FUNCTIONAL ANALYTIC FRAMEWORK FOR MODEL SELECTION

Masashi Sugiyama *,1

* *Tokyo Institute of Technology, Tokyo, Japan*

Abstract: Model selection is one of the most important tasks in the identification of black-box systems. In this paper, we give a novel model selection method from the viewpoint of functional analysis. We formulate the system identification problem as a function approximation problem in a reproducing kernel Hilbert space (RKHS), where the approximation error is measured by the RKHS norm. Within this framework, we derive an estimator of the approximation error called the subspace information criterion (SIC) and show its properties.

Keywords: functional analysis, model selection, reproducing kernel Hilbert space, subspace information criterion.

1. INTRODUCTION

Model selection is one of the most important tasks in the identification of black-box systems. The goal of model selection is to determine the model such that the approximation error between an estimated system and the true system is minimized. However, the approximation error usually depends on the unknown system so it can not be directly calculated. One of the general approaches to model selection is to derive an estimator of the approximation error and then to determine the model such that the estimator is minimized. So far, a number of methods for estimating the approximation error have been proposed from various different standpoints, e.g., methods based on the asymptotic statistics (Akaike, 1974; Murata *et al.*, 1994), the Vapnik-Chervonenkis (VC) theory (Vapnik, 1995), resampling techniques (Efron, 1979; Wahba, 1990; Efron and Tibshirani, 1993), and the Bayesian statistics (Schwarz, 1978; Akaike, 1980).

In this paper, we give a novel model selection method from the viewpoint of functional analysis and show its properties.

[1] The author would like to thank Prof. Hidemitsu Ogawa for his valuable comments and discussions. This research is partially supported by MEXT, Grants-in-Aid for Scientific Research, 14380158 and 14780262.

2. PROBLEM FORMULATION

Let us regard a black-box system as a real-valued function $f(\boldsymbol{x})$ of d variables defined on a subset $\mathcal{D}$ of the d-dimensional Euclidean space $\mathbb{R}^d$. We would like to identify the function $f(\boldsymbol{x})$ from a set of n input-output samples. A sample consists of an input value $\boldsymbol{x}_i$ in $\mathcal{D}$ and a corresponding output value y_i in $\mathbb{R}$. We assume that the output value y_i is degraded by the additive noise ϵ_i with mean zero. That is, the set of samples are expressed as

$$\{(\boldsymbol{x}_i, y_i) \mid y_i = f(\boldsymbol{x}_i) + \epsilon_i\}_{i=1}^n. \tag{1}$$

We consider the case where the unknown function $f(\boldsymbol{x})$ belongs to a specified *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$. The *reproducing kernel* of a functional Hilbert space $\mathcal{H}$ is a bivariate function defined on $\mathcal{D} \times \mathcal{D}$. Let us denote the reproducing kernel of $\mathcal{H}$ by $K(\boldsymbol{x}, \boldsymbol{x}')$. The reproducing kernel of $\mathcal{H}$ satisfies the following conditions (Aronszajn, 1950):

- For any fixed $\boldsymbol{x}'$ in $\mathcal{D}$, $K(\boldsymbol{x}, \boldsymbol{x}')$ is a function of $\boldsymbol{x}$ in $\mathcal{H}$.
- For any function f in $\mathcal{H}$ and for any $\boldsymbol{x}'$ in $\mathcal{D}$, it holds that

$$\langle f(\cdot), K(\cdot, \boldsymbol{x}') \rangle_{\mathcal{H}} = f(\boldsymbol{x}'), \tag{2}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ stands for the inner product in $\mathcal{H}$.

Let $\hat{f}(\boldsymbol{x})$ be an estimate of the function $f(\boldsymbol{x})$. The goal of system identification is to find $\hat{f}(\boldsymbol{x})$ such that it is as 'close' to $f(\boldsymbol{x})$ as possible. We shall measure the closeness between $\hat{f}(\boldsymbol{x})$ and $f(\boldsymbol{x})$ by the expected squared norm in the RKHS $\mathcal{H}$:

$$\mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f\|_{\mathcal{H}}^2, \tag{3}$$

where $\mathrm{E}_{\boldsymbol{\epsilon}}$ denotes the expectation over the noise $\{\epsilon_i\}_{i=1}^n$ and $\|\cdot\|_{\mathcal{H}}$ denotes the norm in the RKHS $\mathcal{H}$. That is, the goal is to find $\hat{f}$ from $\mathcal{H}$ such that

$$\min_{\hat{f}\in\mathcal{H}} \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f\|_{\mathcal{H}}^2. \tag{4}$$

