate: $\hat{D}(\theta) = \frac{1}{K}\sum_{k=1}^{K}\left(E^{k}_{new\,world}(\theta) - E^{k}_{new\,sample}(\theta)\right)$

$$\longrightarrow \boxed{\hat{E}_{gen}(\theta) = E_{sample}(\theta) + \frac{1}{K}\sum_{k=1}^{K}\left(E^{k}_{new\,world}(\theta) - E^{k}_{new\,sample}(\theta)\right)}$$

Bootstrap

$$\hat{E}_{gen}(\theta) = E_{sample}(\theta) + \frac{1}{K}\sum_{k=1}^{K}\left(E^{k}_{new\,world}(\theta) - E^{k}_{new\,sample}(\theta)\right)$$



k	linear model	quadratic model	quadratic model without independent term	5^{th}-order model
10	0.0384	0.0209	0.0155	8.3898
100	0.0345	0.0301	0.0128	703.74
1000	0.0325	0.0807	0.0112	7846.5
10000	0.0333	0.1267	0.0118	6422.5

Bootstrap 632 and 632+

- Improvements on bootstrap
- 632+: better for Nearest Neighbour problems
  - lazy learning
  - vector quantization
  - *k*-NN classification
  - …

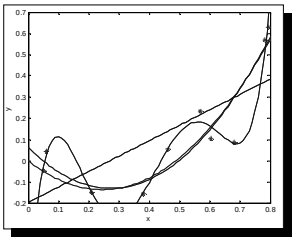
Some theoretical results

- All methods are consistent ???
-

	Bias	Variance
LOO and CV	no	high
Bootstrap	yes	low
Bootstrap 632	No	low

- Boostrap valid for model selection

Boostrap 632



k	linear model	quadratic model	quadratic model without independent term	5^{th}-order model
10	0.0348	0.0118	0.0115	0.0090
100	0.0342	0.0152	0.0128	9.3315
1000	0.0331	0.0243	0.0134	4.1025
10000	0.0339	0.0244	0.0137	3.3666

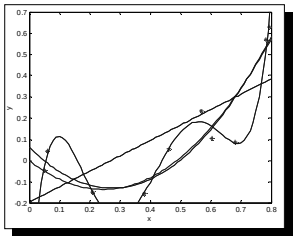
AIC and BIC

$$\hat{E}_{gen}(\theta) = \hat{D}(\theta) + E^{I}(\theta)$$

AIC $\hat{D}(\theta) = \frac{2p\hat{\sigma}}{N}$

BIC $\hat{D}(\theta) = \frac{\ln(N)}{N} p\hat{\sigma}$

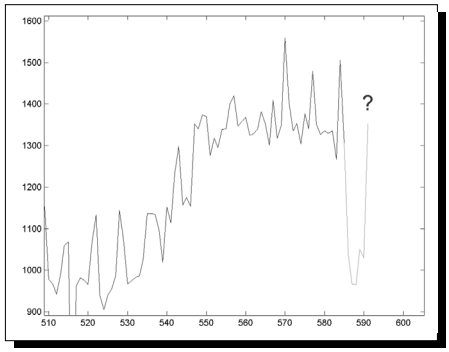
$$\hat{\sigma} = \frac{\sum_{t=1}^{N} (\hat{y}_t - y_t)^2}{N - p}$$



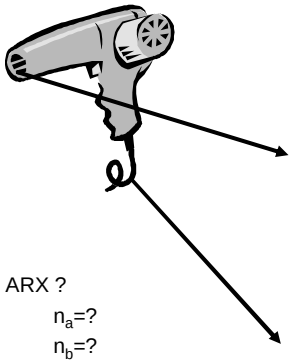
k	linear model	quadratic model	quadratic model without independent term	5th-order model
AIC	0.1752	0.0858	0.0570	0.0319
BIC	0.1973	0.0974	0.0641	0.0366

Time series prediction (NAR)

$$\hat{x}_{t+1} = f(x_t, x_{t-1}, \ldots, x_{t-n}, u_t, u_{t-1}, \ldots, u_{t-n}, \theta)$$



