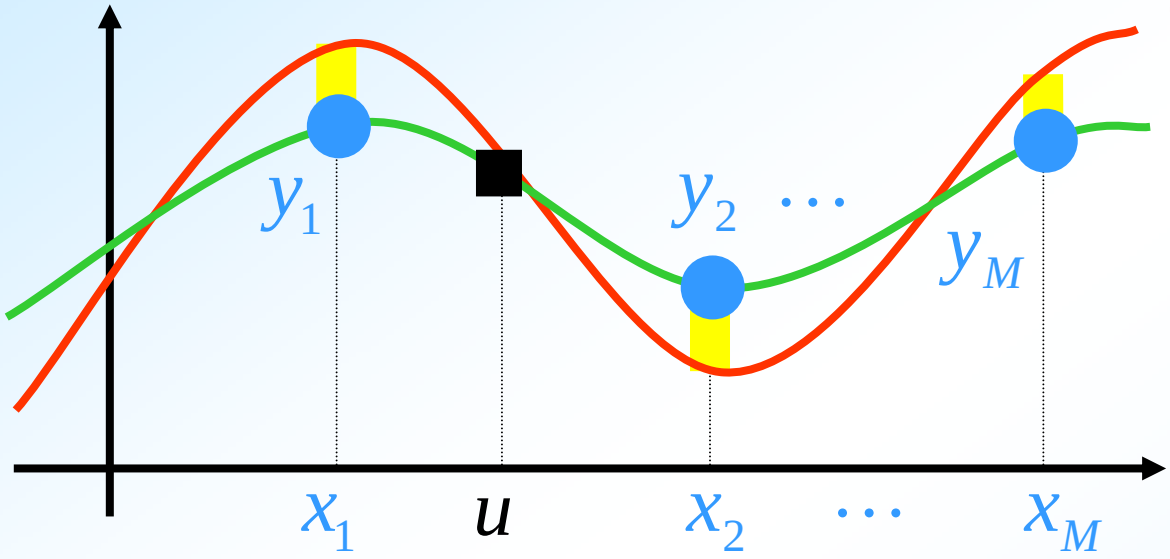
Model Selection with Small Samples

Department of Computer Science,
Tokyo Institute of Technology, Japan.

Masashi Sugiyama
Hidemitsu Ogawa

Supervised Learning



$$y_m = f(x_m) + \varepsilon_m$$

ε_m : mean 0
variance σ^2

From training examples $\{x_m, y_m\}_{m=1}^{M}$, obtain $\hat{f}(x)$ that minimizes generalization error J_G :

$$J_G = E\int \left[\hat{f}(u) - f(u)\right]^2 p(u)du$$

E : Expectation over noise

Future, test input points $u \sim p(u)$

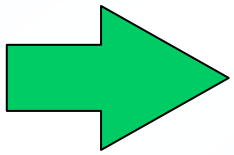
For the Time Being, We Assume…

- Target function $f(x)$ is linear combination of specified basis functions $\{\varphi_i(x)\}_{i=1}^{\mu}$:

$$f(x) = \sum_{i=1}^{\mu} \theta_i \varphi_i(x)$$

- Correlation matrix U of future input u is known.

$$U_{ij} = \int \varphi_i(u)\varphi_j(u)p(u)du$$



Later, we will discuss the case when these assumptions do not hold.

Subset Regression Models

$$\hat{f}_S(x) = \sum_{i \in S} [\hat{\theta}_S]_i \varphi_i(x)$$

S : Subset of indices $\{1,2,\ldots,\mu\}$

μ :# of basis functions

$\hat{\theta}_S$ is determined so that

training error $\sum_{m=1}^{M} \left(\hat{f}_S(x_m) - y_m \right)^2$ is minimized.

$$\hat{\theta}_S = X_S y$$

$$X_S = (A_S^T A_S)^{-1} A_S^T$$

$$[A_S]_{mi} = \begin{cases} \varphi_i(x_m) : i \in S \\ 0 \quad : i \notin S \end{cases}$$

$$y = (y_1, y_2, \ldots, y_M)^T$$

Model Selection

Select the best subset of basis functions
so that generalization error J_G is minimized:

$$J_G = E\int\left[\hat{f}(u) - f(u)\right]^2 p(u)du$$

However, J_G includes unknown target function $f(x)$.