Note that we do not take the expectation over input points $\{\boldsymbol{x}_i\}_{i=1}^n$, as is done in some statistical model selection theories (Akaike, 1974; Murata *et al.*, 1994). Therefore, our approach may be more data-dependent. Since $\|f\|_{\mathcal{H}}^2$ does not depend on $\hat{f}$, we subtract it and use the following measure as the approximation error.

$$\begin{aligned} J &= \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f\|_{\mathcal{H}}^2 - \|f\|_{\mathcal{H}}^2 \\ &= \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f}\|_{\mathcal{H}}^2 - 2\mathrm{E}_{\boldsymbol{\epsilon}}\langle\hat{f}, f\rangle_{\mathcal{H}}, \end{aligned} \tag{5}$$

where $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ denotes the inner product in the RKHS $\mathcal{H}$. The approximation error J defined by Eq.(5) can not be directly calculated since it includes the unknown function $f(\boldsymbol{x})$. The aim of this paper is to derive an estimator of the approximation error J.

3. ESTIMATING APPROXIMATION ERROR J

In this section, we derive an estimator of the approximation error J called the subspace information criterion (SIC) [2].

3.1 *Preliminary*

Our key idea for estimating the approximation error J is to use a linear unbiased estimate $\hat{f}_u$ of the unknown function f, instead of f itself. Here let us assume that we have a linear operator X_u such that

$$\hat{f}_u = X_u\boldsymbol{y}, \quad \mathrm{E}_{\boldsymbol{\epsilon}}\hat{f}_u = f, \tag{6}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$. We will discuss how to obtain the linear operator X_u in the following sections.

Letting $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^\top$, we have the following lemma.

[2] The name 'subspace information criterion' came from the fact that in our early work (Sugiyama and Ogawa, 2001), the criterion was derived for the purpose of selecting subspace models in linear regression.

Lemma 1. The approximation error J is expressed as

$$J = \mathrm{E}_{\boldsymbol{\epsilon}}\left(\|\hat{f}\|_{\mathcal{H}}^2 - 2\langle\hat{f}, \hat{f}_u\rangle_{\mathcal{H}} + 2\langle\hat{f}, X_u\boldsymbol{\epsilon}\rangle_{\mathcal{H}}\right). \tag{7}$$

Based on the above lemma, let us define 'preSIC' by

$$\mathrm{preSIC} = \|\hat{f}\|_{\mathcal{H}}^2 - 2\langle\hat{f}, \hat{f}_u\rangle_{\mathcal{H}} + 2\mathrm{E}_{\boldsymbol{\epsilon}}\langle\hat{f}, X_u\boldsymbol{\epsilon}\rangle_{\mathcal{H}}. \tag{8}$$

The above quantity is named preSIC because SIC will be derived based on this quantity. It is clear from Lemma 1 that preSIC satisfies

$$\mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{preSIC} = J. \tag{9}$$

The third term in preSIC is expected over the noise, and it can not be directly calculated. In the following, we shall give methods of calculating or approximating the third term in preSIC under some conditions.

3.2 *SIC for Linear Estimates*

Let us consider the case where $\hat{f}$ is a linear estimate, i.e., with a linear operator X, $\hat{f}$ is given by

$$\hat{f} = X\boldsymbol{y}. \tag{10}$$

Eq.(10) includes, for example, least-squares or ridge estimation (Hoerl and Kennard, 1970) for linear or kernel regression models. A particular form of the Gaussian process regression (Williams and Rasmussen, 1996) and the least-squares support vector machines (Suykens *et al.*, 2002) are also included.

Let $\boldsymbol{Q}$ be the noise covariance matrix, $\mathrm{tr}\,(\cdot)$ be the trace of an operator, and X_u^* be the adjoint of X_u. Then we have the following lemma.

