Efficient Linear Regression by Minimum Message Length

Enes Makalic and Daniel F. Schmidt

Abstract

This paper presents an efficient and general solution to the linear regression problem using the Minimum Message Length (MML) principle. Inference in an MML framework involves optimising a two-part costing function that describes the trade-off between model complexity and model capability. The MML criterion is integrated into the orthogonal least squares algorithm (MML-OLS) to improve both speed and numerical stability. This allows for the message length to be iteratively updated with the selection of each new regressor, and for potentially problematic regressors to be rejected. The MML-OLS algorithm is subsequently applied to function approximation with univariate polynomials. Empirical results demonstrate superior performance in terms of mean squared prediction error in comparison to several well-known benchmark criteria.

I. Introduction

Linear regression is the process of modelling data in terms of linear relationships between a set of independent and dependent variables. An important issue in multi-variate linear modelling is to determine which of the independent variables to include in the model, and which are superfluous and contribute little to the power of the explanation. This paper formulates the regressor selection problem in a Minimum Message Length (MML) [1] framework which offers a sound statistical basis for automatically discriminating between competing models. In particular, we combine the MML criterion with the well-known Orthogonal Least Squares (OLS) [2] algorithm to produce a search (MML-OLS) that is both numerically stable and computationally efficient. To evaluate the performance of the proposed algorithm

Enes Makalic and Daniel Schmidt are with Monash University
Clayton School of Information Technology
Clayton Campus Victoria 3800, Australia.
Telephone:+61 3 9905 5056, Fax: +61 3 9905 5146
Email: {Enes.Makalic, Daniel.Schmidt}@csse.monash.edu.au

we apply it to the problem of univariate polynomial function approximation in the face of noisy data. The ability of the MML-OLS algorithm to infer sound models is directly tested against several other well established statistical inference criteria. Experiments performed on approximating both true polynomial functions, and non-polynomial functions demonstrate that the MML criterion performs at least as well, and in most cases better than the competition.

This paper is organised as follows: Section II presents a review of previous work in the area, Section III introduces the Minimum Message Length criterion, and Sections IV, V and VI cover the formulation of the linear regression problem in the MML framework and detail the algorithm and its application to polynomial regression. The testing procedure and subsequent analysis of results is covered in Section VII, with Section VIII presenting the concluding remarks. Additionally, the Appendix has details on recreating the experiments.

II. Literature Review

The problem of selecting which regressors are significant to the linear explanation of data remains an open issue. There has been a great deal of research on this topic over the past few decades, which has produced a range of different solutions that have performed with varying levels of success. Amongst the most influential of these has been the Orthogonal Least Squares algorithm [2] which attempts to analyse the effect of each regressor upon the overall explanation independently of all other regressors. Even though the OLS algorithm manages to decouple the effect of each regressor on the modelling error, the issue of determining when to stop including new regressors is not clearly addressed and is based upon an arbitrary selection of the stopping conditions.

The task of determining when a model has become too 'complex' has also been widely studied. As early as 1968, Wallace and Boulton proposed a model selection criterion [3] based on information theory [4] and Bayesian statistics [5]. This work presented a two-part costing function, which optimised a trade-off between the complexity of the model and the ability of the model to describe the data. Subsequent generalisation of this information theory based criterion led to the development of the Minimum Message Length framework which offers a general solution to the problem of parameter estimation and model discrimination. The MML criterion possesses many desirable theoretical properties including data and model invariance, efficiency and unbiasedness. The concept of trade-offs between model complexity and model capability has also motivated the development of a range of other so called 'information criteria', including the well known Akaike Information Criterion (AIC) [6], the Bayesian Information Criterion (BIC) [7] and the Minimum Description Length (MDL) criterion [8], [9]. The poor performance of AIC

in the face of small sample sizes led to the development of a modified form known as the corrected AIC criterion (AICc) [10]. More recently, Djurić [11] introduced the asymptotic *maximum a posteriori* (AMAP) criterion. In contrast to the AIC and BIC criteria, the form of the AMAP penalty term depends on the structure and type of model under consideration.

The MML criterion has been previously applied to the problem of polynomial inference by Viswanathan and Wallace [12] and Fitzgibbon et al [13] with promising results. However, the problem of optimal regressor selection was replaced by one of inferring the required order of the polynomial. Given a particular order of a polynomial, it was assumed that all lower order terms would be included regardless of their individual contribution to the power of the explanation. In addition to producing a potentially less compact model, this naive implementation can also lead to numerical issues due to regressor correlation resulting in ill-conditioned design matrices present in the MML87 message length equations. To partially overcome the problem of numerical instabilities, Viswanathan and Wallace utilised orthonormal polynomial basis functions. However, with small amounts of data the regressors are still not guaranteed to be completely uncorrelated. Fitzgibbon tackled this issue by avoiding MML87 and using the MMLD [14] message length approximation that is based on posterior importance sampling. Although this method circumvents numerical issues, the sampling process is slow and does not scale to higher dimensional models as well as the MML87 criterion. In contrast, the algorithm presented in this paper is a more general solution to the linear regression problem that also remains numerically stable even in the face of correlated regressors. Additionally, the MML-OLS algorithm does not require the message length to be completely recalculated with the addition of new regressors, and is thus computationally efficient. As our algorithm is designed to be applied to a wider scope of problems than merely polynomial regression, we have also proposed a more general set of priors required for message length construction.

III. The Minimum Message Length (MML) Principle

Minimum Message Length (MML) [1], [3], [15], [16], [17] is an objective costing function that encapsulates parameter estimation and model discrimination in one general framework. The problem of inference is interpreted as finding a model that minimises the length of a message that exactly communicates some data to an imaginary 'receiver'. The message is divided into two parts: the first part, or 'assertion', states the model to be used, while the second part, or 'detail', encodes the data using the stated model. Simple models yield short assertions but may not encode the data efficiently, leading to longer details. Conversely, complex models require longer assertions, but generally result in shorter details. MML seeks to minimise this trade-off between model complexity and goodness of fit to the data.

The well-known Minimum Description Length criterion also follows the same principle, and differences between MDL and MML are detailed in the work by Baxter and Oliver [18].

