IMS Collections
Borrowing Strength: Theory Powering Applications –
A Festschrift for Lawrence D. Brown
Vol. 6 (2010) 166–183

DOI: 10.1214/10-IMSCOLL605

Model selection error rates in nonparametric and parametric model comparisons

Yongsung Joo[*,1], **Martin T. Wells**[†,2] **and George Casella**[‡,3]

Dongguk University, Cornell University, and University of Florida

Abstract: Since the introduction of Akaike's information criteria (*AIC*) in 1973, numerous information criteria have been developed and widely used in model selection. Many papers concerning the justification of various model selection criteria followed, particularly with respect to model selection error rates (the probability of selecting a wrong model). A model selection criterion is called consistent if the model selection error rate decreases to zero as the sample size increases to infinity. Otherwise, it is inconsistent. In this paper, we explore sufficient consistency conditions for information criteria in the nonparametric (logspline) and parametric model comparison setting, and discuss finite sample model selection error rates.

Contents

1. Introduction

In past decades, there have been numerous papers addressing the consistency of model selection criteria in various settings; see [15, 16, 26, 27] and [33, 34] and

[*]Supported by KIEST grant 091-091089.
[†]Supported by National Science Foundation Grants DMS-9971586 and 06-12031 and NIH Grant R01-GM083606-01.
[‡]Supported by National Science Foundation Grants DMS-04-05543, DMS-0631632, SES-0631588 and NIH Grant 1R01GM081704.
[1]Dongguk University-Seoul, Seoul, Korea 100-715, e-mail: yongsungjoo@dongguk.edu
[2]Department of Statistical Science, Cornell University, Ithaca NY 14853, e-mail: mtw1@cornell.edu
[3]Department of Statistics, University of Florida, 102 Griffin-Floyd Hall, Gainesville, FL 32611, e-mail: casella@ufl.edu
Keywords and phrases: consistent model selection, log spline model, spline regression, nonparametric regression.
AMS 2000 subject classifications: Primary 62G20; secondary 62F99, 62G08.

the references contained therein. The classical results for finite dimensional models show that leave-n_v-out cross validation [29], BIC [39] and Bayes factors [14, 6] are consistent, while AIC [1], C_p [24], the jackknife, the bootstrap [11], and leave-one-out cross validation are asymptotically equivalent and inconsistent [29]. All of these articles, except [34] (see also [25]), assume that the number of available models (or parameters) is finite. However, in many cases, the analyst wants to include more parameters in the model as the sample size increases, when the true model is in an infinite parameter space. The logspline model is one of the richest nonparametric model families in this category [37, 38] and [20]. In this article our interest is to examine error rates of various model selection criteria for comparing nonparametric logspline models to parametric models.

Let y_i be the random variable of interest for the i^{th} observation and $\psi_{[k]}$ be the parameter in the logspline model $\mathcal{M}_k$ [37]. In this paper, subscript $[k]$ will be used as a general notation to indicate a parameter, constant or a value of model selection criterion in model $\mathcal{M}_k$. One version of the logspline model $\mathcal{M}_k$ refers to a model with the probability density function, $f_{[k]}(y_i|\psi_{[k]})$, that approximates or estimates the true probability density function of the response variable y_i, and does not contain covariates [37]. As an extension of this model, the doubly flexible logspline response model, $f_{[k]}(y_i|\psi_{[k]}(x_i;\theta_{[k]}))$, was introduced in [38] to approximate or estimate the true probability density function of y_i, $f(y_i|x_i)$, that depends on *fixed* predictor variable(s) x_i. Obviously, $f_{[k]}(y_i|\psi_{[k]})$ is a special case of $f_{[k]}(y_i|\psi_{[k]}(x_i;\theta_{[k]}))$ when $\psi_{[k]}(x_i;\theta_{[k]}) = \psi_{[k]}$. In this paper, $f_{[k]}(y_i|\psi_{[k]}(x_i;\theta_{[k]}))$ will be called the logspline model and be of our interest in model selection.

The asymptotics of this family are well studied in [37, 38] and research on other aspects are in [7, 8, 9, 10, 2, 21, 22, 23, 35], and [30]. In most of these papers, the authors propose a data-driven technique to address problems in model selection, and most use AIC [1] or BIC [39] as the evaluation criterion.

In this paper, we will discuss the consistency of model selection criteria based on the relationship among three types of models. First, consider the unknown underlying true model, $\mathcal{M}_T$, which generates the data; let θ^* and Θ^* be the parameter vector and parameter space in $\mathcal{M}_T$, respectively. Secondly, consider the candidate models $\mathcal{M}_k$, which are the models under consideration. Here we let $\theta_{[k]}$ and $\Theta_{[k]}$ be the parameter vector and the parameter space of $\mathcal{M}_k$, respectively. Here the true model is the same as or nested in one of candidate models or the true model does not have a finite parameter space, a nonparametric candidate model is often constructed based on assumed smoothness and other properties of the true model [37, 38]. As in many previous studies [5, 29, 30], we consider consistent model selection between two candidate models, $\mathcal{M}_1$ and $\mathcal{M}_2$. The third model we consider is the encompassing model $\mathcal{M}_\cup$, whose parameter vector $\theta_{[\cup]}$ consists of all parameters in candidate models [4]. Let $J_{[\cup]}$ be the dimension of $\theta_{[\cup]}$. Note that $J_{[\cup]} \geq J_{[k]}$ for any k and any n. Denote the parameter spaces of $\mathcal{M}_\cup$ by $\Theta_{[\cup]}$. The following example is provided for better understanding of this notation.

As a nonparametric candidate model, we consider the logspline model with the number of parameters $J_{[k]}$ increasing with n. Then we assume that the parameter space of $\mathcal{M}_k$ expands cumulatively with the sample size n. In other words, for any $n' > n$, a candidate model $\mathcal{M}_k$ for the sample size n is the same as or nested in $\mathcal{M}_k$ for the sample size n'. As a parametric candidate model, we consider a model that has the same probability density function as the logspline model, but with a finite and fixed number of parameters for any n. The class of generalized linear models are included in this family.

Example 1 (True, candidate, and encompassing models). Suppose that there is a small number of observations of interest from the "true" model $\mathcal{M}_T$, $y_i = \theta_0^* + \theta_1^* \exp(u_i) + \theta_2^* \sin(v_i) + \epsilon_i$ where ϵ_i is independently and identically distributed (iid) as $N(0,1)$, y_i is a response variable and $x_i = (u_i, v_i)$ is a predictor vector of the i^{th} observation. Assume that the variance of ϵ_i is known. Remember that, in this paper, $\mathcal{M}_T$ is assumed to be the same as or nested in one of candidate models. An analyst may consider two "candidate" models:

- $\mathcal{M}_1 : y_i = \theta_{0\,[1]} + \theta_{1\,[1]} u_i + \theta_{2\,[1]} u_i^2 + \theta_{3\,[1]} u_i^3 + \theta_{4\,[1]} (u_i - 1)_+^3 + \theta_{5\,[1]} (u_i - 2)_+^3 + \theta_{6\,[1]} v_i + \theta_{7\,[1]} v_i^2 + \theta_{8\,[1]} v_i^3 + \theta_{9\,[1]} (v_i - 1.5)_+^3 + \epsilon_i$, where where $(u_i - t_j)_+ = max(0, u_i - t_j)$ and t_j's are knots in the spline (nonparametric bivariate regression spline model without interaction terms, Ruppert, Wand and Carroll 2003).
- $\mathcal{M}_2 : y_i = \theta_{0\,[2]} + \theta_{1\,[2]} \exp(u_i) + \theta_{2\,[2]} \sin(v_i) + \epsilon_i$ (parametric model).

Then, the encompassing model $\mathcal{M}_\cup$ is the sum of $\mathcal{M}_1$ and $\mathcal{M}_2$. Because the intercept term is included in both $\mathcal{M}_1$ and $\mathcal{M}_2$, $J_{[1]} = 10$, $J_{[2]} = 3$ and $J_{[\cup]} = 12$.