Lemma 2. When $\hat{f}$ is a linear estimate, it holds that

$$\mathrm{E}_{\boldsymbol{\epsilon}}\langle\hat{f}, X_u\boldsymbol{\epsilon}\rangle_{\mathcal{H}} = \mathrm{tr}\,(X\boldsymbol{Q}X_u^*). \tag{11}$$

Based on the above lemma, we define SIC for linear estimates as follows [3].

$$\mathrm{SIC} = \|\hat{f}\|_{\mathcal{H}}^2 - 2\langle\hat{f}, \hat{f}_u\rangle_{\mathcal{H}} + 2\mathrm{tr}\,(X\boldsymbol{Q}X_u^*). \tag{12}$$

It is clear that the above SIC is an unbiased estimator of the approximation error J.

$$\mathrm{E}_{\boldsymbol{\epsilon}}\mathrm{SIC} = J. \tag{13}$$

[3] In our early work (Sugiyama and Ogawa, 2001), SIC is defined as

$$\|\hat{f} - \hat{f}_u\|_{\mathcal{H}}^2 - \mathrm{tr}\,((X - X_u)\boldsymbol{Q}(X - X_u)^*) + \mathrm{tr}\,(X\boldsymbol{Q}X^*),$$

which is an unbiased estimator of Eq.(3). In the current paper, we ignored some constant terms that correspond to an estimate of $\|f\|_{\mathcal{H}}^2$ thus do not depend on X.

3.3 *SIC for Non-Linear Differentiable Estimates*

Here let us consider the case where $\hat{f}$ is a smooth non-linear estimate, i.e., with a twice almost differentiable (Stein, 1981) non-linear operator X, $\hat{f}$ is given by

$$\hat{f} = X(\boldsymbol{y}). \tag{14}$$

For example, some of the M-estimators such as Huber's robust estimation (Huber, 1981) for linear or kernel regression models are expressed by Eq.(14).

When $X(\boldsymbol{y})$ is almost differentiable, $[X_u^* X](\boldsymbol{y})$ is also almost differentiable since X_u^* is linear. Note that $[X_u^* X](\boldsymbol{y})$ is a vector-valued function from $\mathbb{R}^n$ to $\mathbb{R}^n$. Then we have the following lemma.

Lemma 3. When $\hat{f}$ is a smooth non-linear estimate and the noise $\{\epsilon_i\}_{i=1}^n$ is independently and identically drawn from the normal distribution with mean 0 and variance σ^2, it holds that

$$\mathrm{E}_{\boldsymbol{\epsilon}} \langle \hat{f}, X_u \boldsymbol{\epsilon} \rangle_{\mathcal{H}} = \sigma^2 \mathrm{E}_{\boldsymbol{\epsilon}} \sum_{i=1}^{n} \frac{\partial [X_u^* X]_i(\boldsymbol{y})}{\partial y_i}, \tag{15}$$

where $[X_u^* X]_i(\boldsymbol{y})$ is the i-th output of the vector-valued function $[X_u^* X](\boldsymbol{y})$.

Based on the above lemma, we define SIC for smooth non-linear estimates as follows.

$$\begin{aligned} \mathrm{SIC} = \|\hat{f}\|_{\mathcal{H}}^2 - 2\langle \hat{f}, \hat{f}_u \rangle_{\mathcal{H}} \\ + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial [X_u^* X]_i(\boldsymbol{y})}{\partial y_i}. \end{aligned} \tag{16}$$

Note that even for smooth non-linear estimates, SIC is an unbiased estimator of J, i.e., Eq.(13) holds. It is easy to confirm that Eq.(16) agrees with Eq.(12) when $\hat{f}$ is a linear estimate. Therefore, Eq.(16) can be regarded as a natural extension of Eq.(12).

3.4 *Bootstrap Approximation of SIC for Non-Linear Estimates*

Finally, let us consider the case where $\hat{f}$ is a general non-linear estimate, i.e., with a general non-linear operator X, $\hat{f}$ is given by

$$\hat{f} = X(\boldsymbol{y}). \tag{17}$$

This includes, for example, ℓ_1-norm regularized estimation for linear or kernel regression models (Williams, 1995; Tibshirani, 1996; Chen *et al.*, 1998) or the support vector regression (Vapnik, 1995; Schölkopf and Smola, 2002). For general non-linear estimates, we shall approximate the third term in preSIC by the bootstrap method (Efron, 1979; Efron and Tibshirani, 1993). We define the bootstrap approximation of SIC (BASIC) as follows.

$$\begin{aligned} \mathrm{BASIC} = \|\hat{f}\|_{\mathcal{H}}^2 - 2\langle \hat{f}, \hat{f}_u \rangle_{\mathcal{H}} \\ + 2\mathrm{E}_{\boldsymbol{\epsilon}}^b \langle \hat{f}^b, X_u \hat{\boldsymbol{\epsilon}}^b \rangle_{\mathcal{H}}. \end{aligned} \tag{18}$$

More specifically, we calculate the third term $\mathrm{E}_{\boldsymbol{\epsilon}}^b \langle \hat{f}^b, X_u \hat{\boldsymbol{\epsilon}}^b \rangle_{\mathcal{H}}$ by bootstrapping residuals as follows.

