Fast Bootstrap for model selection

M. Verleysen
Université catholique de Louvain (Belgium)
Machine Learning Group
http://www.dice.ucl.ac.be/mlg/

Outline

- Model selection
  - estimation of generalization error
  - resampling methods
- Boostrap
  - plug-in principle
  - computational load
- Fast Bootstrap
  - idea and hypothesis
  - reduced number of experiments
- Experiments
  - artificial regression example
  - Santa Fe A time series prediction

Outline

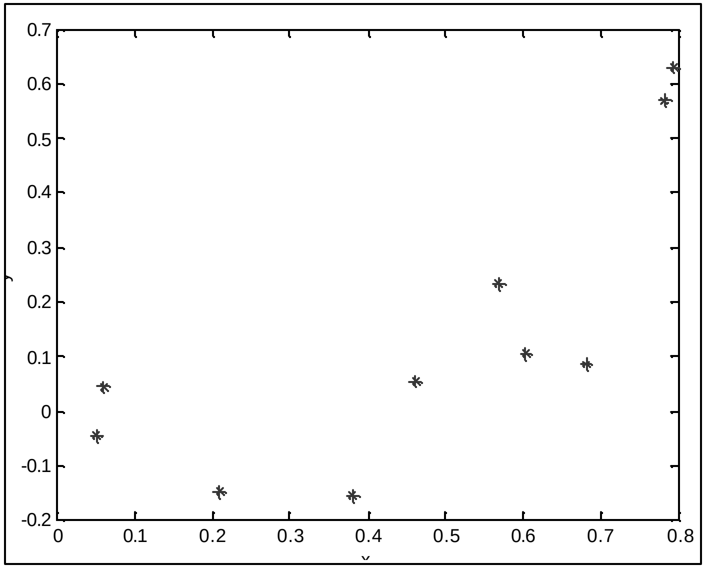
- Model selection
  - estimation of generalization error
  - resampling methods
- Boostrap
  - plug-in principle
  - computational load
- Fast Bootstrap
  - idea and hypothesis
  - reduced number of experiments
- Experiments
  - artificial regression example
  - Santa Fe A time series prediction

Model selection is necessary

- Database to model

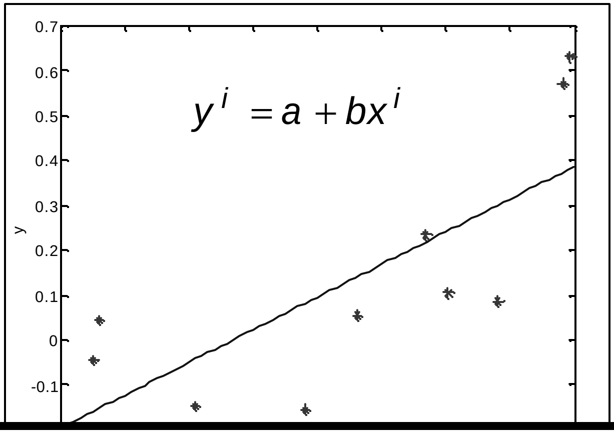
$$\left\{\mathbf{x}^i, y^i\right\}, \text{ with } \mathbf{x}^i \in \Re^D,\ y^i \in \Re,\ 1 \le i \le N$$

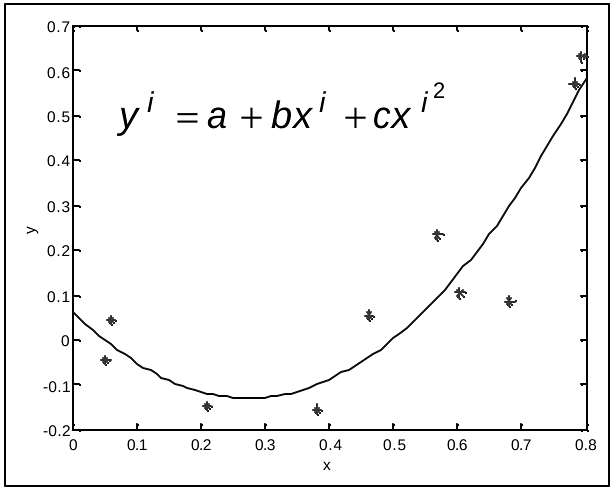
- Example (d=1):



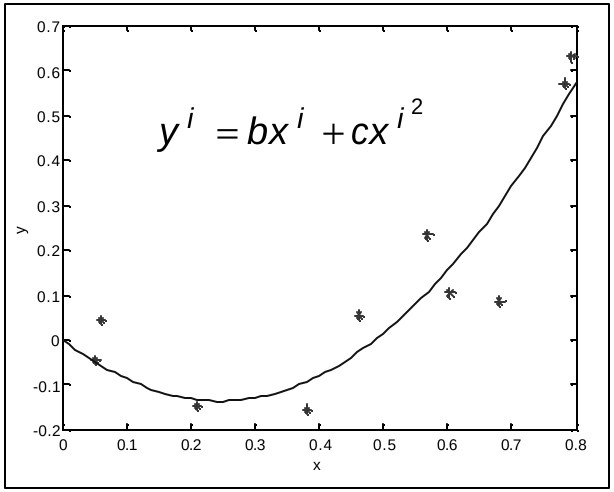
Which (polynomial) model structure to use?

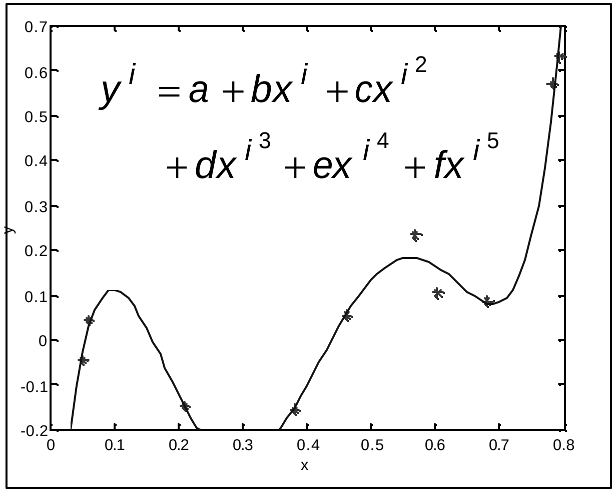
Some possible model structures...





The model structure used to generate the data (with noise)





Model structure selection is performed according to an error criterion

- Notations

$$\mathbf{x}_t \in R^d, y_t \in R$$

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

- Generalization error

$$E_{gen}(q) = \lim_{N \to \infty} \sum_{t=1}^{N} \frac{(g(\mathbf{x}_t, q) - y_t)^2}{N}$$

$$\downarrow$$

$$\hat{E}_{gen}(q) = \sum_{t=1}^{N} \frac{(\hat{y}_t - y_t)^2}{N} \quad (= \hat{E}_{gen})$$

How to estimate the generalization error ?

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

- Problems:
  - The test sample cannot be used for learning
  - Need for 3 sets:
    - Learning
    - Validation (model selection)
    - Test (estimate of model performances)
  - The available sample is always too small in practise...
- Validation is learning (of hyperparameters), not test!

