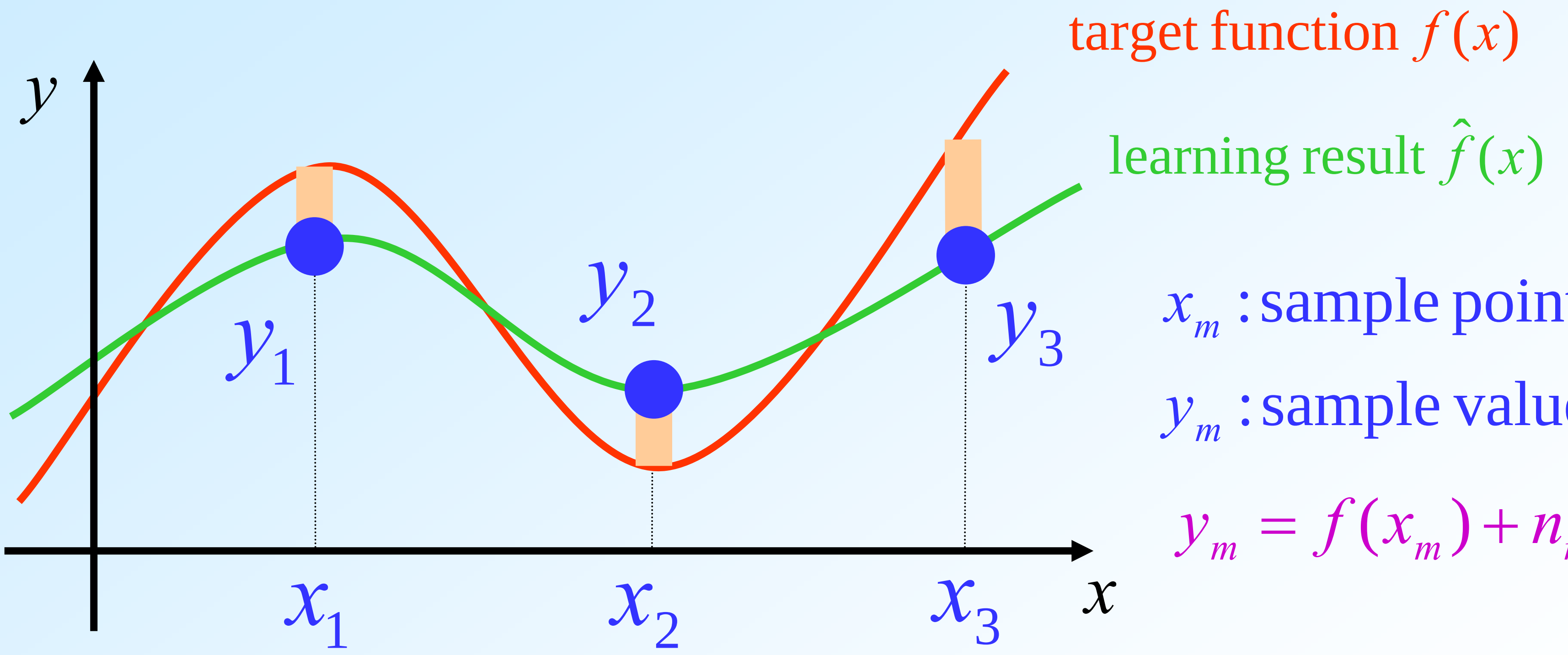
A New Information Criterion for the Selection of Subspace Models

Department of Computer Science,
Tokyo Institute of Technology, Japan

Masashi Sugiyama
Hidemitsu Ogawa

Function Approximation



x_m : sample point

y_m : sample value

$$y_m = f(x_m) + n_m$$

Obtain the optimal approximation $\hat{f}(x)$ to $f(x)$ by using the training examples $\{x_m, y_m\}_{m=1}^{M}$.

Model

Generally, function approximation is performed by estimating parameters of a prefixed set of functions called a model.

