Minimum Message Length and Classical Methods for Model Selection in Univariate Polynomial Regression

Murlikrishna Viswanathan, Young-kyu Yang, and Taeg Keun Whangbo

The problem of selection among competing models has been a fundamental issue in statistical data analysis. Good fits to data can be misleading since they can result from properties of the model that have nothing to do with it being a close approximation to the source distribution of interest (for example, overfitting). In this study we focus on the preference among models from a family of polynomial regressors. Three decades of research has spawned a number of plausible techniques for the selection of models, namely, Akaike's Finite Prediction Error (FPE) and Information Criterion (AIC), Schwartz's criterion (SCH), Generalized Cross Validation (GCV), Wallace's Minimum Message Length (MML), Minimum Description Length (MDL), and Vapnik's Structural Risk Minimization (SRM). The fundamental similarity between all these principles is their attempt to define an appropriate balance between the complexity of models and their ability to explain the data. This paper presents an empirical study of the above principles in the context of model selection, where the models under consideration are univariate polynomials. The paper includes a detailed empirical evaluation of the model selection methods on six target functions, with varying sample sizes and added Gaussian noise. The results from the study appear to provide strong evidence in support of the MML- and SRM- based methods over the other standard approaches (FPE, AIC, SCH and GCV).

Keywords: Polynomial model selection, regression, penalization, minimum message length, MML.

Manuscript received Dec. 4, 2004; revised Aug. 26, 2005.

This research was supported by the Ministry of Information and Communication (MIC), Korea, under the Information Technology Research Center (ITRC) support program supervised by the Institute of Information Technology Assessment (IITA).

Murlikrishna Viswanathan (phone: +82 31 750 5595, email: murli@kyungwon.ac.kr), Young-Kyu Yang (email: ykyang@kyungwon.ac.kr), and Taeg Keun Whangbo (email: tkwhangbo@kyungwon.ac.kr) are with the College of Software, Kyungwon University, Gyeonggi-Do, Korea.

I. Introduction

The most essential question in applied modeling is to determine the true mechanism or distribution that has generated a given set of data. However, answering this question for real world data is illusional. Instead, the question is often reformulated as choosing that model from a set of available models that is closest to these true processes. The challenge of model selection is that there is no clear and universal criterion that defines which model is most suitable. Goodness-of-fit of a model is one aspect that should be considered, but it has to be balanced with the model parameters' variability. An absurd and false model may fit perfectly if the model has enough complexity by comparison to the amount of data available. The price paid is a high sensitivity to small changes in the data, because at least a part of the coincidental noise was misinterpreted as a relevant underlying mechanism. In the limit, the model would simply store the data values and be nothing more than a data replicator. In this view, the problem of model selection can be reformulated as determining what part of the measurements are due to stochastic variations, which should be ignored, in order to capture only those variations caused by the processes of interest. The model that disentangles signal from noise most effectively is then closest to the true model and should be selected as the best [1].

In this study, we focus on the preference among competing models from a family of polynomial regressors. Classical statistics offers a number of well-known techniques for the selection of models in polynomial regression, namely, Finite Prediction Error (FPE) [2], Akaike's Information Criterion (AIC) [3], Schwartz's criterion (SCH) [4], and Generalized Cross Validation (GCV) [5]. Wallace's Minimum Message Length (MML) principle [6]-[8] and Vapnik's Structural Risk

Minimization (SRM) [9], [10] – based on the classical Vapnik-Chervonenkis (VC) theory of VC-dimensionality – are plausible additions to this family of model-selection principles. SRM and MML are generic in the sense that they can be applied to any family of models, and similar in their attempt to define a trade-off between the complexity of a given model and its *goodness of fit* to the data under observation – although they do use different trade-offs, with MML's being Bayesian and SRM's being non-Bayesian as but one of the differences.

Early measures of generalizability such as AIC and SCH addressed the most salient differences among models: the number of free parameters. As is generally well known, a model with many free parameters can provide a better fit to a data sample than a model with few parameters, even if the latter generated the data. AIC and SCH penalize a model more as the number of parameters increases. To be selected, the model with more parameters must overcome this penalty and provide a substantially better fit than a model with fewer parameters. That is, the superior fit obtained with the extra parameters must justify the necessity of those parameters in fully capturing the cognitive process. An equally salient but much less tangible dimension along which models also differ is in their functional form, which refers to the way in which the parameters are combined in a model. More sophisticated selection methods, such as GCV, MML, and SRM are sensitive to a model's functional form as well as to the model complexity (number of parameters). Conceptually, complexity refers to that characteristic of a model that makes it flexible and easily able to fit diverse patterns of data, usually by a small adjustment in one of its parameters.

In this study, we consider a simple domain where the x data are randomly selected from the interval [-1, 1], and the y values are derived from a univariate function, $y = t(x)$, corrupted with Gaussian noise having a zero mean. Least squares approximations of polynomials of degrees of up to 20 are derived from the data generated. These univariate polynomials of varying orders are then offered to the five model selection methods, and the performance of the preferred polynomials are evaluated by their predictive accuracy on test data, similarly generated. This work extends the recent studies by Wallace and Viswanathan [11], [12] by including an extensive empirical evaluation of five polynomial selection methods – FPE [2], SCH [4], GCV [5], MML and SRM (in unreported results, we have found that AIC [3] performs almost identically to FPE). We aim to analyze the behavior of these five methods with respect to variations in the number of training examples and the level of noise in the data. In summary, the primary contributions of this research work are twofold. First, we develop the MML principle for inductive inference in its application to model selection. Second, we present an extensive empirical evaluation of the MML method with other well-known approaches.

II. Background

Polynomial models are among the most frequently used empirical models for fitting functions. These models are popular for several reasons. Polynomial models have a simple form and well-understood properties. Polynomial models have moderate flexibility of shapes, and changes of location and scale in the raw data result in a polynomial model being mapped to a polynomial model. Polynomial models are computationally easier, and univariate functions are useful for modeling in their own right as they can serve as the basic building blocks for functions of higher dimensions.

Least squares is a mathematical optimization technique that attempts to find a "best fit" to a set of data by attempting to minimize the sum of the squares of the differences (called residuals) between the fitted function and the data. An implicit requirement for the least squares method to work is that errors in each measurement be randomly distributed (ideally they should come from Gaussian distribution). It is also well-known that least-squares fitting is a maximum likelihood estimation of the fitted parameters if the measurement errors are independent and normally distributed with constant standard deviation.

As suggested earlier in this paper, we consider the problem of the selection of a univariate polynomial regression function from competing models. In this section, we aim to summarize the two major approaches based on the MML and SRM principles. Let us assume that we have a finite number N of observations of a function $t(x)$ corrupted with additive noise ε,

$$y_i = t(x_i) + \varepsilon_i \quad \text{for} \quad i = 1, \ldots, N.$$

Our approximation of the target function is based on the training set of N observations, where the values, x_i, of the independent variable x are Independently and uniformly distributed in the interval [-1, 1], and the noise values, $\varepsilon_i = y_i - t(x_i)$, are independently and identically distributed by a Normal density with zero mean and unknown variance. The framework of the problem follows Cherkassky and others [13], and has been repeated using MML in Wallace [11] and Viswanathan and others [12]. The values, x_i, of the independent variable x are randomly selected from the uniform distribution on the interval [-1, 1].