Resampling methods

- Small sample: asymtotic results are difficult to use
- **Empirical** model structure selection !
- Possible methods
  - Validation
  - *K*-fold cross-validation
  - Leave-One-Out
  - Bootstrap

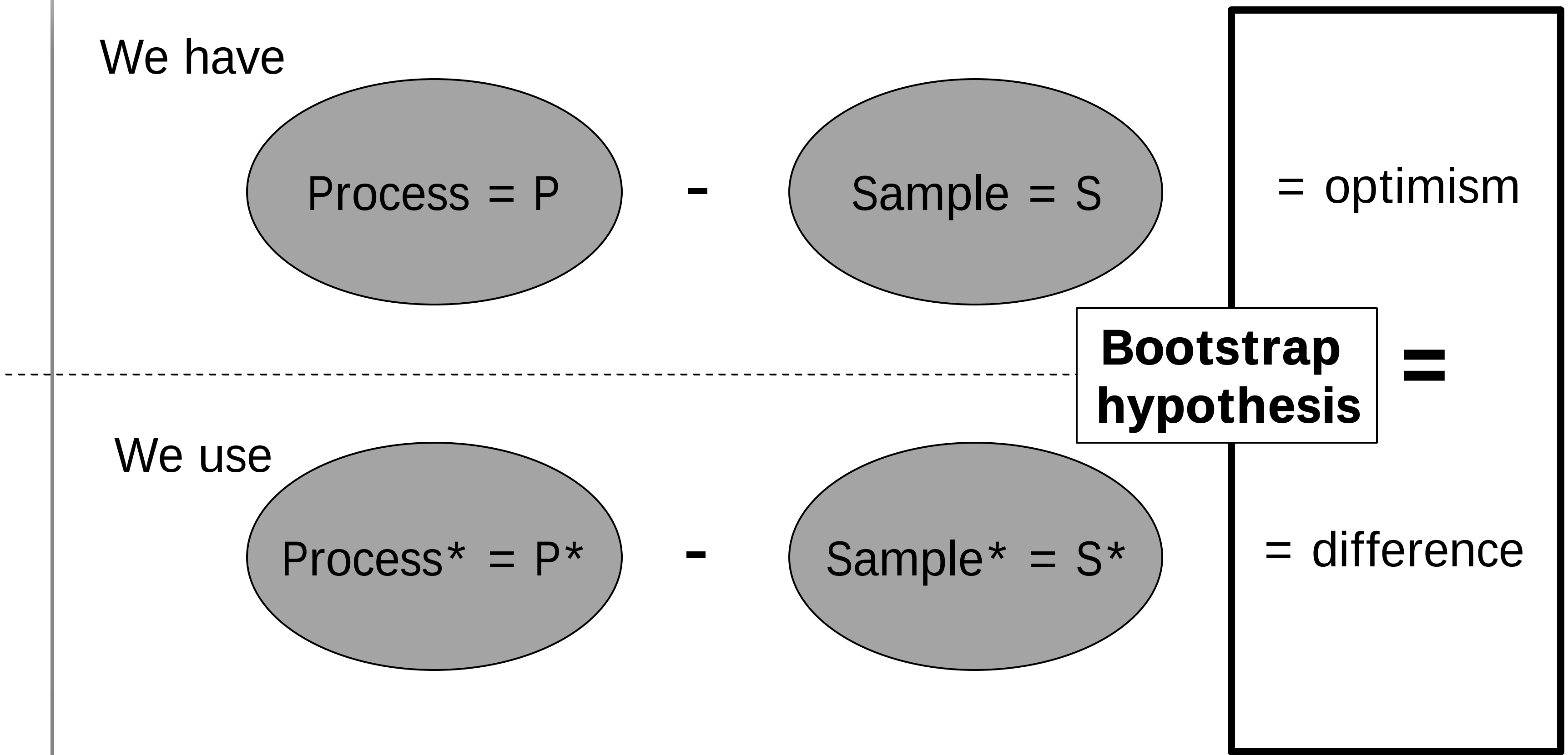
Outline

- Model selection
  - estimation of generalization error
  - resampling methods
- Boostrap
  - plug-in principle
  - computational load
- Fast Bootstrap
  - idea and hypothesis
  - reduced number of experiments
- Experiments
  - artificial regression example
  - Santa Fe A time series prediction

Why the bootstrap ?

- Experimentally: its variance is lower, for a fixed number of experiments
- It is sound on a statistical point of view (both other methods are too!)
- It makes the following approximation possible...