We derive an estimate of J_G called
the subspace information criterion (SIC),
and model is determined so that SIC is minimized.

Key Idea: Unbiased Estimate

$$\hat{f}_u(x) = \sum_{i=1}^{\mu} [\hat{\theta}_u]_i \varphi_i(x)$$

Largest model

$\hat{\theta}_u$: Minimum training error estimate

$$\hat{\theta}_u = X_u y$$

$$X_u = (A^T A)^{-1} A^T$$
$$[A]_{mi} = \varphi_i(x_m)$$
$$y = (y_1, y_2, \ldots, y_M)^T$$

$\hat{\theta}_u$ is an unbiased estimate of true parameter θ :

$$E\,\hat{\theta}_u = \theta$$

E : Expectation over noise

$\hat{\theta}_u$ is used for estimating generalization error of $\hat{\theta}_S$.

Bias / Variance Decomposition

$$J_G = E\left\|\hat{\theta}_S - \theta\right\|_U^2 \qquad \left(= E\int\left[\hat{f}_S(u) - f(u)\right]^2 p(u)du\right)$$

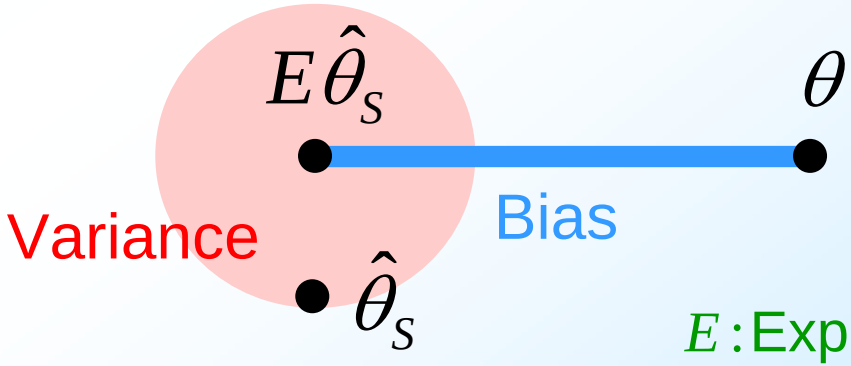
$$= \underbrace{\left\|E\hat{\theta}_S - \theta\right\|_U^2}_{\text{Bias}} + \underbrace{E\left\|\hat{\theta}_S - E\hat{\theta}_S\right\|_U^2}_{\text{Variance}}$$

$$\|\theta\|_U^2 = \theta^T U \theta$$

$$U_{ij} = \int \varphi_i(u)\varphi_j(u)p(u)du$$

Bias $J_B[S]$

Variance $J_V[S]$



E : Expectation over noise

Unbiased Estimate of Variance

$$J_V[S] = E\left\|\hat{\theta}_S - E\hat{\theta}_S\right\|_U^2$$

$$= \sigma^2 \,\mathrm{trace}\left(X_S X_S{}^T\right)$$

$$\hat{\theta}_S = X_S y$$

$$y = (y_1, y_2, \ldots, y_M)^T$$

$$\sigma^2 \to \hat{\sigma}^2 = \left\|A\hat{\theta}_u - y\right\|^2 \Big/ (M - \mu)$$

$$\hat{J}_V[S] = \hat{\sigma}^2 \,\mathrm{trace}\left(X_S X_S{}^T\right)$$

$$E\,\hat{J}_V = J_V$$

E : Expectation over noise

Unbiased Estimate of Bias

$$J_B[S] = \left\| E\hat{\theta}_S - \theta \right\|^2$$

$$= \left\| \hat{\theta}_S - \hat{\theta}_u \right\|^2 - 2\left\langle X_0 z, X_0 \varepsilon \right\rangle - \left\| X_0 \varepsilon \right\|^2$$

