Selecting Ridge Parameters in Infinite Dimensional Hypothesis Spaces

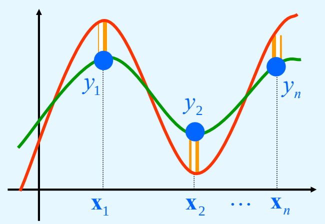
Masashi Sugiyama
Tokyo Institute of Technology, Tokyo, Japan

Klaus-Robert Müller
Fraunhofer FIRST and University of Potsdam, Berlin, Germany

Introduction

- Properly tuning ridge parameter is crucial for better generalization
- Usually generalization error estimator is first derived, and ridge parameter is tuned so that estimated generalization error is minimized
- Most of generalization error estimators proposed so far are derived within asymptotic setting (i.e., large sample assumption)
- However, small sample case is of high practical importance
- We derive an exact unbiased estimator of generalization error: SIC
- Unbiasedness of SIC is guaranteed even with small sample cases

Function Approximation Problem



$f(\mathbf{x})$: Learning target

$\hat{f}(\mathbf{x})$: Learning result

$\{(\mathbf{x}_i, y_i) \mid y_i = f(\mathbf{x}_i) + \varepsilon_i\}_{i=1}^n$: Training examples

$\{\mathbf{x}_i\}_{i=1}^n$: fixed

$\varepsilon_i \overset{i.i.d.}{\sim}$ mean 0, variance σ^2

Assumption

$f \in$ Specified reproducing kernel Hilbert space H

Previous work [1]: $\dim H \le n$ → Current work: No restriction on $\dim H$

[1] Sugiyama, M. & Ogawa, H. : Subspace information criterion for model selection. *Neural Computation*, vol.13, no.8, pp.1863-1889, 2001.

Generalization Measure

$$J_G = E\|\hat{f} - f\|^2$$

E: Expectation over noise

$\|\cdot\|$: Norm in RKHS H

Kernel Ridge Regression

Regression model : $\hat{f}(\mathbf{x}) = \sum_{i=1}^n \alpha_i K(\mathbf{x}, \mathbf{x}_i)$

$$\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^n \left(\sum_{j=1}^n \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) - y_i \right)^2 + \lambda \sum_{i=1}^n \alpha_i^2 \right]$$

λ: Ridge parameter

$$\boldsymbol{\alpha} = \mathbf{X}\mathbf{y}$$
$$\mathbf{X} = (\mathbf{K}^2 + \lambda \mathbf{I})^{-1}\mathbf{K}$$

$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)^T$ $\quad \mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$

$\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ $\quad \mathbf{I}$: Identity matrix

Extracting Essential Part of Generalization Error

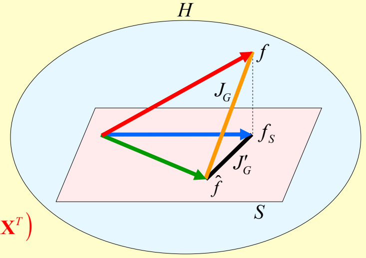
$$J_G = E\|\hat{f} - f\|^2 = \underbrace{E\|\hat{f} - f_S\|^2}_{\text{Essential } = J'_G} + \underbrace{\|f_S - f\|^2}_{\text{Constant}}$$

$$J'_G = E\|\hat{f} - f_S\|^2 = \underbrace{\|E\hat{f} - f_S\|^2}_{\text{Bias}} + \underbrace{E\|\hat{f} - E\hat{f}\|^2}_{\text{Variance}}$$

Variance $= \sigma^2 \,\mathrm{trace}(\mathbf{K}\mathbf{X}\mathbf{X}^T)$

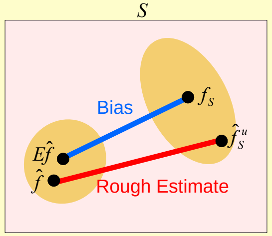
S: Subspace spanned by $\{K(\mathbf{x}, \mathbf{x}_i)\}_{i=1}^n$

f_S: Orthogonal projection of f onto S



Estimation of Bias

If unbiased estimate $\hat{f}_S^u$ of f_S is available, (i.e., $E\hat{f}_S^u = f_S$)



$g = \hat{f} - \hat{f}_S^u$

$$\text{Bias} = \|E\hat{f} - f_S\|^2 = \|g\|^2 + 2\langle Eg, Eg - g\rangle - \|Eg - g\|^2$$