The task then is to find some polynomial function, $f(x) = \hat{t}(x)$, of degree d that may be used to predict the value of $t(x)$

in the interval, $-1 \le x \le 1$. In our evaluation, we only consider polynomials $f(\cdot)$ of degrees up to 20, and for any given degree d, we select the polynomial $f(d, x)$ that minimizes the squared error *SE* on the training data,

$$SE(f(d,x)) = \sum_{i=1}^{N} (y_i - f(d,x_i))^2 \ . \quad (1)$$

The performance or *prediction risk* is the expected performance of the chosen polynomial for new (future) samples. This is measured by its squared prediction error (SPE), which is estimated using a simple Monte Carlo method:

$$SPE(f(d,x)) = \frac{1}{m} \sum_{i=1}^{m} (f(d,x_i) - t(x_i))^2 \ , \quad (2)$$

where $t(x)$ is the target function, m is max(N, 50), and the test data (i = 1 to m) are randomly selected from a uniform distribution in [-1,1].

1. Classical Methods for Model Selection

Among the classical methods compared in this paper, two general approaches can be observed. While GCV [5] is based on data re-sampling, FPE [2] and SCH [4] attempt to penalize model complexity. The use of FPE as opposed to the AIC [3] is justified since FPE is specially derived under the assumption that the distributions of the predictors used in learning and prediction are identical. Furthermore, FPE and AIC give almost the same inference for this class of problem [14], as borne out by some unreported results of our own. As described in Wallace [11] (replicating the problem framework from Cherkassy and others [13]), the selection process for these classical methods is to choose the polynomial that minimizes

$$g(p, N) * SE(f(d,x)),$$

where $p = (d + 1)/N$, and $g(\cdot, \cdot)$ is a function characteristic of the given method of inference. The function $g(\cdot, \cdot)$ is known as a penalty function since it inflates the training error (average residual sum of squares). The following characteristic penalty functions are derived from the classical approaches: FPE, $g(p, N) = (1 + p) / (1 - p)$ SCH, $g(p, N) = 1 + 0.5 \log(n) * p/(1 - p)$ GCV, $g(p, N) = 1/(1-p)^2$

2. VC Dimension and SRM

As defined earlier in section II, the *prediction risk* is the expected performance of an estimator on new samples. The Vapnik-Chervonenkis theory [9] provides non-asymptotic "guaranteed" bounds for the prediction risk of a model based on the concept of the VC-dimension [10]. Generally speaking, the VC-dimension [9], [10] is a measure of model complexity. For a given set of functions, the VC-dimension is the number of instances that can be "shattered" – that is, all possible subsets of the instances are partitioned from their complement subset by functions from this set. For example, in the binary classification case, the VC-dimension is the maximum number of instances m that can be separated into two classes in all possible 2^m ways by using functions from the hypothesis space. The VC-dimension for the set of polynomial functions of degree d can be shown [10] to be equal to $(d + 1)$.

The SRM principle [10], [15] is based on the well-known assumption that, in order to infer models with high generalization ability, we need to define a trade-off between the model complexity and goodness of fit to the data. Employing the VC-dimension as the measure of model complexity, the SRM principle attempts to achieve this trade-off and avoid *over-fitting*.

According to Vapnik [10], in order to choose the polynomial $f(d, x)$ of the best degree d, one can minimize the following function based on the SRM principle:

$$R(\underline{a}, d) \le \frac{\frac{1}{N} \Sigma (y_i - f(d,x_i))^2}{(1 - c\sqrt{\xi N})} \ , \quad (3)$$

where

$$\xi N = \frac{V(\log \frac{N}{V} + 1) - \log \eta}{N}$$

In the expression above, $R(\underline{a}, d)$ is the estimate of the prediction risk of a polynomial of degree d and coefficients $\underline{a}$. The numerator of the right-hand side of (3) is the average squared error, namely $(1/N) \sum_{i=1}^{N} (y_i - f(d,x_i))^2$, achieved by a polynomial of degree d and set of coefficient $\underline{a} = < a_0, \ldots, a_d >$ on N training examples, and the denominator is $1 - c\sqrt{\xi N}$. Term ξN is the error term where N is the number of examples, c is a constant that reflects the tails of the training error distribution, and V is the VC dimension for $f(d, x)$. This approach also takes into account the confidence interval for the prediction. Specifically, it provides an upper bound for the estimate, $R(\underline{a}, d)$, of the prediction risk. The inequality in (3) then holds with probability (1- η), where η represents a confidence interval for the prediction [10]. In our empirical evaluation we have used $V = (d + 1)$, since the VC dimension for a polynomial of degree d is $(d + 1)$ and, as suggested in [15], we have elected to use c = 1. In this evaluation, we employ the error term ξN as described in (3) with $\eta = 0.125$ (this value of η was derived from Vapnik's book [10]). An older error term was investigated in empirical comparisons done by Wallace [11] where the confidence interval η was implemented as a function of the sample size ($\eta = \frac{1}{\sqrt{N}}$). The recent version [15]

from (3) employs a fixed user-defined value. This application of a fixed confidence interval improves the predictive performance of the older approach [10]-[12] at least in specific cases.

3. MML Principle

Minimal Length Encoding techniques like the MML [6]- [8] and Minimum Description Length (MDL) [16]-[18] principles have been popular due to their successful application in model selection. MML is an invariant Bayesian principle [7], [8], [19] based on the information-theoretic assertion that the best model for a given set of data is one that allows the shortest joint encoding of the model and the data, given the model. MML seeks to minimize the length of a "message" that encodes the data (in this problem, the training y-values) by first stating a probabilistic model for the data, and then encoding the data using a code that would be optimal were the stated model true. MML has been shown to provide robust and highly competitive model selection in comparison to various classical model selection criteria (including AIC) within the polynomial degree selection framework [20], although we improve this even further in section II.3.1 by our use of an orthonormal basis. A comparison with the MDL [18] principle of Rissanen, with A.N. Kolmogorov's notion of complexity [21], and with Chaitin's notion of algorithmic information theory [22], is given in Wallace and Dowe [8]. Recent MML work includes single factor analysis [23], multiple factor analysis [24], von Mises circular distributions [25], [26], spherical Fisher distributions [27], mixture modeling [26], [28], [29], segmentation of a binary sequence [30], and general surveys [7], [8].

In our case, since the model is assumed to be a polynomial function with Gaussian noise, the model description need only specify the degree of the polynomial, the coefficient of the polynomial, and the estimated variance (or, equivalently, SD) of the noise. In the general case, the best MML polynomial for any degree is the one that has the shortest two-part message length, of which the first part describes the polynomial in terms of its degree d, coefficient $\underline{a}$, and estimated variance v, while the second part describes the data (via the ε_i) using the given polynomial. An important point to note is that our encoding system uses the degree of the polynomial rather than the number of non-zero coefficients. One reason for this is that the number of non-zero coefficients will depend upon whether we choose basis $1, x, x^2, x^3, \ldots$, the orthonormal basis for integration over the region [-1, 1] or another basis. However, the degree of the polynomial will remain the same regardless of this choice of basis.