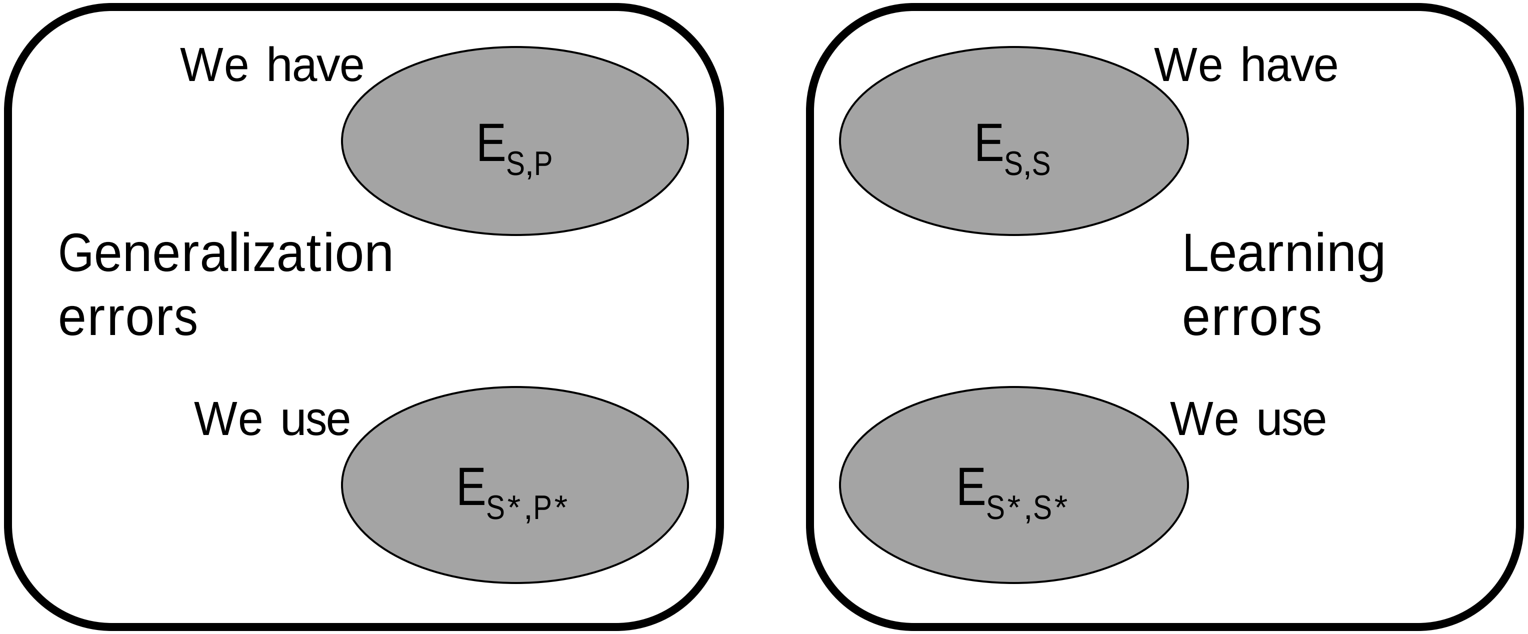
The bootstrap is rather intuitive



Key ideas: plug-in principle and drawing with replacement

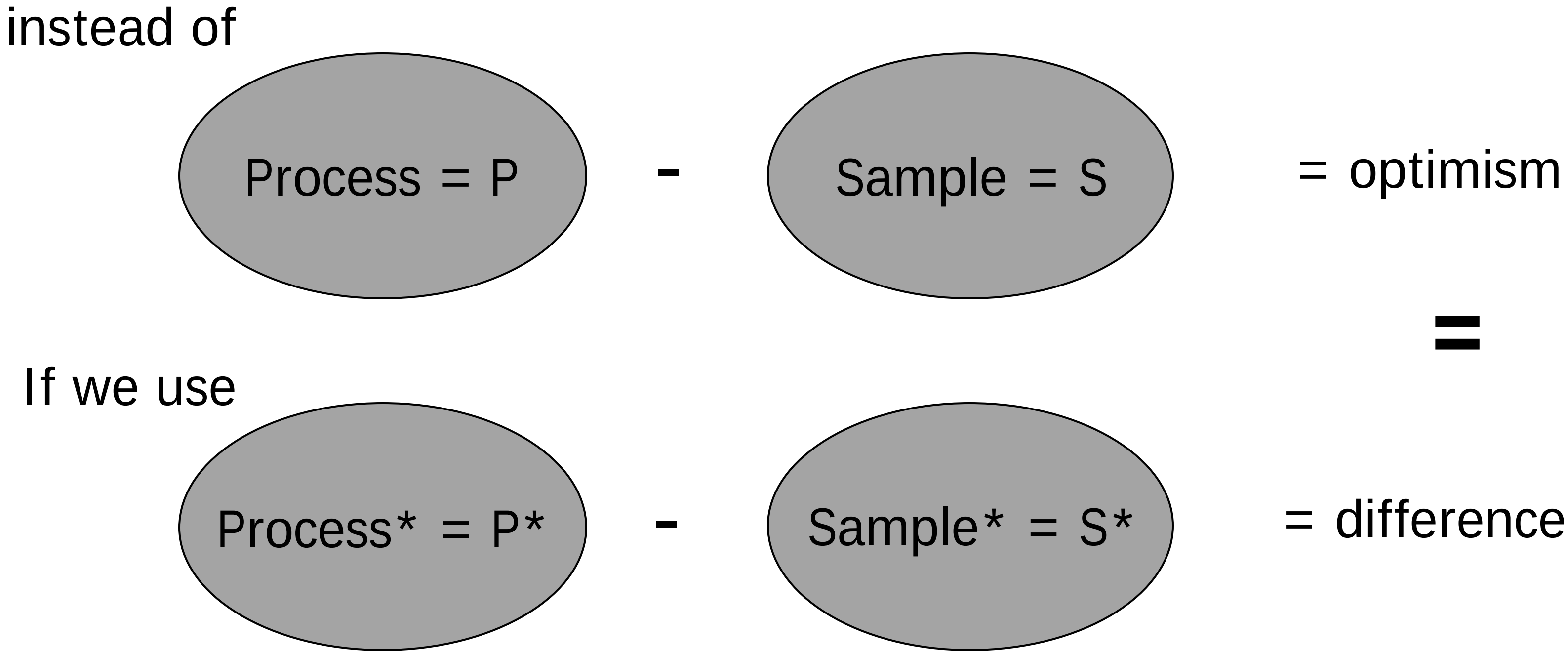
The plug-in principle is the main idea of the bootstrap

- The plug-in principle is often used in statistics
- Here, this principle gives: Process* = Sample
- All errors ate denoted here as $E_{\text{learning set,test set}}$
  - $E_{S,S}$ is a **learning error**
  - $E_{S,P}$ is a **generalization error** ($S \cap P = \emptyset$)



Bootstrap – plug-in

What are the samples in the bootstrap ?



then we must have S* and P* !

- Bootstrap: P* = S (this is all what we know...)
- What about S* ?

What about the new learning sample S* ?

- The new learning sample S* must be randomly drawn from what we know (i.e. P*=S)
- To keep the same size: draw with replacement

P* = S		S*
1		4
2		6
3		3
4		7
5		3
6	Same size !	10
7		5
8		3
9		9
10		10

The intuitive idea can be easely rewritten

We have $E_{S,P}$ - $E_{S,S}$ = optimism

=

We use E_{S^*,P^*} - E_{S^*,S^*} = difference

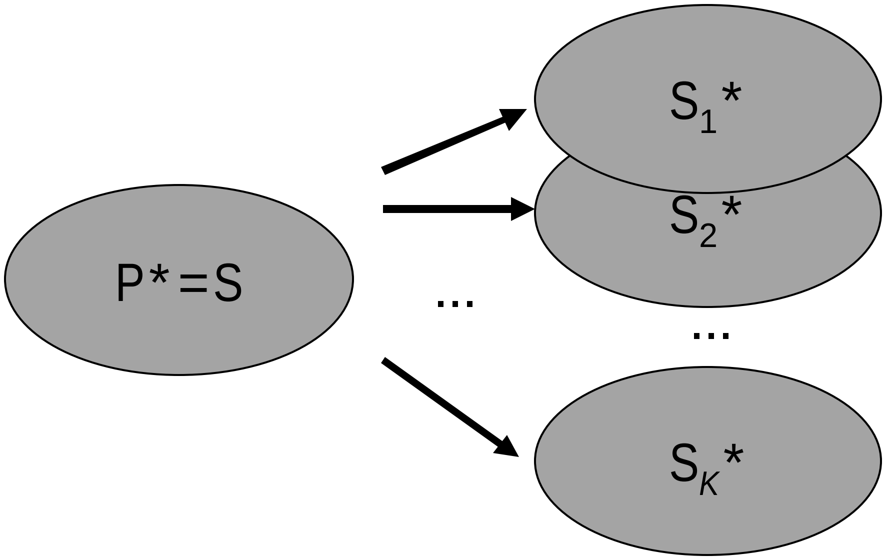
$$
\begin{aligned}
\hat{E}_{gen} &= E_{S,P} + E_{S,S} - E_{S,S} \\
&= E_{S,S} + \left(E_{S,P} - E_{S,S}\right) \\
&\overset{\text{Plug-in}}{=} E_{S,S} + \left(E_{S^*,P^*} - E_{S^*,S^*}\right) \\
&= E_{S,S} + \widehat{optimism}
\end{aligned}
$$

