SPACE KERNEL ANALYSIS

Liuling Gong, Dan Schonfeld

University of Illinois at Chicago, Dept. of Electrical and Computer Engineering
851 S Morgan St, Chicago, IL 60607

ABSTRACT

In this paper, we propose a novel nonparametric modeling technique, namely Space Kernel Analysis (SKA), as a result of the definition of the space kernel. We analyze the uncertainty of SKA and show that SKA is subjected to the bias/variance dilemma. Nevertheless, we demonstrate that, by a proper choice of the space kernel matrix, SKA is able to balance between the robustness and accuracy and hence outperforms other kernel-based learning methods. The cost function of SKA is derived, and it proves that SKA minimizes the Weighted Least Squared cost function whose weight matrix is diagonal and determined by the space kernel matrix. The parallels between SKA and several other nonparametric modeling techniques are examined. Study shows that the traditional Kernel Regression, General Regression Neural Network, Similarity Based Modeling and Radial Basis Function Network are examples of SKA with specified space kernel matrices.

Index Terms— kernels, nonparametric methods, uncertainty analysis, cost function

1. INTRODUCTION

Nonparametric methods provide an explanatory and diagnostic tool to study the association between covariates and responses in complex data sets. As distribution free methods, they do not rely on assumptions that the data are drawn from a given probability distribution. In the context of condition based monitoring and fault detection, a number of nonparametric methods have been applied successfully, and among which the kernel-based learning methods have been studied extensively due to their outstanding performance in real world applications [1].

The most widely studied problems attached to kernel-based learning methods are the identification of appropriate kernel functions and the bandwidth choice. A key paper is [2], in which data-driven bandwidth selectors are discussed. As a result of the issue, one well-known limitation of the kernel-based learning methods is the bias/variance dilemma [3]. Another limitation due to the extension of kernel theory in multivariate regression is the "curse of dimensionality" phenomena [4], which makes the estimate of high-dimensional regression function notoriously difficult. Because of the presumption of smoothness in the data most realizations of Kernel Regression (KR) made, KR is used to produce a smooth estimate of the regression surface, and it's hard to choose a global optimal width in KR involving high-dimensional data sets.

To balance the tradeoff between bias and variance, we propose a novel kernel-based learning method, namely Space Kernel Analysis (SKA). Other than the traditional kernel which operates between two vectors, the space kernel in SKA is defined between a vector and a space. We analyze the uncertainty of SKA and show that SKA is also subjected to the bias/variance dilemma. However, by a proper choice of the space kernel matrix, SKA is able to outperform other kernel-based learning methods. We further study the cost function of SKA, which in the end indicates that SKA produces a Weighted Least Squared (WLS) estimate. Several nonparametric techniques, including KR, General Regression Neural Network (GRNN), Similarity Based Modeling (SBM) and Radial Basis Function Network (RBNF), are examined in this paper, which turn out to be examples of SKA with specified space kernel matrices. The rest of this paper is organized as follows. We first describe the mathematics behind SKA in Section 2. Subsequently, we illustrate SKA with some well-understood nonparametric techniques in Section 3. Finally, we conclude our work in Section 4.

2. SPACE KERNEL ANALYSIS

2.1. A Mathematical Framework for Space Kernel Analysis

A typical learning problem involves an input vector X and a response vector Y, where the pair (X, Y) obeys some unknown joint probability distribution. A training set $\{(X_1, Y_1), \cdots, (X_L, Y_L)\}$ is a collection of observed (X, Y) pairs. The signal model which has m data sources is,

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \cdots, L, \tag{1}$$

where $f = [f_1, \cdots, f_m]^T : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is the underlying dependency, ε_i is the zero-mean noise sample and $\mathrm{E}[\varepsilon_i \cdot \varepsilon_j^T] =$

The authors would like to thank VG Bioinformatics for their funding and support through this project.