In the current experiment, the coefficients are estimated using the maximum likelihood technique, since the difference from MML estimation is small for this problem, and all the other methods advocate using the maximum likelihood estimate. For examples of problems where MML and maximum likelihood estimation are substantially different, see [23] and [24]. The following sections provide details of the actual MML encoding scheme.

A. Encoding the Model with Prior Belief

MML is a Bayesian principle and thus requires the formal specification of prior beliefs. We start by considering the degree of our polynomial models. All degrees from 0 to 20 are considered equally likely a priori, so each degree is coded with a code word of length log(21) *nits*, or natural bits. The coding of the model degree has no influence on the choice of model as all degrees have the same code length.

In coding estimates of the noise variance v, and the polynomial coefficients $\underline{a}$, some scale of magnitude may be assumed. Here, we use the second order sample moment of the given y-values to determine such a scale by defining

$$\vartheta = \frac{1}{N}\sum_{i=1}^{N}(y_i)^2 \,. \tag{4}$$

In encoding a polynomial model of degree d, we suppose that the noise and each of the $(d + 1)$ degrees of freedom of the polynomial may be expected a priori to contribute equally (in very rough terms) to the observed variance ϑ of the y-values. Defining

$$u = \sqrt{\left(\frac{\vartheta}{d+2}\right)} = \sqrt{\frac{\sum_{i=1}^{N} y_i^2}{N(d+2)}}\,, \tag{5}$$

we assume the standard deviation s of the 'noise' v, where $s = \sqrt{v}$, has a negative exponential prior density with mean u, and that each of the coefficients $(a_0, \ldots, a_j, \ldots, a_d)$ of the polynomial has independently an $N(0, u^2)$ prior density. If the coefficients were the usual coefficients of the successive powers of x, the latter assumption would be unreasonable and highly informative. Instead, we represent a d-degree polynomial as

$$f(d, x) = \sum_{j=0}^{d}(a_j Q_j(x))\,, \tag{6}$$

where the set $Q_j(\cdot) : j = 0, \ldots, d$ is a set of polynomials, $Q_j(\cdot)$ being of degree j, which are orthonormal under integration on the interval [-1, 1]. The orthonormal polynomials represent effectively independent modes of contributing to the variance of $f(d, \cdot)$, and therefore it seems reasonable to assume independent prior densities for the coefficient $\{a_j : j = 0, \ldots, d\}$. With these assumptions, the overall prior density for the unknown parameters $\{s, \{a_j\}\}$ is

$$h(s,\underline{a}) = (1/u)^{(-s/u)} * \prod_{j=0}^{d} \left(\frac{1}{\sqrt{2\pi}u}\right) e^{(-a_j^2/(2u^2))} . \quad (7)$$

The amount of information needed to encode the estimates s and $\underline{a}$ depends on the precision to which these are stated in the "message." Specifying the estimates with high precision leads to an increase in the model part of the message length while lower precision leads to lengthening of the data part [7], [8], [25]. The optimum precision, as described in [7], is inversely proportional to the square root of the *Fisher information* $F(s, \underline{a})$, which in this case is given as in [11] and [12] by

$$F(s,\underline{a}) = 2\left(\frac{N}{s^2}\right)^{(d+2)} |M| , \quad (8)$$

where M is the co-variance matrix of the orthonormal polynomials evaluated at the given x-values:

$$M[j,k] = (1/N)\sum_{i=1}^{N} Q_j(x_i)Q_k(x_i) . \quad (9)$$

B. Encoding the Data

Once the model polynomial $f(d, .)$ and the noise standard deviation s have been stated, the given y-values can be coded simply by their differences from $f(d, .)$, these differences being coded as random values from the Normal density $N(0, v)$. The message length required for this is then given as in [7], and [11] by

$$DataMessLen = \left(\frac{N}{2}\right)\log(2\pi v) + (1/2v) * SE(f(d,\cdot)) . \quad (10)$$

C. The Total Message Length

The total message length is approximately given as in [7] and [8] by

$$\begin{aligned} MessLen = &-\log h(s,\underline{a}) + 0.5\log F(s,\underline{a}) + DataMessLen \\ &-((d+2)/2)\log(2\pi) + 0.5\log((d+2)\pi), \end{aligned} \quad (11)$$

where the last two terms arise from the geometry of an optimum quantizing lattice in $(d+2)$-space. The noise variance $v = s^2$ is estimated as

$$\hat{v}_{MML} = \left(\frac{1}{(N-d-1)}\right)\sum_{i=1}^{N}(f(d,x_i)-y_i)^2 , \quad (12)$$

and the coefficient $\underline{a}$ are estimated by conventional least-squares fit. These estimates do not *exactly* minimize the message length *MessLen*, but as in section II.3, for this problem the difference is small. The MML method selects that degree d which minimizes *MessLen* as calculated in (11). A recent theory [31] suggests that the MML estimator will closely approximate the estimator minimizing the expected Kullback-Leibler distance.

III. Experimental Evaluation

The following discussion details the experimental procedure [12] employed in this empirical evaluation. The experimental framework is comprehensive as it considers a variety of scenarios in the generation of data and the evaluation of models. The model selection is performed on varied families of target function used in the data generation. For example, this in essence shows us in an interesting case how a polynomial function preferred by a specific model selection method tries to model data generated by a discontinuous function. Furthermore, each experiment undergoes a form of 1000-fold cross-validation, and therefore the results being averaged are stable with low variance.

For each experiment, a target function is selected in the required interval [-1,1]. The noise is defined in terms of the signal-to-noise ratio (SNR), where SNR is defined as the second moment of the target function, $t(x)$, about zero, divided by the noise variance v. The number of training data points N and the number of evaluations (training and test runs) are specified.

An experiment consists of (averaging over) 10,000 evaluations. In each evaluation or case, N training examples and $m = \max(N, 50)$ test examples are generated. For each case, all least-squares polynomial approximations of degrees up to 20 are found by standard techniques and the training error, $SE(d)$, computed for each of the 21 polynomials. These training error values are then given to each of the model selection methods being compared, and each method selects its preferred choice among the polynomials.

The prediction risk for a method in the current case is then the average squared prediction error (SPE) achieved by the polynomial chosen by the method on the test data. Note again that all selection methods must choose from the same set of 21 polynomials, and their choices are evaluated on the same set of test data. Thus, if two selection methods choose the same degree in some case, they are choosing the same polynomial and will incur the same SPE for this case.

After the 10,000 cases of an experiment have been completed, we obtain for each selection method its SPE, averaged over all cases (10,000), and the standard deviation of its SPE.

