A Comparative Study on Model Selection and Multiple Model Fusion*

Huimin Chen
University of New Orleans
Dept. of Electrical Engineering
New Orleans, LA 70148, USA

Shuqing Huang
Tulane University
Electrical Eng. & Computer Science Dept.
New Orleans, LA 70118, USA

Abstract – There exist quite a few criteria for penalty-based model selection. Although they have various justifications for large sample problems, their performance under small or moderate sample size is unclear which hinders the development of model combination methods using the appropriate penalty term. In this paper, we assess the performance of seven model selection criteria based on linear regression models with unknown noise variance. We set the true data generation mechanism to be within the model set as well as outside the model set. In the latter case, soft model selection through multiple model fusion is proposed and its difference from Bayesian model averaging is highlighted. The penalty term used in each model selection criterion provides a natural link to estimate the model probability without assuming any prior knowledge of the unknown parameter. An important question is whether the estimated model probabilities are consistent when multiple models are fused for prediction or interpolation. We argue that strong consistency only holds under large sample regime while soft model selection can still be better than choosing a single model with small sample size. Our numerical results using different model selection criteria for polynomial fitting indicate that the conditional model estimator (CME) has the best performance in selecting the correct model order and fusing multiple models for prediction and interpolation. The minimum description length (MDL) based criteria are next to CME and outperform Bayesian information criterion (BIC) and Akaike information criterion (AIC) significantly.

Keywords: Model selection, multiple model inference, estimation fusion, linear regression, time series analysis.

1 Introduction

Statistical inference seeks to accurately estimate the unknown parameter or state with the observations dependent upon the unknown quantity. The description of statistical dependence is often called a model or, in some restrictive situations, a hypothesis. The selection of one model among a set of candidates can be interpreted as knowledge discovery from the observed data. Because of the concern with useful knowledge on true data generation mechanism, model selection research places a high value on an algorithm that produces comprehensible output rather than simply the accurate prediction or interpolation. There are also practical reasons for this. In many applications, it is not enough for a selected model to be accurate; it also needs to be understood by human users if they deem it acceptable. The class of models with a fully specified likelihood function of the observations conditioned on the unknown parameter is considered in this paper. Such a setting differs from the data-oriented models widely used in machine learning area where one allows nonparametric relation between the observations and the unknown truth. Overfitting avoidance is often considered the central problem in model selection where one needs to guard against selecting a model that fits the available data well but captures the underlying phenomenon poorly. Current methods to address this problem fall largely into two broad categories. The data-oriented approaches use separate data sets to train and validate models. Model selection relies on the partition of the data set which includes methods such as cross validation [19], bootstrap [6], boosting [7], bagging [4] and their variants. Data-oriented model selection has several disadvantages: unreliable model validation when the data set is small, self prone to overfitting if a large number of models is compared, and the intensive computational burden. Penalty-based approaches use the same data for train-

*Research supported in part by ARO grant W911NF-04-1-0274 and NSF/LEQSF-Pfund-38.
ing and validation but penalize models which are likely to overfit. Bayesian model selection [13], structural risk minimization [20], and PAC learning [10] fall into this category where the major issue is to design an appropriate penalty term for each model.

There exist quite a few criteria for penalty based model selection. Although they have various justifications for large sample problems, their performance under small or moderate sample size is unclear which hinders the development of model combination methods using the appropriate penalty term. In this paper, we assess the performance of seven model selection criteria based on linear regression models with unknown noise variance. We set the true data generation mechanism to be within the model set and outside the model set. In the latter case, soft model selection through multiple model fusion is proposed and its difference from Bayesian model averaging is highlighted. The penalty term used in each model selection criterion provides a natural link to estimate the model probability without assuming any prior knowledge of the unknown parameter. An important question is whether the estimated model probabilities are consistent when multiple models are combined for the prediction or interpolation purposes. We argue that strong consistency only holds under large sample regime while soft model selection can still be better than choosing a single model with small sample size if one uses multiple models for prediction and interpolation. Applications of multiple model fusion to sequential state estimation are briefly presented.