1. Obtain an approximation $\hat{f}$ with samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ as usual.
2. Estimate the noise by $\hat{\epsilon}_i = y_i - \hat{f}(\boldsymbol{x}_i)$.
3. Create bootstrap noise samples $\{\hat{\epsilon}_i^b\}_{i=1}^n$ by sampling with replacement from $\{\hat{\epsilon}_i\}_{i=1}^n$.
4. Obtain an approximation $\hat{f}^b$ with the bootstrap samples $\{(\boldsymbol{x}_i, y_i^b) \mid y_i^b = \hat{f}(\boldsymbol{x}_i) + \epsilon_i^b\}_{i=1}^n$.
5. Calculate $\langle \hat{f}^b, X_u \hat{\boldsymbol{\epsilon}}^b \rangle_{\mathcal{H}}$.
6. Repeat 3. to 5. for a number of times and output the mean of $\langle \hat{f}^b, X_u \hat{\boldsymbol{\epsilon}}^b \rangle_{\mathcal{H}}$.

4. WHEN UNBIASED ESTIMATE OF f IS AVAILABLE

In the previous section, we derived SIC and its approximation for linear, smooth non-linear, and general non-linear estimates. In their derivations, we assumed that a linear unbiased estimate $\hat{f}_u$ of the learning target function f is available. In this section, we show how to obtain $\hat{f}_u$.

4.1 *Existence Condition for Unbiased Estimate of f*

The following theorem shows the existence condition for $\hat{f}_u$.

Theorem 4. A linear unbiased estimate $\hat{f}_u$ of the learning target function f exists if and only if the functions $\{K(\boldsymbol{x}, \boldsymbol{x}_i)\}_{i=1}^n$ span the whole RKHS $\mathcal{H}$.

Now we shall show how to obtain $\hat{f}_u$ under the situation where the condition in the above theorem is fulfilled. To this end, let us introduce the notion of the *Neumann-Schatten product* (Schatten, 1970). For any fixed g in a Hilbert space $\mathcal{H}_1$ and any fixed f in a Hilbert space $\mathcal{H}_2$, the Neumann-Schatten product of f and g, denoted by $(f \otimes \overline{g})$, is an operator from $\mathcal{H}_1$ to $\mathcal{H}_2$ that satisfies for any h in $\mathcal{H}_1$

$$(f \otimes \overline{g})\, h = \langle h, g \rangle f. \tag{19}$$

When both $\mathcal{H}_1$ and $\mathcal{H}_2$ are the Euclidean spaces, $(f \otimes \overline{g})$ is simply expressed as $(f \otimes \overline{g}) = f g^\top$. Using the Neumann-Schatten product, let us define the linear operator A by

$$A = \sum_{i=1}^{n} \left(\boldsymbol{e}_i \otimes \overline{K(\cdot, x_i)} \right), \tag{20}$$

where $\boldsymbol{e}_i$ is the i-th standard basis in $\mathbb{R}^n$, i.e., it is the n-dimensional vector with the i-th element 1

and others 0. The property of the reproducing kernel implies that

$$Af = (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_n))^\top. \quad (21)$$

For this reason, A is called the sampling operator.

Let $A^\dagger$ be the Moore-Penrose generalized inverse of A (Albert, 1972). Then we have the following theorem.

Theorem 5. If the functions $\{K(\boldsymbol{x}, \boldsymbol{x}_i)\}_{i=1}^n$ span the whole RKHS $\mathcal{H}$, a linear operator X_u that provides an unbiased estimate $\hat{f}_u$ is given by

$$X_u = A^\dagger. \quad (22)$$

It can be confirmed that $A^\dagger$ provides the best linear unbiased estimate of f (Albert, 1972).

4.2 *SIC for Linear Regression Models*

Here we consider a standard linear regression problem, and show how SIC can be applied.