- Last step: to estimate

Estimation of the optimism

$$\hat{optimism} = E_{S^*,P^*} - E_{S^*,S^*}$$

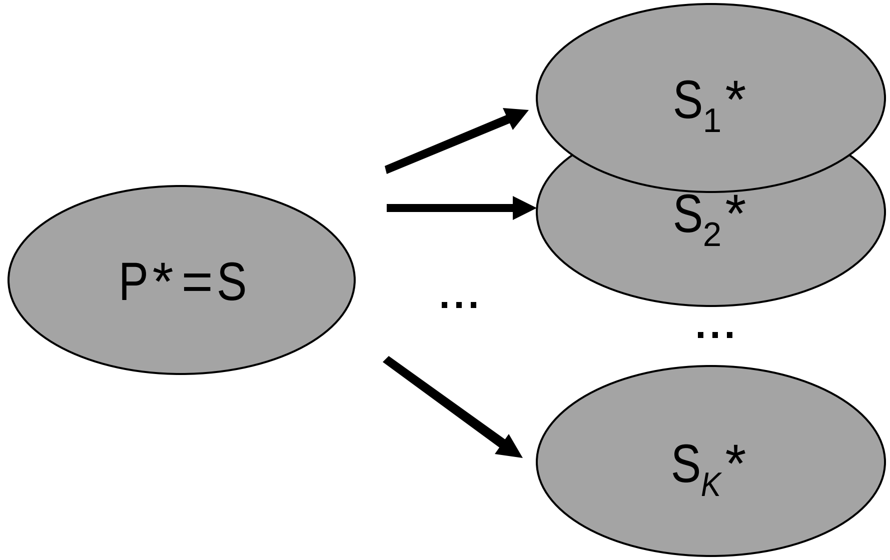
- One estimation is not enough; fluctuations due to
  - sample S^*
  - learning (initial conditions, local minima,...)
  - other numerical problems
- Then: repeat !



$$\hat{optimism} = \frac{1}{K}\sum_{k=1}^{K}\left(E_{S^{*k},P^*} - E_{S^{*k},S^{*k}}\right)$$

In summary...

- This is what we know:



- We draw
- Then we compute

K+1 learnings

$$\hat{E}_{gen} = E_{S,S} + \frac{1}{K}\sum_{k=1}^{K}\left(E_{S_k^*,P^*} - E_{S_k^*,S_k^*}\right)$$

2K+1 evaluations

It seems very nice, but...

- The number of experiments grows dramatically!

- Context: model selection (*Q* models to test, $1=q=Q$)

Model selection:
Several model structures to test

Bootstrap:
Several evaluations of the optimism

Nonlinear models:
Necessity to repeat learning
R repetitions

K repetitions

Q repetitions

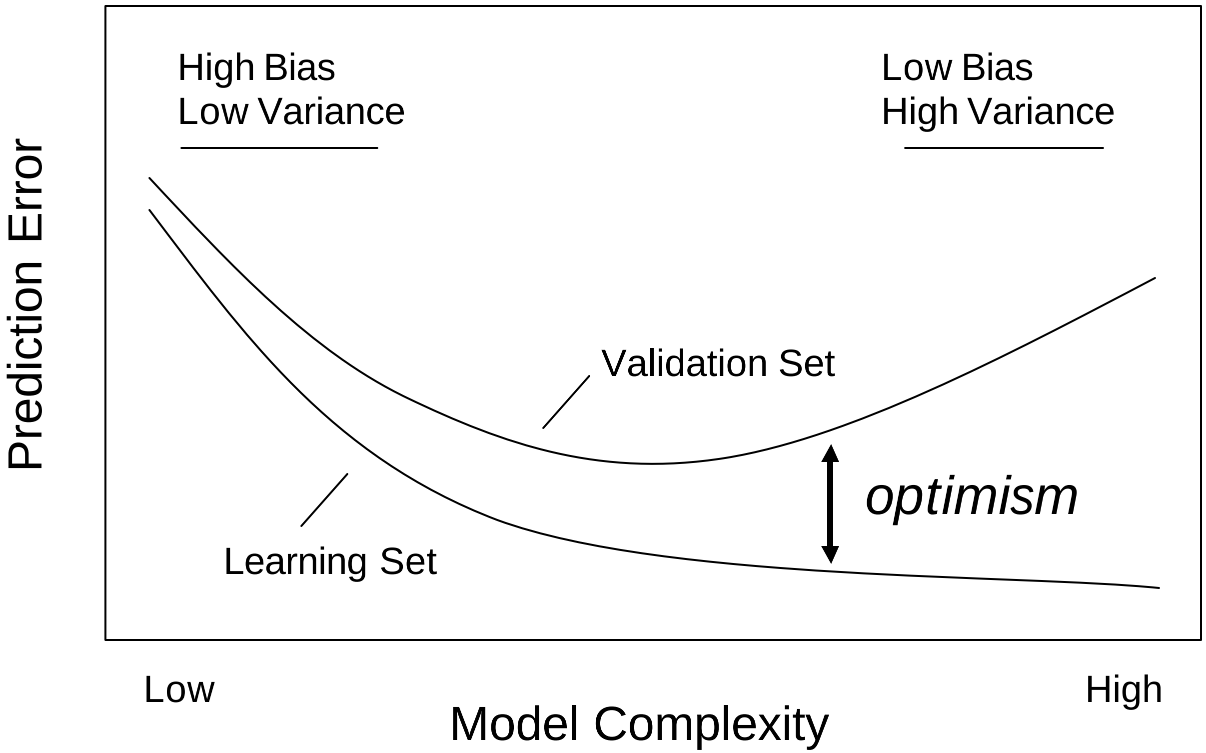
- *R* x *K* x *Q* : too much!

Outline

- Model selection
  - estimation of generalization error
  - resampling methods
- Boostrap
  - plug-in principle
  - computational load
- Fast Bootstrap
  - idea and hypothesis
  - reduced number of experiments
- Experiments
  - artificial regression example
  - Santa Fe A time series prediction

Fast bootstrap

- Attempt to decrease K x Q



- Idea: *optimism* is a very smooth function
  – A polynom of order *o*

Is it justified ?

- Yes, experimentally
- Yes, see AIC and BIC:

  – AIC: $\hat{E}_{gen}(q) = \sum_{t=1}^{N} \frac{(g(x_t, q) - y_t)^2}{N} + \frac{2}{N} \dim(q)$

