ESANN2004 Apr. 28-30, 2004

Regularizing Generalization Error Estimators: A Novel Approach to Robust Model Selection

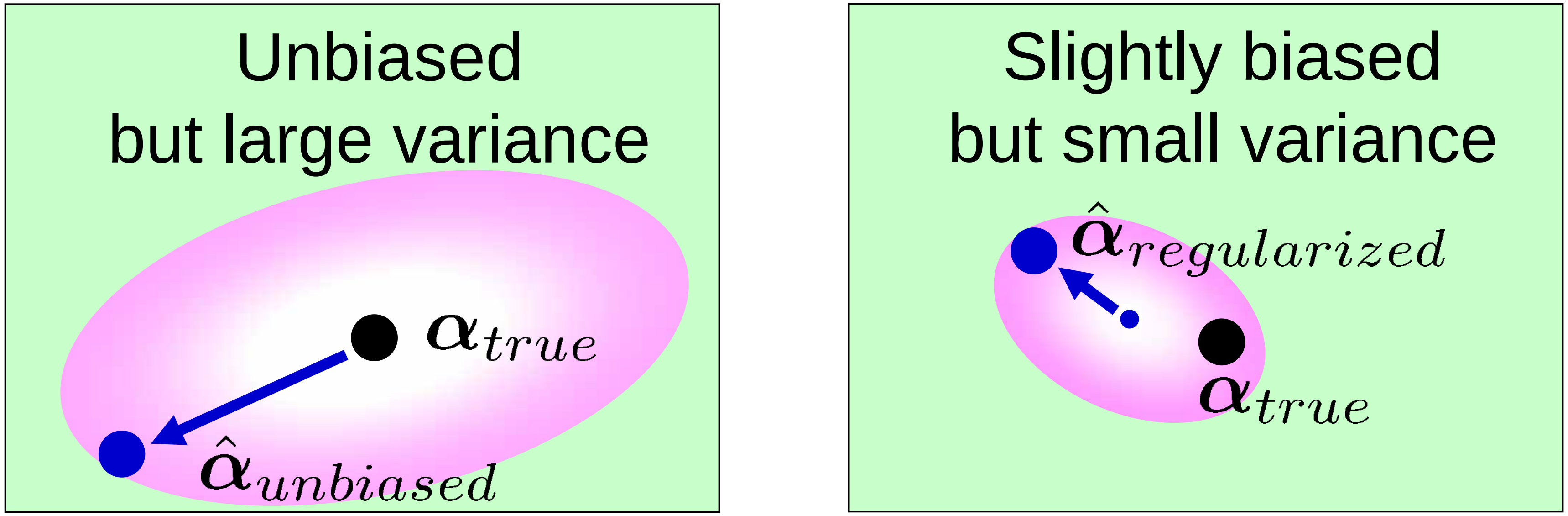
Masashi Sugiyama [1,2]
Motoaki Kawanabe [1]
Klaus-Robert Müller [1,3]

(1) Fraunhofer FIRST, Germany
(2) Tokyo Institute of Technology, Japan
(3) University of Potsdam, Germany