Let us consider the case where the unknown function f is of the form

$$f(\boldsymbol{x}) = \sum_{i=1}^p \alpha_i^* \varphi_i(\boldsymbol{x}), \quad (23)$$

where $\{\varphi_i(\boldsymbol{x})\}_{i=1}^p$ are the specified basis functions and $\{\alpha_i^*\}_{i=1}^p$ are unknown. We estimate $f(\boldsymbol{x})$ by the following linear regression model.

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^p \alpha_i \varphi_i(\boldsymbol{x}). \quad (24)$$

$\{\alpha_i\}_{i=1}^p$ are parameters estimated by

$$\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_p)^\top = \boldsymbol{X}(\boldsymbol{y}), \quad (25)$$

where $\boldsymbol{X}$ is a vector-valued function from $\mathbb{R}^n$ to $\mathbb{R}^p$. The approximation error is measured by the expected weighted distance in the input domain $\mathcal{D}$.

$$\mathrm{E}_{\boldsymbol{\epsilon}} \int_{\mathcal{D}} \left(\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 w(\boldsymbol{x}) d\boldsymbol{x}, \quad (26)$$

where $w(\boldsymbol{x})$ is a specified weight function. Let $\boldsymbol{A}$ be the so-called *design matrix* whose (i, j)-th element is given by $\varphi_j(\boldsymbol{x}_i)$.

This setting corresponds to the case where the RKHS $\mathcal{H}$ is spanned by $\{\varphi_i(\boldsymbol{x})\}_{i=1}^p$ and the inner product is defined by

$$\langle f, g \rangle_{\mathcal{H}} = \int_{\mathcal{D}} f(\boldsymbol{x}) g(\boldsymbol{x}) w(\boldsymbol{x}) d\boldsymbol{x}. \quad (27)$$

Indeed, the reproducing kernel is given by

$$K(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^p \varphi_i(\boldsymbol{x}) \tilde{\varphi}_i(\boldsymbol{x}'), \quad (28)$$

where $\tilde{\varphi}_i(\boldsymbol{x})$ is the dual of $\varphi_i(\boldsymbol{x})$ (Ogawa, 1998). When $\{\varphi_i(\boldsymbol{x})\}_{i=1}^p$ is the orthonormal basis in the RKHS $\mathcal{H}$, $\tilde{\varphi}_i(\boldsymbol{x})$ simply agrees with $\varphi_i(\boldsymbol{x})$.

Then we have the following theorem.

Theorem 6. When the rank of $\boldsymbol{A}$ is p, the functions $\{K(\boldsymbol{x}, \boldsymbol{x}_i)\}_{i=1}^n$ always span the whole RKHS $\mathcal{H}$.

Therefore, when the rank of $\boldsymbol{A}$ is p, an unbiased estimate $\hat{f}_u$ of the learning target function f exists. In this case, SIC can be calculated as follows.

- When $\boldsymbol{X}$ is linear,

$$\mathrm{SIC} = \hat{\boldsymbol{\alpha}}^\top \boldsymbol{U} \hat{\boldsymbol{\alpha}} - 2\hat{\boldsymbol{\alpha}}^\top \boldsymbol{U} \boldsymbol{A}^\dagger \boldsymbol{y} + 2\mathrm{tr}\left(\boldsymbol{U}\boldsymbol{X}\boldsymbol{Q}(\boldsymbol{A}^\top)^\dagger\right), \quad (29)$$

  where $\boldsymbol{U}$ is the p-dimensional matrix whose (i, j)-th element is given by

$$\boldsymbol{U}_{ij} = \int_{\mathcal{D}} \varphi_i(\boldsymbol{x}) \varphi_j(\boldsymbol{x}) w(\boldsymbol{x}) d\boldsymbol{x}. \quad (30)$$

- When $\boldsymbol{X}$ is smooth non-linear,

$$\mathrm{SIC} = \hat{\boldsymbol{\alpha}}^\top \boldsymbol{U} \hat{\boldsymbol{\alpha}} - 2\hat{\boldsymbol{\alpha}}^\top \boldsymbol{U} \boldsymbol{A}^\dagger \boldsymbol{y} + 2\sigma^2 \sum_{i=1}^n \frac{\partial [(\boldsymbol{A}^\top)^\dagger \boldsymbol{U} \boldsymbol{X}]_i(\boldsymbol{y})}{\partial y_i}. \quad (31)$$