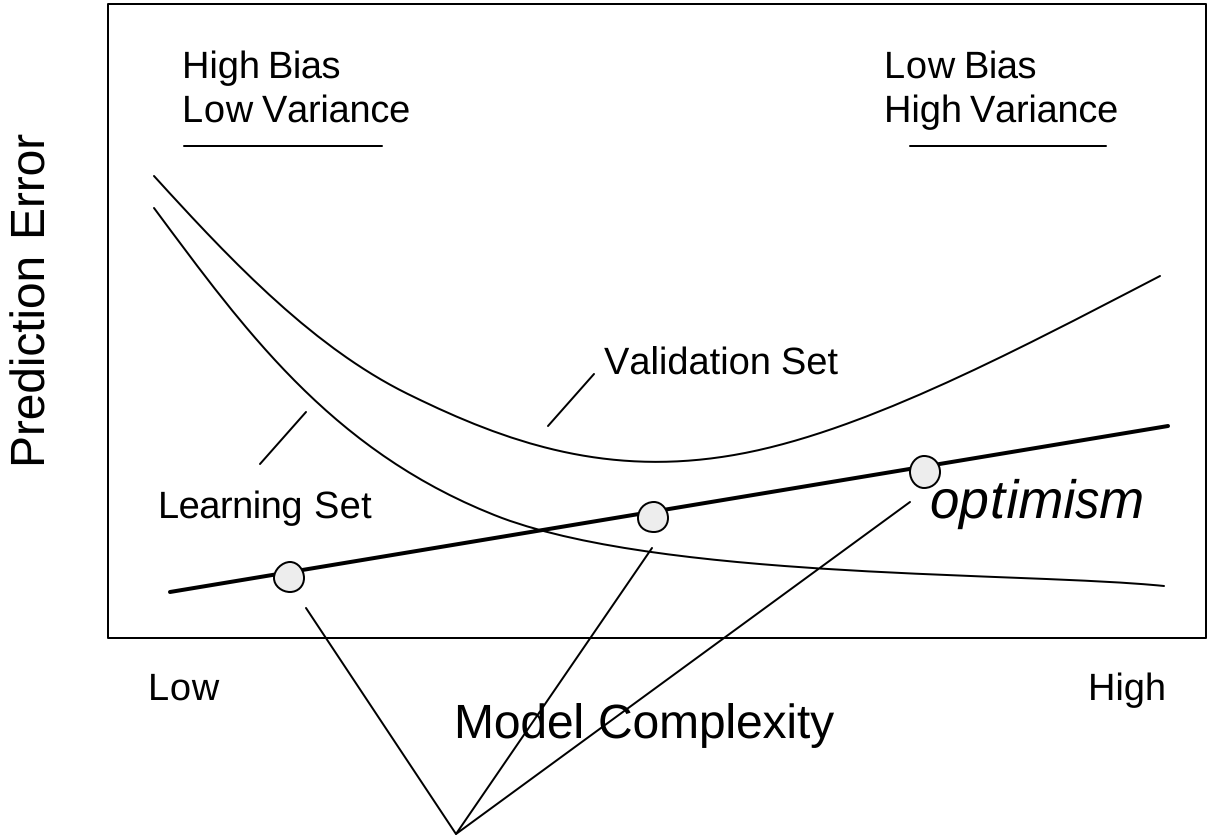
  – BIC: $\hat{E}_{gen}(q) = \sum_{t=1}^{N} \frac{(g(x_t, q) - y_t)^2}{N} + \frac{\ln(N)}{N} \dim(q)$

  – Extension to nonlinear models

  $\dim(q)$ is replaced by $\dim(q)\hat{s} = \dim(q) \frac{\sum_{t=1}^{N} (\hat{y}_t - y_t)^2}{N - \dim(q)}$

- In all cases: *optimism* proportional to dim(θ)
  – Polynom of order *o*=1 !

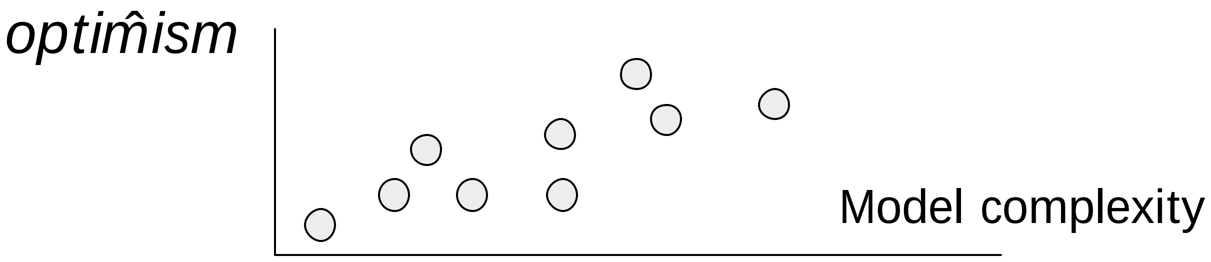
Where is the lower number of experiments ?



- Lower number Q of models to test
- Each model to test: possibility to decrease K (evaluation noise averaged in fitting)

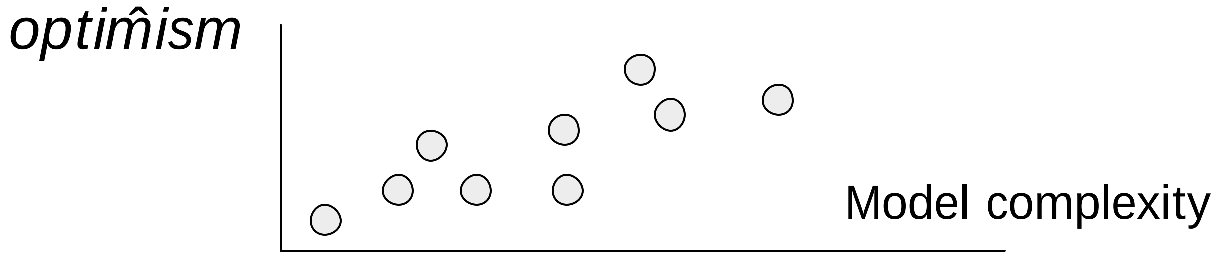
Yes, but... how many experiments ?

- How to choose Q x K in practise?
- Try... and test
- Example:
  - Q models are tested (each one needs K bootstrap samples)



  - Is it linear ? Use Fisher's statistics !

Yes, but… how many experiments ? 2/



- Test o_1=1 (linear) against o_2=*2*

$$F_{o_2-o_1,Q-o_2-1} = \frac{\dfrac{SR_1 - SR_2}{o_2 - o_1}}{\dfrac{SR_2}{Q - o_2 - 1}} \qquad \left(SR = \sum_{t=1}^{N} \left(\hat{y}_t - y_t\right)^2 \right)$$

- If test passes:
  - It is linear
  - We have enough experiments !

Outline

- Model selection
  - estimation of generalization error
  - resampling methods
- Boostrap
  - plug-in principle
  - computational load
- Fast Bootstrap
  - idea and hypothesis
  - reduced number of experiments
- Experiments
  - artificial regression example
  - Santa Fe A time series prediction

And now… the experiments

- Two datasets:
  - Artificial function approximation problem
  - Santa Fe A time series forecasting

- Three models
  - Multi-Layer Perceptron
  - Radial-Basis Function Network
  - Least-Square Support Vector Machine

The model structure parameter

- Multi-Layer Perceptron

$$\hat{y}_t = \sum_{i=1}^{M} w_i \tanh\left(\sum_{j=1}^{D} w_{ij} x_{tj}\right)$$

- Radial-Basis Function Networks

$$\hat{y}_t = \sum_{i=1}^{M} w_i \exp\left(-\frac{\left\|\mathbf{x}_t - \mathbf{c}_i\right\|}{2s_i^2}\right)$$

Model structure parameters

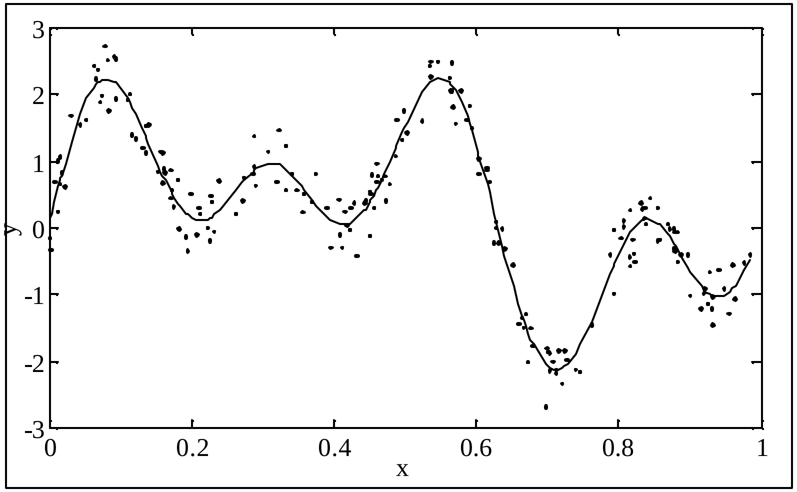
- Least-Square Support Vector Machine

$$\hat{y}_t = \mathbf{?}^T j\left(\mathbf{x}_t\right)$$

  – But optimization is regularized

$$\min_{\mathbf{?},\mathbf{e}} J\left(\mathbf{?},\mathbf{e}\right) = \mathbf{?}^T \mathbf{?} + g\sum_{i=1}^{N}\left(y_t - \hat{y}_t\right)$$