$\Theta \cdot \delta_{ij}$ where δ_{ij} is the Kronecker delta function. All vectors in this paper are column vectors unless otherwise specified. The standard regression task is to estimate the unknown function $f(\cdot)$ based solely on the training set. Typically, $f(\cdot)$ is chosen to minimize some loss function, e.g., the sum of observed squared errors $\sum_{i=1}^{L} \|Y_i - f(X_i)\|^2$.

The idea of estimating $f(\cdot)$ applying a locally weighted average can be traced back to the regressogram proposed by Tukey [5], which partitions the training set into several subsets and then averages the response vectors Y inside each subset. The regressogram produces a quite rough estimate due to its stepwise nature. A natural extension of the regressogram is the moving window estimate, which averages Y based on a centered neighborhood of X [Chapter 3][6]. For further development, we generalize the estimate as a weighted average, where the weights are determined by some kernel function K and space kernel matrix A. The kernel K is a bounded and integrable real-valued function, and the space kernel matrix A is an $L \times L$ matrix defined on the training set. Usually, $K(X_1, X_2)$ is taken to be a positive symmetric function which achieves its maximum when $X_1 = X_2$ and monotonically decreases with $\|X_1 - X_2\|$. Without loss of generality, we assume that the maximum value of K is 1. Matrix A retrieves information from the training set and hence enables the learning process adaptable to various regression surfaces.

Denote $[X_1, \cdots, X_L]$ and $[Y_1, \cdots, Y_L]$ by X_{tr} and Y_{tr} respectively, which are $m \times L$ matrices. The output of SKA at a given input X_{n} is

$$Y_{\text{n}} = \hat{f}_{\text{SKA}}(X_{\text{n}}) = \frac{\sum_{i=1}^{L} Y_i \sum_{j=1}^{L} A_{ij} K(X_j, X_{\text{n}})}{\sum_{i=1}^{L} \sum_{j=1}^{L} A_{ij} K(X_j, X_{\text{n}})}, \quad (2)$$

which can be rewritten in the compact form

$$Y_{\text{n}} = \frac{Y_{\text{tr}} \cdot A \cdot (X_{\text{tr}}^T \otimes X_{\text{n}})}{\mathbf{1}^T \cdot A \cdot (X_{\text{tr}}^T \otimes X_{\text{n}})} = \frac{Y_{\text{tr}} \cdot K_S(X_{\text{tr}}, X_{\text{n}})}{\mathbf{1}^T \cdot K_S(X_{\text{tr}}, X_{\text{n}})}. \quad (3)$$

Here $\otimes$ is the similarity operator defined as $(X_{\text{tr}}^T \otimes X_{\text{n}}) = [K(X_1, X_{\text{n}}), \cdots, K(X_L, X_{\text{n}})]^T$, $\mathbf{1}$ is an $L \times 1$ vector with all elements being 1, and

$$K_S(X_{\text{tr}}, X_{\text{n}}) = A \cdot (X_{\text{tr}}^T \otimes X_{\text{n}}), \quad (4)$$

is the space kernel which is an $L \times 1$ vector giving the similarity between X_{n} and X_{tr}. Notice that, while the traditional kernel K operates between two vectors, the space kernel K_S operates between a vector and a space.

2.2. Bias/Variance Dilemma

A measure of the effectiveness of $f(\cdot)$ as a predictor of Y_{n} at a future input X_{n} is the mean squared error, which can be decomposed into bias and variance components [3],

$$\begin{aligned} \text{E}_{\text{D}}[\|Y_{\text{n}} - f(X_{\text{n}})\|^2] =& \text{E}_{\text{D}}[\|Y_{\text{n}} - \text{E}_{\text{D}}[Y_{\text{n}}]\|^2] + \\ & \text{E}_{\text{D}}[\|\text{E}_{\text{D}}[Y_{\text{n}}] - f(X_{\text{n}})\|^2] \\ =& \text{Var} + \text{Bias}^T \cdot \text{Bias}, \end{aligned} \quad (5)$$

where $\text{E}_{\text{D}}[\cdot]$ represents expectation over the ensemble of possible X_{tr} for a fixed sample size L.