- When $\boldsymbol{X}$ is general non-linear,

$$\mathrm{BASIC} = \hat{\boldsymbol{\alpha}}^\top \boldsymbol{U} \hat{\boldsymbol{\alpha}} - 2\hat{\boldsymbol{\alpha}}^\top \boldsymbol{U} \boldsymbol{A}^\dagger \boldsymbol{y} + 2\mathrm{E}_{\boldsymbol{\epsilon}}^b \hat{\boldsymbol{\epsilon}}^{b\top} (\boldsymbol{A}^\top)^\dagger \boldsymbol{U} \hat{\boldsymbol{\alpha}}^b. \quad (32)$$

4.3 *Estimating Prediction Error and Test Error*

One of the common approximation error measures may be the prediction error (or the expected test error) defined by

$$\int_{\mathcal{D}} \left(\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 p(\boldsymbol{x}) d\boldsymbol{x}, \quad (33)$$

where $p(\boldsymbol{x})$ is the probability density function from which the (future) test input points are drawn. Letting $w(\boldsymbol{x}) = p(\boldsymbol{x})$, we can use SIC for estimating the prediction error.

However, $p(\boldsymbol{x})$ is often unknown so the matrix $\boldsymbol{U}$ can not be calculated. One of the options is to use the empirical distribution of $\{\boldsymbol{x}_i\}_{i=1}^n$ instead of $p(\boldsymbol{x})$ under the assumption that $\{\boldsymbol{x}_i\}_{i=1}^n$ are independently and

identically drawn from $p(\boldsymbol{x})$. That is, $\boldsymbol{U}$ is estimated by

$$\boldsymbol{U}_{ij} \approx \frac{1}{n}\sum_{k=1}^{n}\varphi_i(\boldsymbol{x}_k)\varphi_j(\boldsymbol{x}_k). \tag{34}$$

In this case, it can be confirmed that SIC for linear estimates essentially agrees with Mallows's C_L (Mallows, 1973) and SIC for smooth non-linear estimates essentially agrees with Stein's unbiased risk estimator (Stein, 1981).

When input points without output values (which is often referred to as unlabeled samples) are available, another option comes in handy. That is, we estimate $p(\boldsymbol{x})$ by the empirical distribution of the unlabeled samples. Then $\boldsymbol{U}$ is estimated by

$$\boldsymbol{U}_{ij} \approx \frac{1}{n'}\sum_{k=1}^{n'}\varphi_i(\boldsymbol{x}'_k)\varphi_j(\boldsymbol{x}'_k), \tag{35}$$

where $\{\boldsymbol{x}'_i\}_{i=1}^{n'}$ are the unlabeled samples. Note that ordinary samples $\{\boldsymbol{x}_i\}_{i=1}^{n}$ can also be included in the set of unlabeled samples.

In some cases, test input points $\{\boldsymbol{x}''_i\}_{i=1}^{n''}$ are known in advance, and the goal is to estimate the output values $\{f(\boldsymbol{x}''_i)\}_{i=1}^{n''}$ corresponding to the test input points (which is often referred to as the transductive inference). In such cases, SIC can be used for estimating the error at the test points $\{\boldsymbol{x}''_i\}_{i=1}^{n''}$ by defining $\boldsymbol{U}$ as

$$\boldsymbol{U}_{ij} = \frac{1}{n''}\sum_{k=1}^{n''}\varphi_i(\boldsymbol{x}''_k)\varphi_j(\boldsymbol{x}''_k). \tag{36}$$

5. WHEN UNBIASED ESTIMATE OF f IS NOT AVAILABLE

We showed in Section 4 that a linear unbiased estimate $\hat{f}_u$ of the unknown function f exists if and only if the functions $\{K(\boldsymbol{x},\boldsymbol{x}_i)\}_{i=1}^{n}$ span the whole RKHS $\mathcal{H}$. In this section, we consider the case where the functions $\{K(\boldsymbol{x},\boldsymbol{x}_i)\}_{i=1}^{n}$ do not span the whole RKHS $\mathcal{H}$.

5.1 *Existence Condition for Unbiased Estimate of Projection of f*

So far, we searched the approximation $\hat{f}$ in the whole RKHS $\mathcal{H}$. Here we restrict ourselves to searching the approximation $\hat{f}$ within a subspace $\mathcal{S}$ of the RKHS $\mathcal{H}$.