A brief review of the logspline model is given in Section 2. In Section 3 we consider the case when a nonparametric model ($J_{[1]} \to \infty$) and a parametric model ($J_{[2]} < \infty$) are compared. In contrast, selection between parametric models ($J_{[k]} < \infty$ for k=1,2) is most frequently studied in other model selection literature ([14, 29], *etc.*). Also, we give the needed definitions and the sufficient conditions for particular classes of model selection procedures to be consistent. As applications of results in Section 3, the consistency of AIC, BIC, RIC [12], HQ [17] and leave-one-out [29] are examined in Section 3. When n is finite, the error rates of model selectors are often used to evaluate the performance of model selectors. However, in Section 4, we show that they are not sufficient by themselves because they depend on the relationship between the true model and the candidate models.

2. Log spline models

Consider a random response variable y_i with the unknown true probability density function $f(y_i|x_i)$, with fixed predictor(s) x_i. Assume that $f(y_i|x_i)$ is continuous and positive for any real numbers x_i and y_i. A logspline model, $\mathcal{M}_k$, that approximates $f(y_i|x_i)$ is defined [38] by

$$(1) \quad f_{[k]}(y_i|\psi_{[k]}(x_i;\theta_{[k]})) = \exp\left(\sum_{i=1}^{p_{[k]}} \psi_{i\,[k]}(x_i;\theta_{[k]}) B_{i\,[k]}(y_i) - c_{[k]}(\psi_{[k]}(x_i;\theta_{[k]}))\right),$$

where

$$\psi_{i\,[k]}(x_i;\theta_{[k]}) = \sum_{j=1}^{q_{i\,[k]}} \theta_{ij\,[k]} A_{ij\,[k]}(x_i),$$

$$c_{[k]}(\psi_{[k]}(x_i;\theta_{[k]})) = \log\left\{\int \exp\left(\sum_{i=1}^{p_{[k]}} \psi_{i\,[k]}(x_i;\theta_{[k]}) B_{i\,[k]}(y_i)\right) dy\right\},$$

and $A_{ij\,[k]}(x_i)$ and $B_{i\,[k]}(y_i)$ are spline basis functions. The total number of parameters for estimation is $J_{[k]} = \sum_{i=1}^{p_{[k]}} q_{i\,[k]}$. From now on, let $c_{[k]}(\theta_{[k]}) = c_{[k]}(\psi_{[k]}(x_i;\theta_{[k]}))$ for notational simplicity. [38] gives the various regularity conditions for this model and [37, 38] studies the asymptotic properties of logspline models. The basis,

$A_{ij\,[k]}(x_i)$, in the logspline model can be a multivariate spline basis. The following example shows that the normal bivariate (cubic) regression spline model [28] is a special case of a logspline model. This implies that the normal multiple regression can be also expressed with the probability density function of the logspline model.

Example 2 (Normal bivariate regression splines). Consider two predictors, u_i and v_i, and suppose the relationship between $x_i = (u_i, v_i)$ and y_i is explored with a normal bivariate regression spline ($\mathcal{M}_k$) without interaction terms,

$$y_i = \alpha_{0\,[k]} + \alpha_{1\,[k]}u_i + \alpha_{2\,[k]}u_i^2 + \alpha_{3\,[k]}u_i^3 + \sum_{j=1}^{q_{1[k]}-4} \alpha_{j+3\,[k]}(u_i - t_j)_+^3$$
$$+\beta_{1\,[k]}v_i + \beta_{2\,[k]}v_i^2 + \beta_{3\,[k]}v_i^3 + \sum_{j=1}^{q_{2[k]}-3} \beta_{j+3\,[k]}(v_i - t_j)_+^3 + \epsilon_{i\,[k]},$$

where $(u_i - t_j)_+ = max(0, u_i - t_j)$, t_j's are knots in the spline and $\epsilon_{i\,[k]} \overset{iid}{\sim} N(0, \sigma^2_{[k]})$. Here, the number of knots may or may not increase with the sample size n.

Let

$$\begin{aligned} A_{[k]}(x_i) &= (1, u_i, u_i^2, u_i^3, (u_i - t_1)_+^3, \ldots, (u_i - t_{q_{1[k]}-4})_+^3, \\ &\qquad v_i, v_i^2, v_i^3, (v_i - t_1)_+^3, \ldots, (v_i - t_{q_{2[k]}-3})_+^3), \\ \theta_{[k]} &= ({\alpha_{0\,[k]}}^T, \ldots, {\alpha_{q_{1[k]}-1\,[k]}}^T, {\beta_{1\,[k]}}^T, \ldots, {\beta_{q_{2[k]}\,[k]}}^T, \sigma^2_{[k]})^T, \\ q_{[k]} &= q_{1[k]} + q_{2[k]}. \end{aligned}$$

Also, let $A_{j[k]}(x_i)$ and $\theta_{j[k]}$ be the j-th element of $A_{[k]}(x_i)$ and $\theta_{[k]}$. The probability density function of the regression spline model is

$$f_{[k]}(y_i|\psi_{[k]}(x_i;\theta_{[k]})) = \frac{1}{\sqrt{2\pi\sigma^2_{[k]}}} \exp\left[-\frac{\{y_i - \sum_{j=1}^{q_{[k]}} \theta_{j\,[k]}A_{j\,[k]}(x_i)\}^2}{2\sigma^2_{[k]}}\right]$$
$$= \exp\left[\sum_{j=1}^{q_{[k]}} \frac{\theta_{j\,[k]}A_{j\,[k]}(x_i)}{\sigma^2_{[k]}} y_i - \frac{y_i^2}{2\sigma^2_{[k]}} - \frac{\{\sum_{j=1}^{q_{[k]}} \theta_{j\,[k]}A_{j\,[k]}(x_i)\}^2}{2\sigma^2_{[k]}} + \log\left(\frac{1}{\sqrt{2\pi\sigma^2_{[k]}}}\right)\right],$$

which has the form of the logspline model (1) with

$$\psi_{1\,[k]}(x_i;\theta_{[k]}) = \sum_{j=1}^{q_{[k]}} \frac{\theta_{j\,[k]}A_{j\,[k]}(x_i)}{\sigma^2_{[k]}}, \qquad B_{1\,[k]}(y_i) = y_i,$$
$$\psi_{2\,[k]}(x_i;\theta_{[k]}) = -\frac{1}{2\sigma^2_{[k]}}, \qquad B_{2\,[k]}(y_i) = y_i^2,$$

and

$$c_{[k]}(\theta_{[k]}) = \frac{\left\{\sum_{j=1}^{q_{[k]}} \theta_{j\,[k]}A_{j\,[k]}(x_i)\right\}^2}{2\sigma^2_{[k]}} - \log\left(\frac{1}{\sqrt{2\pi\sigma^2_{[k]}}}\right).$$

This model has $J_{[k]} = q_{[k]} + 1$ parameters. In a similar way, it can be easily shown that the normal multivariate spline model (with or without interaction terms) belongs to the logspline model.

3. Consistent model selection

In this section, we will define a general form of information criteria, $IC_{[k]}$, and find the conditions when $IC_{[k]}$ is consistent.

Define the model selection criteria $IC_{[k]}$ for model $\mathcal{M}_k$ with the sample size of n as

$$IC_{[k]} = \sup_{\theta_{[k]} \in \Theta_{[k]}} \ell_{[k]}(\theta_{[k]}) - a(n)\, J_{[k]}, \tag{2}$$

where $\Theta_{[k]}$ is the parameter space of model $\mathcal{M}_k$, $\ell_{[k]}(\theta_{[k]})$ is the log-likelihood, $a(n)$ is a positive non-decreasing function of n and $J_{[k]}$ is the number of parameters in model $\mathcal{M}_k$. In our paper, we assume $J_{[k]} = o(n^{0.5-\delta})$ for some $\delta \in (0, 0.5)$ for the convergence of the MLE [38]. As examples of (2), there are:

- $AIC_{[k]} = \sup_{\theta_{[k]} \in \Theta_{[k]}} \ell_{[k]}(\theta_{[k]}) - J_{[k]}$, which has $a(n) = 1$ [1].
- $BIC_{[k]} = \sup_{\theta_{[k]} \in \Theta_{[k]}} \ell_{[k]}(\theta_{[k]}) - \frac{\log(n)}{2} J_{[k]}$, which has $a(n) = \frac{\log(n)}{2}$ [39].
- $RIC_{[k]} = \sup_{\theta_{[k]} \in \Theta_{[k]}} \ell_{[k]}(\theta_{[k]}) - \log(J_{[\cup]})\, J_{[k]}$, which has $a(n) = \log(J_{[\cup]})$ [12].
- $HQ_{[k]} = \sup_{\theta_{[k]} \in \Theta_{[k]}} \ell_{[k]}(\theta_{[k]}) - \log(\log(n))\, J_{[k]}$, which has $a(n) = \log(\log(n))$ [17].