polynomial

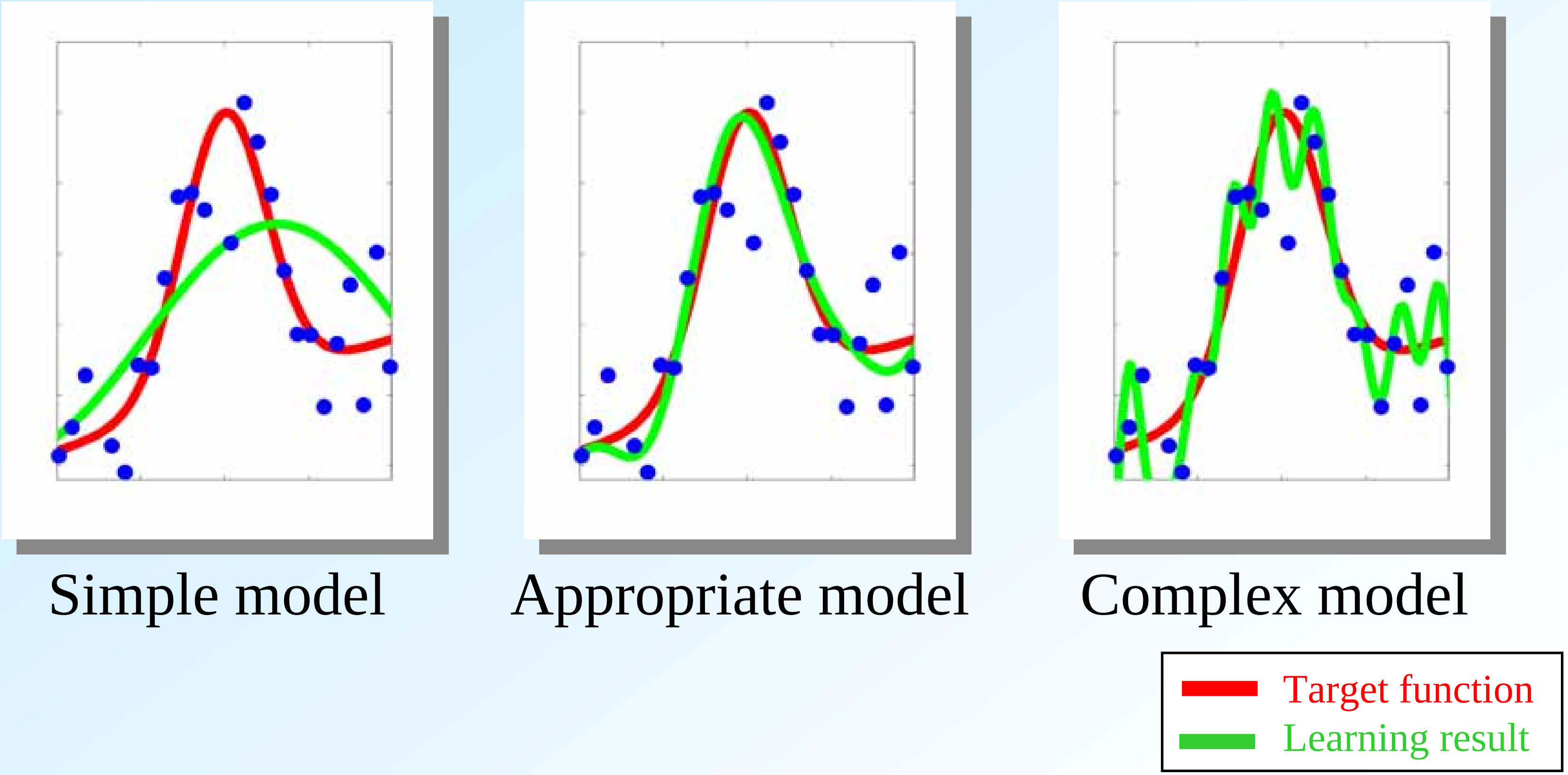
$$\hat{f}(x) = \sum_{n=0}^{N} a_n x^n$$

3-layer neural networks

$$\hat{f}(x) = \sum_{n=1}^{N} a_n \sigma(x; b_n)$$

The choice of the model complexity
(e.g. order of polynomial, number of units)
is crucial for optimal generalization.

Model Selection



Select the best model providing the optimal generalization capability.

Motivation and goal

Most of the traditional model selection criteria
do not work well
when the number of training examples is small.

e.g. AIC (Akaike, 1974),
BIC (Schwarz, 1978),
MDL (Rissanen, 1978),
NIC (Murata, Yoshizawa, & Amari, 1994)

Devise a model selection criterion which works well even when the number of training examples is small.

Setting

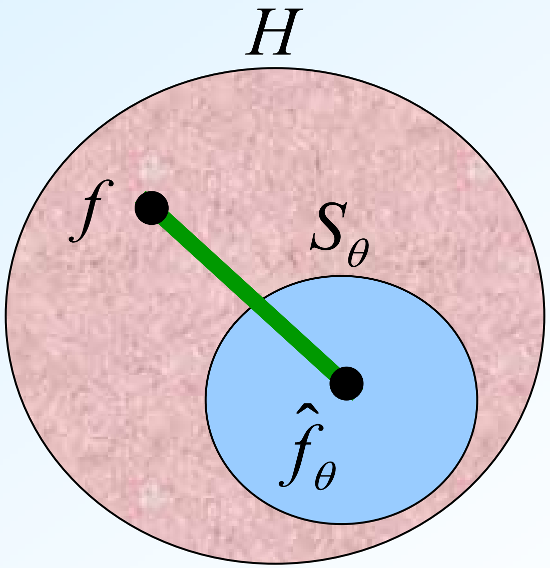
f : learning target function

θ : model

S_θ : family of functions indicated by model θ

$\hat{f}_\theta$: learning result function by model θ

H : Hilbert space including f, S_θ, and $\hat{f}_\theta$



Select, from a set of models, the model minimizing

$$E_n \left\| \hat{f}_\theta - f \right\|^2$$

E_n : expectation over noise

Least mean squares (LMS) learning

LMS learning is aimed at minimizing the training error

$$\sum_{m=1}^{M} \left| \hat{f}_\theta(x_m) - y_m \right|^2$$

The LMS learning result function $\hat{f}_\theta$ is given as

$$\hat{f}_\theta = X_\theta y \quad : \quad X_\theta = \left(\sum_{m=1}^{M} \left(e_m \otimes \overline{K_\theta(x, x_m)} \right) \right)^+$$

$y = (y_1, y_2, \cdots, y_M)$

e_m : m-th standard basis in C^M

$K_\theta(x, x')$: reproducing kernel of S_θ

$+$: Moore – Penrose generalized inverse

$(f \otimes \overline{g})$: Neumann – Schatten product

$$(f \otimes \overline{g}) h = \langle h, g \rangle f$$

Assumptions (1)

The mean noise is zero.

The noise covariance matrix is given as $\sigma^2 I$.

σ^2 is generally unknown.