$$z = (f(x_1), f(x_2), \ldots, f(x_M))^T$$

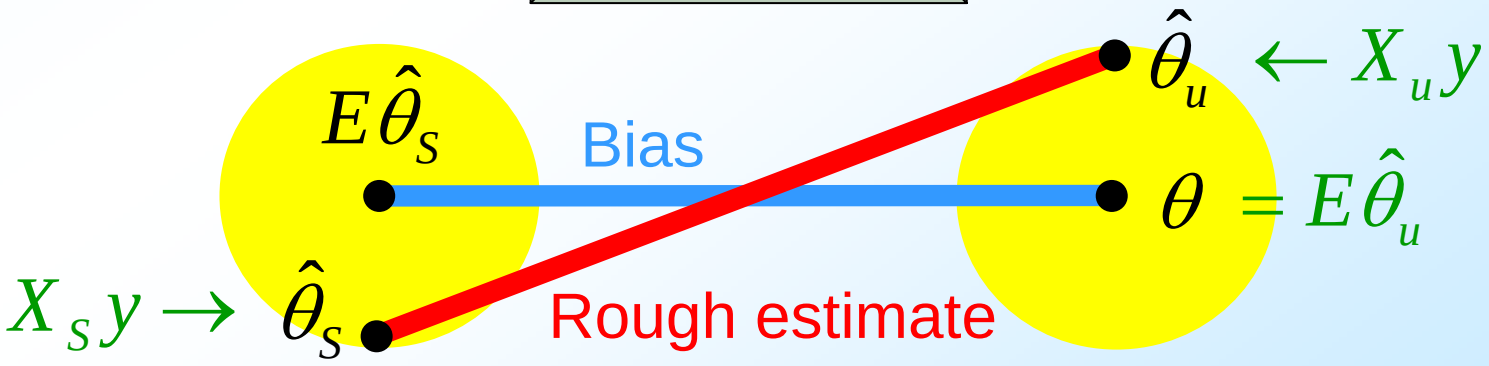
$$\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_M)^T$$

$$X_0 = X_S - X_u$$

E $\qquad$ $E, \sigma^2 \to \hat{\sigma}^2$

$$\hat{J}_B[S] = \left\| \hat{\theta}_S - \hat{\theta}_u \right\|^2 - \quad 0 \quad - \hat{\sigma}^2 \operatorname{trace}\left(X_0 X_0^T \right)$$

$$E\,\hat{J}_B = J_B$$



Subspace Information Criterion (SIC)

$$SIC[S] = \hat{J}_B[S] + \hat{J}_V[S] \qquad X_0 = X_S - X_u$$

$$= \left\| \hat{\theta}_S - \hat{\theta}_u \right\|^2 - \hat{\sigma}^2 \,\mathrm{trace}\left(X_0 X_0^T\right) + \hat{\sigma}^2 \,\mathrm{trace}\left(X_S X_S^T\right)$$

SIC is an unbiased estimate of generalization error J_G with finite samples.

$$E\ SIC[S] = J_G[S]$$

$$J_G[S] = E\int \left[\hat{f}_S(u) - f(u)\right]^2 p(u)du$$

When Assumptions Do Not Hold (1)

When target function $f(x)$ is not included in $L\{\varphi_i(x)\}_{i=1}^{\mu}$ …

$$J_G = E\int\left[\hat{f}_S(u) - f(u)\right]^2 p(u)du$$

$$= E\left\|\hat{\theta}_S - \theta^*\right\|_U^2 + \text{const.}$$

θ^* : Best estimate in $L\{\varphi_i(x)\}_{i=1}^{\mu}$

SIC is an asymptotic unbiased estimate of $E\left\|\hat{\theta}_S - \theta^*\right\|_U^2$:

$$E\ SIC \rightarrow E\left\|\hat{\theta}_S - \theta^*\right\|_U^2 \quad \text{as} \quad M \rightarrow \infty.$$

M :# of training samples

When Assumptions Do Not Hold (2)

When correlation matrix U is not available…

$$U_{ij} = \int \varphi_i(u)\varphi_j(u)p(u)du$$

- If unlabeled samples $\{u_m\}_{m=1}^{M'}$ are available, (samples without output values)

$$\hat{U}_{ij} = \frac{1}{M'}\sum_{m=1}^{M'} \varphi_i(u_m)\varphi_j(u_m)$$

- Training samples $\{x_m\}_{m=1}^{M}$ are used instead, SIC essentially agrees with Mallows's C_P.