Institut Rechnerarchitektur und Softwaretechnik

Abstract

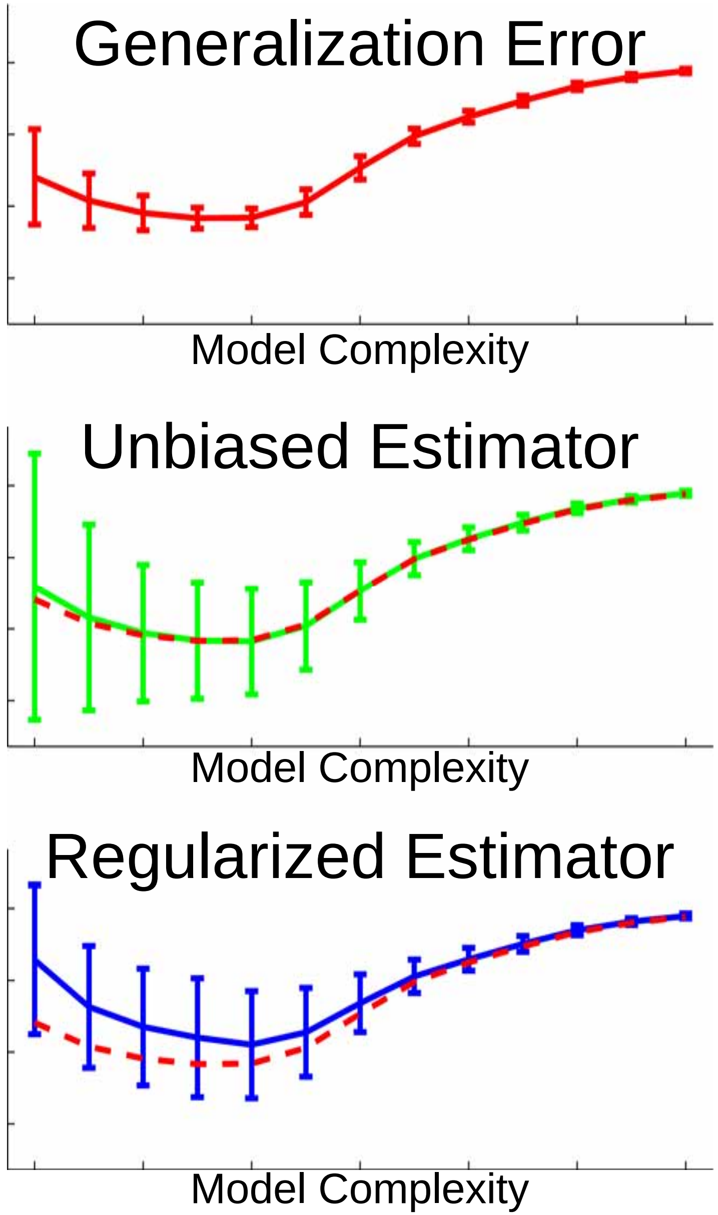
- Model selection is the key to successful learning.
- Which is better?



- We use the spirit of Stein's idea
  for constructing better model criterion:
  Regularized subspace information criterion

Abstract (cont.)

- Unbiased generalization error estimators are often used for model selection, e.g., AIC, CV, SIC...
- However, unbiased estimators can have large variance, which causes unstable model selection.
- We propose regularizing unbiased generalization error estimators.



Kernel Ridge Regression

- Learn $f(\boldsymbol{x})$ from $\{(\boldsymbol{x}_i, y_i) \mid y_i = f(\boldsymbol{x}_i) + \epsilon_i\}_{i=1}^n$
- Kernel regression:

$\epsilon_i \overset{i.i.d.}{\sim} \text{mean } 0, \text{ variance } \sigma^2$

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \hat{\alpha}_i K(\boldsymbol{x}, \boldsymbol{x}_i)$$

$\hat{\alpha}_i$:Parameters to be learned

$K(\boldsymbol{x}, \boldsymbol{x}')$:Kernel function (e.g., Gaussian)

- Ridge estimation:

$$\hat{\boldsymbol{\alpha}}_\lambda = \underset{\hat{\boldsymbol{\alpha}}}{\text{argmin}} \left[\sum_{i=1}^{n} \left(\hat{f}(\boldsymbol{x}_i) - y_i \right)^2 + \lambda \sum_{i=1}^{n} \hat{\alpha}_i^2 \right]$$