Let $f_{\mathcal{S}}$ be the orthogonal projection of f onto the subspace $\mathcal{S}$. Recalling that $\langle\hat{f},f\rangle_{\mathcal{H}} = \langle\hat{f},f_{\mathcal{S}}\rangle_{\mathcal{H}}$ for $\hat{f}\in\mathcal{S}$, the approximation error J defined by Eq.(5) is expressed as

$$J = \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f}\|_{\mathcal{H}}^2 - 2\mathrm{E}_{\boldsymbol{\epsilon}}\langle\hat{f},f_{\mathcal{S}}\rangle_{\mathcal{H}}, \tag{37}$$

where f in Eq.(5) is simply replaced by $f_{\mathcal{S}}$. Therefore, if a linear unbiased estimate of the projection $f_{\mathcal{S}}$ is available, we can make the same discussion as Section 3. Indeed, the following theorem shows the existence condition for a linear unbiased estimate of the projection $f_{\mathcal{S}}$.

Theorem 7. (Sugiyama and Müller, 2002) A linear unbiased estimate of the projection $f_{\mathcal{S}}$ exists if and only if the subspace $\mathcal{S}$ is included in the span of the functions $\{K(\boldsymbol{x},\boldsymbol{x}_i)\}_{i=1}^{n}$.

This theorem means that a linear unbiased estimate of the projection $f_{\mathcal{S}}$ exists if $\hat{f}$ is searched in the span of the functions $\{K(\boldsymbol{x},\boldsymbol{x}_i)\}_{i=1}^{n}$, i.e., $\hat{f}$ is searched of the form

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n}\alpha_i K(\boldsymbol{x},\boldsymbol{x}_i), \tag{38}$$

where $\{\alpha_i\}_{i=1}^{n}$ are parameters to be estimated.

The following theorem shows how to obtain a linear unbiased estimate of the projection $f_{\mathcal{S}}$.

Theorem 8. If the subspace $\mathcal{S}$ is included in the span of the functions $\{K(\boldsymbol{x},\boldsymbol{x}_i)\}_{i=1}^{n}$, a linear operator X_u that provides an unbiased estimate of the projection f is given by Eq.(22).

Theorems 5 and 8 show that the same $A^{\dagger}$ gives an unbiased estimate of f or the projection of f.

When parameters $\{\alpha_i\}_{i=1}^{n}$ in the kernel regression model (38) are estimated by

$$\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1,\hat{\alpha}_2,\ldots,\hat{\alpha}_n)^{\top} = \boldsymbol{X}(\boldsymbol{y}), \tag{39}$$

where $\boldsymbol{X}$ is a vector-valued function from $\mathbb{R}^n$ to $\mathbb{R}^n$, SIC can be calculated as follows.

- When $\boldsymbol{X}$ is linear,

$$\mathrm{SIC} = \hat{\boldsymbol{\alpha}}^{\top}\boldsymbol{K}\hat{\boldsymbol{\alpha}} - 2\hat{\boldsymbol{\alpha}}^{\top}\boldsymbol{y} + 2\mathrm{tr}\left(\boldsymbol{X}\boldsymbol{Q}\right). \tag{40}$$

- When $\boldsymbol{X}$ is smooth non-linear,

$$\mathrm{SIC} = \hat{\boldsymbol{\alpha}}^{\top}\boldsymbol{K}\hat{\boldsymbol{\alpha}} - 2\hat{\boldsymbol{\alpha}}^{\top}\boldsymbol{y} + 2\sigma^2\sum_{i=1}^{n}\frac{\partial[\boldsymbol{X}]_i(\boldsymbol{y})}{\partial y_i}. \tag{41}$$

- When $\boldsymbol{X}$ is general non-linear,

$$\mathrm{BASIC} = \hat{\boldsymbol{\alpha}}^{\top}\boldsymbol{K}\hat{\boldsymbol{\alpha}} - 2\hat{\boldsymbol{\alpha}}^{\top}\boldsymbol{y} + 2\mathrm{E}_{\boldsymbol{\epsilon}}^{b}\hat{\boldsymbol{\epsilon}}^{b\top}\hat{\boldsymbol{\alpha}}^{b}. \tag{42}$$

5.2 *Restriction on Generalization Measure*

As shown above, even when the functions $\{K(\boldsymbol{x},\boldsymbol{x}_i)\}_{i=1}^{n}$ do not span the whole RKHS $\mathcal{H}$, SIC can be applied if the kernel regression model (38) is used.