The supremum of the likelihood, $\sup_{\theta_{[k]} \in \Theta_{[k]}} \ell_{[k]}(\theta_{[k]})$, is a measure of how well model $\mathcal{M}_k$ fits the data and $a(n)\ J_{[k]}$ is a penalty for overfitting the model. A model that explains the data well and is parsimonious should have a high $IC_{[k]}$ value. In the comparison of two models $\mathcal{M}_1$ and $\mathcal{M}_2$, we choose $\mathcal{M}_2$ over $\mathcal{M}_1$ if $IC_{[2]} > IC_{[1]}$.

In evaluating the performance of the model selection criteria in terms of the model selection error rate, two approaches are frequently used: (1) consistency of the model selection criteria assuming a sufficiently large sample size, which we will focus on in this section, and (2) estimation of the model selection error rate using Monte Carlo simulations for small samples, which will be discussed in Section 4.

In this section, we set $\mathcal{M}_1$ to be a nonparametric model and $\mathcal{M}_2$ to be a parametric model without loss of generality. Also, we assume that the regularity condition (the σ-quasiuniform condition on the knot sequence, Stone [38]) is satisfied so that nonparametric candidate models converge to the true model. A model selection criteria is *consistent* if

$$P\,[\text{Choose the better model}] \to 1, \text{ as } n \to \infty.$$

Equivalently, if the error rate of the model selector goes to zero, then it is called a consistent model selection criterion. The following two points highlight issues of model selection.

- *Case (i) when the true model $\mathcal{M}_T$ is not nested in the parametric model $\mathcal{M}_2$:* For example, when the true regression model has an exponential curve, a cubic regression spline ($\mathcal{M}_1$) and a cubic regression ($\mathcal{M}_2$) can be considered as candidate models. Even with large n, $\mathcal{M}_2$ cannot explain the data properly, but $\mathcal{M}_1$ can approximate the true model with large n ($\mathcal{M}_1 \to \mathcal{M}_T$). Therefore, $\mathcal{M}_1$ is the better model in this case.
- *Case (ii) when the true model $\mathcal{M}_T$ is nested in the parametric model $\mathcal{M}_2$:* For example, when the true regression model has an exponential curve, a cubic

regression spline ($\mathcal{M}_1$) and a regression with an exponential curve ($\mathcal{M}_2$) can be considered as candidate models. Because $\mathcal{M}_1 \to \mathcal{M}_T$, both models will be the same as the true model with large n. Because $J_{[1]} > J_{[2]}$, $\mathcal{M}_2$ is a better model because of parsimony.

In general, $\mathcal{M}_1$ can be a multivariate spline model and $\mathcal{M}_2$ can be a multiple regression model. The discussions in this paper are applicable for comparisons of multivariate splines and multiple regressions. Aside from Case (i) and (ii), it is difficult to discuss consistency because it is not clear which candidate model is better than the other. Similar arguments have appeared in many other papers to prove consistency when the number of parameters is finite [5, 29, 30, 32] and references contained therein). The consistency conditions for these two cases are given in the following theorem.

Theorem 1. *Let $y_1, \ldots, y_n$ be iid random variables from the logspline family (1). Also let $J_{[1]}$ and $J_{[2]}$ be the number of parameters to be estimated in a nonparametric logspline model $\mathcal{M}_1$ and a parametric model $\mathcal{M}_2$. A model selection criterion, $IC_{[k]}$, is consistent if*

$$J_{[1]} = o(n^{0.5-\delta}) \text{ for some } \delta \in (0, 0.5), \quad \frac{a(n)J_{[1]}}{n} \to 0 \text{ and } a(n) \to \infty,$$

as $n \to \infty$.

Proof. The proof is summarized as follows (Appendix A presents a detailed proof). First of all, $J_{[1]} = o(n^{0.5-\delta})$ for some $\delta \in (0, 0.5)$ is needed for the MLE convergence in $\mathcal{M}_1$ [37, 38]. In Case (i), a model selection criterion chooses the better model $\mathcal{M}_1$ consistently if

$$\frac{a(n)J_{[1]}}{n} \to 0.$$

In Case (ii), consistency requires $a(n) \to \infty$. □

In addition to the nonparametric and parametric model comparisons there are two other possible cases of model comparisons-'parametric vs parametric' and 'nonparametric vs nonparametric' model comparisons. Remarks 1 – 3 discuss these comparisons as well as applications of Theorem 1.

Remark 1 (Parametric vs. parametric models). Consider a situation when both candidate models have a finite number of parameters for any sample size (i.e. comparison between linear and quadratic regression models or comparison between a regression model with two predictors, x_1 and x_2, and a model with three predictors, x_1, x_2 and x_3). By setting $J_{[k]} < \infty$ for $k = 1, 2$, the consistency conditions in Theorem 1 may degenerate to

$$\frac{a(n)}{n} \to 0 \text{ and } a(n) \to \infty, \tag{3}$$

which are given in many other papers concerning parametric model comparisons (for example; [5, 29, 31]). For this case, BIC and HQ are consistent, but AIC and RIC are not.

Remark 2 (Nonparametric vs. parametric models). Suppose that one candidate is a nonparametric model and the other is parametric. Then, Theorem 1 shows that BIC, RIC and HQ can be consistent depending on $J_{[1]}$, whereas AIC is inconsistent.

Remark 3 (Nonparametric vs. nonparametric models). Suppose that the candidates are two nonparametric models. A typical example is the knot selection problem ([18, 13] and discussions therein), in this setting there are two nonparametric models with different knots. Where asymptotic model selection error rates are concerned, it becomes a relatively simple and less interesting problem. If the true model has an infinite dimension in terms of the bases, and the number of knots increases properly (satisfying the σ-quasiuniform condition on the knot sequence, [38]), both of the nonparametric candidate models will converge to the true model. Then, any model selection criterion is consistent because any chosen candidate model is asymptotically equivalent to the true model. If the true model is of infinite dimension and the number of knots is fixed, none of candidate models is correct. Then, any model selection criterion has a model selection error rate of 100%. Therefore, the model selection error rate is not a very interesting in the knot selection setting. Typically, knot selection is used to gain a better prediction error. There has been much research that favors AIC-type criteria (AIC and criteria that are asymptotically equivalent to AIC) in terms of the prediction error [40]. Hence, we recommend AIC-type model selection criteria for knot selection problems. Also, see [38] for detailed discussions on the convergence rates of the logspline models.

When the sample size n is infinite, both candidates are equivalent to the true model. Also, it is not practically meaningful to select the better model based on parsimony, because both candidates have infinite numbers of parameters with a large n. Therefore, nonparametric models are better compared based on the convergence rates of nonparametric models as $n \to \infty$. See [38] for detailed discussions on the convergence rates of the logspline models.

It is known that AIC, C_p, jackknife, bootstrap [11] and leave-one-out cross validation are asymptotically equivalent and inconsistent when only parametric models ($J_{[k]} < \infty$) are considered as candidates [29]. As another applications of Theorem 1, Corollary 1 shows the inconsistency of the leave-one-out cross validation ($CV(1)$) for nonparametric vs. parametric model comparisons. $CV(1)$ is defined as

$$\hat{\Gamma}_{[k]}^{CV(1)} = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - X_{i\,[k]}\hat{\theta}_{[k]}^{(i)}\right]^2,$$

where $\hat{\theta}_{[k]}^{(i)}$ is the MLE of $\theta_{[k]}$ without the i^{th} observation. The following Corollary can be established and is proved in Appendix B.