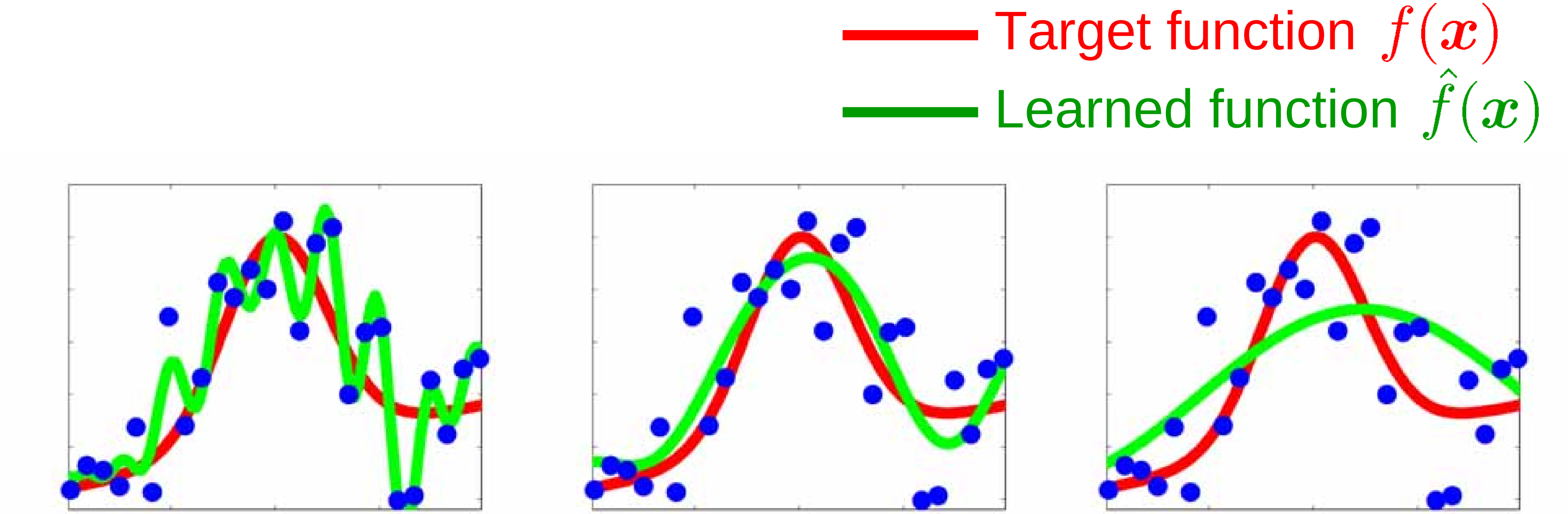
λ :Ridge parameter (model parameter)

$$\hat{\boldsymbol{\alpha}}_\lambda = \boldsymbol{X}_\lambda \boldsymbol{y}$$

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

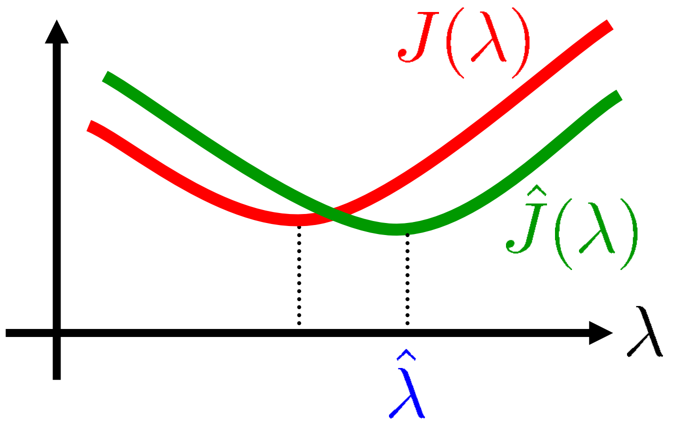
$$\boldsymbol{K}_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

Model Selection



λ is too small　　λ is appropriate　　λ is too large

λ is chosen so that an estimator $\hat{J}$ of generalization error J is minimized.



RKHS-Based Generalization Error

- Assume $f(\boldsymbol{x})$ lies in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with reproducing kernel $K(\boldsymbol{x}, \boldsymbol{x}')$.
- We shall measure the generalization error by the expected RKHS norm:

$$J = \mathbb{E}\|\hat{f}_\lambda - f\|^2 - \underbrace{\|f\|^2}_{\text{Constant}}$$

$$= \mathbb{E}\left[\|\hat{f}_\lambda\|^2 - 2\langle \hat{f}_\lambda, f \rangle\right]$$