- Just $\hat{U} = I$

I : Identity matrix

Computer Simulation

- Target function : $f(x) = \frac{1}{10}\sum_{p=1}^{50}(\sin px + \cos px)$
- x_m : Randomly created in $[-\pi, \pi]$
- $y_m = f(x_m) + \varepsilon_m$: ε_m is subject to $N(0, \sigma^2)$
- Basis functions : $\{1, \sin px, \cos px\}_{p=1}^{100}$ $\mu = 201$
- Compared models : $\{ S_0, S_{10}, \ldots, S_{100} \}$

  S_n : $\{1, \sin px, \cos px\}_{p=1}^{n}$
- Error $= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left| \hat{f}(u) - f(u) \right|^2 du$

Compared Methods

- SIC
- Mallows' C_P
- Leave-one-out cross-validation (CV)
- Akaike's information criterion (AIC)
- Sugiura's corrected AIC (cAIC)
- Schwarz's Bayesian information criterion (BIC)
- Vapnik's measure (VM)

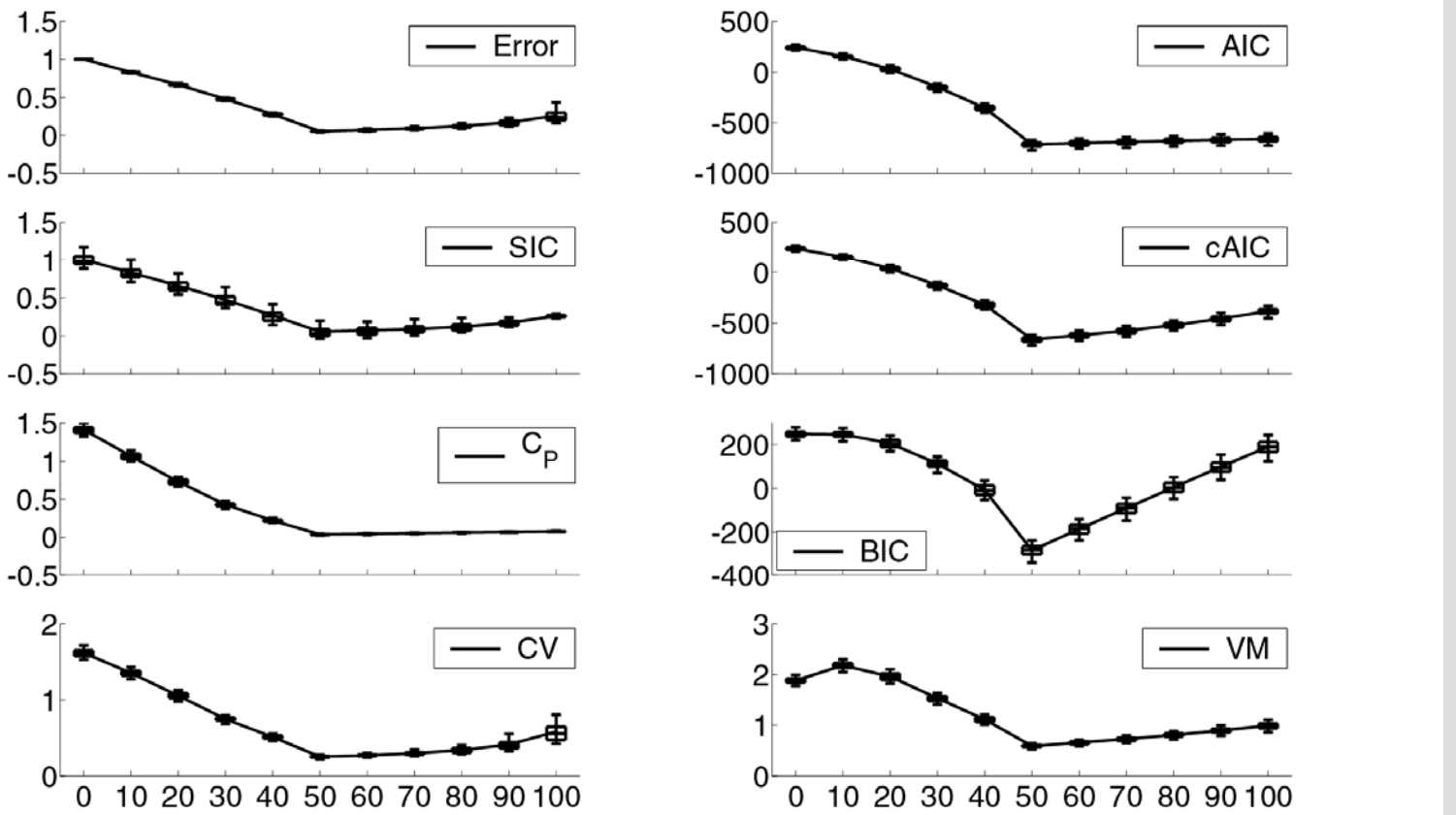
Simulations are performed 100 times with