The Taylor series of $f(X_i)$ at X_{n} is,

$$f(X_i) = f(X_{\text{n}}) + \nabla f(X_{\text{n}}) \cdot (X_i - X_{\text{n}}) + \cdots, \quad (6)$$

where $\nabla f(X_{\text{n}}) = [\nabla f_1(X_{\text{n}}), \cdots, \nabla f_m(X_{\text{n}})]^T$ is an $m \times m$ matrix. Denote $K_S(X_{\text{tr}}, X_{\text{n}})$ by K_S, and let $b_2 = (\mathbf{1}^T \cdot K_S)^{-1}$ be a non-zero scalar. By combination of (1) (3) and (6), we can easily derive that,

$$\text{E}_{\text{D}}[Y_{\text{n}}] = f(X_{\text{n}}) + \nabla f(X_{\text{n}}) \cdot [X_1 - X_{\text{n}}, \cdots, X_L - X_{\text{n}}] \cdot K_S \cdot b_2 + \cdots. \quad (7)$$

Therefore, the bias, which is an $m \times 1$ vector, is

$$\begin{aligned} \text{Bias} =& \text{E}_{\text{D}}[Y_{\text{n}}] - f(X_{\text{n}}) \\ =& \nabla f(X_{\text{n}}) \cdot [X_1 - X_{\text{n}}, \cdots, X_L - X_{\text{n}}] \cdot K_S \cdot b_2 + \cdots. \end{aligned} \quad (8)$$

Assuming the independency of m data sources, i.e., $\nabla^k f(X_{\text{n}}) = \text{diag}\{\frac{\partial^k f_1(X_{\text{n}})}{\partial X_{\text{n},1}^k}, \cdots, \frac{\partial^k f_m(X_{\text{n}})}{\partial X_{\text{n},m}^k}\}$, $k \in \mathbb{Z}^+$, where $X_{\text{n},i}$ is the ith element of X_{n}, the bias for the ith data source is

$$[\text{Bias}]_i = \frac{\partial f_i(X_{\text{n}})}{\partial X_{\text{n},i}} \cdot ([X_{1,i} - X_{\text{n},i}, \cdots, X_{L,i} - X_{\text{n},i}] \cdot K_S \cdot b_2) + \cdots. \quad (9)$$

Notice that the bias increases when the curvature of regression surface, i.e., $\frac{\partial^k f_i(X_{\text{n}})}{\partial X_{\text{n},i}^k}$, increases. This phenomenon has been observed in several kernel-based learning methods [7].

Given that only the first M terms are kept in (9) and $\frac{\partial^k f_i(X_{\text{n}})}{\partial X_{\text{n},i}^k} \neq 0$, to have $[\text{Bias}]_i = 0$, we should have

$$[(X_{1,i} - X_{\text{n},i})^k, \cdots, (X_{L,i} - X_{\text{n},i})^k] \cdot K_S = 0, \ k = 1, \cdots, M, \quad (10)$$

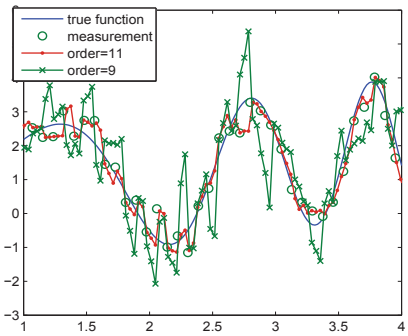
That is, K_S should be orthogonal to vectors $\Delta X_{i,k} = [(X_{1,i} - X_{\text{n},i})^k, \cdots, (X_{L,i} - X_{\text{n},i})^k]^T$, $k = 1, \cdots, M$. Notice that the larger the dimensionality L of K_S, the larger the M for which (10) may hold true, and hence the less the bias. This is intuitive from the point of view that more observations usually results in more accurate estimate.