↓ E $\quad$ ↓ E

$$\widehat{\text{Bias}} = \|g\|^2 - 0 - E\|Eg - g\|^2$$

Theorem

Unbiased estimate $\hat{f}_S^u$ of f_S is given by

$$\hat{f}_S^u(\mathbf{x}) = \sum_{i=1}^n \alpha_i^u K(\mathbf{x}, \mathbf{x}_i) \qquad \boldsymbol{\alpha}^u = (\alpha_1^u, \alpha_2^u, \ldots, \alpha_n^u)^T = \mathbf{K}^+\mathbf{y}$$

$^+$: Generalized Inverse

$$\widehat{\text{Bias}} = \mathbf{y}^T(\mathbf{X} - \mathbf{K}^+)\mathbf{K}(\mathbf{X} - \mathbf{K}^+)\mathbf{y} + \sigma^2 \,\mathrm{trace}(\mathbf{K}(\mathbf{X} - \mathbf{K}^+)^2)$$

Subspace Information Criterion (SIC)

$$\text{SIC} = \widehat{\text{Bias}} + \text{Variance} = \mathbf{y}^T(\mathbf{X} - \mathbf{K}^+)\mathbf{K}(\mathbf{X} - \mathbf{K}^+)\mathbf{y} + \sigma^2 \,\mathrm{trace}(2\mathbf{K}^+\mathbf{K}\mathbf{X} - \mathbf{K}^+)$$

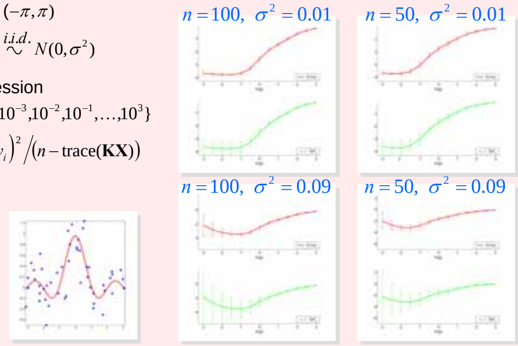
→ SIC is an unbiased estimator of J'_G with finite samples

$$E\,\text{SIC} = J'_G$$

NOTE: Unbiasedness holds even without taking expectation over training input points $\{\mathbf{x}_i\}_{i=1}^n$

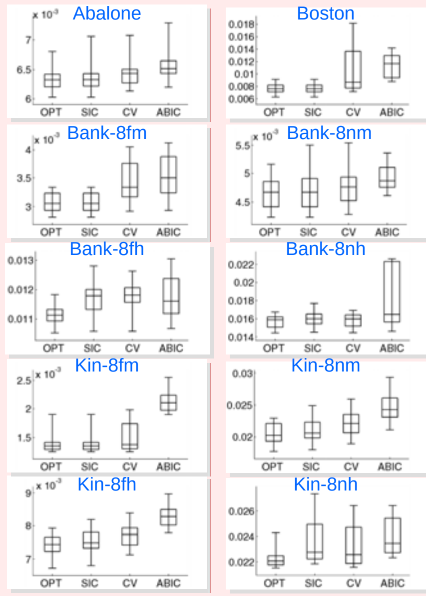
Computer Simulations (Toy datasets)

- Gaussian RKHS: $K(x, x') = \exp(-(x - x')^2/2)$
- $\{x_i\}_{i=1}^n$: Uniform on $(-\pi, \pi)$
- $y_i = f(x_i) + \varepsilon_i$: $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$
- Kernel ridge regression
  $\lambda = \{10^{-3}, 10^{-2}, 10^{-1}, \ldots, 10^3\}$
- $\hat{\sigma}^2 = \sum_{i=1}^n (\hat{f}(x_i) - y_i)^2 / (n - \mathrm{trace}(\mathbf{K}\mathbf{X}))$
- Error $= \|\hat{f} - f_S\|^2$

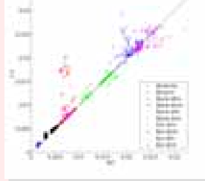


Computer Simulations (DELVE datasets)

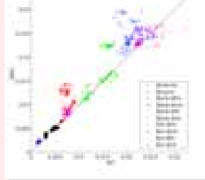
- Training with 100 samples, test with rest (hundreds to thousands)
- Model selection strategies
  - ◆OPT: Minimum test error
  - ◆SIC
  - ◆CV: Leave-one-out cross-validation
  - ◆ABIC: A Bayesian information criterion (an empirical Bayesian method)



SIC vs. CV



SIC vs. ABIC



- SIC works very well for most datasets
- However its performance can be degraded with very large noise (datasets specified by “fh” or “nh”)