The Strict Minimum Message Length [1], [15], [19] criterion partitions a countable data space into regions, each with an associated point estimate. For some partitioning into disjoint regions t_j each with point estimate $\boldsymbol{\theta}_j$, the average message length, $\mathcal{I}_1$, is given by

$$\mathcal{I}_1 = -\sum_{\boldsymbol{\theta}_j \in \Theta^*} \left(\sum_{\mathbf{x}_i \in t_j} r_i \right) \log q_j - \sum_{\boldsymbol{\theta}_j \in \Theta^*} \sum_{\mathbf{x}_i \in t_j} r_i \log f(\mathbf{x}_i | \boldsymbol{\theta}_j) \tag{1}$$

where Θ^* is the model space, $\mathbf{x}_i$ are the possible data values, $r_i \equiv r(\mathbf{x}_i)$ is the marginal probability of data $\mathbf{x}_i$, $q_j = \sum_{\mathbf{x}_i \in t_j} r_i$ is the *a priori* probability of region t_j being asserted, and $f(\mathbf{x}_i|\boldsymbol{\theta}_j)$ is the likelihood of data $\mathbf{x}_i$ given the model estimate $\boldsymbol{\theta}_j$. Minimising the above expression results in the SMML estimator [19]. Despite possessing many strong theoretical properties, the construction of the SMML estimator is impractical for all but the simplest of problems. This is due to the fact that the problem of partitioning the data space into optimal regions is NP-hard. Several approximations to SMML have been devised, the most successful and widely applied being the MML87 approximation [16]. In this framework, the approximation to the message length for a given model estimate, $\hat{\boldsymbol{\theta}}$, is given by

$$\mathcal{I}_1(\mathbf{x}) = -\log h(\hat{\boldsymbol{\theta}}) + \frac{1}{2} \log F(\hat{\boldsymbol{\theta}}) - \log f(\mathbf{x}|\hat{\boldsymbol{\theta}}) + c(D) \tag{2}$$

where $h(\cdot)$ is the prior probability density over Θ^*, $F(\cdot)$ is the determinant of the expected Fisher Information Matrix $I(\hat{\boldsymbol{\theta}})$, D is the dimensionality of the model $\hat{\boldsymbol{\theta}}$, and $c(\cdot)$ is some small constant introduced through the approximation. To perform inference with MML87, we seek the model that minimises (2). The sequel is concerned with the inference of linear regression models in the MML87 framework.

IV. MML Linear Regression

We observe N pairs of measurements $\{\mathbf{u}_i \in [-1,1]^M, y_i \in [-1,1]\}_{i=1}^N$ where $y_i = \mathcal{F}(\mathbf{u}_i) + v_i$ and M is the dimensionality of the input space. The $\mathbf{v} = [v_1, \ldots, v_N]^T$ are assumed to be i.i.d. as $v_i \sim \mathcal{N}(0, \sigma^2)$. The goal is to estimate the best linear explanation of y_i given $\mathbf{u}_i$, and thus indirectly approximate $\mathcal{F}(\cdot)$. Define $\mathbf{X} = [\mathbf{u}_1, ..., \mathbf{u}_N]^T$, $\mathbf{y} = [y_1, ..., y_N]^T$. and column i of $\mathbf{X}$ as candidate regressor $\mathbf{x}_i$. Rather than using the entire set of regressors, $\mathbf{X}$, we select only a subset $\boldsymbol{\Phi} \subset \mathbf{X}$. An index vector $\mathbf{k} \in \{1, ..., M\}^P$ defines which regressors are to be used in matrix $\boldsymbol{\Phi}$ where P is the number of regressors to be used.

The linear regression model of $\mathbf{y}$ given $\mathbf{X}$ and $\mathbf{k}$ is

$$\Phi = [\mathbf{x}_{k_1}, \ldots, \mathbf{x}_{k_P}] \tag{3}$$

$$\mathbf{y} = \mathbf{\Phi}\boldsymbol{\alpha} + \mathbf{v} \tag{4}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^P$ are the deterministic parameters of our linear regression model. The entire parameter set including deterministic parameters, noise model parameters and model structure parameters is $\boldsymbol{\theta} = \{P, \mathbf{k}, \boldsymbol{\alpha}, \sigma\}$. The task is to find a $\hat{\boldsymbol{\theta}}$ that minimises the message length (2); this task includes estimating both $\boldsymbol{\alpha}$ and σ as well as determining an optimal regressor set $\mathbf{k}$.

A. Likelihood Function

Given the assumption that the targets are corrupted by normally distributed noise, the negative log-likelihood of the linear regression is

$$\begin{aligned} -\log f(\mathbf{y}|\mathbf{\Phi}, \boldsymbol{\alpha}, \sigma) &= \frac{N}{2}\log(2\pi) + N\log\sigma \\ &\quad + \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{\Phi}\boldsymbol{\alpha})^T(\mathbf{y} - \mathbf{\Phi}\boldsymbol{\alpha}) \\ &\quad - N\log\epsilon \end{aligned} \tag{5}$$

where σ^2 is the variance of the noise and ϵ is the measurement accuracy of the data. The accuracy to which the data is measured is often used in estimating a lower bound on the accuracy of the model.

B. Priors

The MML87 criterion requires the specification of prior densities over all model parameters $\boldsymbol{\theta} = \{P, \mathbf{k}, \boldsymbol{\alpha}, \sigma\}$. Recall that P is the number of regressors being used, $\mathbf{k}$ is an index vector which lists regressors to be included, $\boldsymbol{\alpha}$ are the deterministic parameters and σ is the standard deviation of the noise. The total prior density of the model $\boldsymbol{\theta}$ is

$$h(\boldsymbol{\theta}) \equiv h(P) \cdot h(\mathbf{k}|P) \cdot h(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \cdot h(\sigma) \tag{6}$$

where $\boldsymbol{\gamma}$ is a hyperparameter that defines the shape of the prior density on $\boldsymbol{\alpha}$, and is discussed further below. We give the number of regressors to be used, P, a uniform prior

$$h(P) = \frac{1}{M}, \ (1 \leq P \leq M) \tag{7}$$

If adequate prior knowledge is available, a more informative prior density such as the geometric distribution may be appropriate. The index vector $\mathbf{k}$ is also given a uniform prior over the range of possible

combinations its members may assume

$$h(\mathbf{k}|P) = \binom{M}{P}^{-1} \tag{8}$$

Each regression coefficient, α_i, is given a normal prior with a standard deviation chosen from a discrete set Γ thus

$$h(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \frac{1}{\dim(\Gamma)^P} \cdot \prod_{i=1}^{P} \mathcal{N}(\alpha_i|0, \gamma_i^2),\ (\gamma_i \in \Gamma) \tag{9}$$

The set Γ is chosen to reflect the expected magnitude of the regression coefficients $\boldsymbol{\alpha}$. The wider the range the α_i cover, the more members will be required in the set Γ. The aim here is to ensure a suitable density is always available for efficient encoding of all members of $\boldsymbol{\alpha}$. The particular $\gamma_i \in \Gamma$ used for each coefficient is therefore chosen to minimise the overall message length (2), and this is similar in concept to regularisation [20], [21]. Alternatively, one may use the g-prior [22] with the scale chosen in a similar fashion from the set Γ. Finally, σ is given a scale invariant prior over the range $[a, b]$ with $\epsilon \leq a < b$

$$\Omega = \log \frac{b}{a} \tag{10}$$

$$h(\sigma) = \frac{1}{\Omega\sigma}, (\sigma \in [a, b]) \tag{11}$$

where Ω is a normalisation constant.