$\|\cdot\|$:Norm in RKHS $\mathcal{H}$

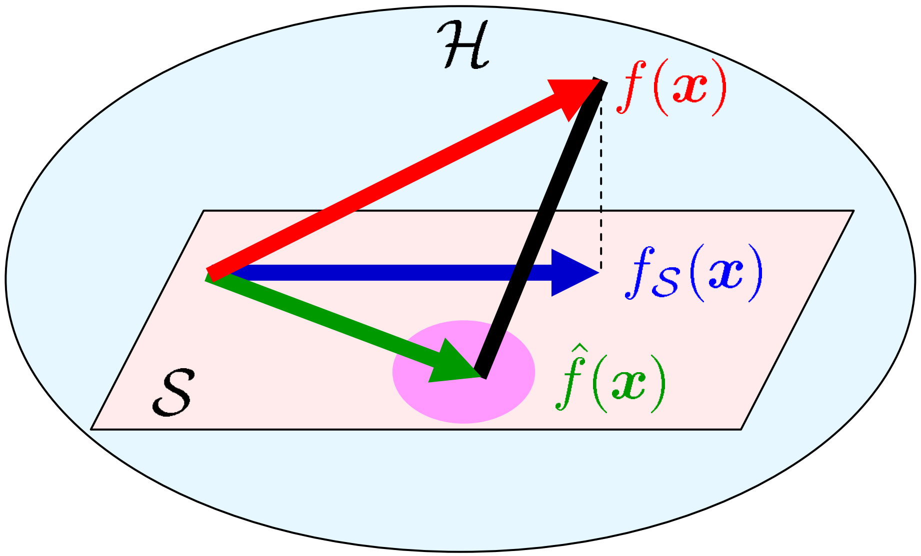
$\mathbb{E}$:Expectation over the noise

Projection of Learning Target

- $f_{\mathcal{S}}$: Projection of f onto $\mathcal{S} = \mathcal{L}(\{K(\boldsymbol{x}, \boldsymbol{x}_i)\}_{i=1}^{n})$

$$f_{\mathcal{S}}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i^* K(\boldsymbol{x}, \boldsymbol{x}_i)$$

α_i^* :Unknown coefficients



- Generalization error J is expressed as

$$J = \mathbb{E}\left[\|\hat{f}_\lambda\|^2 - 2\langle \hat{f}_\lambda, f_{\mathcal{S}}\rangle\right]$$

$$= \mathbb{E}\left[\langle \boldsymbol{K}\hat{\boldsymbol{\alpha}}_\lambda, \hat{\boldsymbol{\alpha}}_\lambda\rangle - 2\langle \boldsymbol{K}\hat{\boldsymbol{\alpha}}_\lambda, \boldsymbol{\alpha}^*\rangle\right]$$

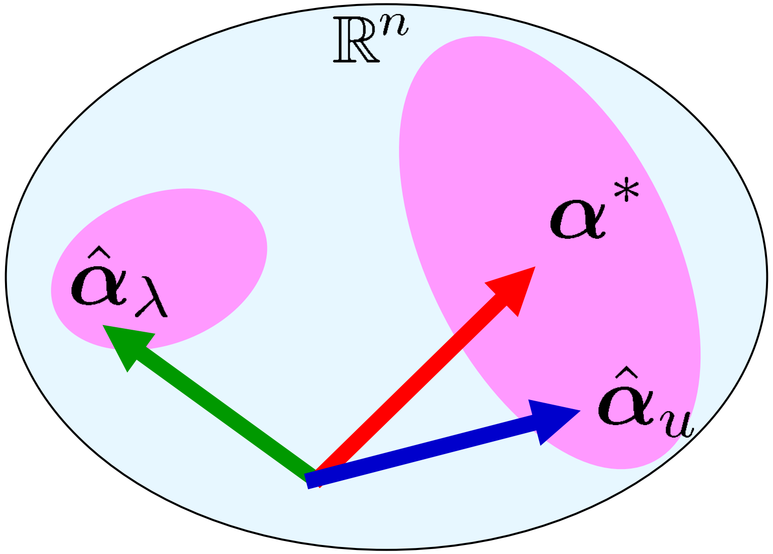
Subspace Information Criterion

Sugiyama & Ogawa (Neural Comp., 2001)
Sugiyama & Müller (JMLR, 2002)

- Replace unknown $\boldsymbol{\alpha}^*$ by its unbiased estimator $\hat{\boldsymbol{\alpha}}_u = \boldsymbol{X}_u \boldsymbol{y}$.