$$(M, \sigma^2) = (250,\ 0.6),\quad (500,\ 0.6),$$
$$(250,\ 0.2),\quad (500,\ 0.2)$$

M :# of training samples σ^2 : Noise variance

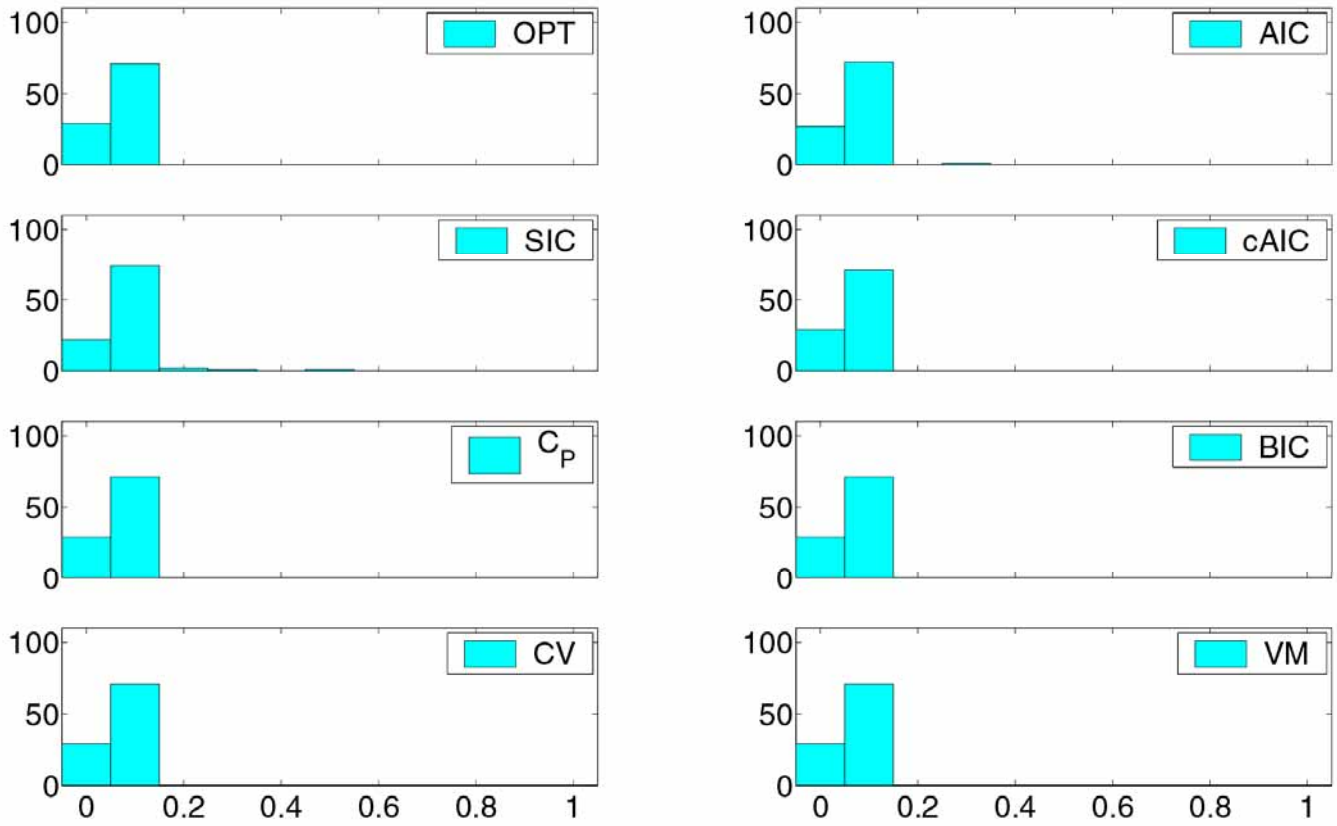
Easiest Case (Curve)