Corollary 1. *In the comparison of a regression spline model ($\mathcal{M}_1$) and a parametric regression model ($\mathcal{M}_2$), the leave-one-out cross validation $CV(1)$ is inconsistent.*

4. Finite Sample model selection error rates

The purpose of this section is to explore error rates of model selection criteria $IC_{[k]}$'s when the sample size is finite. Using simulation studies, we demonstrate that there is no clear-cut preferred model selection criterion with respect to model selection error rates.

4.1. A simulation study with two candidate models

Consider two candidate models:

1. $\mathcal{M}_1$ (the cubic regression spline with two equally-spaced knots)
2. $\mathcal{M}_2$ (the quadratic regression model).

Model $\mathcal{M}_1$ is a flexible nonparametric model that, with large n, can approximate any true regression model. Also, note that $\mathcal{M}_2$ is nested in $\mathcal{M}_1$. Define $\ell_{[1]}$ and $\ell_{[2]}$ as the maximum log-likelihoods for $\mathcal{M}_1$ and $\mathcal{M}_2$, and $J_{[1]}$ and $J_{[2]}$ as the numbers of parameters in these models. For the simulation studies in Table 1, data sets are generated 10,000 times from each true model with the sample sizes $n = 50$ and 100.

Stone [38] shows that the global optimal convergence rate is achieved by setting the number of parameters equal to

$$J_{[k]} \sim n^{\frac{p_{[k]}}{p_{[k]}+q_{[k]}+2p_{[k]}q_{[k]}}},$$

where $a_n \sim b_n$ means that a_n/b_n is bounded away from 0 and infinity. Here, $p_{[k]}$ and $q_{[k]}$ are the number of spline bases, $B_{i\,[k]}(y_i)$ and $A_{ij\,[k]}(x_i)$, as defined in (1). Typically, $p_{[k]}$ and $q_{[k]}$ are the assumed smoothness, which is how many times the function $f_{[k]}(\cdot)$ is differentiable with y_i and x_i, respectively. When the normal distribution is assumed, $p_{[k]} = \infty$ because $f_{[k]}(\cdot)$ is infinitely differentiable with y_i. Therefore,

$$J_{[k]} \sim n^{\frac{q_{[k]}}{1+2q_{[k]}}}.$$

In this example, suppose that we use $q_{[k]} = 2$ for $\mathcal{M}_1$. Then, a reasonable rule is to take the number of knots equal to the closest integer less than $n^{1/5}$. Here, $J_{[k]} = 2$ for both $n = 50$ and 100. Even though there may be a better way of selecting the number of knots, we chose a slowly increasing function of n for $J_{[k]}$ in this simulation study to make the interpretation of simulation study results easier.

The first true model is

$$\mathcal{M}_{T1} : y_i = 1 + \sin(x_i) + 3\cos(x_i) + 4\log(x_i) + \epsilon_i,$$

where $\epsilon_i \overset{iid}{\sim} N(0, 0.15^2)$. Because $\mathcal{M}_{T1}$ is not nested in $\mathcal{M}_2$, this can be an example of Case (i). The predictor vector $x = (x_1, \ldots, x_n)^T$ is constructed with n equally spaced real numbers within a given range. For example, when $x_i \in [1, 3]$, $x = (1, 1 + 1/(n-1), \ldots, 3)^T$. This true model has an infinite dimensional parameter space in terms of regression spline bases. Because $\mathcal{M}_1$ is a nonparametric model, of which the number of parameters increases with n, $\mathcal{M}_1$ can fit the data with a large n as good as the true model $\mathcal{M}_{T1}$ does. But $\mathcal{M}_2$ cannot. Even with a finite n, $\mathcal{M}_1$ can fit a complicated trend in $\mathcal{M}_{T1}$ better than $\mathcal{M}_2$ can. Therefore, $\mathcal{M}_1$ is considered as the better model in this case.

Because

$$\ell_{[1]} - a(n)J_{[1]} < \ell_{[2]} - a(n)J_{[2]} \Leftrightarrow \ell_{[1]} - \ell_{[2]} < 3a(n),$$

the model selection error rate is

$$P(\ell_{[1]} - \ell_{[2]} < 3a(n)). \tag{4}$$

Here, the magnitude of $a(n)$'s at each fixed n determines the error rates of $IC_{[k]}$'s. For example, for $n = 50$ or 100, we have

$$a^{AIC}(n) < a^{HQ}(n) < a^{RIC}(n) < a^{BIC}(n),$$

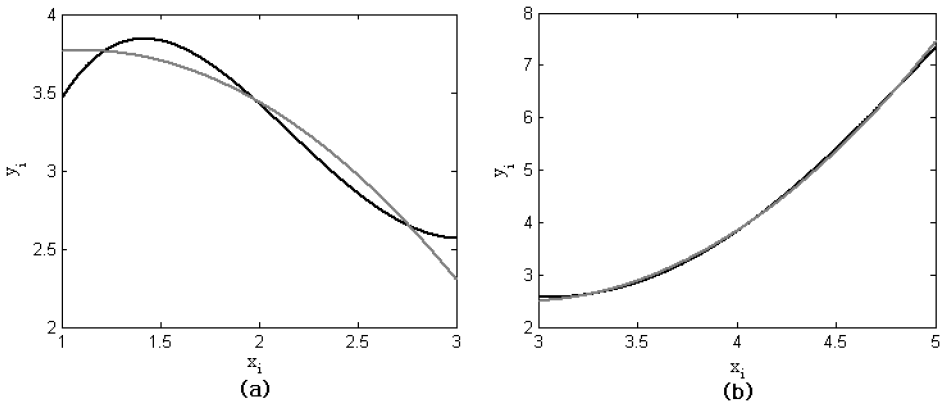

FIG 1. *The true mean function* $y_i = 1 + sin(x_i) + 3cos(x_i) + 4log(x_i)$ *(dark-colored line) and the closest quadratic function (light-colored line) when (a)* $x_i \in [1, 3]$ *and (b)* $x_i \in [3, 5]$.

where

$$\begin{aligned} a^{AIC}(n) &= 1, \\ a^{HQ}(n) &= \frac{3}{2}\log(\log(n)), \\ a^{RIC}(n) &= \log(J_{[\cup]}) = \log(J_{[1]}), \\ a^{BIC}(n) &= \log(n)/2. \end{aligned}$$

Then, the error rate also increases in the order of AIC, HQ, RIC and BIC. Different error rates of $IC_{[k]}$'s are caused by a choice of $a(n)$ or rejection regions of the test.

Figure 1 shows the mean function of $\mathcal{M}_{T1}$ and the closest quadratic function that minimizes

$$\int |(1 + sin(x_i) + 3cos(x_i) + 4log(x_i)) - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)|\, dx_i. \tag{5}$$

The true mean function is closer to the quadratic function when $x_i \in [3, 5]$ (Figure 1-(b)) than when $x_i \in [1, 3]$ (Figure 1-(a)). Therefore, in simulation studies, model selection criteria are expected to have a higher model selection error rate when data are simulated with $x_i \in [3, 5]$ than with $x_i \in [1, 3]$, selecting quadratic model $\mathcal{M}_2$ more often.

Table 1 reports the rates of choosing each candidate model. For example, AIC chooses the spline model $\mathcal{M}_1$ with probability 0.684 when 100 observations are simulated from $\mathcal{M}_{T1}$ with $x_i \in [3, 5]$. Because $\mathcal{M}_1$ is considered as the better model, 0.684 is one minus the model error rate or a successful model selection rate. As expected, the overall performance of the model selection criteria in Table 1 is better when the data are simulated from $\mathcal{M}_{T1}$ with $x_i \in [1, 3]$ than with $x_i \in [3, 5]$. Also, the model selection error rates also increase in the order of AIC, HQ, RIC and BIC. Note that AIC is the best model selection criterion because a small $a^{AIC}(n)$ makes AIC choose a larger model $\mathcal{M}_1$ with higher probability. As the sample size increases, the error rates of all model selection criteria are reduced.