3 Problem Formulation

Consider a class of models \( M_1, \ldots, M_K \) where model \( M_i \) assumes that the observation \( z \) is governed by a likelihood functional \( f_i(z|\theta_i) \) depending on the unknown parameter \( \theta_i \) \((i = 1, \ldots, K)\). The dimension of \( \theta_i \) is denoted by \( p_i \). Denote by \( z^n \) a vector of \( n \) independent observations. Given \( z^n \), one wants to find the best model \( M_i \) among the \( K \) candidates. A model with fully specified likelihood function in a parametric form is often called a statistical hypothesis. Model selection becomes a hypothesis testing problem when the models do not have overlap in their parameter spaces or they have different parametric forms of the likelihood functions. In many engineering applications, the models are nested from a simple to a more complicated one varying in the dimension of the parameter space. In this paper, we will consider the general family of parametric models but use linear regression problem as the illustrative example to show how different model selection criteria behave under small or moderate sample size.

If \( M_i \) is declared as the true data generation mechanism, one may also want to find the best estimate of \( \theta_i \) as a function of \( z^n \). In a Bayesian setting, prior distribution, e.g., a proper probability density function, is assigned to \( \theta_i \) and the marginal likelihood function is obtained as follows.

\[
f_i(z^n) = \int_{\Theta_i} f_i(z^n|\theta_i)p(\theta_i)d\theta_i
\]

Then model \( M_i \) is selected if

\[
i = \arg \max_j f_j(z^n)P(M_j)
\]

The difficulty lies in the specification of appropriate priors over different parameter spaces for different models. Besides, \( f_i(z^n) \) is the average likelihood which may not correspond to any particular value \( \theta_i \).

An alternative approach is to estimate \( \theta_i \) using \( z^n \) and put the estimate \( \hat{\theta}_i \) into the likelihood function. Thus one can select model \( M_i \) if

\[
i = \arg \max_j f_j(z^n|\hat{\theta}_j)
\]

where the maximum likelihood (ML) estimate is often used, i.e.,

\[
\hat{\theta}_j = \arg \max_{\theta_j} f_j(z^n|\theta_j)
\]

The above model selection criterion is a natural extension of the generalized likelihood ratio test to deal with multiple hypotheses [9]. A major issue of this approach is the potential overfit to a complicated model \( M_i \) with limited data. Such a model may not generalize well since the flexibility of the parameter is mainly used to fit the noise portion of the data. Thus one needs to design a penalty term to account for the potential overfit. It is desirable to have a consistent penalty term that guarantees finding the correct model when \( n \) is large enough. Existing model selection criteria can be written in a general form

\[
l_j = -\log f_j(z^n|\hat{\theta}_j) + d_j(z^n), \; j = 1, \ldots, K
\]

being minimized among the \( K \) candidates. The first term of \( l_j \) uses the best estimate of \( \theta_j \) to fit the negative log-likelihood function. The second term \( d_j(z^n) \) is a penalty function that varies for different criteria.
The Akaike information criterion (AIC) uses
\[ d_j(z^n) = 2p_j - 2 \log L_j(z^n) \] [1]. It is an asymptotically unbiased estimator of a variant of Kullback's divergence between the true data generation model and the fitted model under the assumption that the true model is correctly specified or overfitted.

Following a similar spirit to minimizing the distance between the true model and the fitted one, the Kullback information criterion (KIC) uses
\[ d_j(z^n) = 2p_j - 2 \log L_j(z^n) \] to avoid the potential overfitting using AIC [17]. The major difference from AIC lies in that KIC is based on Kullback's symmetric divergence measure between the true data generation model and the fitted model.

The Bayesian information criterion (BIC) uses
\[ d_j(z^n) = \frac{p_j}{2} \log n \] [16]. It can be interpreted as a Bayesian procedure (1) under the assumption that the nuisance parameter \( \theta_j \) can be well approximated by a normal distribution centered at \( \hat{\theta}_j \). In the literature, the penalty \( \frac{p_j}{2} \log n \) is sometimes interpreted as the natural prior over the dimension of parameter space \( p_j \), i.e., the prior model probabilities.