$(M, \sigma^2) =$ (250, 0.6) (500, 0.6)
(250, 0.2) (500, 0.2)



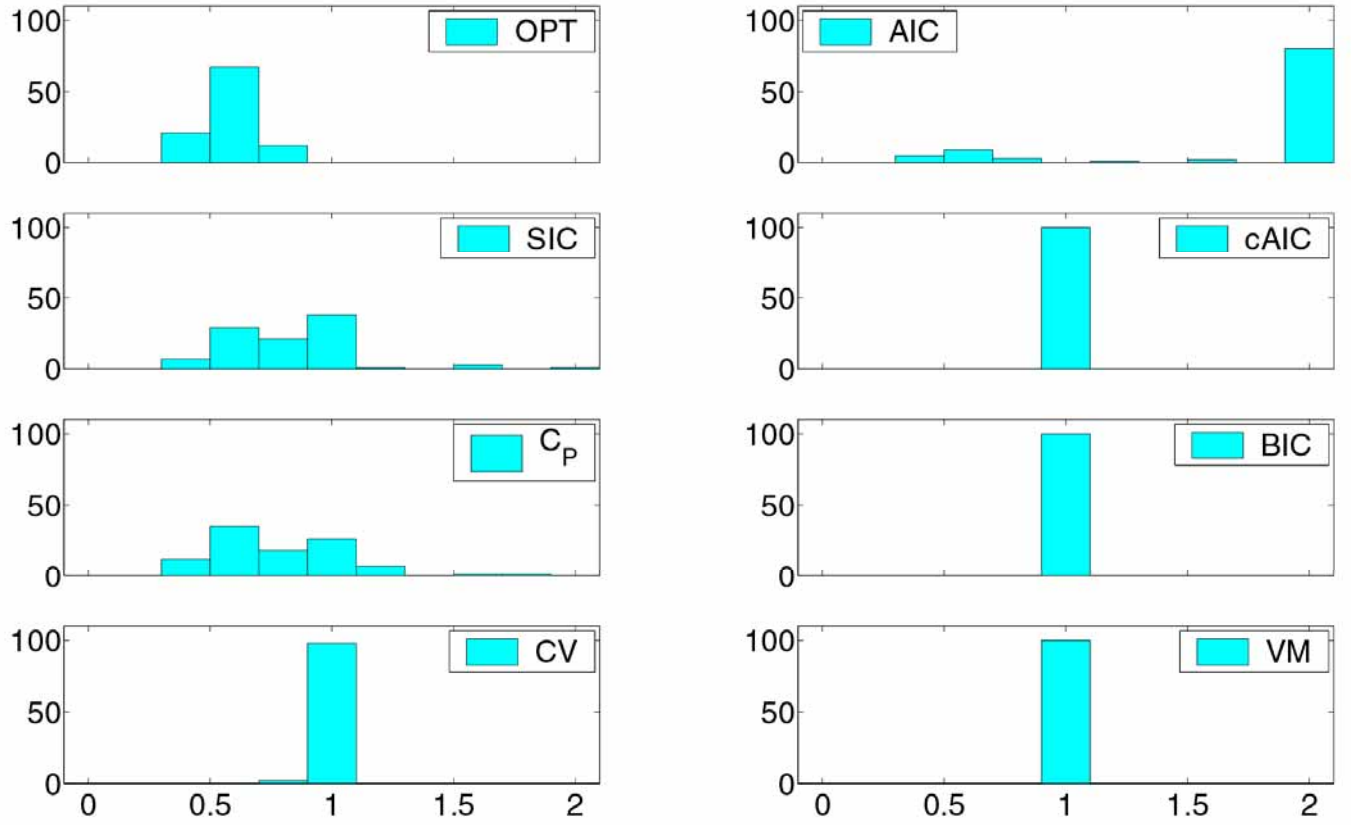
Easiest Case (Error)