The variance, which is a scalar, can be easily derived as

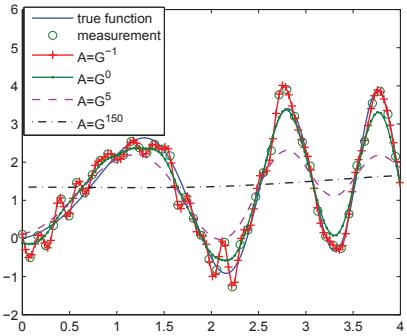
$$\begin{aligned} \text{Var} &= \text{E}_{\text{D}}[\|Y_{\text{n}} - \text{E}_{\text{D}}[Y_{\text{n}}]\|^2] \\ &= \text{E}_{\text{D}}[([\varepsilon_1, \cdots, \varepsilon_L] \cdot K_S \cdot b_2)^T \cdot ([\varepsilon_1, \cdots, \varepsilon_L] \cdot K_S \cdot b_2)] \\ &= \text{Trace}(\Theta) \cdot \left\| K_S / (\mathbf{1}^T \cdot K_S) \right\|^2. \end{aligned} \quad (11)$$

which achieves the minimum when all the elements in K_S are identical. Notice that, when $\mathbf{1}^T \Delta X_{i,k} \neq 0$ which is the most likely case, this condition is incompatible with the condition under which (10) holds. Therefore, SKA is also subjected to the bias/variance dilemma.

An algorithm is implemented such that the bias is minimized to some order M, i.e., matrix A is generated at each X_{n} such that (10) is satisfied. An example is shown in Fig. 1(a) when $y = 2\sin(x^2) + 0.5x$. We observe that the estimates exhibit high variance even when M is relatively large. The reason is that the algorithm is implemented such that only the bias is minimized regardless of the variance.



(a) Bias is minimized to different orders of Taylor series.



(b) Space kernel matrix A is the power of G matrix.

Fig. 1. SKA estimates of $y = 2\sin(x^2) + 0.5x$.

2.3. Cost Function for Space Kernel Analysis

A linear regression model is given by

$$Y = X\beta + \nu, \tag{12}$$

where Y is an $L \times 1$ vector, X is an $L \times m$ design matrix, β is an $m \times 1$ vector of unknown parameters, ν is a zero-mean $L \times 1$ vector and $\mathrm{E}[\nu \cdot \nu^T] = \sigma^2\Lambda$, $\Lambda = \mathrm{diag}(\frac{1}{\lambda_1}, \cdots, \frac{1}{\lambda_L})$. If X is of full rank m, the unique solution to minimizing the cost function

$$\mathrm{E}_{\mathrm{WLS}} = (Y - X\beta)^T W (Y - X\beta) \tag{13}$$

is the Weighted Least Squares (WLS) estimate

$$\hat{\beta}_{\mathrm{WLS}} = (X^T W X)^{-1} X^T W Y, \tag{14}$$

where $W = \Lambda^{-1}$ is the weight matrix [Chapter 4] [8].

Let $Y = Y_{\mathrm{tr}}^T$, $X = \mathbf{1}$, $\beta = Y_{\mathrm{n}}^T$, and

$$W = \mathrm{diag}(K_S(X_{\mathrm{tr}}, X_{\mathrm{n}})) = \mathrm{diag}(A \cdot (X_{\mathrm{tr}}^T \otimes X_{\mathrm{n}})), \tag{15}$$