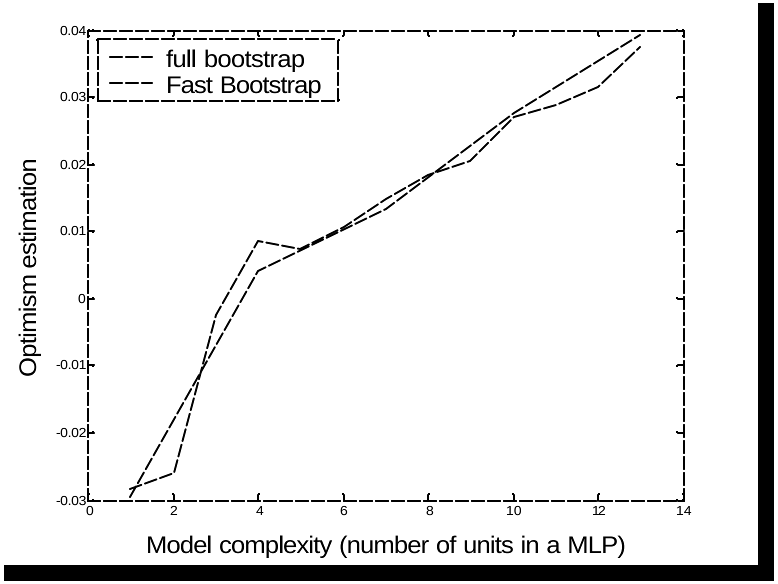
Artificial regression example

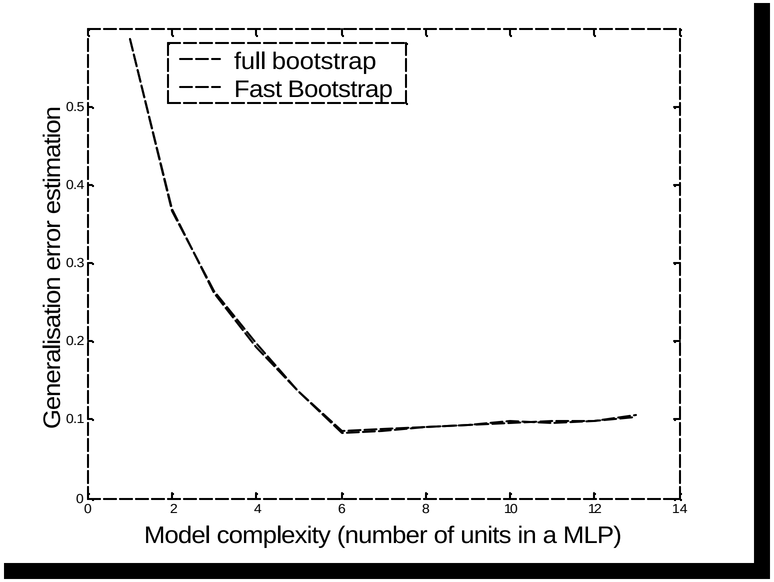


$$y_t = \sin(5x_t) + \sin(15x_t) + \sin(25x_t) + e_t$$

- 200 sample
- distribution of ε_t : i.i.d. uniformly in [-0.5,0.5]

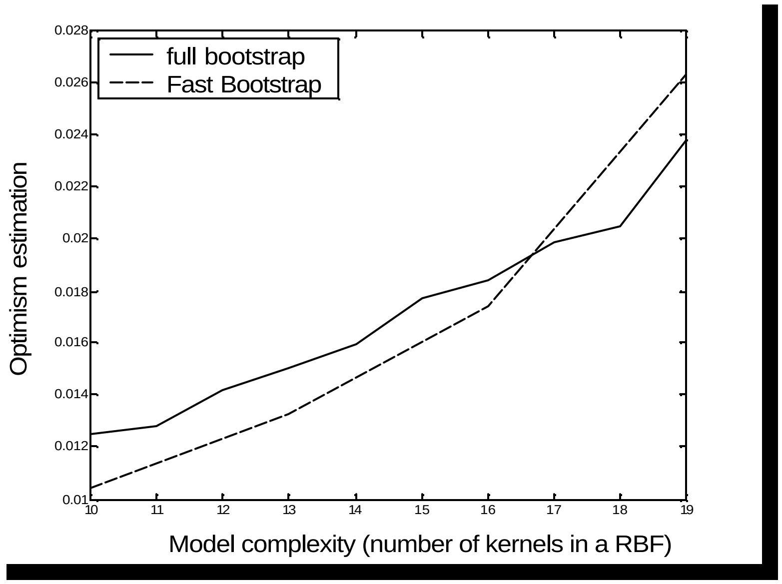
MLP on artificial regression example

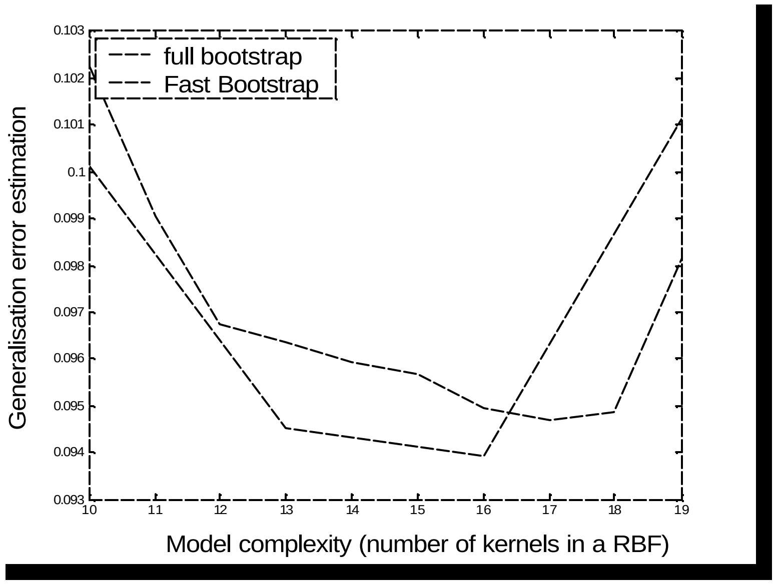




MLP	*Number of hidden neurons*	*Bootstrap replications*	*Number of experiments*	*Gain*	$F_{1,2}$
Bootstrap	1-13 by steps of 1	100	1300		
Fast Bootstrap	1-13 by steps of 3	10	50	96.2%	3.25