The five selection methods – namely, MML, SRM, FPE, SCH and GCV – were evaluated on six target functions $t(x)$ in the interval [-1,1]. The methods were tested under different scenarios. First, the noise level was kept constant with varying numbers of training points, and then the noise levels were varied with a fixed number of training points. For each method evaluated under a particular scenario, comparisons were made on the basis of SPE

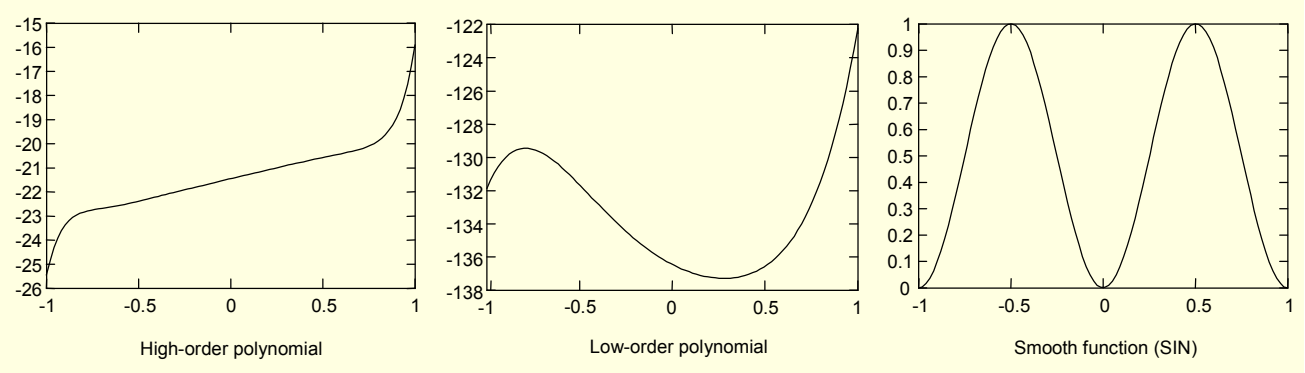


Fig. 1. Plots of smooth functions.

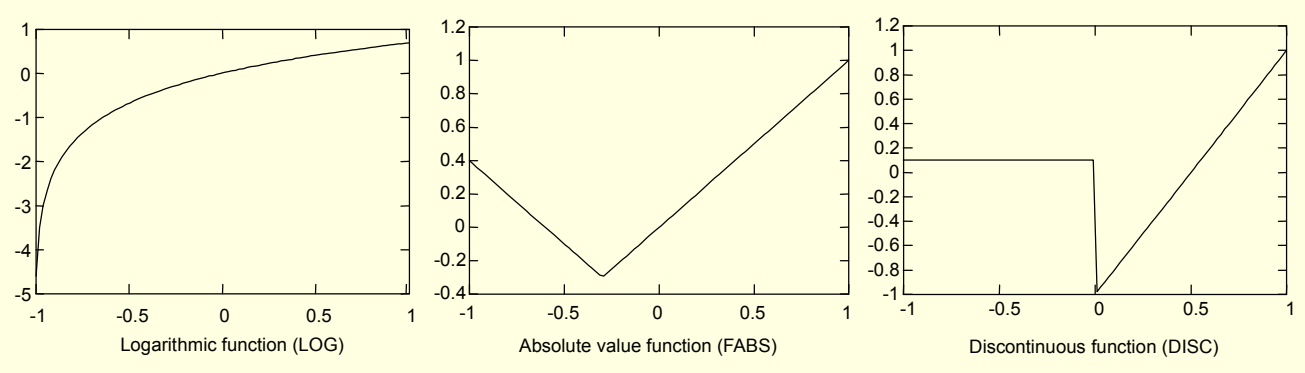


Fig. 2. Plots of other target functions.

and standard deviation of the SPE. As an artificial yardstick, we also include a method called "BEST". The BEST polynomial is obtained by selecting from the 21 candidate least-square fit polynomials (from section I) the polynomial that would give the smallest SPE as calculated on the test data. Thus, BEST shows the best performance that could possibly be obtained in principle; but, these results are of course not realizable, as they are based on knowledge of the noise-free values of the target function at the test x_i's.

IV. Analysis of Results

In the experimental evaluation, we consider two sets of target functions. In the former, the target functions belong to the family of polynomials of maximum degrees of 20, while the latter set consists of non-polynomials. Figures 3 through 8 include plots of results from experiments with six target functions. Each figure presents two scenarios. In the first case, the sample size is increased and the standard deviation of the noise is kept constant; vice versa in the second case. Some of the prediction methods can be seen to have a SPE typically of the order of 100 times that of MML. Due to this relatively poor performance of those estimation techniques, the SPE is plotted on a logarithmic scale. As mentioned in section III, we also plot the standard deviations from some of these simulation runs with a sample size fixed at $N = 40$ and noise varying. Of the six target functions considered, which are described in section IV.1, when $N = 40$ we found the ratio of FPE squared error to MML squared error to be the smallest for the smooth SIN function, closely followed by the discontinuous function. We found the ratio to be largest for the low order polynomial. In Fig. 9, we plot the results comparing the standard deviation of SPE of all methods on the SIN function and the low-order polynomial.

In general, from the examples included in this paper and other tests, it can be observed that MML and SRM give lower errors than other methods in most cases. The results clearly show that none of the methods, FPE, SCH, or GCV, is competitive with SRM or MML, except under conditions of low noise and large N, when all methods give similar results. An interesting observation is that the SRM method is based on a theory that does not assume that the target function belongs to the model family from which the approximation is drawn (in this case, the polynomials). MML, however, is in part motivated by a theory that does make this assumption. It is curious that the only target/test conditions under which SRM performs comparably with MML are ones where this assumption is largely valid. Furthermore, an examination of the percentiles of the error distribution suggests that the MML and SRM-based methods usually have similar median errors. The high

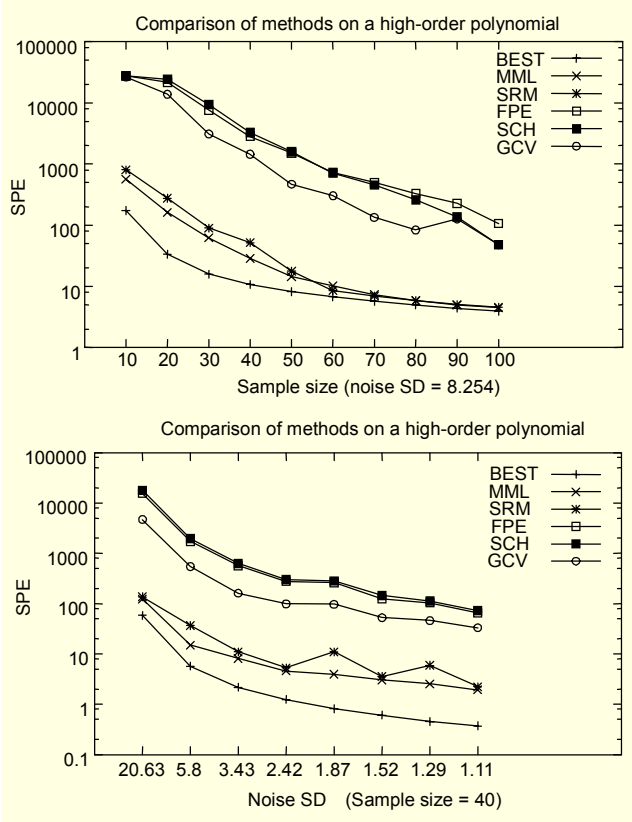


Fig. 3. Comparing SPE on a high-order polynomial.