  $\boldsymbol{X}_u = \boldsymbol{K}^\dagger$

  $\boldsymbol{K}^\dagger$: Generalized inverse



- Adding a modification term, we have an unbiased estimator of generalization error.

$$\mathrm{SIC}(\lambda) = \langle \boldsymbol{K}\hat{\boldsymbol{\alpha}}_\lambda, \hat{\boldsymbol{\alpha}}_\lambda \rangle - 2\langle \boldsymbol{K}\hat{\boldsymbol{\alpha}}_\lambda, \hat{\boldsymbol{\alpha}}_u \rangle + 2\sigma^2 \mathrm{tr}(\boldsymbol{K}\boldsymbol{X}_\lambda \boldsymbol{X}_u^\top)$$

$$\mathbb{E}\,\mathrm{SIC} = J$$

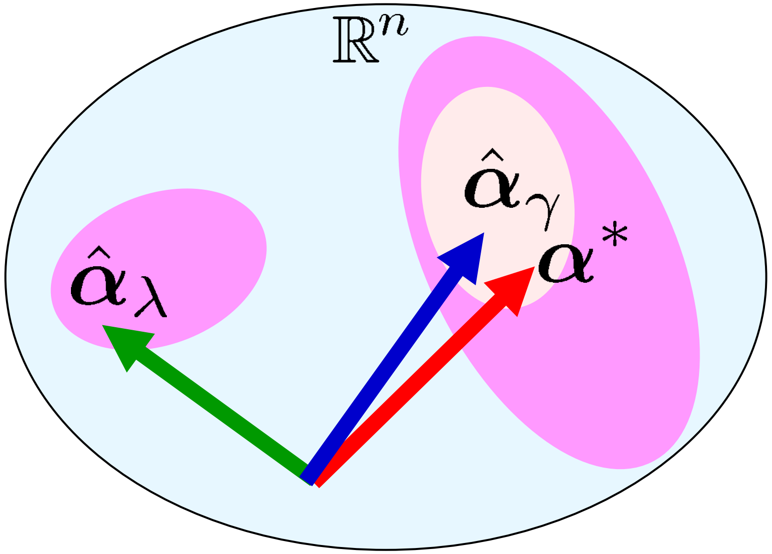
Variance of Unbiased Generalization Error Estimators

- Unbiased generalization error estimators can have large variance, which causes unstable model choice.
- A natural way to circumvent this problem is to allow small bias in order to reduce variance.

We regularize SIC for robust model selection

Regularized SIC

- Unbiased estimator $\hat{\boldsymbol{\alpha}}_u$ can cause large variance of SIC.
- Replace $\hat{\boldsymbol{\alpha}}_u$ by a regularized estimator $\hat{\boldsymbol{\alpha}}_\gamma = \boldsymbol{X}_\gamma \boldsymbol{y}$.



$$\boldsymbol{X}_\gamma = (\boldsymbol{K}^2 + \gamma \boldsymbol{I})^{-1} \boldsymbol{K}$$

γ :Regularization parameter

$$\mathrm{RSIC}(\lambda;\gamma) = \langle \boldsymbol{K}\hat{\boldsymbol{\alpha}}_\lambda, \hat{\boldsymbol{\alpha}}_\lambda \rangle - 2\langle \boldsymbol{K}\hat{\boldsymbol{\alpha}}_\lambda, \hat{\boldsymbol{\alpha}}_\gamma \rangle + 2\sigma^2 \mathrm{tr}(\boldsymbol{K}\boldsymbol{X}_\lambda \boldsymbol{X}_\gamma^\top)$$

- How to choose γ ?

Determining Degree of Regularization in RSIC

- Expected squared error of RSIC.

$$\mathrm{ESE}(\gamma;\lambda) = \mathbb{E}\left[\mathrm{RSIC}(\gamma;\lambda) - J(\lambda)\right]^2$$

- We can obtain an unbiased estimator of ESE.

$$\mathbb{E}\,\widehat{\mathrm{ESE}}(\gamma;\lambda) = \mathrm{ESE}(\gamma;\lambda)$$