in model (12). The minimization to (13) results in

$$\begin{aligned} Y_{\mathrm{n}}^T &= \frac{\mathbf{1}^T \cdot \mathrm{diag}(A \cdot (X_{\mathrm{tr}}^T \otimes X_{\mathrm{n}}))}{\mathbf{1}^T \cdot \mathrm{diag}(A \cdot (X_{\mathrm{tr}}^T \otimes X_{\mathrm{n}})) \cdot \mathbf{1}} \cdot Y_{\mathrm{tr}}^T \\ &= \left(\frac{Y_{\mathrm{tr}} \cdot A \cdot (X_{\mathrm{tr}}^T \otimes X_{\mathrm{n}})}{\mathbf{1}^T \cdot A \cdot (X_{\mathrm{tr}}^T \otimes X_{\mathrm{n}})} \right)^T, \end{aligned} \tag{16}$$

which is equivalent to SKA shown in (3). As a result of (15), we are able to adjust the weights in (13) by handling the matrix A, which allows the adaptability and flexibility of SKA.

The square matrix $G = (X_{\mathrm{tr}}^T \otimes X_{\mathrm{tr}})$ is commonly designated as the similarity matrix. We will study the role of G in SKA when $A = G^m$, $m \in \mathbb{Z}$ in the following.

Lemma 1: Given a semi-positive matrix $W_0 = \mathrm{diag}(w_1^{(0)}, \cdots, w_L^{(0)}) \succeq 0$, if $\frac{w_i^{(0)}}{w_j^{(0)}} \geq 1$ holds for some $i \neq j$, $1 \leq i, j \leq L$, we have the following inequality hold for matrix $W_t = \mathrm{diag}(G^t \cdot W_0 \cdot \mathbf{1}) = \mathrm{diag}(w_1^{(t)}, \cdots, w_L^{(t)}) \succeq 0$, $t \in \mathbb{Z}^+$:

$$1 \leq \frac{w_i^{(t)}}{w_j^{(t)}} \leq \frac{w_i^{(t-1)}}{w_j^{(t-1)}} \leq \cdots \leq \frac{w_i^{(0)}}{w_j^{(0)}}. \tag{17}$$

Proof: This lemma can be proved by induction.

(i) When $t = 1$, we have

$$W_1 = \mathrm{diag}(G \cdot W_0 \cdot \mathbf{1}) = \mathrm{diag}(w_1^{(1)}, \cdots, w_L^{(1)}), \tag{18}$$

where $w_i^{(1)} = \sum_{k=1}^{L} (X_i^T \otimes X_k) \cdot w_k^{(0)} \geq 0$. Denote $\frac{w_i^{(0)}}{w_j^{(0)}}$ by p such that $p \geq 1$ and $p = \infty$ when $w_j^{(0)} = 0$. Therefore,

$$\frac{w_i^{(1)}}{w_j^{(1)}} = \frac{p \cdot w_j^{(0)} + (X_i^T \otimes X_j) \cdot w_j^{(0)} + \sum_{k \neq i,j} (X_i^T \otimes X_k) \cdot w_k^{(0)}}{w_j^{(0)} + p \cdot (X_j^T \otimes X_i) \cdot w_j^{(0)} + \sum_{k \neq i,j} (X_j^T \otimes X_k) \cdot w_k^{(0)}}. \tag{19}$$

Denote the denominator in (19) by Ψ, and we have,

$$\begin{aligned} \frac{w_i^{(1)}}{w_j^{(1)}} - \frac{w_i^{(0)}}{w_j^{(0)}} &\cong \frac{(1-p) \cdot \sum_{k \neq i,j} (X_i^T \otimes X_k) \cdot w_k^{(0)}}{\Psi} \\ &\quad + \frac{(1-p^2) \cdot (X_i^T \otimes X_j) \cdot w_j^{(0)}}{\Psi} \qquad (20) \\ &\leq 0, \qquad (21) \end{aligned}$$

where (20) follows from the assumption that $\sum_{k \neq i,j} (X_i^T \otimes X_k) \cdot w_k^{(0)} \cong \sum_{k \neq i,j} (X_j^T \otimes X_k) \cdot w_k^{(0)}$ and (21) follows from the fact that $p \geq 1$, $w_i^{(0)} \geq 0$, $(X_i^T \otimes X_j) \geq 0$, $1 \leq i, j \leq L$. Equality holds if and only if $p = 1$. Furthermore,