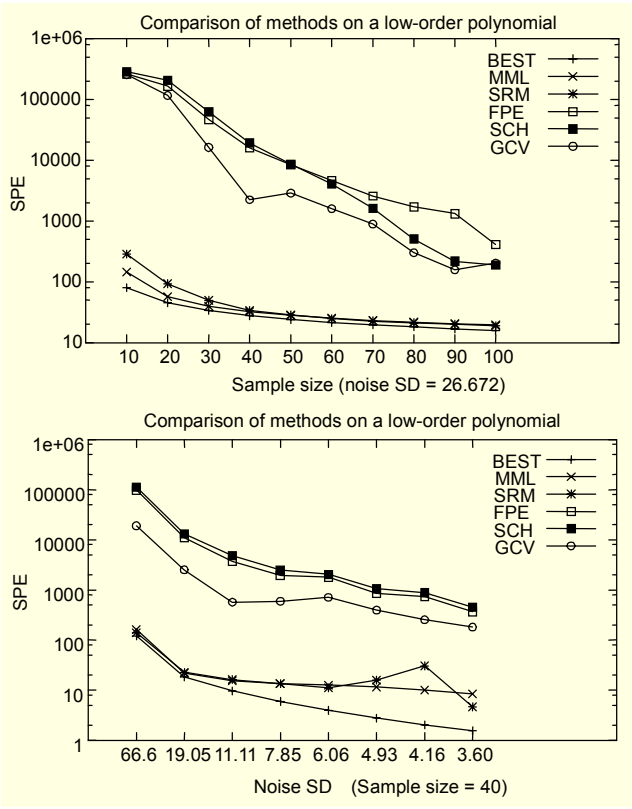


Fig. 4. Comparing SPE on a low-order polynomial.

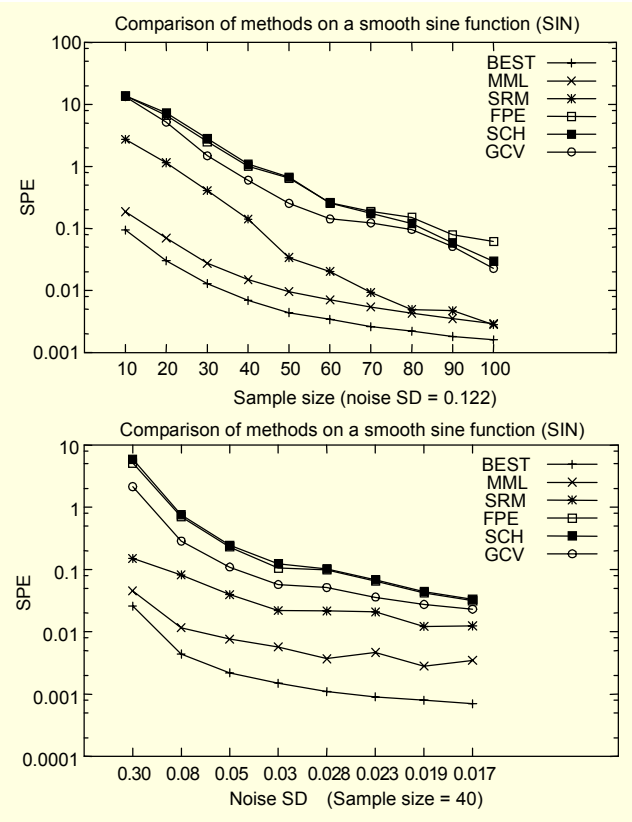


Fig. 5. Comparing SPE on a smooth function.

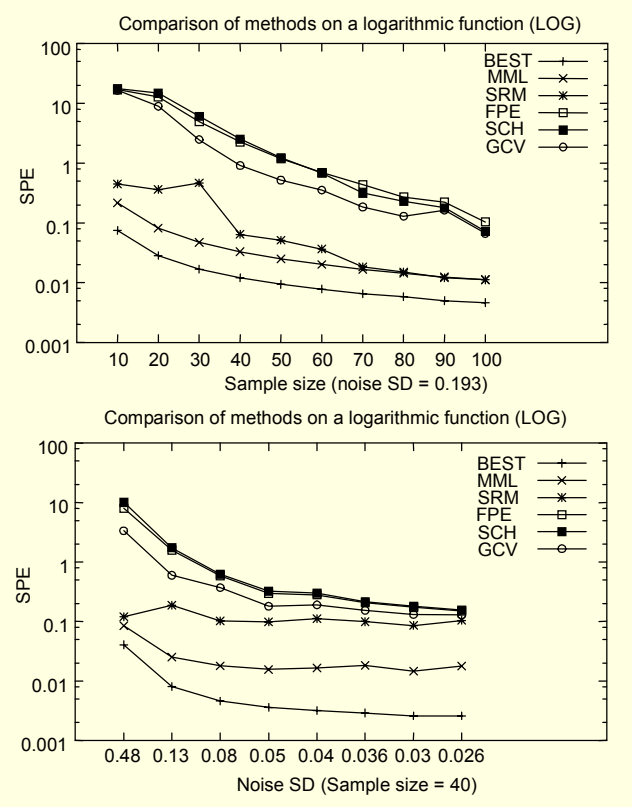


Fig. 6. Comparing SPE on a logarithmic function.

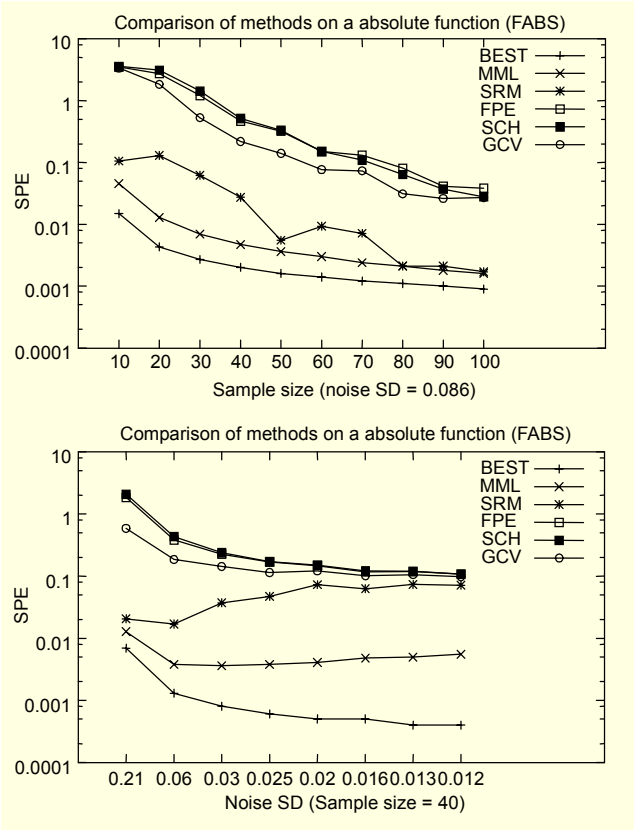


Fig. 7. Comparing SPE on a absolute function.