TABLE 1

The rate of choosing either quadratic regression (Quad) or regression spline (Spline) model. The number of observations simulated from the true model is n, and a(n) are the respective penalty terms. ASE is the average squared error defined at (6)

Selector			True Model: Spline $= \mathcal{M}_{T1}^{\dagger}$					
			$x_i \in [1,3]$			$x_i \in [3,5]$		
	n	**a(n)**	**Spline**$^{\sharp}$	**Quad***	***ASE***	**Spline**	**Quad**	***ASE***
AIC	50	1.00	0.999	0.001	0.003	0.489	0.511	0.003
BIC	50	1.96	0.976	0.024	0.003	0.137	0.863	0.003
HQ	50	1.36	0.997	0.003	0.003	0.317	0.683	0.003
RIC	50	1.79	0.986	0.014	0.003	0.175	0.825	0.003
AIC	100	1.00	1.000	0.000	0.001	0.684	0.316	0.002
BIC	100	2.30	1.000	0.000	0.001	0.192	0.808	0.002
HQ	100	1.53	1.000	0.000	0.001	0.453	0.547	0.002
RIC	100	1.79	1.000	0.000	0.001	0.349	0.651	0.002

Selector			True Model: Quad $= \mathcal{M}_{T2}^{\ddagger}$		
			$x_i \in [3,5]$		
	n	**a(n)**	**Spline**	**Quad**	***ASE***
AIC	50	1.00	0.148	0.852	0.082
BIC	50	1.96	0.015	0.985	0.064
HQ	50	1.36	0.063	0.937	0.071
RIC	50	1.79	0.022	0.978	0.065
AIC	100	0.129	0.871	0.040	
BIC	100	2.30	0.004	0.996	0.031
HQ	100	1.53	0.034	0.966	0.034
RIC	100	1.79	0.017	0.983	0.032

† $\mathcal{M}_{T1} : y_i = 1 + \sin(x_i) + 3\cos(x_i) + 4\log(x_i) + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0, 0.15^2)$.

‡ $\mathcal{M}_{T2} : y_i = 1 + x_i + x_i^2 + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0,1)$.

$^{\sharp}$ Probability of selecting the spline model $\mathcal{M}_1 : y_i = \sum_{j=0}^{3} \beta_{j\,[1]} x_i^j + \beta_{4\,[1]}(x_i - t_1)_+^3 + \beta_{5\,[1]}(x_i - t_2)_+^3 + \epsilon_{i\,[\cup]}$, where t_1 and t_2 are equally spaced knots within the range of x_i.

* Probability of selecting the quadratic model $\mathcal{M}_2 : y_i = \beta_{0\,[2]} + \beta_{1\,[2]} x_i + \beta_{2\,[2]} x_i^2 + \epsilon_{i\,[2]}$.

Now consider the quadratic model $\mathcal{M}_{T2}$: $y_i = 1 + x_i + x_i^2 + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0,1)$, as the true model that will be used in generating data. In this case, $\mathcal{M}_2$ is the better model because of parsimony. Because $\mathcal{M}_{T2}$ is nested in $\mathcal{M}_2$, this can be an example of Case (ii). Similar patterns are observed as in the previous simulations with $\mathcal{M}_{T1}$, except that the order of the $IC_{[k]}$'s is reversed for error rates (in the second last column of Table 1). Error rates (ASE) increase in the order of BIC, RIC, HQ and AIC when $n = 50$ or 100. Note that BIC is the best model selection criterion because a large $a^{BIC}(n)$ makes BIC choose a smaller model $\mathcal{M}_2$ with a higher probability.

The simulation results can be summarized as follows. When two candidate models $\mathcal{M}_1$ and $\mathcal{M}_2$ are considered, the magnitudes of $a(n)$'s determine which model selection criterion performs best in terms of model selection error rates. When the true model is not nested in $\mathcal{M}_2$, the $IC_{[k]}$ with the smallest $a(n)$ is the best for any true model and any sample size. When the true model is nested in $\mathcal{M}_2$, the $IC_{[k]}$ with the largest $a(n)$ is the best for any true model and any sample size. If more than two candidates are compared and the true model is neither the smallest or the largest model, the closeness (5) between the true and candidate models becomes another important factor in determining the order of the model selection criterion in terms of error rates.

4.2. A simulation study with three candidate models

In many papers (for example; [29, 30]), simulation studies consider more than two candidate models. As an example of this subsection, consider simulated data from $\mathcal{M}_{T1}$ and three competing candidate models-$\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$. Candidate models $\mathcal{M}_1$ and $\mathcal{M}_2$ are the same as defined in the previous simulation studies and $\mathcal{M}_3$ is the candidate model with exactly the same parametrization as the true model $\mathcal{M}_{T1}$: $y_i = \beta_0 + \beta_1 sin(x_i) + \beta_2 cos(x_i) + \beta_3 \log(x_i) + \epsilon_i$. This will be called the exact model, distinguishing from the true model with known parameter values. Obviously, $\mathcal{M}_3$ is the best candidate model in this case. Candidates $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$, have 6, 3 and 4 parameters in their mean functions, respectively. Simulation studies for the sample size 50 and 100 are conducted with these models and results are given in Table 2.

As discussed previously, the true mean function is close to the quadratic function when $x_i \in [3, 5]$ (Figure1-(b)). This makes the competition between the quadratic model $\mathcal{M}_2$ and the exact model $\mathcal{M}_3$ tense when $x_i \in [3, 5]$. Although the spline model $\mathcal{M}_1$ can also generate a mean function as $\mathcal{M}_2$ does, $\mathcal{M}_1$ has a higher number of parameters, which is penalized by $a(n)$ in $IC_{[k]}$. In this case, the $IC_{[k]}$ with a small $a(n)$ may perform better because a small $a(n)$ makes $IC_{[k]}$ choose $\mathcal{M}_3$ with a large number of parameters instead of $\mathcal{M}_2$. In Table 2, AIC has the lowest model selection error rate 0.426(=0.147+0.279) and 0.261(=0.146+0.115), or equivalently

TABLE 2

The rate of choosing either the quadratic regression (Quad), exact (Exact) or regression spline (Spline) model: True model is $y_i = 1 + \sin(x_i) + 3\cos(x_i) + 4\log(x_i) + \epsilon_i$ *where* $\epsilon_i \overset{iid}{\sim} N(0, 0.15^2)$. *The number of observations simulated from the true model is n, and a(n) are the respective penalty terms. ASE is the average squared error defined at (6)*

Selector			$\boldsymbol{x_i \in [1,3]}$			
	n	**a(n)**	**Spline**$^\sharp$	**Quad***	**Exact**†	***ASE***
AIC	50	1.00	0.172	0.000	0.828	0.002
BIC	50	1.96	0.031	0.001	0.969	0.002
HQ	50	1.36	0.091	0.000	0.909	0.002
RIC	50	1.79	0.041	0.000	0.959	0.002
AIC	100	1.00	0.147	0.000	0.854	0.001
BIC	100	2.30	0.012	0.000	0.988	0.001
HQ	100	1.53	0.053	0.000	0.947	0.001
RIC	100	1.79	0.033	0.000	0.967	0.001

Selector			$\boldsymbol{x_i \in [3,5]}$			
	n	**a(n)**	**Spline**	**Quad**	**Exact**	***ASE***
AIC	50	1.00	0.147	0.279	0.574	0.002
BIC	50	1.96	0.023	0.531	0.446	0.002
HQ	50	1.36	0.071	0.387	0.542	0.002
RIC	50	1.79	0.032	0.496	0.471	0.002
AIC	100	1.00	0.146	0.115	0.739	0.001
BIC	100	2.30	0.011	0.350	0.640	0.001
HQ	100	1.53	0.050	0.207	0.740	0.001
RIC	100	1.79	0.028	0.256	0.716	0.001

$^\sharp$ Probability of selecting the spline model $\mathcal{M}_1$: $y_i = \sum_{j=0}^{3} \beta_{j\,[1]} x_i^j + \beta_{4\,[1]}(x_i - t_1)_+^3 + \beta_{5\,[1]}(x_i - t_2)_+^3 + \epsilon_{i\,[\cup]}$, where t_1 and t_2 are equally spaced knots within the range of x_i.

* Probability of selecting the quadratic model $\mathcal{M}_2$: $y_i = \beta_{0\,[2]} + \beta_{1\,[2]} x_i + \beta_{2\,[2]} x_i^2 + \epsilon_{i\,[2]}$.