C. Fisher Information Matrix

The Fisher Information Matrix, $I(\boldsymbol{\theta})$, is defined as a matrix of expected second derivatives of the negative log-likelihood function with respect to all continuous parameters of the model, with entry (i, j) given by

$$I(\boldsymbol{\theta})_{i,j} = -\int_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}|\boldsymbol{\theta}) \frac{\partial^2}{\partial\theta_{i,j}} \log f(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} \tag{12}$$

where $\mathcal{X}$ is the entire data space. It is used to estimate the precision to which the continuous parameters are to be stated in the message assertion. The determinant of $I(\boldsymbol{\theta})$ for the linear regression problem is given by

$$F(\boldsymbol{\theta}) = |I(\boldsymbol{\theta})| = \left(\frac{2N}{\sigma^2}\right)\left(\frac{|\boldsymbol{\Phi}^T\boldsymbol{\Phi}|}{\sigma^{2P}}\right) \tag{13}$$

where $\boldsymbol{\Phi}^T\boldsymbol{\Phi}$ is the design matrix, and $|\cdot|$ denotes the determinant operator. We observe that $I(\boldsymbol{\theta})$ is positive semi-definite, and thus $|I(\boldsymbol{\theta})| \geq 0$.

D. MML87 Estimators

Substituting (5), (6) and (13) into (2) yields the message length expression for a linear regression model. We differentiate the full message length expression to obtain MML estimators for the continuous model parameters:

$$\hat{\boldsymbol{\alpha}} = \left(\Phi^T\Phi + \sigma^2 \mathrm{diag}(\boldsymbol{\gamma})^{-2}\right)^{-1} \Phi^T \mathbf{y} \tag{14}$$

$$\hat{\sigma}^2 = \frac{1}{N-P} \left(\mathbf{y} - \Phi\boldsymbol{\alpha}\right)^T \left(\mathbf{y} - \Phi\boldsymbol{\alpha}\right) \tag{15}$$

Points of interest:

1) Due to the fact that the Fisher Information (13) has no dependency upon the regression parameters $\boldsymbol{\alpha}$, the MML estimates $\hat{\boldsymbol{\alpha}}$ are in this case identical to the *maximum a posteriori* (MAP) estimates.
2) The estimate $\hat{\sigma}$ is observed to be unbiased, unlike the corresponding maximum likelihood estimator [1].
3) While the estimates $\hat{\boldsymbol{\alpha}}$ are the same as the MAP estimates, the addition of the Fisher Information term may lead to a different selection of model order.

V. Algorithm

A naive method for inferring an optimal regressor set $\mathbf{k}$ involves starting with $\mathbf{k} = \emptyset$, and at each step adding the unused regressor which reduces the message length by the greatest amount. This requires a complete recomputation of the message length for each candidate model at each step. There are several problems with this approach: not only is it slow, as the design matrix must be computed at every step for every possible model, but if the regressors are correlated the Fisher Information Matrix can become ill-conditioned leading to incorrect message lengths.

To counter both of these concerns, we combine the message length criterion with the orthogonal least squares estimator [2]. This leads to the following improvements over the naive approach:

1) The OLS algorithm ensures that the regressors are orthogonal to each other, and thus that the design matrix will be diagonal. This allows us to determine the contribution of a regressor by examining its autocorrelation, and enables us to avoid numerical ill-conditioning of the design matrix by rejecting those regressors that contribute little to the error reduction.
2) The contribution of each regressor to the message length can be calculated independently of the contribution of all other regressors; specifically, the determinant of the Fisher Information Matrix is reduced to a product of its diagonal terms. Similar conditions apply to the likelihood and prior.

These properties result in a more stable message length approximation and a dramatic improvement in the speed of the search for the optimal $\mathbf{k}$. Since MML87 is invariant under one-to-one data and model transformations [1], the inferences performed in the orthogonalised space are guaranteed to be the same as those performed in the original space, assuming suitable prior transformations.

A. The MML-OLS Algorithm

This section is loosely based on the algorithm presented by Billings et al [2]. Begin by defining

$$\mathbf{y} = \mathbf{W}\mathbf{g} + \mathbf{v} \tag{16}$$

where

$$\mathbf{W} = \mathbf{\Phi}\mathbf{A}^{-1} \tag{17}$$

$$\mathbf{g} = \mathbf{A}\boldsymbol{\alpha} \tag{18}$$

The matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_P]$ is the regressor matrix $\mathbf{\Phi}$ in an orthogonalised space, $\mathbf{A}$ is a transformation matrix, $\mathbf{g}$ are the transformed parameters and v_i are i.i.d. as $v_i \sim \mathcal{N}(0, \sigma^2)$. Additionally, define $\mathbf{e}(i)$ as the total error of the model, $\mathbf{k}(i)$ as the set of regressors being used, $\Xi(i)$ as the contribution of the Fisher term to the message length, $\rho(i)$ as the contribution of the prior terms to the message length, and $\boldsymbol{\gamma}(i)$ as the standard deviations for the priors on $\mathbf{g}(i)$, all at step i. Initialise these as

$$\mathbf{e}(0) = \mathbf{y} \tag{19}$$

$$\Xi(0), \rho(0) = 0 \tag{20}$$

$$\mathbf{k}(0), \mathbf{g}(0), \mathbf{W}(0), \mathbf{A}(0), \boldsymbol{\gamma}(0) = \emptyset \tag{21}$$

The algorithm works as follows: at each step, we consider the effect of adding one of the remaining regressors to the model. We therefore produce a set of candidate models from which we select the 'best' one to become the i-th order regression model. This process is repeated until we have either included all possible regressors, i.e. $P = M$, or have determined that all the remaining regressors contribute so little to the explanation that including them causes numerical issues. We use the following notation in the subsequent sections: $\mathbf{e}(i, j)$ denotes the error term of the model at step (i) which has been modified to include regressor j, $\mathbf{e}(i) \equiv \mathbf{e}(i, j^*(i))$ where $j^*(i)$ is the index of the candidate model that yields the lowest criterion score at step i, and $\mathbf{e}_k(i)$ denotes the k-th element of the vector $\mathbf{e}(i)$. This notation clearly extends to all quantities and not merely the error term.

B. Regressor Selection

1) Each step begins by determining

$$\Psi = \{1, \ldots, M\} - \mathbf{k}(i-1) \tag{22}$$

as the set of unused regressors under consideration.