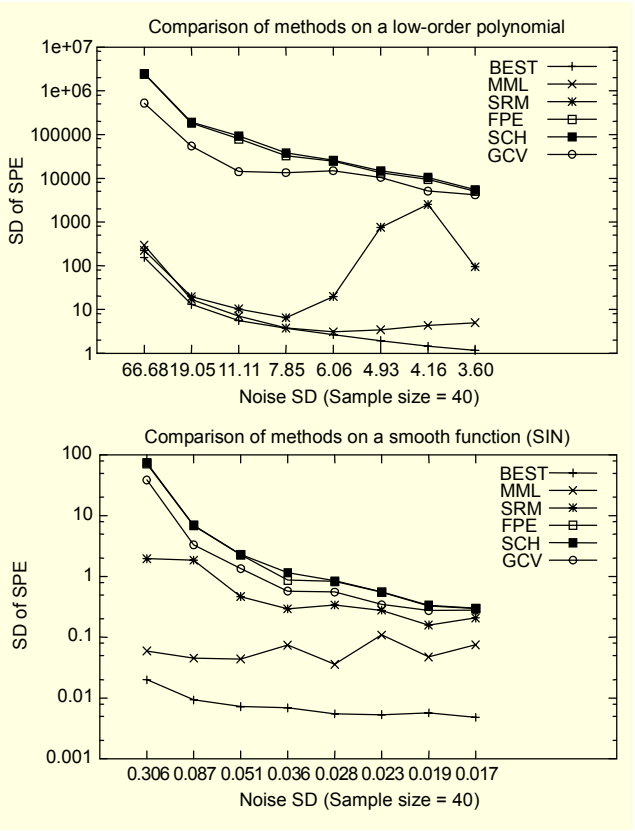


Fig. 9. Standard deviation of SPE

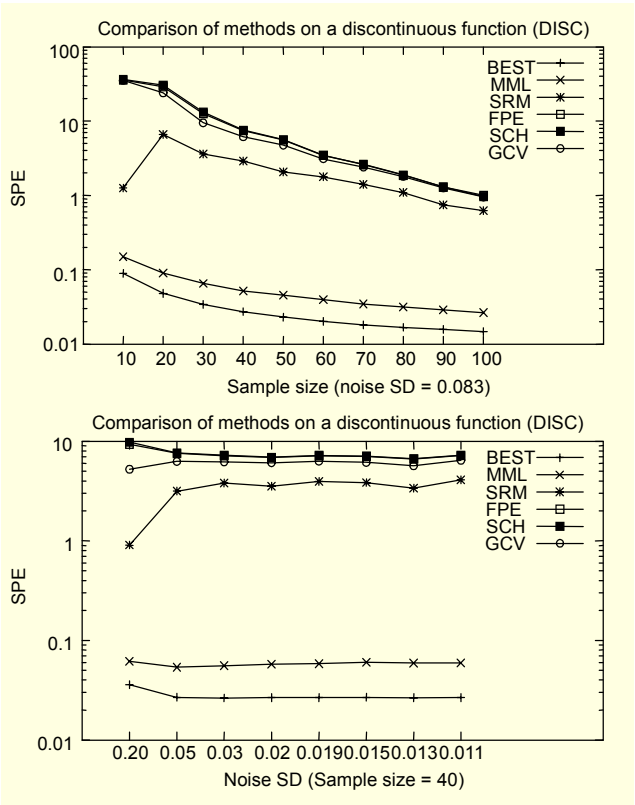


Fig. 8. Comparing SPE on a discontinuous function.

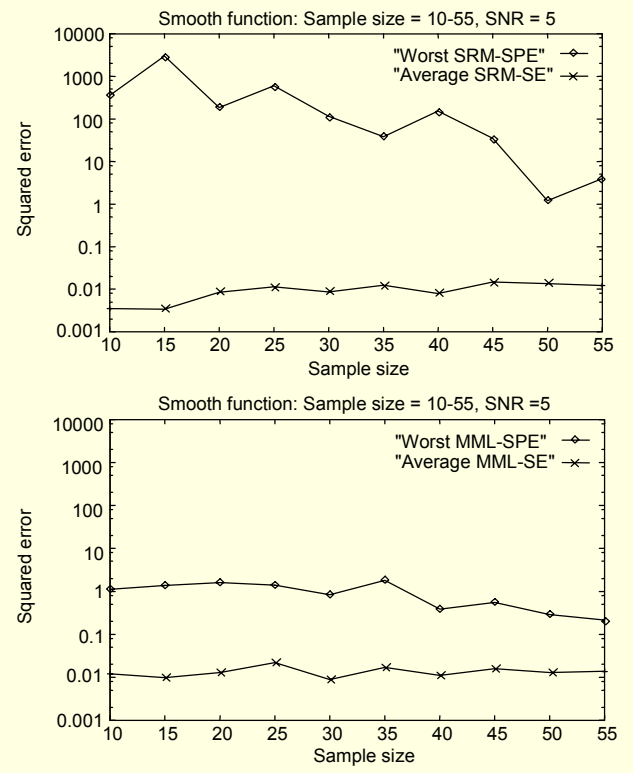


Fig. 10. A comparison of the SPE (test data) and SE (training data) of the worst models selected by SRM and MML.

SPE of the SRM method seems to be the result of occasional very large errors using the selected model. Evidence of this is given in Fig. 10.

1. Target Functions

In evaluating the five polynomial selection methods, we consider the following two polynomial and four nonpolynomial target functions. The following polynomial target functions are presented in Fig. 1.

A higher-order polynomial:

$$y = 0.623x^{18} - 0.72x^{15} - 0.801x^{14} + 9.4x^{11} - 5.72x^{9} + 1.873x^{6} - 0.923x^{4} + 1.826x - 21.45$$

A lower-order polynomial:

$$y = 9.72x^{5} + 0.801x^{3} + 9.4x^{2} - 5.72x - 136.45$$

Non-polynomial target functions:

It is interesting to observe the performance of the different methods when the target functions are not polynomials. The task here is challenging as we seek the best polynomial approximation. The following target functions were utilized in the comparison, with $-1 \le x \le 1$, and are plotted in Figs. 1 and 2.

SIN: $y = (\sin(\pi (x + 1)))^2$;
LOG: $y = \log(x + 1.01)$;
FABS: $y = | x + 0.3 | - 0.3$; and,
DISC: if $(x < 0.0)$ $y = 0.1$; else $y = 2x - 1.0$.

2. The "Guaranteed" SRM Bound: Is It Loose ?

In this section, we raise some issues about the application of the SRM principle to the current problem framework. The model selection methods based on the SRM principle define a "guaranteed" bound on the prediction risk of a model, where the prediction risk, as defined earlier, is the expected SRM achieved by the selected model. The SRM bound states that the prediction risk for a chosen model will not be exceeded with a probability of at least 1-η. In our case, (3) is the upper bound on the prediction risk and it is "guaranteed" [10] with a probability of $1 - \eta = 1 - 0.125 = 0.875$, as shown in section II.2. Based on our empirical analysis, we find that this bound is not exceeded, with a probability greater than 1-η, for any fixed order of polynomial and any sample size, noise ratio, or target function. In fact, the probability of bound violations is usually much less than η.

However, the polynomial model selected by the SRM method from the 21 polynomial families often exceeds the bound with a probability greater than η. For example, the bound is violated in 96.5% of the cases of an experiment on the smooth target function with a sample size and SNR ratio, as shown in section III, of 10. This is a typical result for a model selection with small sample sizes. Thus, the SRM bound seems to be plausible only when applied to selecting models from the same family. Once this assumption is relaxed, the SRM measure cannot be guaranteed to provide an upper-bound on the prediction risk. In this context, another important observation is that the VC theory does not provide any information on the magnitude of the prediction error when this upper-bound is exceeded.