† Probability of selecting the exact model $\mathcal{M}_3$: $y_i = \beta_0^{\mathcal{M}_3} + \beta_1^{\mathcal{M}_3} \sin(x_i) + \beta_2^{\mathcal{M}_3} \cos(x_i) + \beta_3^{\mathcal{M}_3} \log(x_i) + \epsilon_i^{\mathcal{M}_3}$.

the highest successful model selection rate 0.574 and 0.739 for $n = 50$ and 100. When $x_i \in [1, 3]$, the true mean function is relatively far from the quadratic function (Figure1-(a)). Therefore, the $IC_{[k]}$ can determine easily that $\mathcal{M}_3$ is better than $\mathcal{M}_2$. Table 2 shows that all model selection criteria choose $\mathcal{M}_2$ with probability zero or very close to zero probability. Because the spline model $\mathcal{M}_1$ is more flexible and can fit the data better than $\mathcal{M}_2$, the competition between $\mathcal{M}_1$ and $\mathcal{M}_3$ is a little more tense than between $\mathcal{M}_2$ and $\mathcal{M}_3$. Therefore, $IC_{[k]}$ with higher $a(n)$ should perform better penalizing a high number of parameters in $\mathcal{M}_1$. Because of the highest $a(n)$ value, Table 2 shows that BIC has the lowest model selection error rates 0.031 and 0.012 (or the highest successful model selection rates 0.969 and 0.988) with $n = 50$ and 100. This simulation study shows that the closeness of the true model and candidates is an important factor that controls the model selection error rates. By controlling the closeness of $\mathcal{M}_{T1}$ and $\mathcal{M}_2$, we may also generate examples that has HQ or RIC as the best model selection criterion for small samples.

The previous examples demonstrated that, even though the error rate has been used as an important part of evaluation of the model selection criterion in many papers [5, 19, 29, 30, 41], it is not sufficient by itself to show which model selection criterion is better than others. Therefore, it is necessary for researchers to choose examples very carefully and state the limit of the simulation study for model selection error rates.

In addition to model selection error rates, the average squared error (ASE) is calculated in Table 1 and 2 for each model selection criterion. When a model is chosen using a selection criterion δ, let

$$ASE_\delta = n^{-1} \sum_{i=1}^{n} \{\mu^*(x_i) - \mu_\delta(x_i, \hat{\theta}_\delta)\}^2, \tag{6}$$

where $\mu^*(\cdot)$ is the mean function of the true regression model and $\mu_\delta(\cdot)$ is the mean function of a regression model that is chosen by criterion δ (*i.e.* AIC, BIC, etc). Yang (2005) notes that ASE corresponds to the risk, $R(\delta) = n^{-1} \sum_{i=1}^{n} \{\mu(x_i) - \mu_\delta(x_i, \hat{\theta}_\delta)\}^2$, and shows that AIC makes the minimax risk converge to zero with large n and BIC makes the minimax risk converge to a nonzero constant. This means that, even though $\mu_{BIC}(\cdot)$ converges to $\mu^*(\cdot)$ because of the consistency of the model selector and point estimator, this convergence is not fast enough to make $\lim_{n\to\infty} ASE_{BIC} \to 0$. Even though ASE_{BIC} was usually larger than ASE_{AIC} in our simulation studies, ASE_{BIC} was very small with large n because BIC selects the exact or the better model in almost 100% of the cases, and point estimates of this model converge reasonably fast ($O_p(\sqrt{J/n})$, Theorem 1 in [38]).

5. Discussion

Many new model selection criteria have been developed in past decades, compared with other criteria based on model selection error rates. While many other papers are interested in comparing parametric models, this paper discusses error rates when nonparametric and parametric models are compared. First, consistency conditions of model selection criteria are provided for nonparametric and parametric model comparisons with a large n. When the number of parameters in the nonparametric model is forced to be finite, these conditions may reduce to the conventional consistency conditions, which shows the smooth connection between our and conventional

results. Second, with a small n, error rates are compared using simulation studies. Model selection error rates have been used as one of most important measures in comparing model selection criteria. It is shown that the error rate may not provide strong evidence of the best model selection criterion by itself, because it varies depending on the candidate models and true model.

There have been many studies on the probability of selecting the correct model and the prediction error, particularly, comparing AIC-type and BIC-type model selection criteria [29, 30, 31, 14]. Originally, AIC was derived to minimize the Kullback-Leibler distance between the true model and the estimated candidate model [1] and BIC was derived as an approximation of the posterior model probability [39]. Therefore, we can easily expect that AIC will perform better than BIC in terms of prediction and BIC will perform better than AIC in terms of model selection error rates. Many results match with this expectation [3, 40], and [14]. Therefore, if someone is interested in finding the correct parameter space of the true model, the probability of selecting the correct model should be used in evaluating the performance of a model selection criterion. Then, BIC-type criteria will be preferred because it consistent. However, if good prediction is the goal, the prediction error should be used in evaluating the performance of a model selection criterion. In terms of prediction error, AIC is known as a favorable model selection criterion.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *the Second International Symposium on Information Theory* (B. N. Petrov and F. Czaki, eds.) 267–281. Akademiai Kiado, Budapest.

[2] Barron, A. R. and Sheu, C. (1991). Approximation of density functions by sequences of exponential families. *Ann. Statist.* **19** 1347–1369.

[3] Barron, A. R., Birge, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413.

[4] Berger, J. and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91** 109–122.

[5] Bozdogan, H. (1987). Model selection and Akaike's information criterion: General theory and its analytical extensions. *Psychometrika* **52** 345–370.

[6] Casella, G., Giron, F. J. and Moreno, E. (2009). Consistent model selection in regression. *Ann. Statist.* **37** 1207–1228.

[7] Crain, B. R. (1974). Estimation of distributions using orthogonal expansions. *Ann. Statist.* **2** 454–463.

[8] Crain, B. R. (1976). Exponential models, maximum likelihood estimation and the Haar conditions. *J. Amer. Statist. Assoc.* **71** 737–740.

[9] Crain, B. R. (1976). More on estimation of distributions using orthogonal expansions. *J. Amer. Statist. Assoc.* **71** 741–745.

[10] Crain, B. R. (1977). An information theoretic approach to approximating a probability distribution. *SIAM J. Appl. Math.* **32** 339–346.

[11] Efron, B. (1983). Estimating error rate of a prediction rule: Improvement on cross validation. *J. Amer. Statist. Assoc.* **78** 316–331.

[12] Foster, D. and George, E. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975.

[13] Friedman, J. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19** 1–67.

[14] Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *J. Roy. Statist. Soc. Ser. B* **56** 501–514.
[15] Haughton, D. (1988). On the choice of a model fit from an exponential family. *Ann. Statist.* **16** 190–195.
[16] Haughton, D. (1994). Consistency of a class of information criteria for model selection in non linear regression. *Theory Probab. Appl.* **37** 47–53.
[17] Hannan, E. P. and Quinn, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B* **41** 190–195.
[18] Hastie, T. (1989). Flexible parsimonious smoothing and additive modeling: Discussion. *Technometrics* **31** 23–29.
[19] Hurvich, C., Shumway, R. and Tsai, C. (1990). Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples. *Biometrika* **77** 709–719.
[20] Kooperberg, C. and Stone, C. (1991). A study of logspline density estimation. *Comput. Statist. Data Anal.* **12** 327–347.
[21] Kooperberg, C. and Stone, C. (1992). Logspline density estimation for censored data. *J. Computat. Graph. Statist.* **1** 301–328.
[22] Kooperberg, C., Stone, C. and Truong, Y. K. (1995). Hazard regression. *J. Amer. Statist. Assoc.* **90** 78–94.
[23] Leonard, T. (1978). Density estimation, stochastic processes and prior information (with discussion). *J. Roy. Statist. Soc. Ser. B* **40** 113–146.
[24] Mallow, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.
[25] Moreno, E., Giron, F. J. and Casella, G. (2010). Consistency of objective Bayes factors as the model dimension grows. *Ann. Statist.* To appear.
[26] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12** 758–765.
[27] Potscher, B. M. (1989). Model selection under nonstationarity: Autoregressive models and stochastic linear regression. *Ann. Statist.* **17** 1257–1274.
[28] Ruppert, D., Wand, M. P. and Caroll, R. J. (2003). *Semiparametric Regression.* Cambridge Univ. Press, New York.
[29] Shao, P. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88** 486–494.
[30] Shao, P. (1996). Bootstrap model selection. *J. Amer. Statist. Assoc.* **91** 655–665.
[31] Shao, P. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7** 221–264.
[32] Shao, P. and Rao, S. (2000). The GIC for model selection: A hypothesis test approach. *J. Statist. Plan. Inf.* **88** 215–231.
[33] Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63** 114–126.
[34] Shibata, R. (1981). An optimal selection of regression variables. *Bimetrika* **68** 45–54.
[35] Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810.
[36] Strawderman, R. L. and Tsiatis, A. A. (1996). On the asymptotic properties of a flexible hazard estimator. *Ann. Statist.* **24** 41–63.
[37] Stone, C. (1990). Large sample inference for log-Spline models. *Ann. Statist.* **18** 717–741.
[38] Stone, C. (1991). Asymptotics for doubly flexible logspline response models. *Ann. Statist.* **19** 1832–1854.