$(M, \sigma^2) =$ (250, 0.6) (500, 0.6)
(250, 0.2) (500, 0.2)



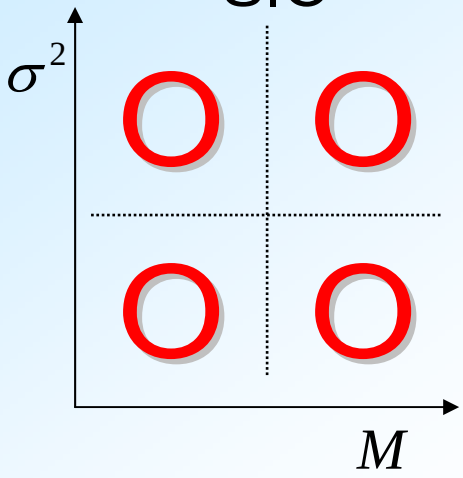
Hardest Case

$(M, \sigma^2) = (250,\ 0.6)$ (500, 0.6)
(250, 0.2) (500, 0.2)

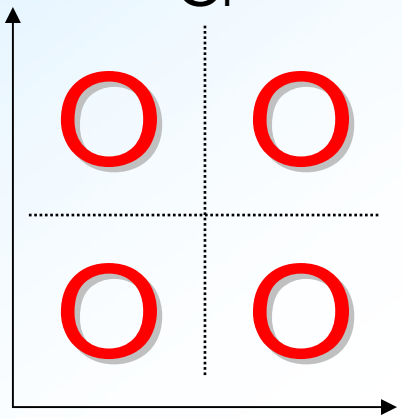


Summary of Simulations

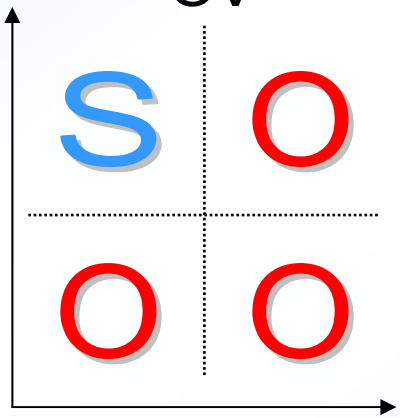
SIC



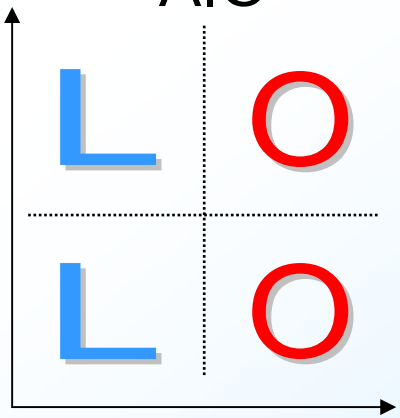
CP



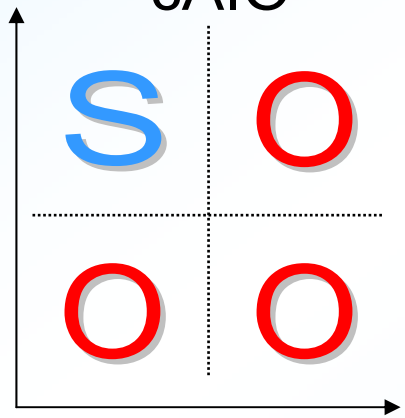
CV



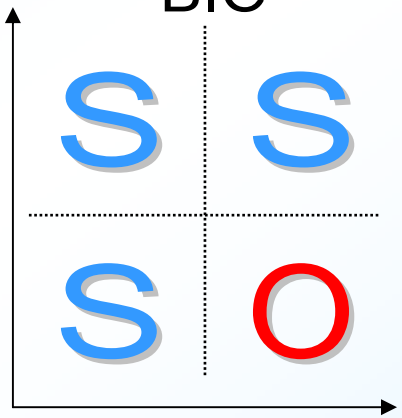
AIC



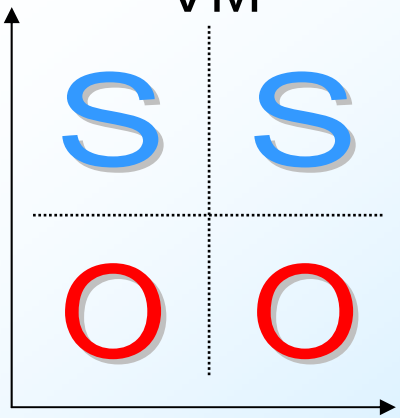
cAIC



BIC



VM