Lyung's hair-dryer

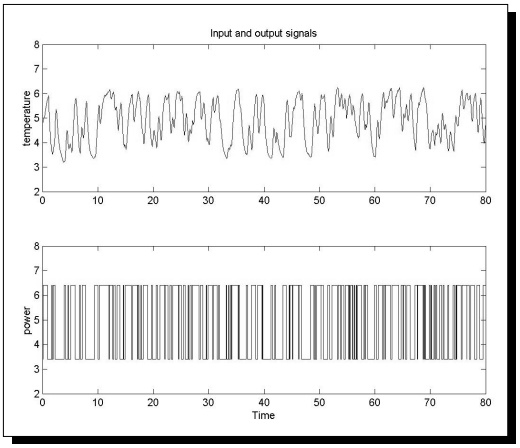


ARX ?

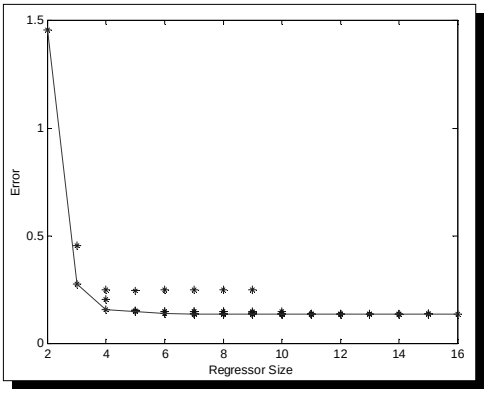
n_a=?

n_b=?