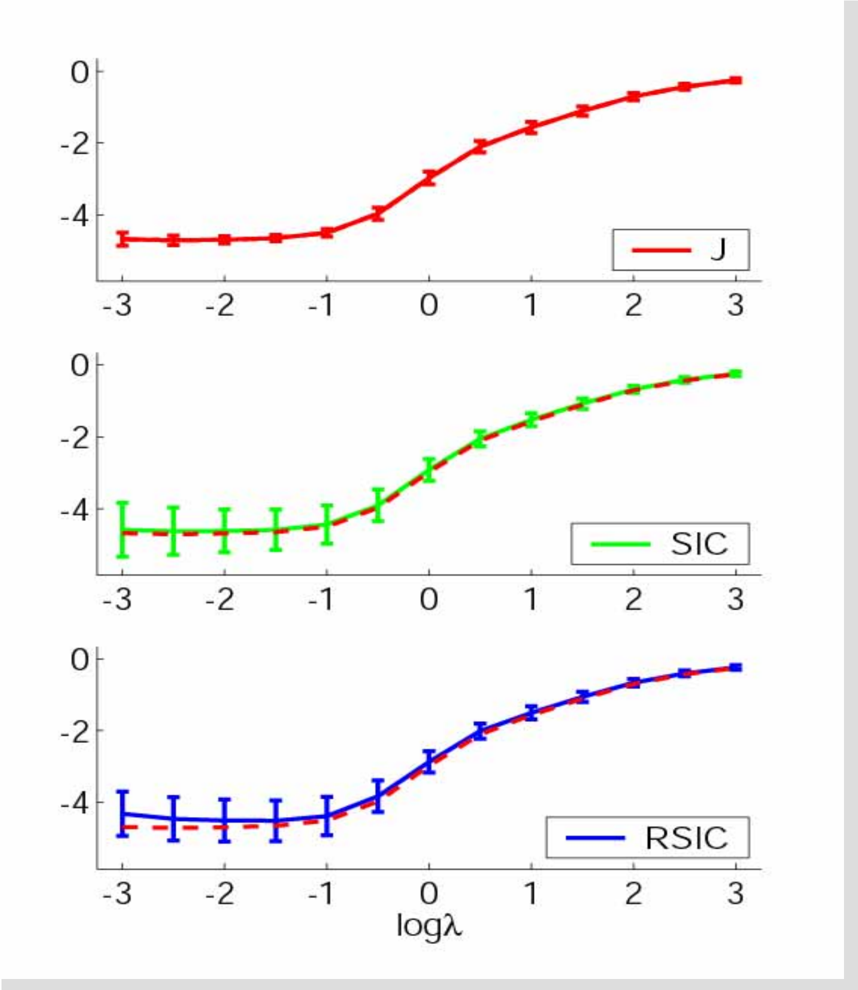
$$\begin{aligned}\widehat{\mathrm{ESE}}(\gamma;\lambda) = &\langle \boldsymbol{B}\boldsymbol{y},\boldsymbol{y}\rangle^2 - \sigma^2\|(\boldsymbol{B}+\boldsymbol{B}^\top)\boldsymbol{y}\|^2 - 2\sigma^2\mathrm{tr}\left(\boldsymbol{B}\right)\langle \boldsymbol{B}\boldsymbol{y},\boldsymbol{y}\rangle\\ &+\sigma^4\mathrm{tr}(\boldsymbol{B}^2+\boldsymbol{B}^\top\boldsymbol{B}) + \sigma^4\mathrm{tr}(\boldsymbol{B})^2\\ &+\sigma^2\|(\boldsymbol{C}+\boldsymbol{C}^\top)\boldsymbol{y}\|^2 - \sigma^4\mathrm{tr}(\boldsymbol{C}^2+\boldsymbol{C}^\top\boldsymbol{C})\end{aligned}$$