[39] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.
[40] Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* **92** 937–950.
[41] Zheng, X. and Loh, W. (1995). Consistent variable selection in linear models. *J. Amer. Statist. Assoc.* **90** 151–156.

Appendix A: Proof of Theorem 1

The following arguments are based on the assumption that MLE converges ($J_{[1]} = o(n^{0.5-\delta})$ for some $\delta \in (0, 0.5)$, [37, 38]).

First, suppose that true model $\mathcal{M}_T$ is not nested in parametric model $\mathcal{M}_2$ as in Case (i). As $n \to \infty$, for any true model $\mathcal{M}_T$, $\mathcal{M}_1$ converges to $\mathcal{M}_T$ ($\mathcal{M}_1 \to \mathcal{M}_T$). Therefore, $\mathcal{M}_1$ is the better model in this case.

$$
\begin{aligned}
&P[\text{selecting the better model}] \\
&= P\left[IC_{[1]} > IC_{[2]}\right] \\
&= P\left[\sup_{\theta_{[1]}\in\Theta_{[1]}} \ell_{[1]}(\theta_{[1]}) - a(n)J_{[1]} > \sup_{\theta_{[2]}\in\Theta_{[2]}} \ell_{[2]}(\theta_{[2]}) - a(n)J_{[2]}\right] \\
&= P\left[\sup_{\theta_{[1]}\in\Theta_{[1]}} \ell_{[1]}(\theta_{[1]}) - \sup_{\theta_{[2]}\in\Theta_{[2]}} \ell_{[2]}(\theta_{[2]}) > a(n)(J_{[1]} - J_{[2]})\right] \\
&= P\left[\sup_{\theta_{[1]}\in\Theta_{[1]}} \left[\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p_{[1]}} \psi_{j\,[1]}(x_i;\theta_{[1]})B_{j\,[1]}(y_i) - c_{[1]}(\theta_{[1]})\right\}\right]\right. \\
&\quad - \sup_{\theta_{[2]}\in\Theta_{[2]}} \left[\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p_{[2]}} \psi_{j\,[2]}(x_i;\theta_{[2]})B_{j\,[2]}(y_i) - c_{[2]}(\theta_{[2]})\right\}\right] \\
&\qquad\qquad \left. > \frac{1}{n}a(n)(J_{[1]} - J_{[2]})\right]
\end{aligned}
$$

In order to show that this probability goes to 1, we need to know the convergence of

$$
\sup_{\theta_{[1]}\in\Theta_{[1]}} \left[\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p_{[1]}} \psi_{j\,[1]}(x_i;\theta_{[1]})B_{j\,[1]}(y_i) - c_{[1]}(\theta_{[1]})\right\}\right] \qquad \text{for} \quad k = 1, 2.
$$

Let $\hat{\theta}_{[1]}$ be the MLE of the parameter $\theta_{[1]}$ in model $\mathcal{M}_1$. By the uniqueness of

the maximum likelihood estimate [38] we have that

$$\begin{aligned}
&\sup_{\theta_{[1]}\in\Theta_{[1]}}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p_{[1]}}\left(\psi_{j\,[1]}(x_i;\theta_{[1]})B_{j\,[1]}(y_i)-c_{[1]}(\theta_{[1]})\right)\right\}\right]\\
&\quad=\quad\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p_{[1]}}\left(\psi_{j\,[1]}(x_i;\hat{\theta}_{[1]})B_{j\,[1]}(y_i)-c_{[1]}(\hat{\theta}_{[1]})\right)\right\}\right]\\
&\quad\to\quad\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p_{[1]}}\left(\psi_{j\,[1]}(x_i;\theta^{*}_{[1]})B_{j\,[1]}(y_i)-c_{[1]}(\theta^{*}_{[1]})\right)\right\}\right],
\end{aligned}$$

where $\psi_{j\,[1]}(x_i;\hat{\theta}_{[1]})\to\psi_{j\,[1]}(x_i;\theta^{*}_{[1]})$. Then, by the weak law of large numbers,

$$\begin{aligned}
&\frac{1}{n}\,\Delta\ell(n)\\
&\overset{def}{=}\sup_{\theta_{[1]}\in\Theta_{[1]}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{p_{[1]}}\psi_{j\,[1]}(x_i;\theta)B_{j\,[1]}(y_i)-c_{[1]}(\theta_{[1]})\right]\right\}\\
&\quad-\sup_{\theta_{[2]}\in\Theta_{[2]}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{p_{[2]}}\psi_{j\,[2]}(x_i;\theta_{[2]})B_{j\,[2]}(y_i)-c_{[2]}(\theta_{[2]})\right]\right\}\\
&\to E\left[\sum_{j=1}^{p_{[1]}}\psi_{j\,[1]}(x;\theta^{*}_{[1]})B_{j\,[1]}(y)-c_{[1]}(\theta^{*}_{[1]})\right]\\
&\quad-E\left[\sum_{j=1}^{p_{[2]}}\psi_{j\,[2]}(x;\theta^{*}_{[2]})B_{j\,[2]}(y)-c_{[2]}(\theta^{*}_{[2]})\right]\\
&>0.
\end{aligned}$$

Hence,

$$P[\text{selecting the better model}]=P\left[\frac{1}{n}\,\Delta\ell(n)-\frac{a(n)(J_{[1]}-J_{[2]})}{n}>0\right]\to 1,$$

if

$$\frac{a(n)(J_{[1]}-J_{[2]})}{n}\to 0\Leftrightarrow\frac{a(n)J_{[1]}}{n}\to 0\text{ as }n\to\infty.$$

Next, suppose that true model $\mathcal{M}_T$ is nested in parametric model $\mathcal{M}_2$ as in Case (*ii*). Also, remind that $J_{[1]}>J_{[2]}$ for any large n. Even though $\mathcal{M}_1\to\mathcal{M}_T$, $\mathcal{M}_2$ is considered as the better model because of parsimony. Assume $\psi_{j\,[\cup]}(x_i;\hat{\theta}_{[\cup]})$ converge to $\psi_{j\,[\cup]}(x_i;\theta^{*\,[\cup]})$. Because $\Theta_{[k]}\subset\Theta_{[\cup]}(n)$ for k=1 and 2,

$$\begin{aligned}
&\sup_{\theta_{[k]}\in\Theta_{[k]}}\left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{p_{[k]}}\psi_{j\,[k]}(x_i;\theta_{[k]})B_{j\,[k]}(y_i)-c_{[k]}(\theta_{[k]})\right\}\right]\\
&\le\sup_{\theta_{[\cup]}\in\Theta_{[\cup]}}\left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{p_{[\cup]}(n)}\psi_{j\,[\cup]}(x_i;\theta_{[\cup]})B_{j\,[\cup]}(y_i)-c_{[\cup]}(\theta_{[\cup]})\right\}\right]
\end{aligned}$$

and

$$\sup_{\theta_{[k]}\in\Theta_{[k]}}\left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{p_{[k]}}\psi_{j\,[k]}(x_i;\theta_{[k]})B_{j\,[k]}(y_i)-c_{[k]}(\theta_{[k]})\right\}\right]$$
$$-\left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{p_{[\cup]}(n)}\psi_{j\,[\cup]}(x_i;\theta^{*\,[\cup]})B_{j\,[\cup]}(y_i)-c_{[\cup]}(\theta^{*\,[\cup]})\right\}\right]\geq 0.$$