2) Next, we look at each unused regressor in turn. For all $j \in \Psi$
   a) Set $k_i(i,j) = j$ (The candidate model now includes the j-th regressor)
   b) Update $\mathbf{W}(i,j)$, $\mathbf{A}(i,j)$ and $\mathbf{g}(i,j)$ as per Billings et al [2].
   c) Compute the autocorrelation of the new regressor as $\tau(i,j) = \mathbf{w}_i(i,j)^T \mathbf{w}_i(i,j)$
   d) Compute the message length, $\mathcal{I}(i,j)$, of the j-th candidate model as per Section V-C.
3) Prune the set Ψ to remove those regressors with autocorrelations below a predetermined threshold, δ

$$\Psi_p = \{j \in \Psi : \tau(i,j) \geq \delta\} \tag{23}$$

4) If $\Psi_p = \emptyset$, set $P = (i-1)$ and terminate the algorithm as all valid regressors have been included. Otherwise, select the best candidate model for step i as

$$j^*(i) = \arg\min_{j \in \Psi_p} \{\mathcal{I}(i,j)\} \tag{24}$$

Alternatively, candidate models may also be selected based on their likelihood scores.

5) If the algorithm has not terminated, return to step 1.

Once the algorithm has terminated, we select the optimal model, $\hat{\boldsymbol{\theta}}$, from our set of best candidates as

$$m^* = \arg\min_{m \in \{1,\ldots,P\}} \{\mathcal{I}(m)\} \tag{25}$$

$$\hat{\boldsymbol{\theta}} = \{m^*, \mathbf{k}(m^*), \mathbf{g}(m^*), \sigma(m^*)\} \tag{26}$$

It remains to select a suitable value of δ. A $\tau(i,j) < 1$ indicates that a regressor is highly correlated with a previously selected regressor and contributes little to the likelihood. Including such a regressor in the non-orthogonalised space could lead to an ill-conditioned design matrix and result in a break down of the MML87 message length approximation. Thus, a conservative choice of $\delta = 1$ is suitable when nothing about the orthogonality of the regressor set is known.

C. Message Length Update Laws

In this section we detail the expressions used to iteratively update the message length of candidate model j at step i. Begin by updating the error term to include the effects of the new regressor j

$$\mathbf{e}(i,j) = \boldsymbol{e}(i-1) - \mathbf{w}_i(i,j)g_i(i,j) \tag{27}$$

Following this, we estimate the standard deviation of the model residuals $\mathbf{e}(i,j)$ as

$$\sigma(i,j) = \left(\frac{\mathbf{e}(i,j)^T\mathbf{e}(i,j)}{N-i}\right)^{\frac{1}{2}} \tag{28}$$

We now have all the required information to compute the message length of candidate model j. The negative log-likelihood term is given by

$$L(i,j) = \frac{N-i}{2} + N\log\sigma(i,j) \tag{29}$$

Next we must update the Fisher Information term. Given the orthogonal nature of the regressor matrix $\mathbf{W}$, the design matrix $\mathbf{W}^T\mathbf{W}$ is guaranteed to be diagonal. As the determinant of a diagonal matrix is merely the product of its diagonal terms, the effect of the new regressor j upon the Fisher Information term is simply given by

$$\Xi(i,j) = \Xi(i-1) + \frac{1}{2}\log\tau(i,j) \tag{30}$$

Following this we update the contribution of the prior terms to the message length. Due to the assumed statistical independence of the regression parameters $\mathbf{g}$ the new prior term is given by

$$\rho(i,j) = \rho(i-1) + \frac{1}{2\gamma_i(i,j)^2}g_i(i,j)^2 + \log\gamma_i(i,j) \tag{31}$$

with $\gamma_i(i,j) \in \Gamma$ chosen to minimise (31). Finally, we arrive at the expression for the complete message length for candidate model j at step i

$$\mathcal{I}(i,j) = L(i,j) + \Xi(i,j) + \rho(i,j) - i\log\sigma(i,j) + Z(i) \tag{32}$$

where $Z(i)$ is constant for all candidate regressors j at step i, i.e. it depends only on the dimensionality of the model

$$\begin{aligned} Z(i) \;=\; & \frac{N}{2}\log(2\pi) - N\log\epsilon + \log\binom{M}{i} \\ & - \log i + \frac{i}{2}\log(2\pi) + \log\Omega + \frac{1}{2}\log(2N) \\ & + i\log\dim(\Gamma) - \frac{(i+1)}{2}\log(2\pi) \\ & + \frac{1}{2}\log((i+1)\pi) - 1 \end{aligned} \tag{33}$$

and need only be computed once per step. This term includes prior terms, normalisation constants, the measurement accuracy of the data and the MML87 approximation constants.

VI. Application to Polynomial Regression

To demonstrate the effectiveness of the MML-OLS algorithm on a real world problem we consider the task of univariate function approximation by polynomials. We observe N pairs of measurements $\{d_i, y_i\}_{i=1}^{N}$, where $y_i = f(d_i) + v_i$ and $v_i \sim \mathcal{N}(0, \sigma^2)$. We wish to approximate the unknown, possibly nonlinear function $f(\cdot)$. A polynomial approximation of $f(\cdot)$ is of the form

$$\hat{y}_i = \sum_{j=1}^{M} \alpha_j \omega_j(d_i) \tag{34}$$

where M is the maximum number of basis functions, and $\omega_j(\cdot)$ is the j-th order polynomial basis function. Legendre polynomials are used for the basis functions $\omega_j(\cdot)$ due to their property of mutual orthogonality [23]. The procedure begins by constructing the regressor matrix $\mathbf{X}$

$$\mathbf{u}_i = [\omega_1(d_i), ..., \omega_M(d_i)]^T \tag{35}$$

$$\mathbf{X} = [\mathbf{u}_1, ..., \mathbf{u}_N]^T \tag{36}$$

as per Section II. The problem is now clearly one of linear regression; the MML-OLS algorithm is subsequently applied to select the pertinent regressors.

VII. Results and Discussion

A. Testing Procedure

1) Test Functions: To test the MML-OLS algorithm on a practical task we applied polynomial regression to data generated from the following functions:

$$\begin{aligned}
\textbf{F1}: \quad y &= 9.72x^5 + 0.801x^3 + 9.4x^2 - 5.72x - 136.45 \\
\textbf{F2}: \quad y &= 0.623x^{18} - 0.72x^{15} - 0.801x^{14} \\
&\quad +9.4x^{11} - 5.72x^9 + 1.873x^6 \\
&\quad -0.923x^4 + 1.826x - 21.45 \\
\textbf{SIN}: \quad y &= (\sin(\pi(x+1)))^2 \\
\textbf{LOG}: \quad y &= \log(x + 1.01) \\
\textbf{ABS}: \quad y &= |x + 0.3| - 0.3
\end{aligned}$$

$$\textbf{DISC}: \quad y \quad = \begin{cases} 0.1 & \text{for } x < 0 \\ 2x - 1 & \text{for } x \geq 0 \end{cases}$$

It should be noted that functions **F1** and **F2** are true polynomials, **SIN** and **LOG** are smooth non-polynomials and **ABS** and **DISC** are piecewise non-polynomial functions.