The minimum description length (MDL) criterion uses
\[ d_j(z^n) = \frac{p_j}{2} \log n - \frac{1}{2} \log I(\theta_j) + \frac{1}{2} \int |I(\theta_j)|^{1/2} d\theta_j \] where \( I(\theta_j) = \lim_{n \to \infty} \frac{1}{n} \mathbb{E} \left[ \frac{\partial^2 \log f_j(z^n|\theta_j)}{\partial \theta_j \partial \theta_j'} \right] \) and the integral in (6) is over an appropriate subset of the parameter space. The MDL criterion intends to minimize the overall code length of a model and the observation described by the model.

There also exist a few variants of the MDL penalty such as stochastic information complexity (SIC) [14] and predictive MDL (pMDL) [2] based on different encoding schemes. Note that the penalty term in AIC does not depend on the observation and it depends on the observation only through its dimension \( n \) in BIC.

In the above criteria, the ML estimate is used for the unknown parameter.

The conditional model estimator (CME) [8] has recently been proposed that avoids estimating \( \theta_j \) by using a conditional density with the sufficient statistic \( T_j(z^n) \) of \( \theta_j \) to select model \( M_i \) if
\[ i = \arg\min_j \log f_j(z^n|T_j(z^n)) \] (8)
The above functional does not depend on \( \theta_j \) and can be viewed as the negative log-likelihood of the ancillary statistic [9]. Note that the sufficient statistic of \( \theta_j \) is not unique. Using the technique developed in [8], it can be shown that the minimum variance unbiased (MVU) estimate of \( \theta_j \) should be used if it exists. In this case, \( d_j(z^n) = \log f_j(\hat{\theta}_j|\theta_j) \). Otherwise, class-specific CME [8] has to be developed to remove the dependence on \( \theta_i \) within a specific class of nested models. The CME may not be applicable to non-nested models.

Given seven renowned model selection criteria, namely, AIC, KIC, BIC, MDL, SIC, pDML, CME, we are interested in the performance of small and moderate sample size where the true data generation mechanism may or may not belong to the set of \( K \) models. We focus on linear regression models of different orders under which all model selection criteria have closed form solutions. In some applications such as prediction and interpolation, multiple models can be used instead of selecting a single model. The main issue is that estimating model probability \( P(M_i|z^n) \) requires the prior density \( p(\theta_j) \) which may not be available for all models. A recent work [18] proposed to estimate \( P(M_i|z^n) \) based on \( l_j \) without assuming the prior density of \( \theta_j \).

In particular, the authors recommended using BIC in their multiple model approach. Since models may have overlap with each other, the Bayesian justification of using BIC is questionable and it is not invariant with respect to the location or scale change of the unknown parameter in the model. Thus we want to focus on the estimation accuracy of the model probabilities based on \( l_j \) but using different model selection criteria. This has a strong implication to the multiple model estimation and prediction where the outputs from different models are combined via the weighted sum of their model probabilities. Our numerical study intends to provide a general guideline for choosing a sound criterion in model selection and model probability estimation.

### 3 Model Selection Criteria for Linear Regression Models

In this section, analytical forms of (5) are obtained for the seven model selection criteria which will be used in the simulation study. For a set of \( K \) models, if the observation \( y \) is assumed to be generated by model \( M_i \), then we have the linear regression \( y = H_i \theta_i + w \), where \( H_i \) is a known \( n \times p_i \) matrix; \( \theta_i \) is an unknown \( p_i \times 1 \) vector; and \( w \sim N(0, \sigma^2 I) \) is the Gaussian noise vector with unknown variance \( \sigma^2 \). The MVU estimate of \( \theta_i \) is \( \hat{\theta}_i = (H_i^T H_i)^{-1} H_i^T y \). The residue sum squares (RSS) of \( \hat{\theta}_i \) is \( R_i = \| y - H_i \hat{\theta}_i \|^2 \). The MVU estimate of \( \sigma^2 \) is \( \hat{\sigma}^2 = R_i/(n-p_i) \), which is different from the ML estimate given by \( R_i/n \).