Assumptions (2)

One of the models gives an unbiased learning result $\hat{f}_u$.

$$E_n \hat{f}_u = f \quad : \quad \hat{f}_u = X_u y$$

If $\{K_H(x, x_m)\}_{m=1}^M$ span H, then $X_u = \left(\sum_{m=1}^M \left(e_m \otimes \overline{K_H(x, x_m)} \right) \right)^+$

$K_H(x, x')$: reproducing kernel of H

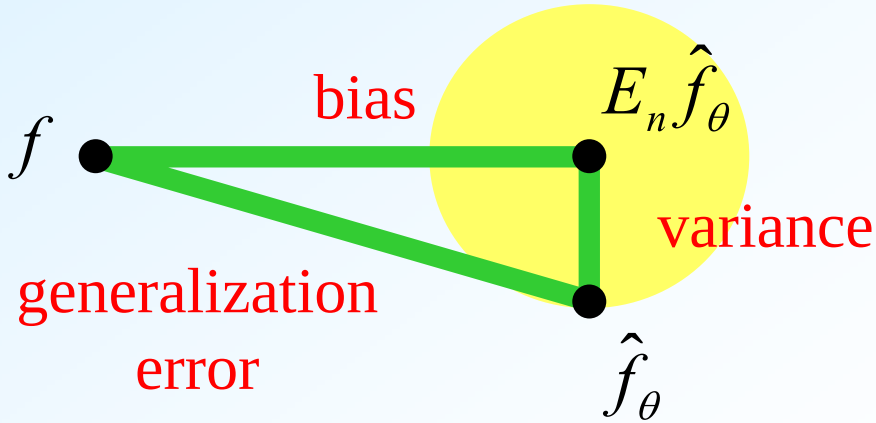
Roughly speaking, $\{K_H(x, x_m)\}_{m=1}^M$ span H if $M \geq \dim(H)$

M : the number of training examples

Generalization error and bias/variance

$$\underbrace{E_n\left\|\hat{f}_\theta - f\right\|^2}_{\text{generalization error}} = \underbrace{\left\|E_n\hat{f}_\theta - f\right\|^2}_{\text{bias}} + \underbrace{E_n\left\|\hat{f}_\theta - E_n\hat{f}_\theta\right\|^2}_{\text{variance}}$$

E_n : expectation over noise



Estimation of bias

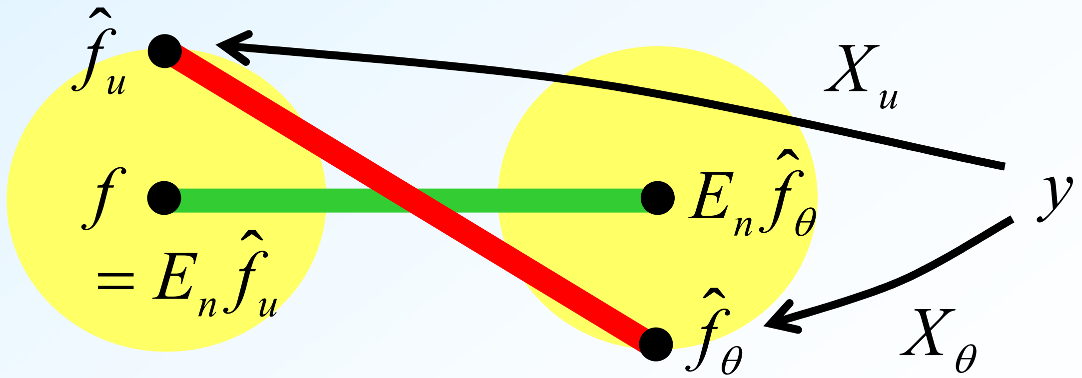
POINT!

$$\underbrace{\left\|E_n\hat{f}_\theta - f\right\|^2}_{\text{bias}} = \left\|\hat{f}_\theta - \hat{f}_u\right\|^2 - 2\,\mathrm{Re}\left\langle E_n\hat{f}_\theta - f,\, X_0 n\right\rangle - \left\|X_0 n\right\|^2$$

$$E_n \Downarrow \qquad\qquad E_n \Downarrow$$

$$\approx \left\|\hat{f}_\theta - \hat{f}_u\right\|^2 - \quad 0 \quad - \sigma^2\,\mathrm{tr}\left(X_0 X_0^*\right)$$

$X_0 = X_\theta - X_u$, $n = (n_1, n_2, \cdots, n_M)^T$, σ^2 : noise variance, X_0^* : adjoint operator of X_0



Estimation of noise variance

$$\underbrace{E_n\left\|\hat{f}_\theta - f\right\|^2}_{\text{generalization error}} \approx \underbrace{\left\|\hat{f}_\theta - \hat{f}_u\right\|^2 - \sigma^2\,\mathrm{tr}\left(X_0 X_0^*\right)}_{\text{bias estimate}} + \underbrace{\sigma^2\,\mathrm{tr}\left(X_\theta X_\theta^*\right)}_{\text{variance}}$$

σ^2 : noise variance, $X_0 = X_\theta - X_u$, X^* : adjoint operator of X

$$\boxed{\hat{\sigma}^2 = \frac{\sum_{m=1}^{M}\left|\hat{f}_u(x_m) - y_m\right|^2}{M - \dim(H)}}$$

$\hat{\sigma}^2$ is an unbiased estimate of σ^2

Subspace Information Criterion (SIC)

From a set of models, select the model minimizing the following SIC.

$$\mathrm{SIC} = \left\| \hat{f}_{\theta} - \hat{f}_{u} \right\|^{2} - \hat{\sigma}^{2}\,\mathrm{tr}\left(X_{0}X_{0}^{*}\right) + \hat{\sigma}^{2}\,\mathrm{tr}\left(X_{\theta}X_{\theta}^{*}\right)$$

POINT!

The model minimizing SIC is called the minimum SIC model (MSIC model).

MSIC model is expected to provide the optimal generalization capability.

Validity of SIC

SIC gives an unbiased estimate
of the generalization error:

$$E_n \mathrm{SIC} = E_n \left\| \hat{f}_\theta - f \right\|^2$$

POINT!