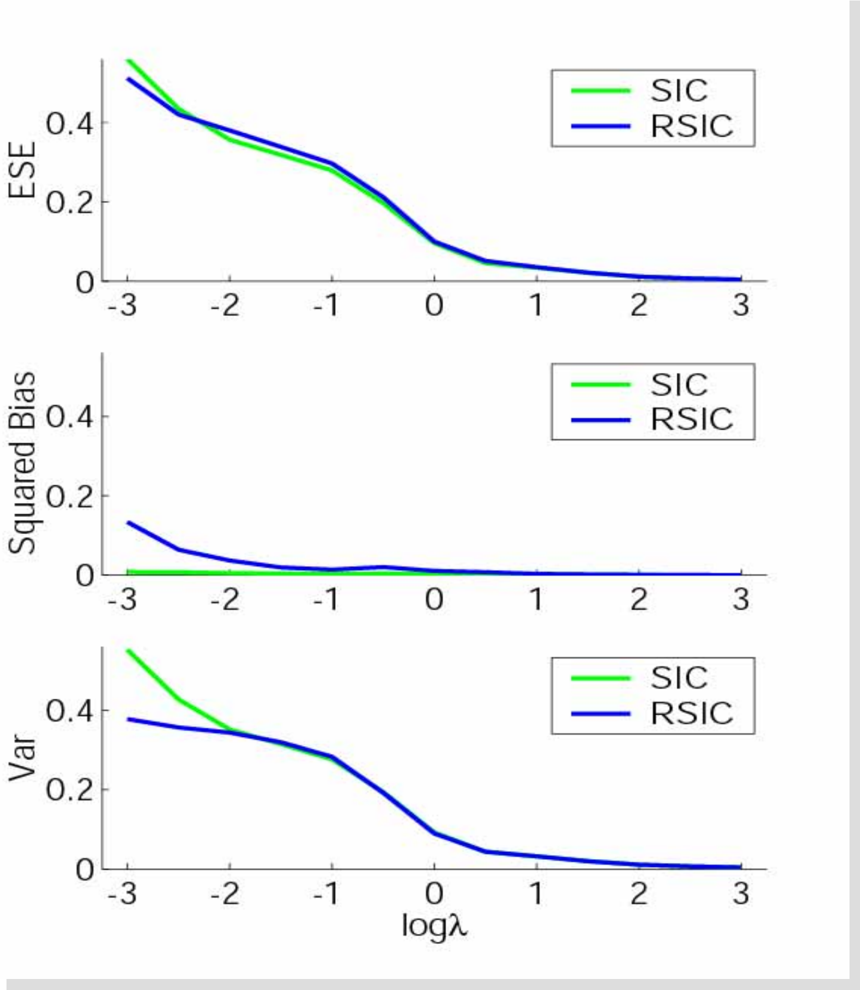
$$\boldsymbol{B} = 2\boldsymbol{X}_u^\top\boldsymbol{K}\boldsymbol{X}_\lambda - 2\boldsymbol{X}_\gamma^\top\boldsymbol{K}\boldsymbol{X}_\lambda$$
$$\boldsymbol{C} = \boldsymbol{X}_\lambda^\top\boldsymbol{K}\boldsymbol{X}_\lambda - 2\boldsymbol{X}_\gamma^\top\boldsymbol{K}\boldsymbol{X}_\lambda$$

- We determine γ so that $\widehat{\mathrm{ESE}}$ is minimized.