For small and moderate sample size, a bias compensated version of the AIC selects the model that mini-
mizes the cost [17]
\[
\text{AIC}(p_i) = \frac{n}{2} \log R_i + \frac{(p_i + 1)n}{n - p_i - 2}
\] (9)

It was found that the bias compensation improves model selection accuracy of the AIC for linear regression problem. Thus we use the bias compensated AIC in the comparative study next. Following a similar justification, the bias compensated KIC minimizes the cost
\[
\text{KIC}(p_i) = \frac{n}{2} \log R_i + \frac{(p_i + 1)n}{n - p_i - 2} - n\psi\left(\frac{n - p_i}{2}\right) + g(n)
\] (10)

where \(g(n) = n \log \frac{n}{2}\) and \(\psi(\cdot)\) is the digamma function [17]. The BIC selects the model that minimizes the cost
\[
\text{BIC}(p_i) = \frac{n}{2} \log R_i + \frac{p_i + 1}{2} \log n
\] (11)

The MDL based on the NML density minimizes the cost [2]
\[
\text{MDL}(p_i) = \frac{n}{2} \log (\hat{\sigma}_i^2) + \frac{p_i + 1}{2} \log F_i + L_i
\] (12)

where \(F_i = (y^T y - R_i)/(p_i \hat{\sigma}_i^2)\) and \(L_i = \frac{1}{2} \log \left(\frac{n - p_i}{p_i}\right)\).

The SIC selects the model that minimizes the cost [14]
\[
\text{SIC}(p_i) = \frac{n - p_i - 2}{2} \log R_i + \frac{p_i + 1}{2} \log n + \frac{1}{2} \log |H_i^T H_i|
\] (13)

The predictive MDL minimizes the cost [2]
\[
\text{pMDL}(p_i) = \frac{n}{2} \log (\hat{\sigma}_i^2) + \frac{p_i + 1}{2} \log F_i + \log n
\] (14)

Finally, by using the sufficient statistic \(\hat{\theta}_i\) and \(\hat{\sigma}_i^2\), the CME minimizes the cost [8]
\[
\text{CME}(p_i) = \frac{n - p_i - 2}{2} \log (\hat{\sigma}_i^2) + \frac{1}{2} \log |H_i^T H_i| + C_i
\] (15)

where \(C_i = \frac{n - p_i}{2} \log \left[\pi(n - p_i)\right] - \log \Gamma\left(\frac{n - p_i}{2}\right)\).

Since the closed forms are available for the above model selection criteria in linear regression, we can have a comparative study of model selection criteria without resorting to any data oriented model validation. Model order selection for autoregressive models has important applications in time series analysis and forecasting and the penalty based criteria are popularly used in this area especially AIC (without bias compensation) and BIC due to the simplicity of their penalty terms. Thus the performance comparison of the above model selection criteria will also help the practitioner gain insights on which criterion to use for problems with small or moderate sample size.

4 Multiple Model Fusion and Model Probability Estimation

Contrary to model selection, multiple model identification tries to find a subset of plausible models from the model set. Dynamically identifying the subset based on the structural difference among the candidate models has been studied in [11, 12] where one cares more about the inference of the unknown state rather than validating a particular model with the given data. The decision problem is converted into an estimation problem where the model probability \(P(M_i|z^n)\) needs to be inferred from the observations for the purposes of prediction or interpolation. The multiple model approach has abandoned the assumption that the true data generation mechanism has to be one of the candidate models. From the Bayesian point of view, selecting a single model to make prediction or interpolation ignores the uncertainty left by finite data as to which model is the correct one. Thus a Bayesian formalism uses all possible models in the model space under consideration when making predictions, with each model weighted by its “posterior” probability being the correct one. The approach is called Bayesian model averaging and widely used in combining different learning algorithms [3]. The major difficulty in applying the Bayesian approach lies in the specification of priors when the candidate models are nested or partially overlapped. To be more specific, assuming that we have the observation vector \(z^n\), the likelihood of a new observation \(y\) can be approximated by
\[
f(y|z^n) \approx \sum_{i=1}^{K} P(M_i|z^n) f_i(y|\hat{\theta}_i)
\] (16)