2) Experimental Design: To rigorously test the MML-OLS algorithm we generated synthetic data from the given test functions. All inputs were uniformly drawn from $[-1, 1]$, and all targets were rescaled to the same range. The targets were then corrupted with normally distributed noise. We examine the effect of varying data sizes and noise on the inferences made by the MML-OLS method. The first test involved using all combinations of data sizes $N = \{25, 50, 75, 100\}$ and Signal-to-Noise ratios $SNR = \{1, 2, 4, 8\}$. The second set of tests involved either holding the data size N constant, and varying the SNR value, or holding the SNR constant and varying the N. The SNR is defined as

$$SNR = \frac{\sigma_f^2}{\sigma_v^2} \tag{37}$$

where σ_f^2 is the variance of the function, and σ_v^2 is the variance of the corrupting noise. For each combination of N and SNR we ran 100 tests and found the mean Squared Prediction Error (SPE) over all test runs. To provide an artificial yardstick by which the performance of the competing model selection criteria can be compared against, we find the model which gives the lowest SPE from all the available models. This is deemed the 'best' model. We compare the MML-OLS criterion against several well known and widely used inference criteria: AIC, AICc, BIC and AMAP. These are defined below

$$\mathcal{AIC}(\mathbf{x}) = 2L + 2k \tag{38}$$

$$\mathcal{AIC}_{\mathcal{C}}(\mathbf{x}) = 2L + 2k + \frac{2k(k+1)}{N-k-1} \tag{39}$$

$$\mathcal{BIC}(\mathbf{x}) = 2L + k\log(N) \tag{40}$$

$$\mathcal{AMAP}(\mathbf{x}) = 2L + k^2\log(N) \tag{41}$$

where L is the negative log-likelihood of the data given the model, k is the number of free parameters in the model, and N is the data size. The MDL78 criterion when applied to linear regression yields the same costing function as the BIC criterion [8]. Ideally, the choice of best candidate model at each step of the search algorithm (Section V-B, Step 4) should be based on message lengths for the MML criterion and on negative log-likelihood scores for the the competing methods. However, we desired that all criteria were to be presented with the same set of models from which they inferred their 'optimal' choice. To this end, log-likelihood scores were chosen to guide the search in all experiments. This is a sound choice, as given the priors used in the MML scheme, the candidate model with the lowest message length will

often be the one with the lowest negative log-likelihood. If anything, this puts the MML criterion at a slight disadvantage in comparison to the competing criteria, and it is conceivable that guiding the search on message lengths would yield some improvement in results.

Finally, we wished to visualise the types of models inferred by the MML-OLS criterion. To this end, we found the average outputs of all inferred models over the input data space and plotted the resulting 'mean' model with the true function. We also plotted the standard deviations of the inferred models from the mean model to capture the variation in polynomials selected by our criterion. The settings of the algorithm parameters used in the experiments are as follows: $\epsilon = 0.01$, $\Gamma = \{0.1, 1\}$, $a = 0.01, b = 2$, $\delta = 0$. In this application $\delta = 0$ was a suitable choice as the basis functions from which the regressors are generated are mutually orthogonal, and therefore the data, while not expected to be orthogonal will have low correlation.

B. Results Analysis

Results for tests on functions **F1**, **F2** and **ABS** are presented in Figures 1, 2 and 3 respectively; these show mean SPE scores for fixed SNR and varying N, and fixed N and varying SNR. Functions **F1** and **F2** were chosen as they are representative of a wide range of polynomial models, and **ABS** was chosen to test the performance on a non-smooth, non-polynomial model. The plots clearly show that MML outperforms all competing criteria in all cases in terms of SPE. In the presence of little data and high noise MML demonstrates a tendency to degrade in a much smoother fashion than the other criteria, in particular AIC.

Tables I-VI present mean SPE scores for combinations of different N and SNR values for all test functions. Due to space constraints, the scores for the AIC and AICc criteria have been omitted, but are discussed in the sequel. An examination of Tables I-V reveals that MML achieves lower mean SPE scores in almost all cases. In those cases where MML does not achieve the best results, the difference in SPE scores is insignificant. In contrast, when MML achieves superior results they are often several orders of magnitude lower than those of the competition. For low amounts of data ($N = 25$) the BIC criterion often degrades very badly, resulting in extremely high SPE scores due to a large amount of overfitting. The AICc and AMAP criteria gracefully handle lower amounts of data, but when a reasonable amount of data is available the BIC criterion tends to outperform AICc and BIC. The AMAP criterion penalises model complexity more strongly than BIC which prevents it from breaking down when small amounts of data are available. However, this same property means that the AMAP criterion will tend to underfit even when more data is available.

Table VI shows a different trend: in this case, MML rarely achieves the lowest SPE score, excepting the cases where the data size is small. This is easily explained by the nature of the target function; polynomials are not the ideal class of models for approximating discontinuous functions. Additionally, the chosen priors are in conflict with the nature of the function and improvements to the performance of the MML criterion can be made by choosing more suitable priors. However, this is considered neither useful nor necessary, as a piecewise-polynomial model is a far better alternative if we believe the function to be discontinuous, and would be expected to yield much lower message lengths than a normal polynomial model.

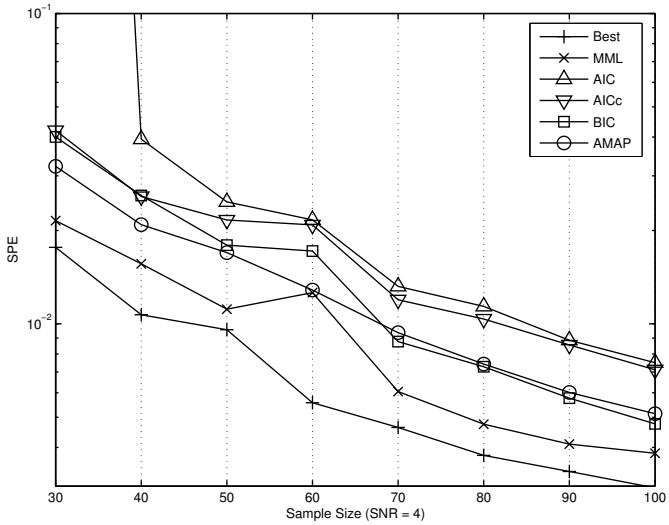
The degradation of the AIC and BIC criteria for small data sizes and low SNR can be attributed to their tendency to overfit and select complex models. In contrast to this, the MML, AICc and AMAP criteria are quite conservative, and select reasonable lower order models even when presented with very high noise. Excluding AICc, the other criteria are all based on assumptions that do not hold for small data sizes. While all unbiased criteria should converge to the 'best' model as $N \rightarrow \infty$, they may select very different models when these assumptions are violated. To summarise, it has been demonstrated that MML performs well for all combinations of tested N and SNR values. In contrast, the performance of all the other criteria under consideration varies significantly depending on data size and noise level.