$$\frac{w_i^{(1)}}{w_j^{(1)}} - 1 \cong \frac{(p-1) \cdot \left(w_j^{(0)} - (X_i^T \otimes X_j) \cdot w_j^{(0)} \right)}{\Psi} \geq 0, \tag{22}$$

which follows from the fact that $(X_i^T \otimes X_j) = (X_j^T \otimes X_i) < 1$, $\forall i \neq j$. Equality holds if and only if $p = 1$.

(ii) When $t = k$, $k \in \mathbb{Z}^+$, we postulate that (17) is true.

(iii) When $t = k + 1$, we have $W_{k+1} = \mathrm{diag}(G^{k+1} \cdot W_0 \cdot \mathbf{1}) = \mathrm{diag}(G \cdot W_k \cdot \mathbf{1})$, where $W_k \succeq 0$ and $\frac{w_i^{(k)}}{w_j^{(k)}} \geq 1$ according to the postulate in (ii). Therefore, (17) holds for W_{k+1} following the similar steps in (i). ■

Lemma 1 indicates that the similarity matrix G smoothes the estimate which converges in the training space when $A = G^t$, $t \to \infty$. An example when $A = G^t$, $t = -1, 0, 5, 150$ is shown in Fig. 1(b), which corroborates with the analysis.

3. EXAMPLES OF SPACE KERNEL ANALYSIS

The Nadaray-Watson Kernel Regression (NW-KR) is the most popular nonparametric estimator and is defined as [6]

$$f_{\text{NW}}(X_{\text{n}}) = \frac{\sum_i Y_i K(X_{\text{n}}, X_i)}{\sum_i K(X_{\text{n}}, X_i)} = \frac{Y_{\text{tr}} \cdot (X_{\text{tr}}^T \otimes X_{\text{n}})}{\mathbf{1}^T \cdot (X_{\text{tr}}^T \otimes X_{\text{n}})}, \quad (23)$$

which is equivalent to (3) when $A = I$. While the General Regression Neural Network (GRNN) proposed by Specht [9] is in fact an example of NW-KR where the kernel is a Gaussian function, it is also an example of SKA when $A = I$.

As an interpolation technique, Similarity Based Method (SBM) is designed to exactly fit the training data [10]. The definition of SBM is

$$f_{\text{SBM}}(X_{\text{n}}) = \frac{Y_{\text{tr}} \cdot (X_{\text{tr}}^T \otimes X_{\text{tr}})^{-1} \cdot (X_{\text{tr}}^T \otimes X_{\text{n}})}{\mathbf{1}^T \cdot (X_{\text{tr}}^T \otimes X_{\text{tr}})^{-1} \cdot (X_{\text{tr}}^T \otimes X_{\text{n}})}, \quad (24)$$

which is equivalent to (3) when $A = G^{-1} = (X_{\text{tr}}^T \otimes X_{\text{tr}})^{-1}$.

The output of a normalized general Radial Basis Function Network (RBFN) is [11],

$$f_{\text{RB}}^{\text{G}}(X_{\text{n}}) = \frac{\sum_{i=1}^L c_i w_i}{\sum_{i=1}^L w_i} = \frac{\sum_{i=1}^L c_i R_i(X_{\text{n}}, X_i)}{\sum_{i=1}^L R_i(X_{\text{n}}, X_i)}, \quad (25)$$

where c_i is the connection weight and $R_i(\cdot)$ is typically a Gaussian function centered at X_i. Comparing (25) with (23), we notice that the general RBFN is actually an NW-KR where the kernel is a Gaussian function, i.e., GRNN.