O Works well

S Selects smaller

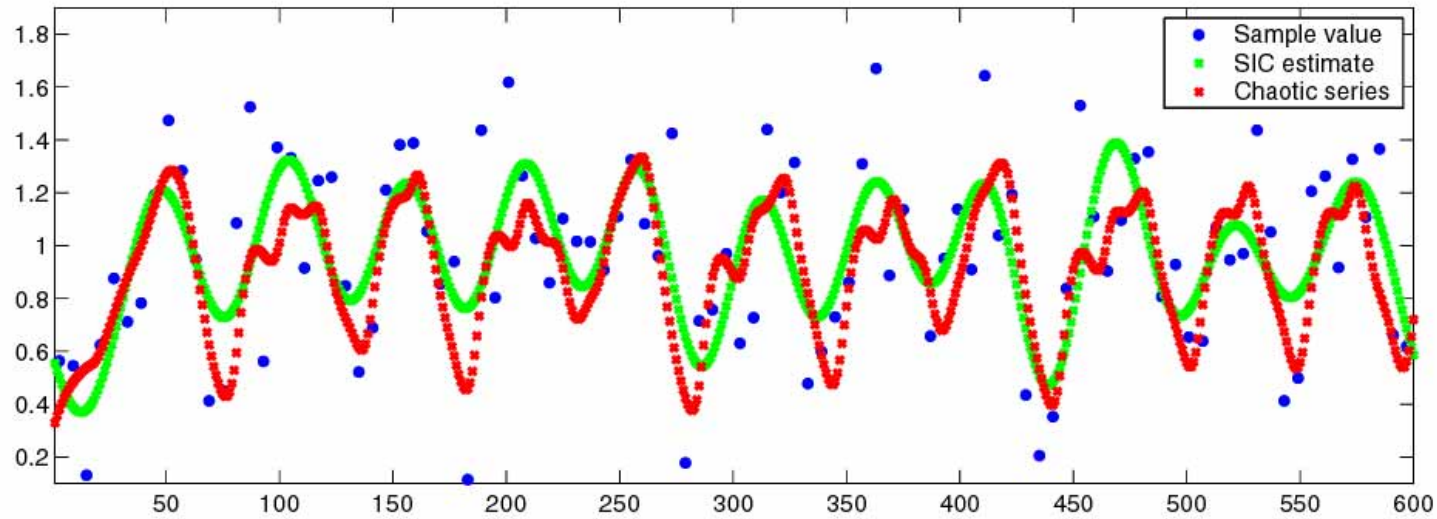
L Selects larger

M :# of training samples

σ^2 : Noise variance

When $f(x)$ is not in $L\{\varphi_i(x)\}_{i=1}^{\mu}$.

Interpolation of chaotic series



Similar results were obtained !!

Conclusions

- We proposed a new model selection criterion called subspace information criterion (SIC).
- SIC gives an unbiased estimate of generalization error.
- Computer simulations showed that SIC works well with small samples and large noise.