using multiple models while the single model likelihood is \(f_i(y|\hat{\theta}_i)\) if model \(M_i\) is selected. Without assigning prior to \(\theta_i\), one can not apply Bayes formula to estimate the model probability.

If one does not stick to the formal Bayesian solution, then the penalty \(l_i\) in the model selection criterion can be used to estimate the model probability. The authors in [18] proposed the following estimator
\[
P(M_i|z^n) = e^{-l_i} \left(\sum_{j=1}^{K} e^{-l_j}\right)^{-1}
\] (17)

and they recommended using BIC for \(l_i\) due to its Bayesian flavor. In our opinion, the model probability can not be interpreted as the posterior since the prior of \(\theta_i\) is unspecified. It represents the self-assessment of how likely \(M_i\) is selected using the criterion \(l_i\). We are interested in the consistency of model probability estimate for all criteria even when the true data generation mechanism does not belong to any single model.
sufficient statistic of information with variable set of models if one can find the outline of (17) is to approximate the Bayesian evaluation, i.e., parameter estimation. The essence is to find the outcome provided via multiple model fusion on a predetermined minimum model probability. Note that the outcome provided via multiple model fusion is valid for any model set, not limited to linear models. Denote by \( \hat{y}_i \) the estimate (prediction or interpolation) of \( y \) using \( M_i \). The single model estimate uses \( \hat{y}_i \) if \( M_i \) is selected as the best model. The multiple model estimate uses

\[
\hat{y} = \sum_{i=1}^{K} P(M_i | z^n) \hat{y}_i
\]

and a subset of the \( K \) models can be identified based on a predetermined minimum model probability. Note that the outcome provided via multiple model fusion is valid for any model set, not limited to linear models. The essence of (17) is to approximate the Bayesian evidence of each model without any dependence on the unknown parameter \( \theta \), so that all data can be used for the inference purpose, i.e., parameter estimation. The method can be extended to sequential state estimation with variable set of models if one can find the sufficient statistic of \( z^{n-1} \) at the time of processing the \( n \)-th observation to compute (17). An approximated sequential algorithm for joint model selection and state estimation is suggested in [5].

The approach above differs from data-oriented multiple model combination where the fusion of multiple model outputs relies on the means of variations in the data set. Indeed, those techniques including cross validation, boosting and bagging can be used to “validate” the correctness of a certain penalty term in estimating the model probability. In this paper we examine the accuracy of model probability estimate via Monte Carlo simulations with finite sample size. For each model selection criterion, it is directly linked to the actual model selection performance rather than certain error bounds [20] being derived using data-oriented multiple model combinations.

### Computer Simulations

We use the polynomial fitting example in [8] to evaluate seven model selection criteria given in Section 3 in terms of the probabilities of choosing the polynomial orders from 1 to 10. Based on each criterion, we also estimate the model probability for each order and the average model probability is obtained via \( 10^4 \) independent runs. The signal is a polynomial (correct order = 3)

\[
s(t) = 0.4t + 0.1t^2, \quad t = 0, 1, ..., n - 1
\]

and the observation \( y(t) \) is generated by the signal plus additive Gaussian noise with variance \( \sigma^2 \). Table 1 shows the actual probability of choosing a particular order polynomial and the corresponding estimated average model probability in parentheses with 30 observations and large noise variance \( \sigma^2 = 10^2 \). Similar comparison for small noise variance \( \sigma^2 = 10 \) is shown in Table 2. When noise power is large, CME has the largest probability of choosing the correct order. Its model probability estimate is also more consistent than the rest of the criteria. SIC and MDL based criteria tend to underestimate the order while AIC, BIC and KIC tend to overestimate the order. Similar observations were reported in [8] where CME was compared with BIC only (denoted by MDL in [8]). When noise power decreases, CME and MDL based criteria find the correct order almost perfectly while AIC, BIC and KIC still overestimate the model order. AIC is much worse