Learning Sinc Function

Noise level: Small, Kernel: Gaussian

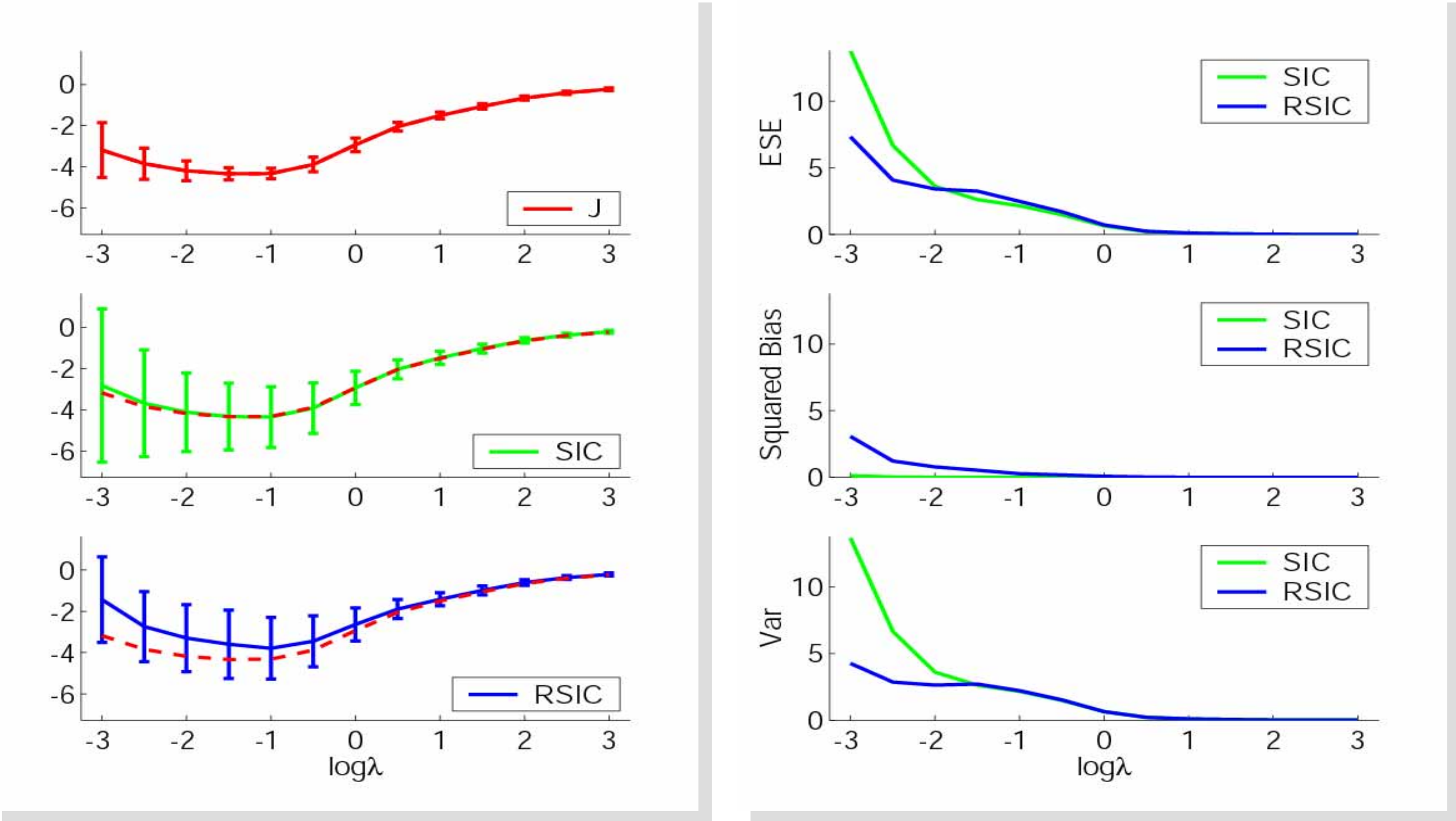




RSIC maintains good performance of SIC!

Learning Sinc Function

Noise level: Large,   Kernel: Gaussian



RSIC improves over unbiased SIC!

Test Errors for DELVE Data Sets

Normalized test error

(Test error obtained with best ridge parameter is normalized to 1)

Data	SIC	RSIC	Cross Validation	Empirical Bayes
Abalone	1.0131 ± 0.0002	1.0144 ± 0.0002	1.0146 ± 0.0002	1.0204 ± 0.0003
Boston	1.0001 ± 0.0007	1.0016 ± 0.0007	1.0071 ± 0.0007	1.1406 ± 0.0008
Bank-8fm	1.0001 ± 0.0001	1.0703 ± 0.0001	1.0708 ± 0.0001	1.0030 ± 0.0001
Bank-8nm	1.0001 ± 0.0004	1.0002 ± 0.0004	1.0461 ± 0.0005	1.0477 ± 0.0005
Bank-8fh	1.0604 ± 0.0004	1.0025 ± 0.0003	1.0026 ± 0.0003	1.0003 ± 0.0003
Bank-8nh	1.0987 ± 0.0004	1.0028 ± 0.0005	1.2177 ± 0.0008	1.4200 ± 0.0008
Kin-8fm	1.0000 ± 0.0001	1.0000 ± 0.0001	1.0010 ± 0.0001	1.4548 ± 0.0004
Kin-8nm	1.0104 ± 0.0011	1.0097 ± 0.0010	1.0241 ± 0.0007	1.0371 ± 0.0006
Kin-8fh	1.1103 ± 0.0002	1.0021 ± 0.0003	1.0057 ± 0.0003	1.2025 ± 0.0001
Kin-8nh	1.1015 ± 0.0008	1.0451 ± 0.0009	1.0017 ± 0.0004	1.0361 ± 0.0004

Best and comparable methods by t-test are shown by red.

Conclusions and Outlook

- We proposed regularizing model selection criteria for stabilization.
- Simulation showed that model selection performance is improved especially when noise level is large.
- Improving accuracy of $\widehat{\mathrm{ESE}}$.
- Theoretically investigate model selection performance.
- Applying the same idea to choosing the number of folds in the cross-validation score.