E_n : expectation over noise

cf. AIC gives an asymptotic unbiased estimate
of the generalization error.

SIC will work well even when
the number of training examples is small.

Illustrative Simulation

$$f(x) = \sqrt{2}\sin x + 2\sqrt{2}\cos x - \sqrt{2}\sin 2x - 2\sqrt{2}\cos 2x + \sqrt{2}\sin 3x$$
$$- \sqrt{2}\cos 3x + 2\sqrt{2}\sin 4x - \sqrt{2}\cos 4x + \sqrt{2}\sin 5x - \sqrt{2}\cos 5x$$

$$x_m = -\pi - \frac{\pi}{M} + \frac{2\pi m}{M}, \qquad y_m = f(x_m) + n_m$$

n_m : subject to $N(0,3)$

$$\boxed{\text{compared models} : \{ S_1, S_2, \cdots, S_{20} \}}$$

S_N : Hilbert space spaned by $\{1, \sin nx, \cos nx\}_{n=1}^{N}$ defined on $[-\pi, \pi]$

Compared model selection criteria

- SIC

$$H = S_{20} \quad : \quad \dim(H) = 41$$

- Network information criterion (NIC)

(Murata, Yoshizawa, & Amari, 1994)

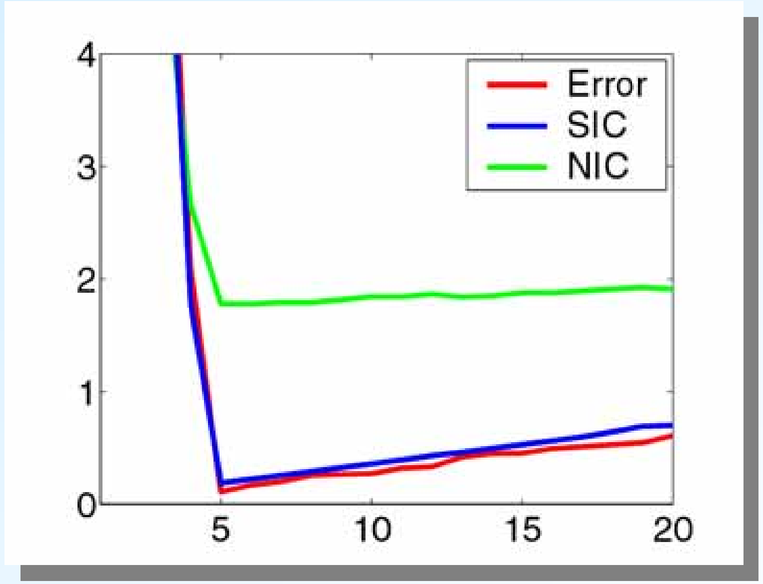
A generalized AIC

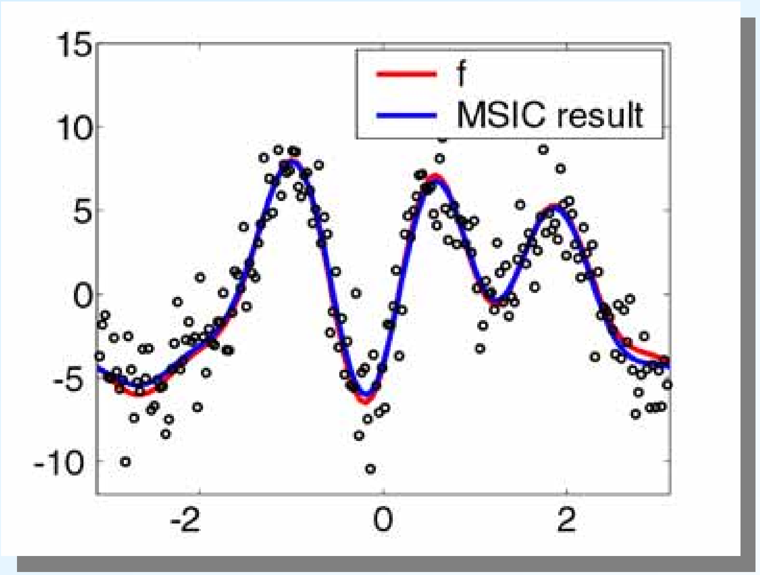
In this simulation, SIC and NIC are fairly compared.

$$\text{Error} = \left\| \hat{f} - f \right\|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \hat{f}(x) - f(x) \right|^2 dx$$

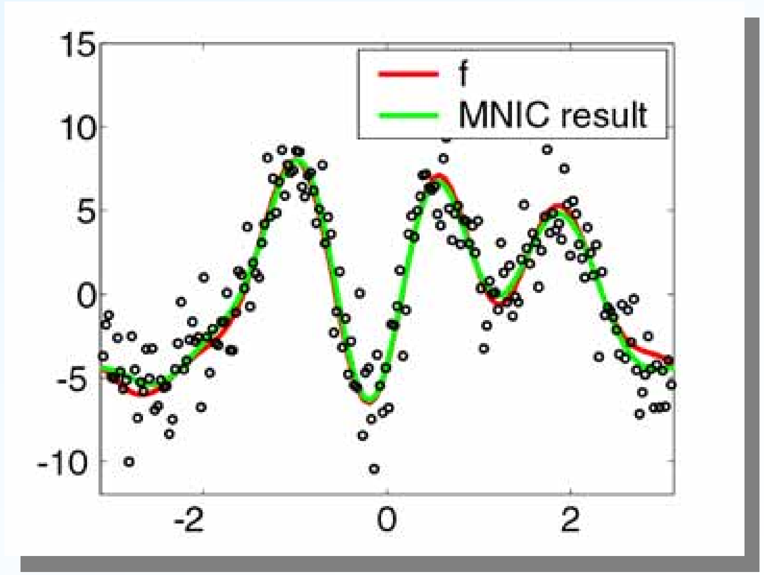
$M = 200$

Optimal model S_5 (Error = 0.11)