Figure 4 shows the average inferred models for functions **F1** and **ABS**. In both cases 100 tests were performed with $N = 40$ and $SNR = 2$. It is clear that even in the presence of high noise and with little data available, the MML criterion on average selects models that are close to the true underlying function. Additionally, the models selected do not deviate greatly from the mean and would all be considered plausible given the low amount of data.

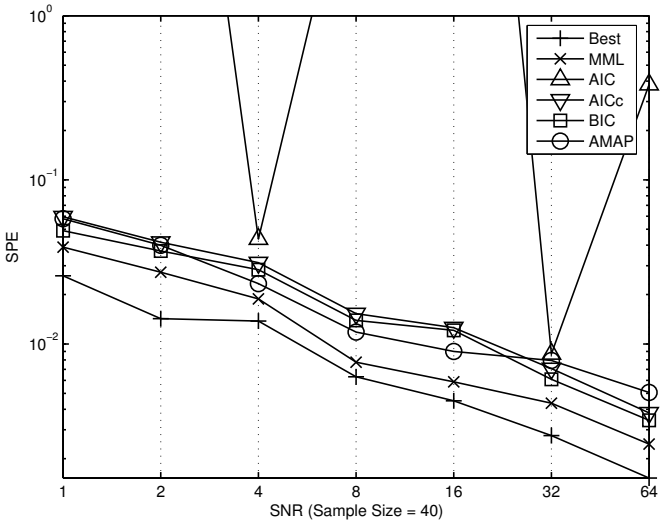
It is interesting to note that MML outperforms the competing criteria in terms of SPE even though the MML criterion is designed to select the most plausible model structure given the data, rather than attempting to explicitly minimise the model prediction error.

VIII. Conclusion and Future Work

We have presented an efficient algorithm for performing linear regression with the MML criterion based around the Orthogonal Least Squares (OLS) procedure. The MML-OLS algorithm allows for fast and numerically stable selection of regressors while using the MML criterion to determine the optimal regressor set. This algorithm has been applied to polynomial regression, and has been compared to other well known model selection criteria such as AIC and BIC. In terms of squared prediction error (SPE) the MML-OLS algorithm significantly outperformed the competing criteria, while remaining of the same