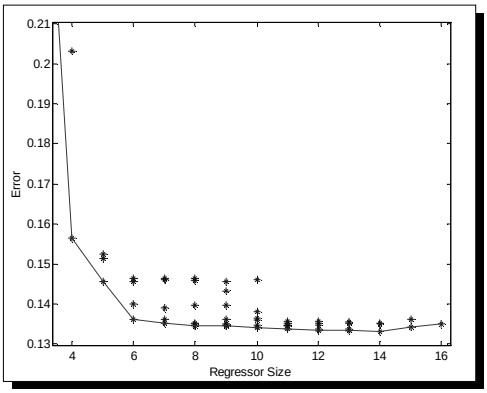
n_k=3



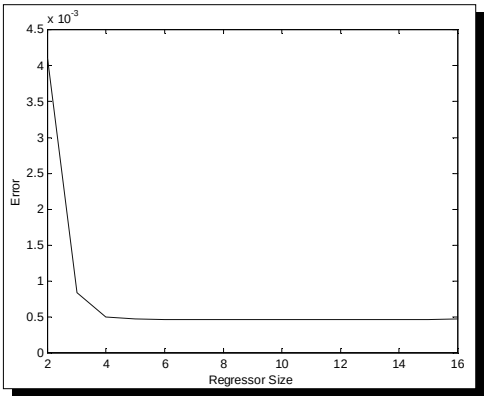
Lyung's hair-dryer: validation



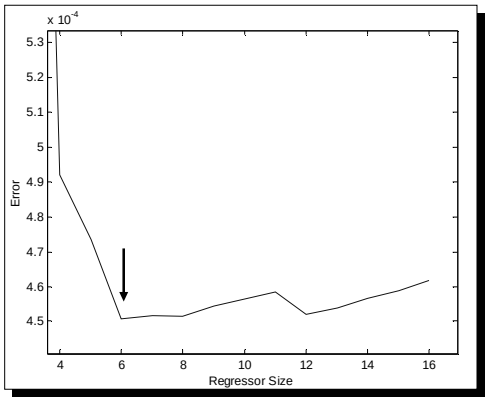
Lyung's hair-dryer: validation



Lyung's hair-dryer: bootstrap

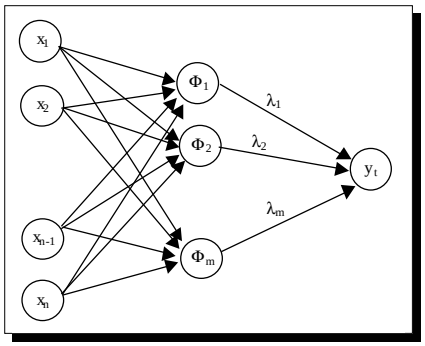


Lyung's hair-dryer: bootstrap



- Compared to best ARX model found by Lyung: n_a=3, n_b=3, n_k=3

RBFN as NAR models

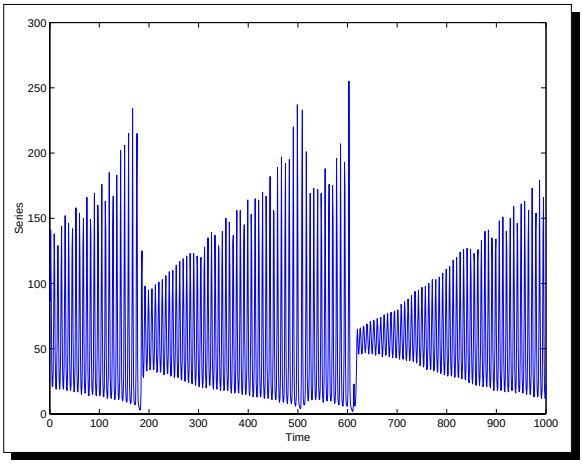


$$x_t = \begin{bmatrix} x_{t1} \\ x_{t2} \\ \dots \\ x_{tn} \end{bmatrix}$$

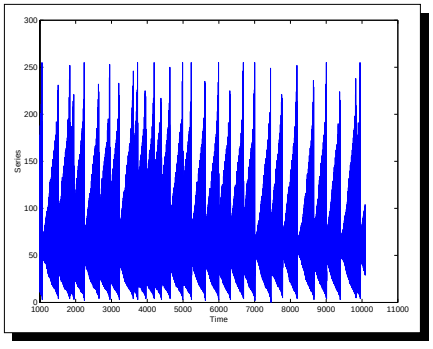
$$\hat{y}_t = \sum_{i=1}^{m} \lambda_i \Phi_i(x_t, c_i, \sigma_i)$$

$$\Phi_i(x_t, c_i, \sigma_i) = e^{-\left(\frac{\|x_t - c_i\|^2}{\sqrt{2}\sigma_i}\right)^2}$$

Santa Fe A



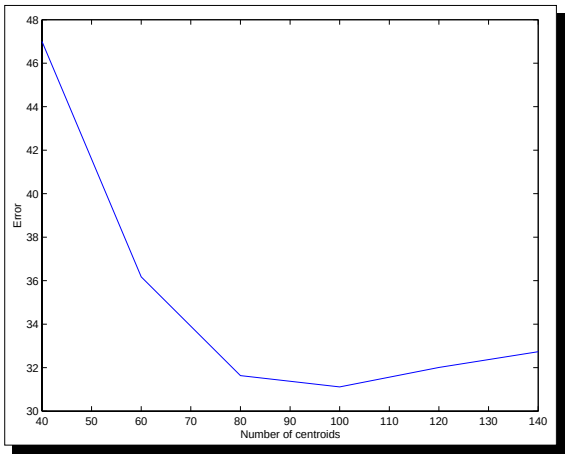
Santa Fe A: test



$$E_{gen}(\theta) = \lim_{T\to\infty} \sum_{t=1}^{T} \frac{(\hat{x}_t(\theta) - x_t)^2}{T} \approx \hat{E}_{gen}(\theta) = \sum_{t=1001}^{10000} \frac{(\hat{x}_t(\theta) - x_t)^2}{9000}$$