However, in this case, we should care about the fact that the shape of the kernel function $K(\boldsymbol{x}, \boldsymbol{x}')$ and the definition of the norm in the RKHS $\mathcal{H}$ relate each other. That is, if we use a desired kernel function, then we can no longer define the approximation error measure as desired. Conversely, if we use the desired approximation error measure, then the shape of the kernel function can no longer be chosen as desired.

For example, if the following Gaussian kernel is used

$$K(\boldsymbol{x}, \boldsymbol{x}') = g(\boldsymbol{x} - \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2}\right), \quad (43)$$

then the norm in the Gaussian RKHS is given by

$$\|\hat{f} - f\|_{\mathcal{H}}^2 = \int \frac{\left(F[\hat{f}](\boldsymbol{\omega}) - F[f](\boldsymbol{\omega})\right)^2}{F[g](\boldsymbol{\omega})} d\boldsymbol{\omega}, \quad (44)$$

where $F[\cdot]$ denotes the Fourier transform. This norm has a property that high frequency components are strongly penalized.

6. CONCLUSIONS

In this paper, we formulated the system identification problem as a function approximation problem in a reproducing kernel Hilbert space (RKHS), and derived an estimator of the approximation error defined by the RKHS norm called the subspace information criterion (SIC). When the approximation function is estimated in a linear or smooth non-linear fashion, SIC has an analytic form and is an unbiased estimator of the true approximation error. When the approximation function is estimated in a general non-linear fashion, we proposed approximating SIC by the bootstrap method. SIC can be applied when a linear unbiased estimate of the unknown target function is available. We provided the necessary and sufficient condition for the existence of the linear unbiased estimate. For the cases where such a linear unbiased estimate does not exist, we further showed that SIC can be still applied if the approximation function is a kernel regression model.

7. REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **AC-19**(6), 716–723.

Akaike, H. (1980). Likelihood and the Bayes procedure. In: *Bayesian Statistics* (N. J. Bernardo et al. Eds.). University Press. Valencia. pp. 141–166.

Albert, A. (1972). *Regression and the Moore-Penrose Pseudoinverse*. Academic Press. New York and London.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society* **68**, 337–404.

Chen, S. S., D. L. Donoho and M. A. Saunders (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* **20**(1), 33–61.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* **7**(1), 1–26.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. Chapman & Hall. New York.

Hoerl, A.E. and R.W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**(3), 55–67.

Huber, P. J. (1981). *Robust Statistics*. John Wiley. New York.

Mallows, C. L. (1973). Some comments on C_P. *Technometrics* **15**(4), 661–675.

Murata, N., S. Yoshizawa and S. Amari (1994). Network information criterion — Determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks* **5**(6), 865–872.

Ogawa, H. (1998). Theory of pseudo biorthogonal bases and its application. In: *Research Institute for Mathematical Science, RIMS Kokyuroku, 1067, Reproducing Kernels and their Applications*. number 1067. pp. 24–38.

Schatten, R. (1970). *Norm Ideals of Completely Continuous Operators*. Springer-Verlag. Berlin.

Schölkopf, B. and A. J. Smola (2002). *Learning with Kernels*. MIT Press. Cambridge, MA.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* **9**(6), 1135–1151.

Sugiyama, M. and H. Ogawa (2001). Subspace information criterion for model selection. *Neural Computation* **13**(8), 1863–1889.

Sugiyama, M. and K.-R. Müller (2002). The subspace information criterion for infinite dimensional hypothesis spaces. *Journal of Machine Learning Research* **3**(Nov), 323–359.

Suykens, J. A. K., T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle (2002). *Least Squares Support Vector Machines*. World Scientific Pub. Co.. Singapore.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**(1), 267–288.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag. Berlin.

Wahba, H. (1990). *Spline Model for Observational Data*. Society for Industrial and Applied Mathematics. Philadelphia and Pennsylvania.

Williams, C. K. I. and C. E. Rasmussen (1996). Gaussian processes for regression. In: *Advances in Neural Information Processing Systems* (D. S. Touretzky, M. C. Mozer and M. E. Hasselmo, Eds.). Vol. 8. The MIT Press. pp. 514–520.

Williams, P. M. (1995). Bayesian regularization and pruning using a Laplace prior. *Neural Computation* **7**(1), 117–143.