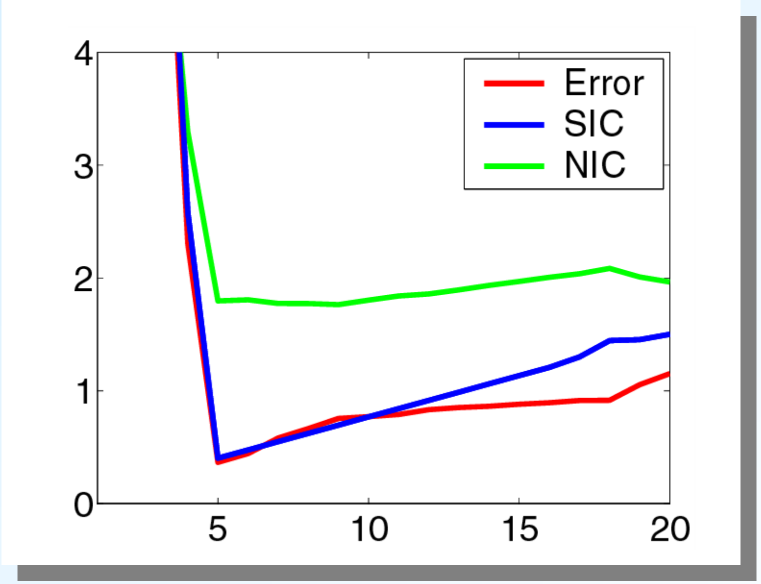
MSIC model S_5 (Error = 0.11)

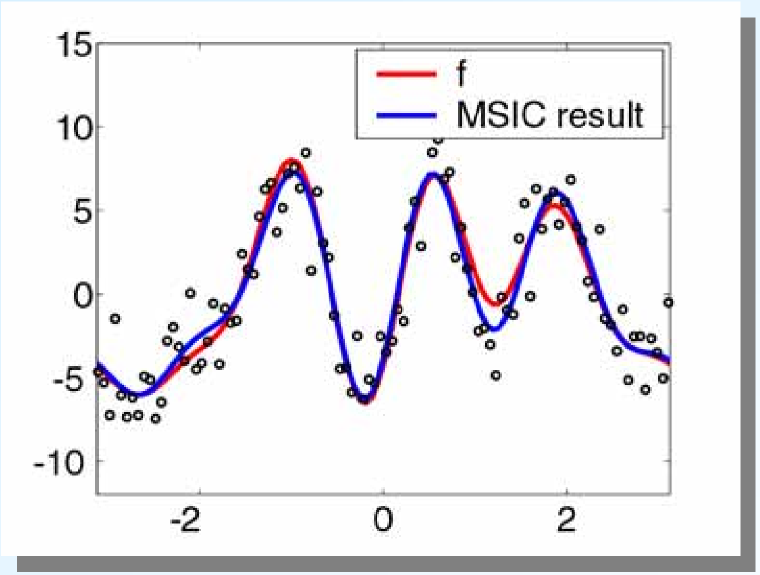


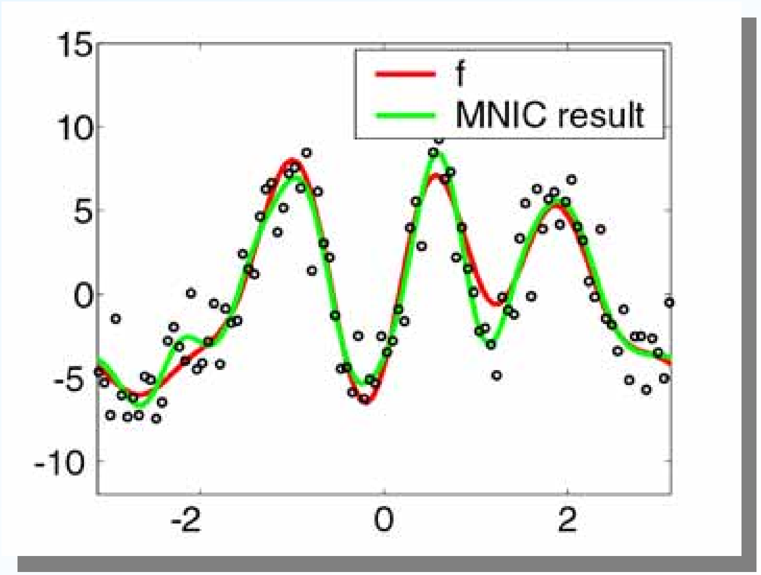
MNIC model S_6 (Error = 0.17)

$M = 100$

Optimal model S_5 (Error = 0.37)







MSIC model S_5 (Error = 0.37)