Finally, we find that, in its application to the current model selection framework, the SRM principle is not paying any attention to the variance of the selected model. On an examination of the cases where the performance of the SRM-based models is worse than usual, we find that the estimated variance of the approximating polynomial is orders of magnitude larger than the second moment of the training samples. Thus, the SRM principle is accepting some clearly unreasonable models. As an example, the plots in Fig. 10 present a comparison of the variance of the worst model (in terms of SPE) selected by the SRM and MML methods over the 10,000 runs with their average training errors. The model selection methods were evaluated on the SIN function from section IV.1 with a fixed signal-to-noise ratio (SNR) and variable sample size.

3. Comparison with Bayesian Piecewise Polynomial Fitting

Denison and others [32] present a Bayesian approach to estimating a variety of curves using piecewise polynomials. Although their approach is far more flexible due to the generality of the model class and the resulting variety of functions that their system can fit, we offer a limited comparison with our MML-based orthonormal polynomial approach. The comparisons are not equivalent since in our approach the SPE is computed from test data, while Denison and others take a posterior sampling approach.

In Fig. 11, we compare MML with the piecewise polynomial fitting (PP) on the low order polynomial. The samplesize is $N = 200$, the SNR is 150, as shown in section III, and the standard deviation of the noise is $s = 0.317$. The plot presents the SPE

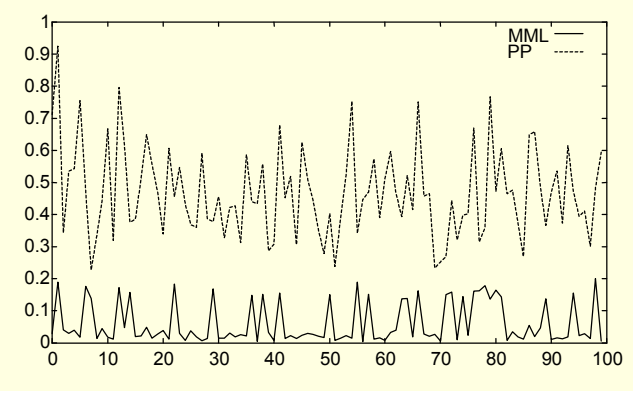


Fig. 11. Comparing MML and piecewise polynomial fitting on a low order polynomial. The X axis represents the run count (up to 100), while on the Y axis we have the SPE.

achieved by both methods for each of 100 runs of the experiment. Figure 12 presents a similar scenario for the high-order polynomial. The sample size is 200, the SNR is 100, and the standard deviation of the noise $s = 0.724$. The MML fit was clearly better than the PP fit for these two polynomial cases, especially for the high-order polynomial, where the SPE was more than a hundredfold worse for PP than for MML.

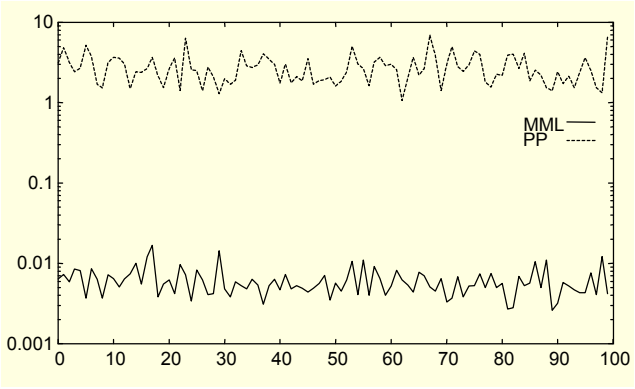


Fig. 12. Comparing MML and piecewise polynomial fitting on a high order polynomial. The X axis represents the run count (up to 100), while on the Y axis we have the SPE.

V. Conclusions

The primary aim of this research endeavour is to demonstrate the plausible performance of the Minimum Message Length principle in comparison to other methods. Based on our empirical evaluation, we find that the MML approach in general provides more robust and accurate model selection than the other methods in the current problem framework, especially FPE (and AIC), SCH, and GCV. Finally, we have provided some analysis of the different model selection methods under varying scenarios. It is essential to note that the comparison with the polynomial splines (Denison and others) is not exhaustive in any sense. The polynomial spline system offers superior performance in comparison to MML for a number of complex target functions.

VI. Future Work

SRM and MML have been shown to be plausible inductive principles. The support vector machine (SVM) is a universal constructive learning procedure based on the statistical learning theory of Vapnik [9], [10]. The term "universal" implies that the SVM can be used to learn a variety of representations including polynomial estimators, radial basis functions, decision trees, neural networks, and splines [33], [34]. MML is also universal in this sense [8]. One of our research efforts in progress is to extend the current univariate regression problem presented here to a multivariate design, thus enabling a direct comparison of MML with SVMs and classical multivariate approaches [35], [36]. A possibly implicit suggestion from an anonymous reviewer is the extension of our MML univariate polynomial model here to an MML piecewise model. Wallace had been pursuing work in progress to define the Vapnik-Chervonenkis (VC) theory as a special case of the MML principle.

Acknowledgements

We would like to express our gratitude to Professor Chris Wallace, Assoc. Professor David Dowe, and Assoc. Professor Kevin Korb for their useful discussions and the encouragement to submit this empirical work. We would also like to thank Dr. Guoqi Qian, Dr. Rohan Baxter, Dean McKenzie and anonymous reviewers for useful comments on the paper. The evaluation of the probability on the SRM bound in section IV.2 was initially suggested by Dr. Kevin Korb, School of Computer Science, Monash University.

References

[1] Anouk de Brauwere, Fjo De Ridder, Rik Pintelon, Marc Elskens, Johan Schoukens, and Willy Baeyens, Model Selection through a Statistical Analysis of the Minimum of a Weighted Least Squares Cost Function, *Chemometrics and Intelligent Laboratory Systems*, vol. 76, no. 2, 2005, pp. 163-173.

[2] H. Akaike, Fitting Autoregressive Models for Prediction, *Annals of the Institute of Statistical Mathematics*, vol. 21, 1969, pp. 243-247.

[3] H. Akaike, Statistical Predictor Information, *Annals of the Institute of Statistical Mathematics*, vol. 22, 1970, pp. 203-217.

[4] G. Schwarz, Estimating the Dimension of a Model, *Ann. Stat.*, vol. 6, 1978, pp. 461-464.

[5] P. Craven and G. Wahba, Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Crossvalidation, *Numerical Math.*, vol. 31, 1979, pp. 377-403.

[6] C.S. Wallace and D.M. Boulton, An Information Measure for Classification, *Computer Journal*, vol. 11, no. 2, 1968, pp. 195-209.

[7] C.S. Wallace and P.R. Freeman, Estimation and Inference by Compact Coding, *J. R. Statist. Soc B*, vol. 49, no. 3, 1987, pp. 240-265.

[8] C.S. Wallace and D.L. Dowe, Minimum Message Length and Kolmogorov Complexity, *Computer J.*, vol. 42, no. 4, 1999, pp. 270-283.

[9] V.N. Vapnik, *Estimation of Dependencies Based on Empirical Data*, Springer-Verlag, New York, 1982.

[10] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

[11] C.S. Wallace, On the Selection of the Order of a Polynomial Model,

Technical Report, Royal Holloway College, London, 1997.