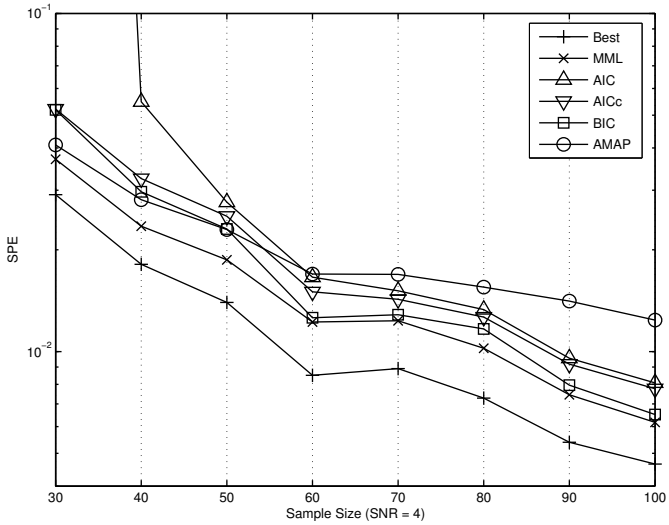


(a) Varying N and fixed SNR for function **F1**

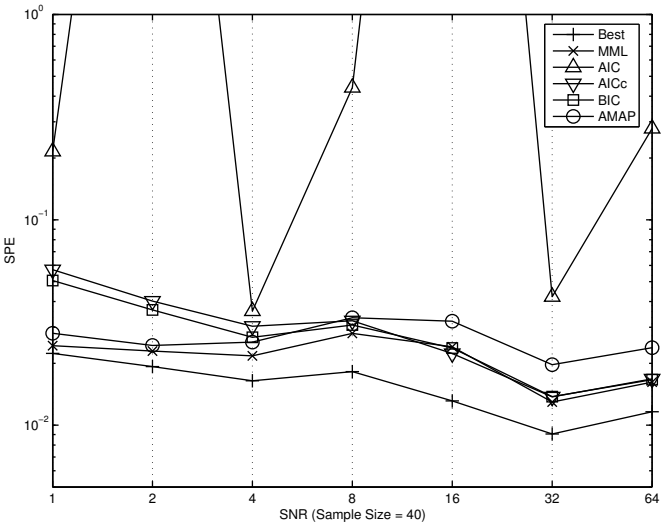


(b) Varying SNR and fixed N for function **F1**

Fig. 1. Results for varying N and SNR for function **F1**

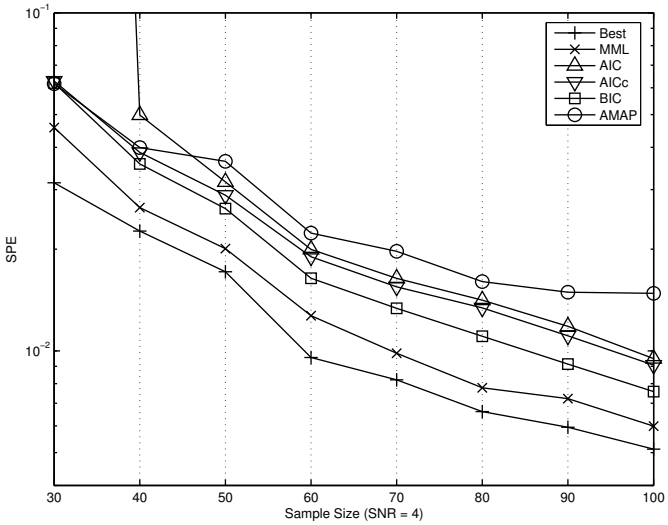


(a) Varying N and fixed SNR for function **F2**

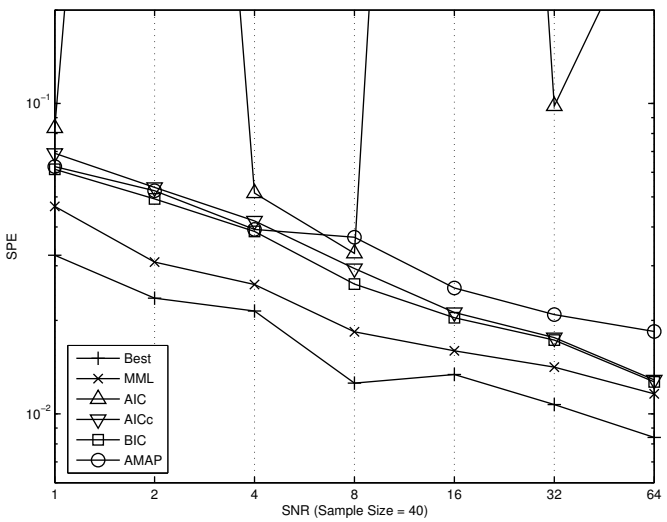


(b) Varying SNR and fixed N for function **F2**

Fig. 2. Results for varying N and SNR for function **F2**

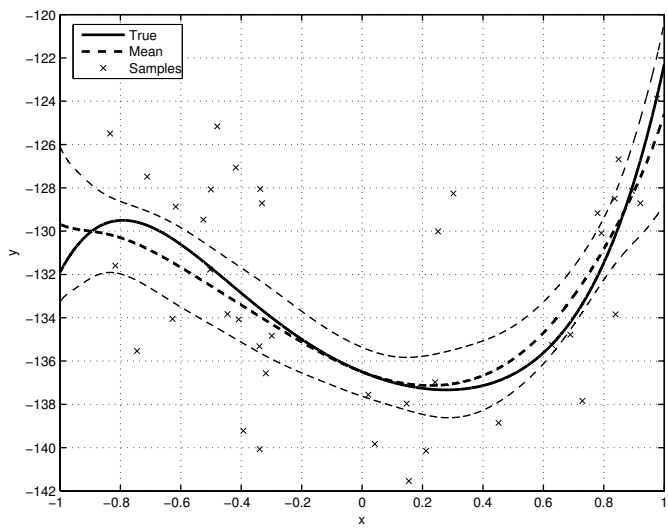


(a) Varying N and fixed SNR for function **ABS**

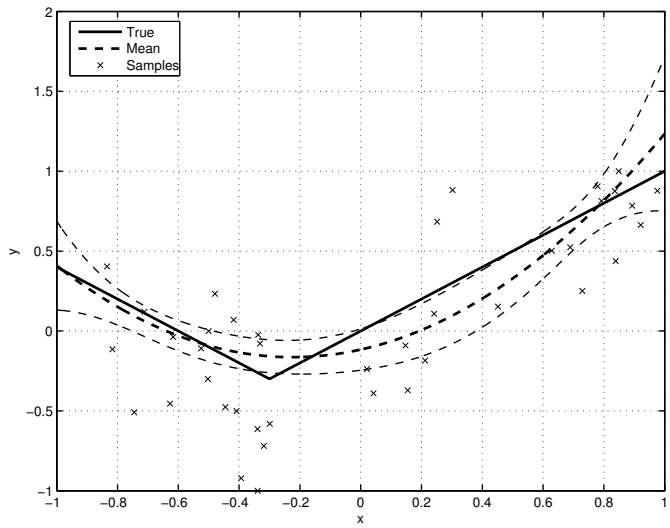


(b) Varying SNR and fixed N for function **ABS**

Fig. 3. Results for varying N and SNR for function **ABS**



(a) Function **F1**



(b) Function **ABS**

Fig. 4. Average of Models Inferred from functions **F1** and **ABS**

TABLE I

FUNCTION **F1** MEAN SPE SCORES

	SNR=1			SNR=2			SNR=4			SNR=8		
N	MML87	BIC	AMAP	MML87	BIC	AMAP	MML87	BIC	AMAP	MML87	BIC	AMAP
25	**9.03e-02**	9.83e+14	1.01e-01	**6.65e-02**	8.49e+06	7.82e-02	**5.32e-02**	1.22e+07	6.30e-02	**2.45e-02**	2.69e+10	3.91e-02
50	**2.58e-02**	3.31e-02	4.33e-02	**1.64e-02**	2.54e-02	2.75e-02	**1.42e-02**	2.18e-02	1.93e-02	**1.32e-02**	1.94e-02	1.59e-02
75	**1.23e-02**	1.72e-02	2.14e-02	**6.96e-03**	1.14e-02	1.07e-02	**5.71e-03**	8.20e-03	1.02e-02	**3.77e-03**	5.00e-03	5.74e-03
100	**5.18e-03**	8.44e-03	1.58e-02	**5.12e-03**	7.18e-03	8.82e-03	**3.86e-03**	5.51e-03	5.79e-03	**2.47e-03**	3.41e-03	4.20e-03

TABLE II

FUNCTION **F2** MEAN SPE SCORES

	SNR=1			SNR=2			SNR=4			SNR=8		
N	MML87	BIC	AMAP	MML87	BIC	AMAP	MML87	BIC	AMAP	MML87	BIC	AMAP
25	**6.10e-02**	6.49e+10	6.98e-02	**6.09e-02**	7.09e+05	6.98e-02	**4.79e-02**	9.62e+09	5.54e-02	**4.37e-02**	6.27e+12	4.93e-02
50	**2.24e-02**	3.75e-02	2.71e-02	**2.41e-02**	3.01e-02	2.74e-02	**2.34e-02**	2.73e-02	3.13e-02	1.29e-02	**1.21e-02**	2.18e-02
75	**1.39e-02**	1.79e-02	1.64e-02	**1.42e-02**	1.65e-02	1.75e-02	**1.09e-02**	1.10e-02	1.55e-02	**5.80e-03**	6.26e-03	1.34e-02
100	**1.18e-02**	1.45e-02	1.39e-02	**1.02e-02**	1.15e-02	1.50e-02	7.22e-03	**7.13e-03**	1.35e-02	**4.15e-03**	4.40e-03	1.07e-02

TABLE III

FUNCTION **SIN** MEAN SPE SCORES

	SNR=1			SNR=2			SNR=4			SNR=8		
N	MML87	BIC	AMAP	MML87	BIC	AMAP	MML87	BIC	AMAP	MML87	BIC	AMAP
25	**1.12e-01**	5.05e+10	1.14e-01	**1.08e-01**	1.27e+06	1.19e-01	**8.54e-02**	1.36e+10	1.06e-01	**5.97e-02**	6.21e+12	7.90e-02
50	**4.60e-02**	4.94e-02	6.07e-02	**2.71e-02**	3.23e-02	5.34e-02	**1.98e-02**	2.33e-02	3.97e-02	**1.34e-02**	1.47e-02	2.29e-02
75	**2.24e-02**	2.36e-02	3.97e-02	**1.29e-02**	1.50e-02	3.22e-02	**1.06e-02**	1.24e-02	2.39e-02	**7.02e-03**	7.50e-03	1.69e-02
100	1.38e-02	**1.32e-02**	3.30e-02	**9.32e-03**	1.01e-02	2.33e-02	**6.57e-03**	7.16e-03	1.51e-02	**3.77e-03**	4.12e-03	1.19e-02