A Gaussian interpolation RBFN, which yields exact desired outputs for all training data, is defined as [Chapter 9][12]

$$f_{\text{RB}}^{\text{I}}(X) = \sum_{i=1}^{L} c_i \exp\left[\frac{-\|X_i - X\|^2}{2\sigma_i^2}\right] = \sum_{i=1}^{L} c_i K(X, X_i), \quad (26)$$

where σ_i is the given width parameter and c_i is the unknown weight coefficient. Rewrite (26) in a compact matrix form when $X = X_i$, $i = 1, \cdots, L$, and we have

$$Y_{\text{tr}} = C \cdot G, \quad (27)$$

where $Y_{\text{tr}} = [f(X_1), \cdots, f(X_L)]$, $C = [c_1, \cdots, c_L]$, and $G = (X_{\text{tr}}^T \otimes X_{\text{tr}})$ is the similarity matrix. When G is nonsingular, we have a unique solution to (27),

$$C = Y_{\text{tr}} \cdot G^{-1}. \quad (28)$$

Therefore, the output of interpolation RNBF at input X_{n} is

$$Y_{\text{n}} = C \cdot (X_{\text{tr}}^T \otimes X_{\text{n}}) = Y_{\text{tr}} \cdot (X_{\text{tr}}^T \otimes X_{\text{tr}})^{-1} \cdot (X_{\text{tr}}^T \otimes X_{\text{n}}), \quad (29)$$

which is equivalent to SBM, i.e., SKA where $A = G^{-1}$, without the coefficients normalization step.

4. CONCLUSION

The definition of space kernel is given in this paper, which gives the similarity between a vector and a space. As a result, a novel nonparametric modeling technique, namely Space Kernel Analysis (SKA), is studied. The uncertainty analysis of SKA demonstrates that SKA is subjected to the bias/variance dilemma like many other kernel-based learning methods. However, by a proper choice of the space kernel matrix, SKA is able to balance between the robustness and accuracy as required. Further study shows that SKA minimizes the cost function in WLS estimate whose weight matrix is determined by the space kernel. The parallels between SKA and several other nonparametric modeling techniques are examined, which show that some well-known nonparametric modeling techniques, including Kernel Regression (KR), General Regression Neural Network (GRNN), Similarity Based Method (SBM) and Radial Basis Function Network (RBFN), are examples of SKA where space kernel matrix is specified with various matrices.

5. REFERENCES

[1] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.

[2] W. Hardle and J. Marron, "Optimal bandwidth selection in nonparametric regression function estimation," *Ann. Statist.*, vol. 13, pp. 1465–1481, 1985.

[3] S. Geman, "Neural networks and the bias/variance dilemma," *Neural Computation*, vol. 4, pp. 1–58, 1992.

[4] R. Bellman, *Adaptive Control Process*, Princeton University Press, Princeton,NJ, 1961.

[5] J. Tukey, "Curves as parameters, and touch estimation," *Proceedings of the 4th symposium on Mathematics, Statistics and Probability*, pp. 681–694, 1961.

[6] M. Schimek, *Smoothing and Regression: Approaches, Computation, and Application*, John Wiley & Sons, New York, 2000.

[7] J. Fox, *Nonparametric Simple Regression: Smoothing Scatterplots*, Sage Publications, Inc., 2000.

[8] G. Arnold R. Brook, *Applied Regression Analysis and Experimental Design*, Marcel Dekker, New York, 1985.

[9] D. Specht, "A general regression neural network," *IEEE Trans. On Neural Networks*, vol. 2, pp. 568–576, 1991.

[10] S. Wegerich and X. Xu, "A performance comparison of similarity based and kernel modeling techniques," in *Proc. of MARCON 2003*, TN, May 2003.

[11] J. Moody and C. Darken, "Learning with localized receptive fields," *Proceedings of the 1988 Connectionist Models Summer School, eds. Touretzky, Hinton, and Sejnowski. Morgan-Kaufmann, Publishers*, 1988.

[12] C. Sun J. Jang and E. Mizutani, *Neuro-fuzzy and Soft Computing*, Prentice Hall, 1997.