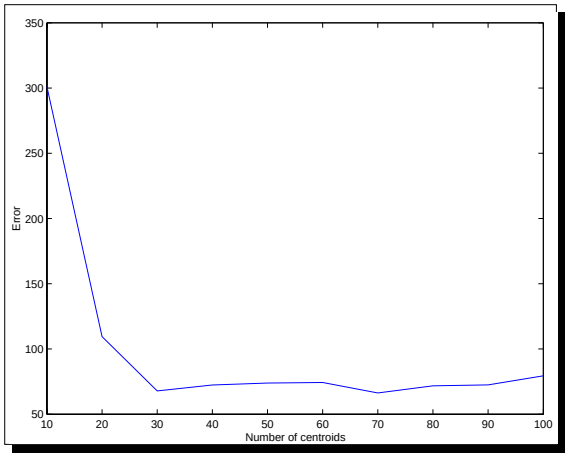
Santa Fe A: test

$\hat{E}_{gen}(\theta)$



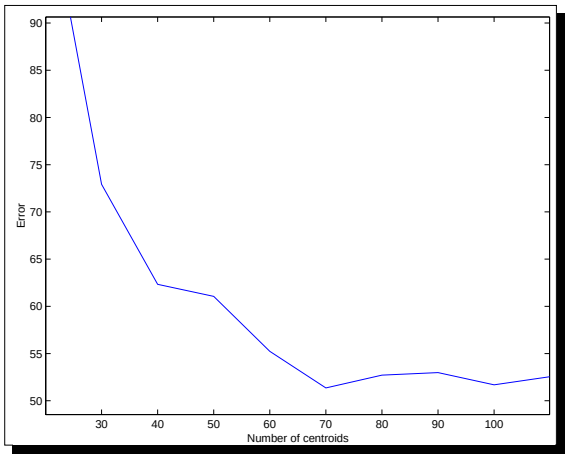
Santa Fe A: cross-validation

$\hat{E}_{gen}(\theta)$



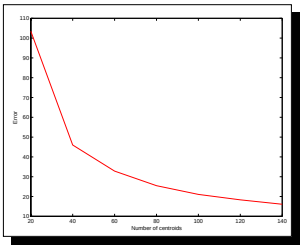
Santa Fe A: leave-one-out

$\hat{E}_{gen}(\theta)$

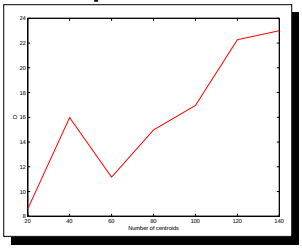


Santa Fe A: bootstrap

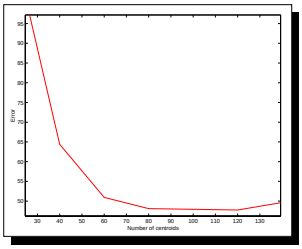
$E_{sample}(\theta)$



$+\hat{D}(\theta)$

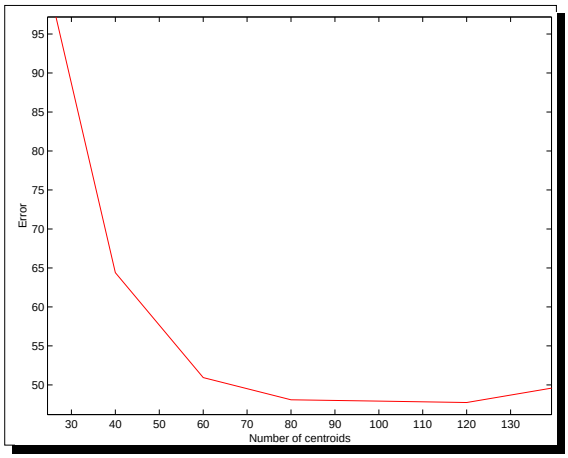


$=\hat{E}_{gen}(\theta)$



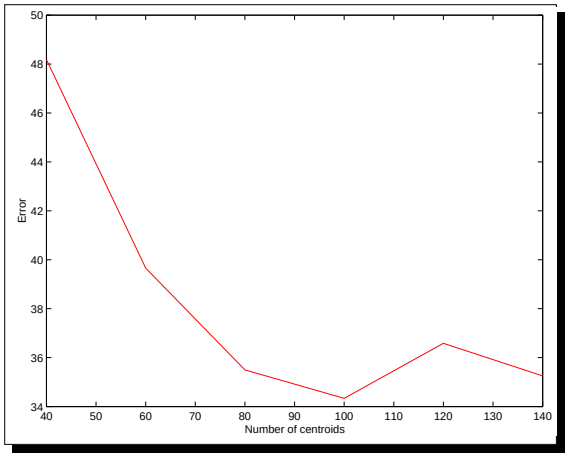
Santa Fe A: bootstrap

$\hat{E}_{gen}(\theta)$



Santa Fe A: bootstrap 632

$\hat{E}_{gen}(\theta)$



Some conclusions…

- Bootstrap: OK for model selection
- Boostrap 632(+): also OK for error estimation
- both require less data than CV and LOO
- for specific cases (nearest neighbour, VQ): bootstraps fail