Let

$$B_{[\cup]}(y_i)=(B_{1\,[\cup]}(y_i),\ldots,B_{p_{[\cup]}(n)}(y_i))^T,$$
$$\psi_{[\cup]}(x;\hat{\theta}_{[\cup]}-\theta^{*\,[\cup]})=(\psi_{1\,[\cup]}(x;\hat{\theta}_{[\cup]}-\theta^{*\,[\cup]}),\ldots,\psi_{p_{[\cup]}(n)}(x;\hat{\theta}_{[\cup]}-\theta^{*\,[\cup]}))^T,$$

and

$$\bigtriangledown c_{[\cup]}(\theta^{*\,[\cup]})=\left[\frac{dc_{[\cup]}(\theta_{[\cup]})}{d\theta_{[\cup]}}\right]_{\theta_{[\cup]}=\theta^{*\,[\cup]}}.$$

Then,

$$\sup_{\theta_{[k]}\in\Theta_{[k]}}\left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{p_{[k]}}\psi_{j\,[k]}(x_i;\theta_{[k]})B_{j\,[k]}(y_i)-c_{[k]}(\theta_{[k]})\right\}\right]$$
$$-\left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{p_{[\cup]}(n)}\psi_{j\,[\cup]}(x_i;\theta^{*\,[\cup]})B_{j\,[\cup]}(y_i)-c_{[\cup]}(\theta^{*\,[\cup]})\right\}\right]$$
$$\leq\sup_{\theta_{[\cup]}\in\Theta_{[\cup]}}\left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{p_{[\cup]}(n)}\psi_{j\,[\cup]}(x_i;\theta_{[\cup]})B_{j\,[\cup]}(y_i)-c_{[\cup]}(\theta_{[\cup]})\right\}\right]$$
$$-\left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{p_{[\cup]}(n)}\psi_{j\,[\cup]}(x_i;\theta^{*\,[\cup]})B_{j\,[\cup]}(y_i)-c_{[\cup]}(\theta^{*\,[\cup]})\right\}\right]$$
$$=\sum_{i=1}^{n}\left[\sum_{j=1}^{p_{[\cup]}(n)}\left\{(\psi_{j\,[\cup]}(x_i;\hat{\theta}_{[\cup]})-\psi_{j\,[\cup]}(x;\theta^{*\,[\cup]}))B_{j\,[\cup]}(y_i)\right\}\right.$$
$$\left.-\left\{c_{[\cup]}(\hat{\theta}_{[\cup]})-c_{[\cup]}(\theta^{*\,[\cup]})\right\}\right].$$

Using Lemma 14 in [37], we can show

$$c_{[\cup]}(\hat{\theta}_{[\cup]})-c_{[\cup]}(\theta^{*\,[\cup]})=\bigtriangledown c_{[\cup]}(\theta^{*\,[\cup]})^T\psi_{[\cup]}(x;\hat{\theta}_{[\cup]}-\theta^{*\,[\cup]})+O_p\left(\frac{J_{[\cup]}(n)}{n}\right).$$

Then

$$\sum_{i=1}^{n}\left[\sum_{j=1}^{p_{[\cup]}(n)}\left\{(\psi_{j\,[\cup]}(x_i;\hat{\theta}_{[\cup]})-\psi_{j\,[\cup]}(x;\theta^{*\,[\cup]}))\,B_{j\,[\cup]}(y_i)\right\}\right.$$
$$\left.-\left\{c_{[\cup]}(\hat{\theta}_{[\cup]})-c_{[\cup]}(\theta^{*\,[\cup]})\right\}\right]$$
$$=\sum_{i=1}^{n}\left[\left\{B_{[\cup]}(y_i)-\bigtriangledown c_{[\cup]}(\theta^{*\,[\cup]})\right\}^T\psi_{[\cup]}(x;\hat{\theta}_{[\cup]}-\theta^{*\,[\cup]})+O_p\left(\frac{J_{[\cup]}(n)}{n}\right)\right]$$
$$=O_p(J_{[\cup]}(n))+O_p(J_{[\cup]}(n))\ ([38],\text{ Lemma 13 and (21)})$$
$$=O_p(J_{[\cup]}(n)).$$

Therefore, the difference of the *sup*'s in the following equation is bounded by $O_p(J_{[\cup]}(n))$.

$$
\begin{aligned}
&P[\text{selecting the better model}] \\
&= P\left[\sup_{\theta_{[1]}\in\Theta_{[1]}} \ell_{[1]}(\theta_{[1]}) - \sup_{\theta_{[2]}\in\Theta_{[2]}} \ell_{[2]}(\theta_{[2]}) - a(n)(J_{[1]} - J_{[2]}) < 0\right] \\
&= P\left[\frac{\sup_{\theta_{[1]}\in\Theta_{[1]}} \ell_{[1]}(\theta_{[1]}) - \sup_{\theta_{[2]}\in\Theta_{[2]}} \ell_{[2]}(\theta_{[2]})}{J_{[\cup]}(n)} - \frac{a(n)(J_{[1]} - J_{[2]})}{J_{[\cup]}(n)} < 0\right] \\
&\to 1
\end{aligned}
$$

if

$$\frac{a(n)(J_{[1]} - J_{[2]})}{J_{[\cup]}(n)} \to \infty.$$

Because $J_{[1]}/J_{[\cup]}(n) \to 1$ and $J_{[2]} < \infty$, this condition is equivalent to $a(n) \to \infty$.

Therefore $IC_{[k]}$ is consistent if

$$J_{[1]} = o(n^{0.5-\delta}) \text{ for some } \delta \in (0, 0.5), \quad \frac{a(n)J_{[1]}}{n} \to 0 \text{ and } a(n) \to \infty, \text{ as n} \to \infty. \quad \blacksquare$$

Appendix B: Proof of Corollary 1

The leave-one-out cross validation, $CV(1)$ of model $\mathcal{M}_k$, is

$$
\begin{aligned}
\Gamma^{CV(1)\,[k]} &= \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - X_{i\,[k]}\hat{\theta}^{(i)}_{[k]}\right\}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - X_{i\,[k]}\hat{\theta}_{[k]}}{1 - h_{ii\,[k]}}\right\}^2
\end{aligned}
$$

where $\hat{\theta}^{(i)}_{[k]}$ is the estimate of $\theta_{[k]}$ without the i^{th} observation, $\hat{\theta}_{[k]}$ is the estimate of $\theta_{[k]}$ using all observations and $h_{i\,[k]}$ is the i^{th} diagonal element of the projection matrix $H_{[k]} = X_{[k]}(X^{[k]\,T}X_{[k]})^{-1}X^{[k]\,T}$. Suppose $\mathcal{M}_T$ is nested in $\mathcal{M}_2$ as in Case (ii). In this case, both candidate models converge to the true model as $n \to \infty$. Because $\left(1 - h_{i\,[k]}\right)^{-2} = 1 + 2h_{i\,[k]} + O\left\{\left(h_{i\,[k]}\right)^2\right\}$,

$$
\begin{aligned}
\Gamma^{CV(1)\,[k]} &= \frac{1}{n}\sum_{i=1}^{n}\left(e_{i\,[k]}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(e_{i\,[k]}\right)^2\left[2h_{i\,[k]} + O\left\{\left(h_{i\,[k]}\right)^2\right\}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(e_{i\,[k]}\right)^2 + \frac{2J_{[k]}\sigma^2}{n} + o_p\left(\frac{1}{n}\right), \qquad (7)
\end{aligned}
$$

where $e_{i\,[k]} = y_i - X_{i\,[k]}\hat{\theta}_{[k]}$, $e_{[k]} = \left(e_{1\,[k]}, \ldots, e_{n\,[k]}\right)^T$ and I_n is the $n \times n$ identity matrix. The leave-one-out cross validation in (7) is asymptotically equivalent to AIC, which has $a(n) = 1$. Therefore, $CV(1)$ is inconsistent by Theorem 1. $\blacksquare$