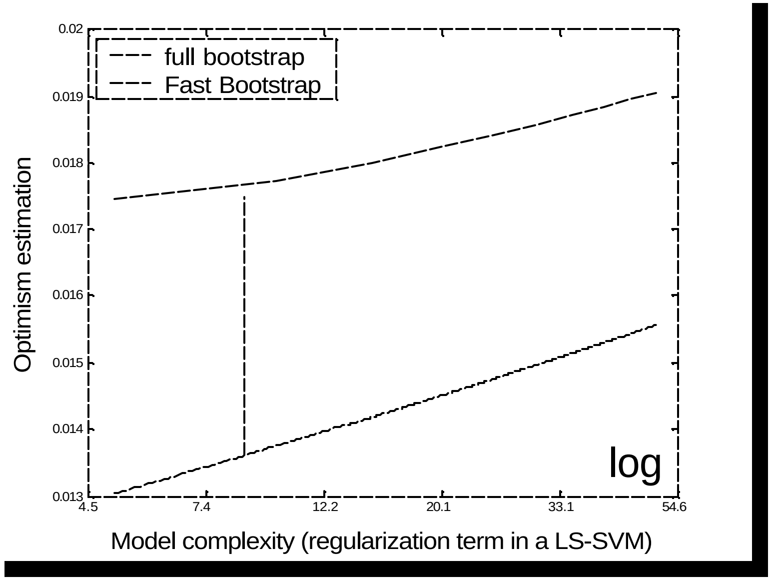
RBFN on artificial regression example

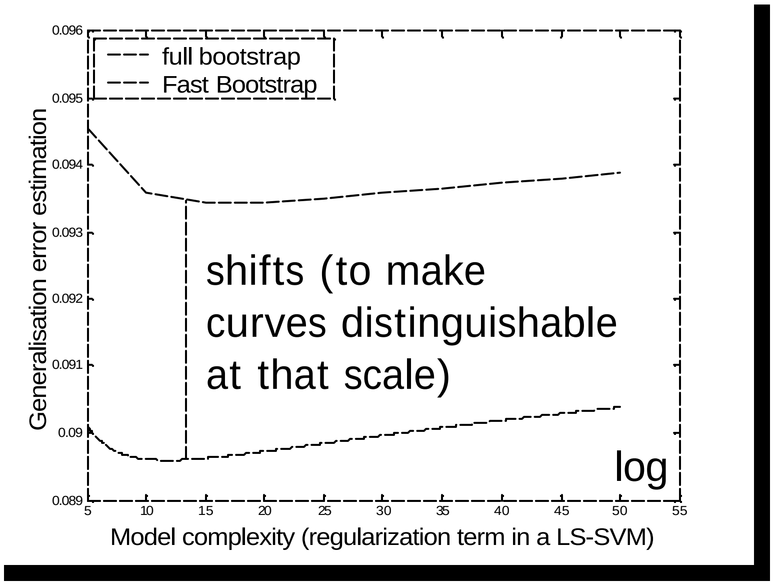




RBFN	*Number of kernels*	*Bootstrap replications*	*Number of experiments*	*Gain*	$F_{1,1}$
Bootstrap	10-19 by steps of 1	100	1000		
Fast Bootstrap	10-19 by steps of 3	10	40	96%	16.31

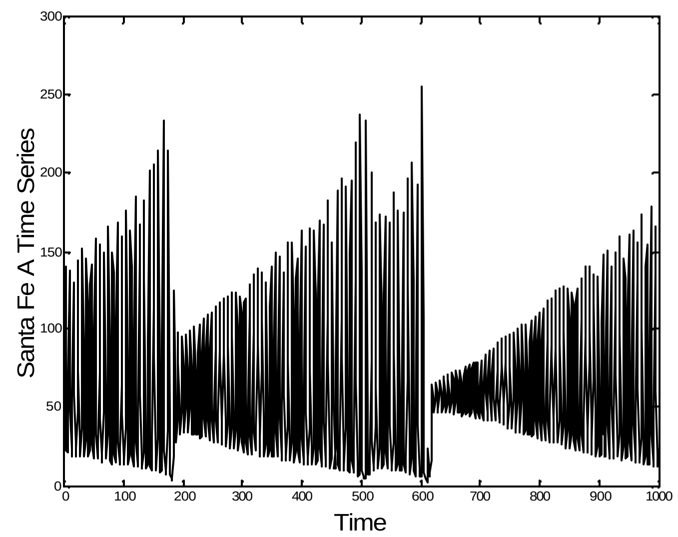
LS-SVM on artificial regression example





LS-SVM	*Regularization g*	*Bootstrap replications*	*Number of experiments*	*Gain*	$F_{1,7}$
Bootstrap	5-50 by steps of 0.1	100	45100		
Fast Bootstrap	5-50 by steps of 5	10	100	99.8%	0

Santa Fe A time series forecasting



- Far-infrared laser in a chaotic state

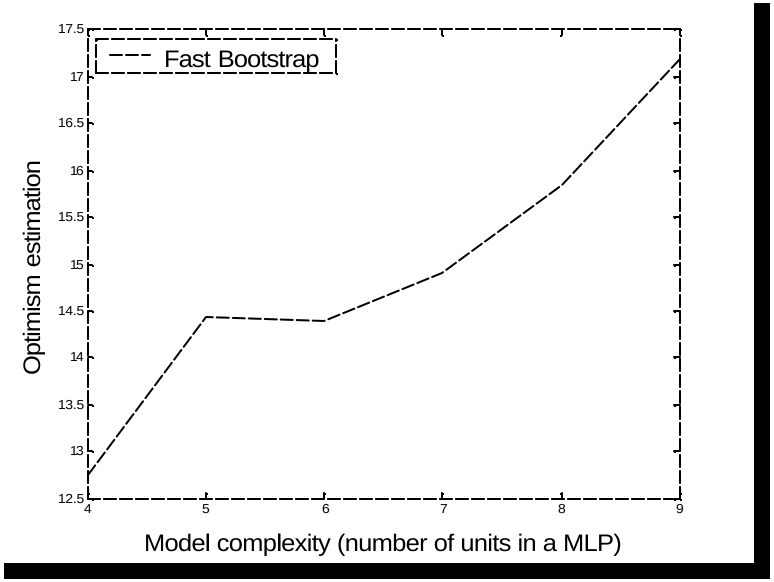
$$\hat{y}_t = g\left(\underbrace{y_t, y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5}}_{\mathbf{x}_t}, q(q)\right)$$