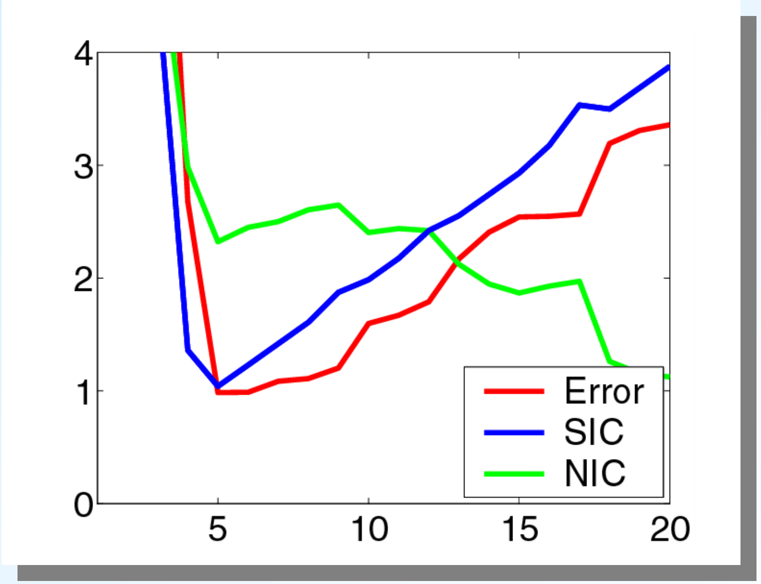
MNIC model S_9 (Error = 0.75)

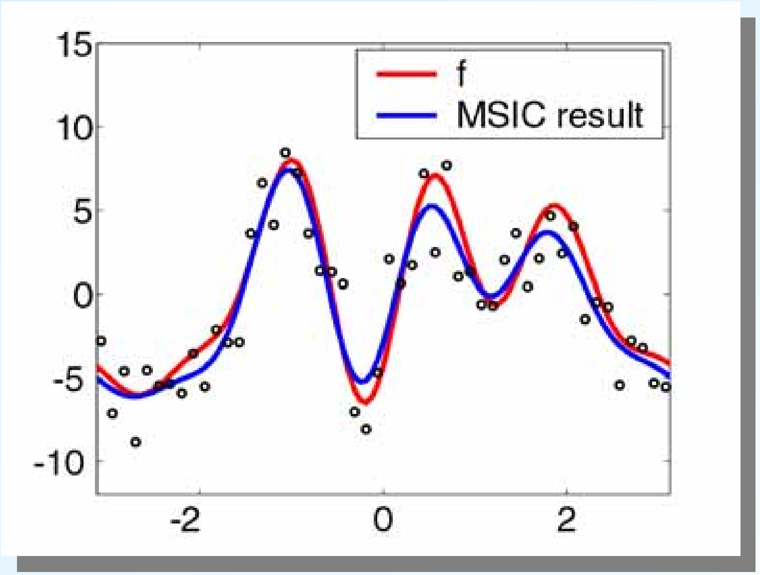
$M = 50$

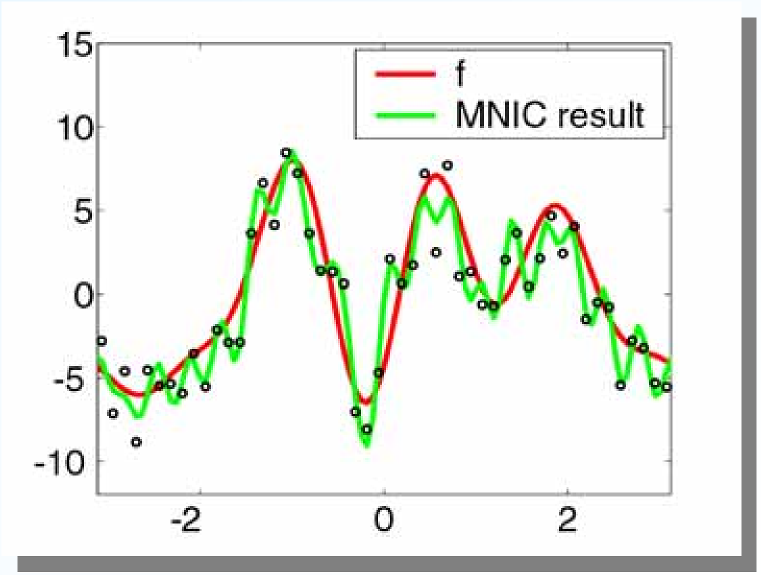
Optimal model S_5 (Error = 0.98)

POINT!

SIC works well even when M is small.