<table>
<thead>
<tr>
<th>order</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>0(0)</td>
<td>0(0)</td>
<td>.58(.32)</td>
<td>.11(19)</td>
<td>.07(12)</td>
<td>.05(09)</td>
<td>.05(07)</td>
<td>.05(06)</td>
<td>.04(06)</td>
<td>.04(06)</td>
</tr>
<tr>
<td>KIC</td>
<td>0(0)</td>
<td>.04(07)</td>
<td>.78(66)</td>
<td>.15(18)</td>
<td>.02(06)</td>
<td>.01(02)</td>
<td>.01(01)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
</tr>
<tr>
<td>BIC</td>
<td>0(0)</td>
<td>.03(05)</td>
<td>.82(58)</td>
<td>.07(19)</td>
<td>.03(08)</td>
<td>.02(04)</td>
<td>.01(02)</td>
<td>.01(01)</td>
<td>.00(01)</td>
<td>.00(01)</td>
</tr>
<tr>
<td>MDL</td>
<td>0(0)</td>
<td>.31(33)</td>
<td>.69(65)</td>
<td>.09(02)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
</tr>
<tr>
<td>SIC</td>
<td>0(0)</td>
<td>.24(28)</td>
<td>.76(72)</td>
<td>.09(07)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
</tr>
<tr>
<td>pMDL</td>
<td>0(0)</td>
<td>.11(15)</td>
<td>.86(75)</td>
<td>.02(08)</td>
<td>.01(01)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
</tr>
<tr>
<td>CME</td>
<td>0(0)</td>
<td>.12(14)</td>
<td>.88(84)</td>
<td>.09(02)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
</tr>
</tbody>
</table>

Table 1: Comparison of model selection probability and estimated model probability (\( p_i = 3; n = 30; \sigma^2 = 10^2 \)).

<table>
<thead>
<tr>
<th>order</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>0(0)</td>
<td>0(0)</td>
<td>.58(.32)</td>
<td>.11(19)</td>
<td>.07(13)</td>
<td>.05(07)</td>
<td>.04(06)</td>
<td>.04(06)</td>
<td>.04(06)</td>
<td>.04(06)</td>
</tr>
<tr>
<td>KIC</td>
<td>0(0)</td>
<td>0(0)</td>
<td>.91(72)</td>
<td>.07(19)</td>
<td>.01(05)</td>
<td>.01(02)</td>
<td>.01(01)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
</tr>
<tr>
<td>BIC</td>
<td>0(0)</td>
<td>0(0)</td>
<td>.85(62)</td>
<td>.08(21)</td>
<td>.03(08)</td>
<td>.02(04)</td>
<td>.01(02)</td>
<td>.01(01)</td>
<td>.00(01)</td>
<td>.00(01)</td>
</tr>
<tr>
<td>MDL</td>
<td>0(0)</td>
<td>0(0)</td>
<td>1.99</td>
<td>.01(01)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
</tr>
<tr>
<td>SIC</td>
<td>0(0)</td>
<td>0(0)</td>
<td>1.11</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
</tr>
<tr>
<td>pMDL</td>
<td>0(0)</td>
<td>0(0)</td>
<td>.97(88)</td>
<td>.02(10)</td>
<td>.01(02)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
</tr>
<tr>
<td>CME</td>
<td>0(0)</td>
<td>0(0)</td>
<td>1.11</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
<td>.00(00)</td>
</tr>
</tbody>
</table>

Table 2: Comparison of model selection probability and estimated model probability (\( p_i = 3; n = 30; \sigma^2 = 10 \)).