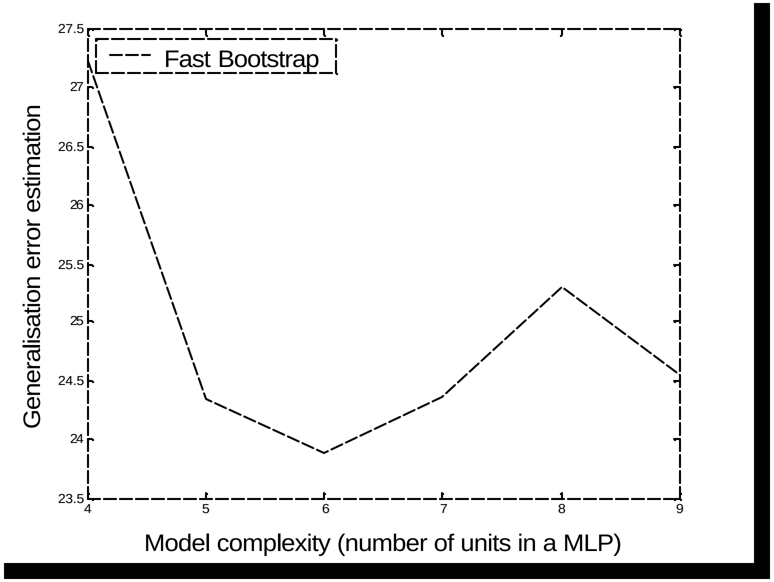
MLP, RBFN, LS-SVM

parameters

model structure parameter

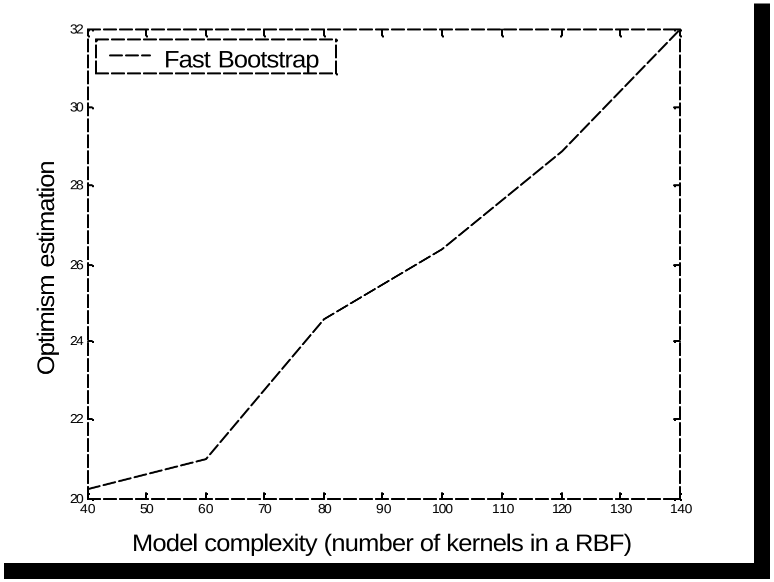
MLP on Santa Fe A time series example

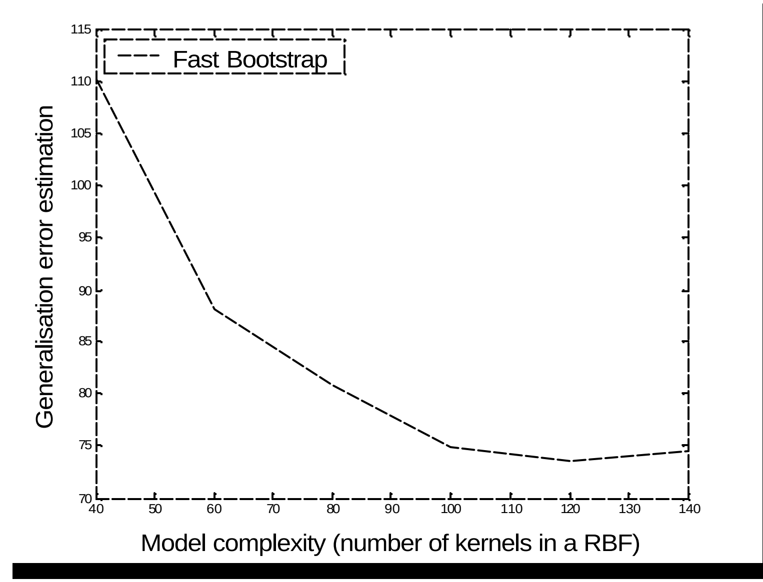




MLP	*Number of hidden neurons*	*Bootstrap replications*	*Number of experiments*	$F_{1,3}$
Fast Bootstrap	4-9 by steps of 1	10	60	0.2002

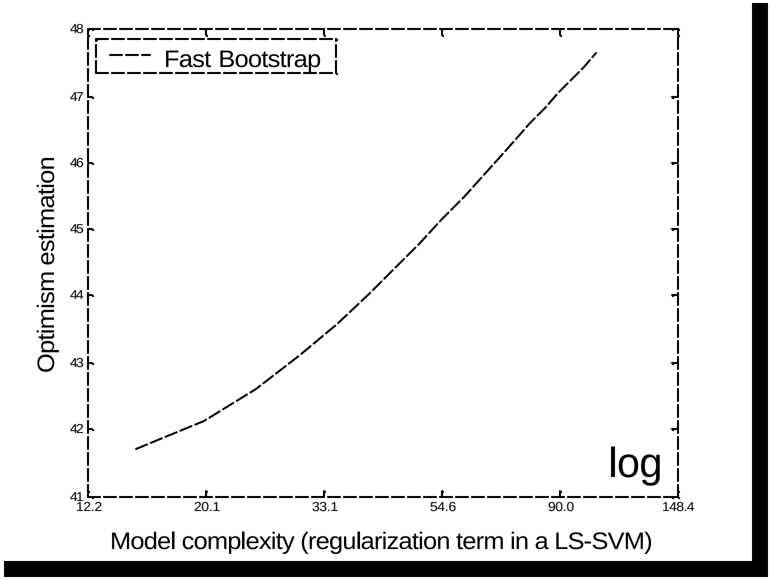
RBFN on Santa Fe A time series example

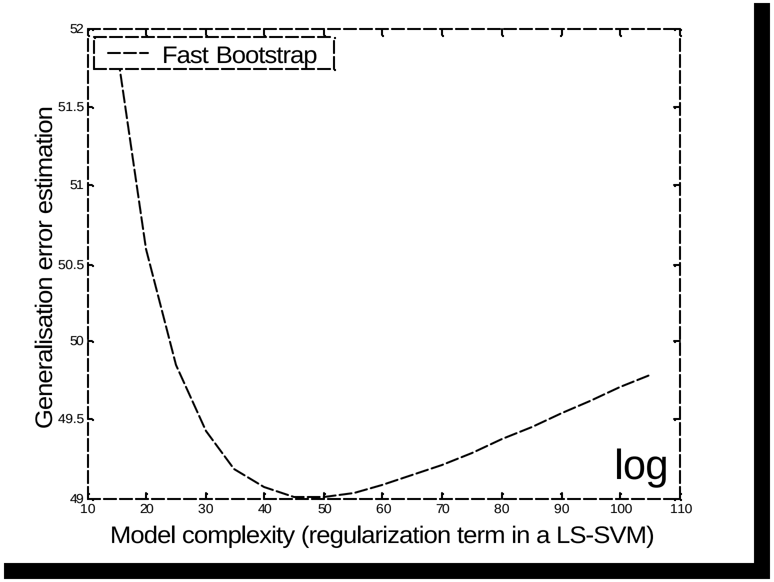




RBFN	*Number of kernels*	*Bootstrap replications*	*Number of experiments*	$F_{1,2}$
Fast Bootstrap	60-140 by steps of 20	20	100	1.6472

LS-SVM on Santa Fe A time series example





LS-SVM	*Regularization g*	*Bootstrap replications*	*Number of experiments*	$F_{1,16}$
Fast Bootstrap	15-105 by steps of 5	10	190	0

To conclude

- Bootstrap: efficient model selection method
- But: computationally intensive!
- Solution: Fast Boostrap
  - uses smoothness hypothesis on *optimism*
  - hypothesis seems reasonable
  - makes the number of experiments decrease by 1-2 orders of magnitude!
- Further work:
  - theoretical insights about hypothesis
  - Moody's Generalized Prediction Error could explain the log in LS-SVM ?

Thanks to...

- Amaury Lendasse and Geoffroy Simon

for

- the initial idea
- the experiments !