[12] M. Viswanathan and C. Wallace, A Note on the Comparison of Polynomial Selection Methods, *Proc. 7th Int. Workshop on Artif. Intell. and Stats.*, Morgan Kauffman, Jan. 1999, pp. 169-177.

[13] V. Cherkassky and M. Mulier. *Learning from Data: Concepts, Theory, and Method*, chapter 4, Wiley and Sons, 1998, pp. 119-127.

[14] Y. Sakamoto et al., *Akaike Information Criterion Statistics*, KTK Scientific Publishers, 1986, pp. 191-194.

[15] V.N. Vapnik, *Computational Learning and Probabilistic Reasoning*, chapter Structure of Statistical Learning Theory, Wiley and Sons, 1996.

[16] J. Rissanen, Modeling by Shortest Data Description, *Automatica*, vol. 14, 1978, pp. 465-471.

[17] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, N.J., 1989.

[18] J. Rissanen, Hypothesis Selection and Testing by the MDL Principle, *The Computer J.*, vol. 42, no. 4, 1999, pp. 260- 269.

[19] C.S. Wallace and D.M. Boulton, An Invariant Bayes Method for Point Estimation, *Classification Society Bulletin*, vol. 3, no. 3, 1975, pp.11-34.

[20] R.A. Baxter and D.L. Dowe, Model Selection in Linear Regression Using the MML Criterion. In J.A. Storer and M. Cohn, editors, *Proc. 4'th IEEE Data Compression Conf.*, Snowbird, Utah, Mar. 1994, p. 498. IEEE Computer Society Press, Los Alamitos, CA. Also TR 276 (1996), Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia.

[21] A.N. Kolmogorov, Three Approaches to the Quantitative Definition of Information, *Problems of Information Transmission*, vol. 1, 1965, pp. 4-7.

[22] G.J. Chaitin, On the Length of Programs for Computing Finite Sequences, *J.A.C.M.*, vol. 13, 1966, pp. 547-549.

[23] C.S. Wallace and P.R. Freeman, Single Factor Analysis by MML Estimation., *J. of the Royal Statistical Society (Series B)*, vol. 54, 1992, pp. 195-209.

[24] C.S. Wallace, Multiple Factor Analysis by MML Estimation, *Technical Report 95/218*, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1995. Accepted, to appear in J. Multiv. Analysis.

[25] C.S. Wallace and D.L. Dowe, MML Estimation of the von Mises Concentration Parameter, *Tech Rept TR 93/193*, Dept. of Computer Science, Monash Univ., Clayton, Victoria 3168, Australia, 1993. prov. accepted, Aust. and N.Z. J. Stat.

[26] C.S. Wallace and D.L. Dowe, MMLclustering of Multistate, Poisson, von Mises Circular and Gaussian Distributions. *Statistics and Computing*, vol. 10, no. 1, 2000, pp. 73-83.

[27] D.L. Dowe, J.J. Oliver, and C.S. Wallace, MML Estimation of the Parameters of the Spherical Fisher Distribution, In A. Sharma et al., editor, *Proc. 7th Conf. Algorithmic Learning Theory (ALT'96), LNAI 1160*, Sydney, Australia, Oct. 1996, pp. 213-227.

[28] C.S. Wallace, Intrinsic Classification of Spatially- Correlated Data, *Computer J.*, vol. 41, no. 8, 1998, pp. 602-611.

[29] T. Edgoose and L. Allison, MML Markov Classification of Sequential Data, *Statistics and Computing*, vol. 9, 1999, pp. 269-278.

[30] M. Viswanathan, C.S. Wallace, D.L. Dowe, and K. Korb, Finding Cutpoints in Noisy Binary Sequences - a Revised Empirical Evaluation, *Proc. 12th Australian Joint Conf. on Artificial Intelligence*, Sydney, Australia, Dec. 1999, pp. 405-416.

[31] D.L. Dowe, R.A. Baxter, J.J. Oliver, and C.S. Wallace, Point Estimation Using the Kullback-Leibler Loss Function and MML, *Proc. 2nd Pacific Asian Conf. on Knowledge Discovery and Data Mining (PAKDD'98)*, Springer Verlag, Melbourne, Australia, Apr. 1998, pp. 87-95.

[32] D.G.T. Denison, B.K. Mallick, and A.F.M. Smith, Automatic Bayesian Curve Fitting, *J. Roy. Statist. Soc. Series B*, vol. 60, 1998, pp. 333-350.

[33] Corinna Cortes and Vladimir Vapnik, Support Vector Networks, *Machine Learning*, 20:273, 1995.

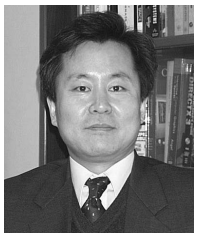
[34] V. Vapnik, S. Golowich, and A. Smola, Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing, In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, MIT Press, Cambridge, MA, 1997, pp. 281-287.

[35] Lee K. and Park H., A New Similarity Measure Based on Intraclass Statistics for Biometric Systems, *ETRI J.*, vol. 25, no. 5, 2003, pp. 401-406.

[36] Rhee M. kil, Function Approximation Based on a Network with Kernel Functions of Bounds and Locality : an Approach of Non-parametric Estimation, *ETRI J.*, vol. 15, no. 2, 1993, pp. 35-51.

Murlikrishna Viswanathan received the BS degree in computer science and information systems from Deakin University, Australia, in 1996 followed by a First Class Honours degree in 1997. In 2002 he completed the PhD in computer science from Monash University, Australia. Since 2004, he has been an IITA Distinguished Professor with the College of Software at Kyungwon University, Korea. He is currently leading projects at Kyungwon University in machine learning, data fusion, image analysis and Bayesian model selection techniques. From 2001 to the middle of 2004, he was a Faculty Member with the Department of Computer Science, University of Melbourne, Australia.

Young-kyu Yang received the BS and MS degrees from Seoul National University in 1972, 1974 respectively. He received the PhD degree in satellite image processing from Texas A&M University in 1985. He joined Korea Institute of Science and Technology (KIST)/Systems Engineering Research Institute (SERI) in 1973 and transferred to ETRI in 1996 where he served as the Dirctor of Artificial Intelligence, Image Processing, Supercomputer Center. Most recently, he worked as the Director of the Spatial Information Technology Center of Computer and Software Research Laboratory. He joined Kyungwon University in 2003 and now is serving as the Dean of the Graduate School of Software. He served as the President of the Korea Society of Remote Sensing (2001-2003) and the President of Korean Geo-Spation Information Society (2003-2005). His research interests include spatial information processing (GIS, RS, LBS), image processing, and telmatics systems.



Taeg Keun Whangbo received the BS from Korea University in 1983, MS in computer science from CUNY, and PhD in computer Science from S.I.T in 1988 and 1995. He had been with Q-Systems as a member of Senior Technical Staff from 1988 to 1993. He has also served as Researcher at Samsung Advanced Institute of Technology from 1995 to 1997. In 1997, he joined, and is currently an Associate Professor in the Department of Internet Media at Kyungwon University. He is also Dean of Computer Information Center at Kyungwon University. His current research interests cover machine learning, computer vision, and computer graphics.