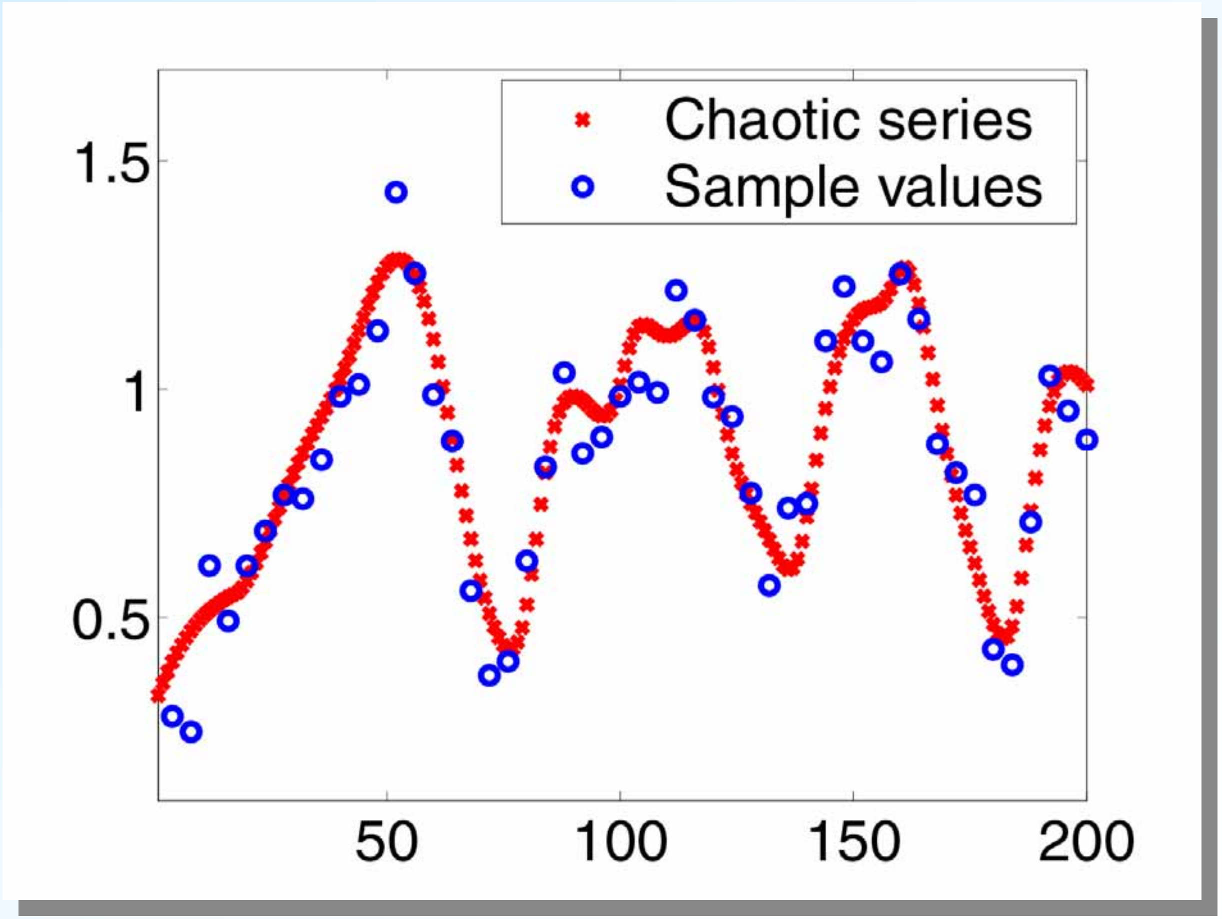
MSIC model S_5 (Error = 0.98)

MNIC model S_{20} (Error = 3.36)

Unrealizable case

Estimate a chaotic series $\left\{ h_p \right\}_{p=1}^{200}$ from M sample values $\left\{ y_m \right\}_{m=1}^{M}$

$M = 100$



Estimation of chaotic series

Consider sample point $x_p = -0.995 + \frac{2}{200}(p-1)$

corresponding to the chaotic series $\{h_p\}_{p=1}^{200}$

$$\hat{h}_p = \hat{f}\left(-0.995 + \frac{2}{200}(p-1)\right)$$ is an estimate of h_p

$$\text{Error} = \sum_{p=1}^{200} \left| \hat{h}_p - h_p \right|^2$$

We perform the simulation 1000 times.

Compared model selection criteria

- SIC

$$H = S_{40} \quad : \quad \dim(H) = 41$$

- NIC

log loss is adopted as the loss function.

$\{ x_m \}_{m=1}^{M}$ are regarded as uniformly distributed.

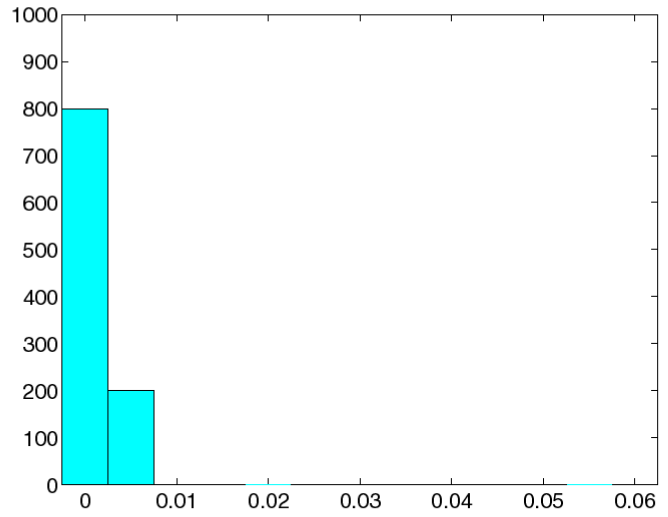
Compared models : $\{ S_{15}, S_{20}, S_{25}, S_{30}, S_{35}, S_{40} \}$