The approach above differs from data-oriented multiple model combination where the fusion of multiple model outputs relies on the means of variations in the data set. Indeed, those techniques including cross validation, boosting and bagging can be used to “validate” the correctness of a certain penalty term in estimating the model probability. In this paper we examine the accuracy of model probability estimate via Monte Carlo simulations with finite sample size. For each model selection criterion, it is directly linked to the actual model selection performance rather than certain error bounds [20] being derived using data-oriented multiple model combinations.
than BIC. KIC performs better than AIC and BIC but worse than CME and MDL based criteria. Note that pMDL has an apparent inconsistency between the actual order selection probability and the average model probability. Numerous other simulation examples with different true model orders also confirm this.

Next, we consider the order selection when the signal is not a polynomial function. In this case, we evaluate the generalization error for the single best model and multiple models. The results shown in Table 3 use

\[ s(t) = 0.4 \cdot 1.2^t + 0.1t^2; \quad t = 0, 1, \ldots, n - 1 \]

(20)
to generate observations. We estimate the mean square prediction/interpolation error at \( t = 0.5, 1.5, \ldots, n + 0.5, n + 1.5, n + 2.5 \). Similar comparison is made in Table 4 with

\[ s(t) = 20 \cos(0.2t), \quad t = 0, 1, \ldots, n - 1 \]

(21)

The variance of generalization error \( \hat{\sigma}_{\text{SM}}^2 \) from the single best model and \( \hat{\sigma}_{\text{MM}}^2 \) from the multiple model estimator are listed in addition to the actual order selection probability and the estimated model probability for each criterion. We can see that the generalization error variance is larger than the noise variance \( \sigma^2 = 10 \) due to model mismatch. In Table 3, CME and MDL based criteria have comparable generalization error while BIC and AIC have much larger error. KIC performs better than AIC and BIC but slightly worse than CME and MDL based criteria. Note that CME and SIC choose order 4 with probability close to 1 while MDL and pMDL also put certain weight on order 5. The multiple model approach yields smaller error for AIC, BIC, KIC, and pMDL even though the model probability estimates are not consistent. In Table 4, CME has the smallest generalization error and the MDL based criteria have comparable performance to CME. KIC is slightly worse but performs much better than AIC and BIC. Note that MDL and pMDL choose order 5 with probability close to 1 while no single order is dominant in other cases. The multiple model estimator has a better performance than the single best model in AIC, BIC and CME but has worse performance in SIC.

In all cases, CME has the best performance and MDL based criteria are next to CME. These four criteria outperform BIC and AIC significantly. KIC has a better performance than BIC and AIC but is slightly worse than CME and MDL based criteria.

### 6 Discussion and Conclusions

In this paper seven penalty-based model selection criteria were compared with a linear regression problem of unknown noise variance. The results in terms of model selection and prediction accuracy provide us a general guideline to choose the appropriate penalty term for model selection. Conditional model estimator (CME) is preferable whenever one has the MVU estimate for the unknown parameter of each model as
a minimum sufficient statistic. When CME is not applicable, minimum description length (MDL) based criteria should be used for small or moderate sample size instead of BIC and AIC. The recently proposed KIC with bias compensation is usually inferior to CME and MDL based criteria but outperforms BIC and AIC. For the prediction and interpolation of unseen data, multiple model fusion has a clear advantage over the single best model especially when the true data generation mechanism does not belong to the model set. Similar observations were also reported in [18], however, the comparison is based on BIC, which is not suitable for small or moderate sample size. CME seems to be more consistent in estimating the model probability. CME and MDL based criteria are preferred when applying the multiple model estimation compared with BIC and AIC. The generalization error may increase when using multiple model with inconsistent model probability estimate. Thus caution has to be exercised when using (18) to combine the outputs from multiple models with non-nested model structure. Future work will extend the penalty-based criteria to the models being specified with semi-parametric distributions. Such a setting has a lot of applications in machine learning especially for model selection among kernel-based learning algorithms.

References