TABLE IV

FUNCTION **LOG** MEAN SPE SCORES

	SNR=1			SNR=2			SNR=4			SNR=8		
N	MML87	BIC	AMAP	MML87	BIC	AMAP	MML87	BIC	AMAP	MML87	BIC	AMAP
25	**9.14e-02**	7.33e+14	1.03e-01	**6.88e-02**	1.29e+07	8.55e-02	**5.47e-02**	7.66e+06	6.61e-02	**4.21e-02**	9.82e+10	5.80e-02
50	**2.63e-02**	3.94e-02	3.31e-02	**2.19e-02**	2.73e-02	3.51e-02	**1.97e-02**	2.16e-02	3.27e-02	**2.24e-02**	2.63e-02	2.94e-02
75	**1.49e-02**	1.86e-02	2.42e-02	**1.26e-02**	1.54e-02	2.41e-02	**8.17e-03**	1.05e-02	1.74e-02	**5.63e-03**	7.44e-03	1.24e-02
100	**1.16e-02**	1.37e-02	1.83e-02	**8.44e-03**	1.01e-02	1.79e-02	**5.95e-03**	7.36e-03	1.37e-02	**4.43e-03**	4.91e-03	1.17e-02

TABLE V

FUNCTION **ABS** MEAN SPE SCORES

	SNR=1			SNR=2			SNR=4			SNR=8		
N	MML87	BIC	AMAP	MML87	BIC	AMAP	MML87	BIC	AMAP	MML87	BIC	AMAP
25	**1.06e-01**	5.23e+10	1.24e-01	**7.67e-02**	1.56e+06	1.14e-01	**6.17e-02**	1.78e+05	7.74e-02	**3.27e-02**	2.78e+10	5.11e-02
50	**3.29e-02**	4.61e-02	5.30e-02	**2.36e-02**	3.27e-02	3.61e-02	**2.41e-02**	3.29e-02	3.15e-02	**1.58e-02**	2.22e-02	3.00e-02
75	**1.58e-02**	2.11e-02	3.31e-02	**1.36e-02**	1.87e-02	2.22e-02	**8.52e-03**	1.22e-02	2.68e-02	**6.78e-03**	8.57e-03	1.93e-02
100	**1.05e-02**	1.40e-02	1.76e-02	**7.68e-03**	9.43e-03	1.58e-02	**6.17e-03**	7.63e-03	1.51e-02	**4.78e-03**	5.36e-03	1.14e-02

TABLE VI

FUNCTION **DISC** MEAN SPE SCORES

	SNR=1			SNR=2			SNR=4			SNR=8		
N	MML87	BIC	AMAP	MML87	BIC	AMAP	MML87	BIC	AMAP	MML87	BIC	AMAP
25	**1.61e-01**	2.86e+10	1.63e-01	**1.67e-01**	5.20e+08	1.74e-01	**1.69e-01**	8.08e+09	1.85e-01	**1.70e-01**	1.07e+11	1.82e-01
50	8.07e-02	**7.96e-02**	9.11e-02	**8.64e-02**	9.06e-02	1.04e-01	7.66e-02	**7.17e-02**	1.01e-01	**7.84e-02**	8.64e-02	1.06e-01
75	5.13e-02	**4.55e-02**	6.68e-02	5.08e-02	**4.45e-02**	7.55e-02	5.91e-02	**4.61e-02**	8.29e-02	2.06e-01	**4.81e-02**	8.32e-02
100	3.41e-02	**3.08e-02**	4.84e-02	4.00e-02	**3.06e-02**	5.71e-02	3.70e-02	**3.01e-02**	6.45e-02	9.87e-02	**3.30e-02**	6.55e-02

order of time complexity as these simpler methods. Future work in consideration includes an application to other more complex model classes such as radial basis function networks.

APPENDIX I

The complete Matlab^{TM} source code for all experiments detailed in this paper can be downloaded from

http://www.csse.monash.edu.au/∼dschmidt/MML-OLS/MML-OLS.zip

ACKNOWLEDGEMENTS

The authors would like to thank Lloyd Allison and Andrew Paplinski for reviewing drafts of the paper and providing insightful comments.

REFERENCES

[1] C. S. Wallace, *Statistical and Inductive Inference by Minimum Message Length*. Berlin, Germany: Springer, 2005.

[2] S. A. Billings, S. Chen, and M. J. Korenberg, "Identification of MIMO non-linear systems using a forward-regression orthogonal estimator," *International Journal of Control*, vol. 49, pp. 2157–2189, 1989.

[3] C. S. Wallace and D. M. Boulton, "An information measure for classification," *Computer Journal*, vol. 11, pp. 185–194, August 1968.

[4] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, July and October 1948.

[5] G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*. Wiley-Interscience, 1992.

[6] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, December 1974.

[7] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[8] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[9] ——, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, 1996.

[10] C. M. Hurvich and C. L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, pp. 297–307, 1989.

[11] P. M. Djurić, "Asymptotic MAP criteria for model selection," *IEEE Transactions on Signal Processing*, vol. 46, no. 10, pp. 2726–2735, October 1998.

[12] M. Viswanathan and C. S. Wallace, "A note on the comparison of polynomial selection methods," in *Proc. 7th Int. Workshop on Artif. Intelligence and Statistics*. Ft. Lauderdale, Florida, U.S.A.: Morgan Kaufmann, 1999, pp. 169–177.

[13] L. J. Fitzgibbon, D. L. Dowe, and L. Allison, "Univariate polynomial inference by Monte Carlo message length approximation," in *Int. Conf. Machine Learning 2002*. Sydney: Morgan Kaufmann, July 2002, pp. 147–154.

[14] ——, "Message from Monte Carlo," School of Computer Science and Software Engineering, Monash University, Australia 3800, Tech Report 2002/107, Decemember 2002.

[15] C. S. Wallace and D. M. Boulton, "An invariant Bayes method for point estimation," *Classification Society Bulletin*, vol. 3, no. 3, pp. 11–34, 1975.

[16] C. S. Wallace and P. R. Freeman, "Estimation and inference by compact coding," *J. Royal Statistical Society B*, vol. 49, pp. 240–252, 1987.

[17] L. Allison, "Models for machine learning and data mining in functional programming," *J. Functional Programming*, vol. 15, no. 1, pp. 15–32, 2005.

[18] R. A. Baxter and J. J. Oliver, "MML and MDL: similarities and differences," Dept. of Computer Science, Monash University, Clayton 3168, Australia, Tech. Rep. 207, 1994.

[19] G. E. Farr and C. S. Wallace, "The complexity of Strict Minimum Message Length inference," *Computer Journal*, vol. 45, no. 3, pp. 285–292, 2002.

[20] D. J. C. MacKay, "A practical Bayesian framework for backpropagation networks," *Neural Computation*, vol. 4, no. 3, pp. 448–472, 1992.

[21] ——, "Comparison of approximate methods for handling hyperparameters," *Neural Computation*, vol. 11, no. 5, pp. 1035–1068, 1999.

[22] A. Agliari and C. C. Parisetti, "A-g reference informative prior: A note on Zellner's g prior," *The Statistician*, vol. 37, no. 3, pp. 271–275, 1988.

[23] I. N. Bronshtein, K. A. Semendyayev, G. Musiol, and H. Muehlig, *Handbook of Mathematics*. Berlin, Germany: Springer, 2003.