S_N : Hilbert space spaned by $\{x\}_{n=0}^{N}$

defined on $[-1,1]$

$M = 250$

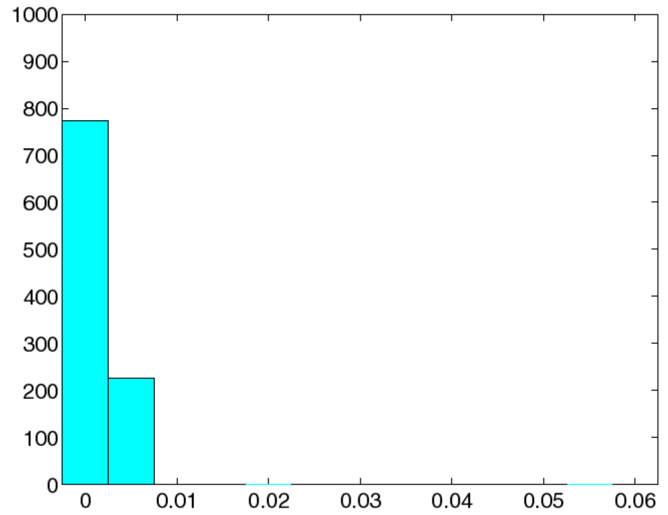
SIC



Mean

0.0021

NIC

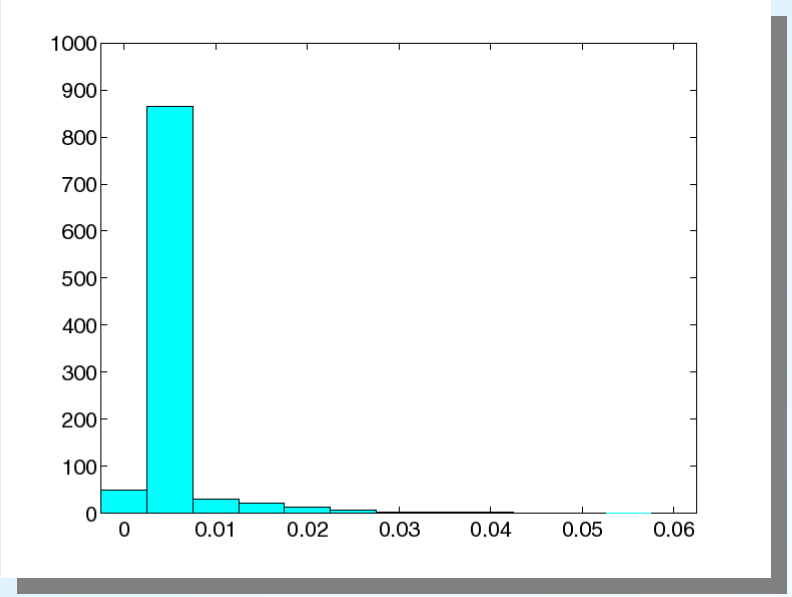


Mean

0.0022

$M = 150$

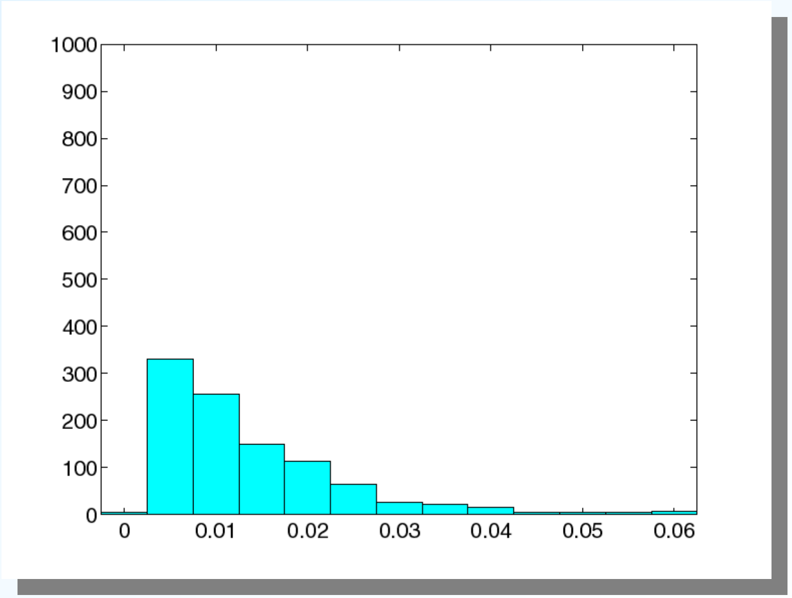
SIC



Mean

0.0058

NIC



Mean

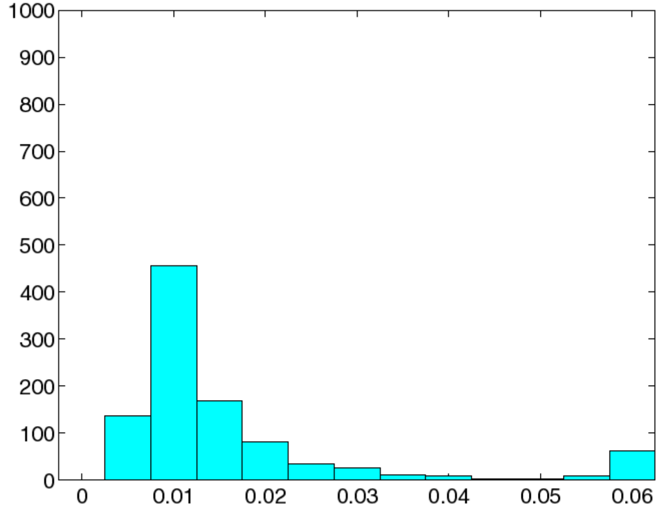
0.013

$$M = 50$$

POINT!

SIC works well even when M is small.

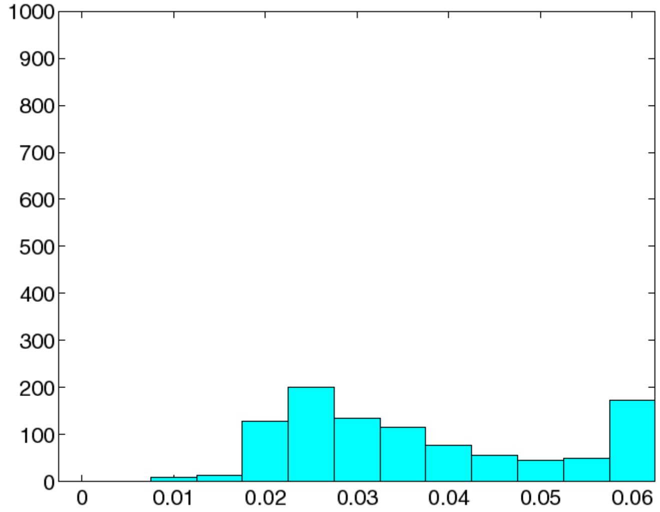
SIC



Mean

0.018

NIC



Mean

0.040

Conclusions

- We proposed a new model selection criterion named the subspace information criterion (SIC) .
- SIC gives an unbiased estimate of the generalization error.
- SIC works well even when the number of training examples is small.