Model Selection Error Rates in Nonparametric vs. Parametric Model Comparisons

Yongsung Joo [*], George Casella [†], and Martin Wells [‡]

September 25, 2006

ABSTRACT

Since the introduction of Akaike's information criteria (AIC) in 1973, many information criteria have been developed and widely used in model selection. Many papers concerning the justification of 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 the consistency conditions of information criteria in nonparametric (logspline) vs. parametric model comparisons, and discuss model selection error rates when the sample size is finite.

Keywords: Consistent model selection criteria, Information criteria, Log spline model, Model selection error rate.

[*] Assistant Professor, Biostatistics, College of Public Health and Health Professions, University of Florida, Gainesville FL 32611 (Correspondence: E-mail yjoo@phhp.ufl.edu, Tel 1-352-273-5822, Fax 1-352-273-5365)

[†] Distinguished Professor, Department of Statistics, College of Arts and Sciences, University of Florida, Gainesville FL 32611 and is supported by National Science Foundation Grant DMS-9971586.

[‡] Professor & Chair, Department of Biological Statistics and Computational Biology, Cornell University, Ithaca NY 14853 and is supported by National Science Foundation Grant DMS-9971586.

1 INTRODUCTION

In past decades, there were many papers that addressed the consistency of model selection criteria in various settings; see Haughton (1988, 1994), Nishii(1984), Potscher (1989) and Shibata (1976, 1981) and the references contained therein. The "classical" results for finite dimensional models show that leave-n_v-out cross validation (Shao 1993), BIC (Schwarz 1978) and Bayes factor (Gelfand and Dey 1994) are consistent, while AIC (Akaike 1973), C_p(Mallow 1973), jackknife, bootstrap (Efron 1983) and leave-one-out cross validation are asymptotically equivalent and inconsistent (Shao 1993). All of these articles, except Shibata (1981), 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, assuming the true model is in an infinite parameter space. The logspline model is one of the largest nonparametric model families in this category (Stone 1990, 1991 and Kooperberg and Stone 1991). Our interest is to examine error rates of model selection criteria in nonparametric logspline model vs. parametric model comparisons.

Let y_i be the random variable of interest for the i^{th} observation and $\psi^{\mathcal{M}_k}$ be the parameter in the logspline model $\mathcal{M}_k$ (Stone 1990). One version of the logspline model $\mathcal{M}_k$ refers to a model with the probability density function (pdf), $f^{\mathcal{M}_k}(y_i|\psi^{\mathcal{M}_k})$, that approximates or estimates the true pdf of the response variable y_i, and does not contain covariates (Stone 1990). As an extension of this model, the doubly flexible logspline response model, $f^{\mathcal{M}_k}(y_i|\psi^{\mathcal{M}_k}(x_i;\theta^{\mathcal{M}_k}))$, was introduced in Stone (1991) to approximate or estimate the true pdf of y_i, $f(y_i|x_i)$, that depends on $fixed$ predictor variable(s) x_i. Obviously, $f^{\mathcal{M}_k}(y_i|\psi^{\mathcal{M}_k})$ is a special case of $f^{\mathcal{M}_k}(y_i|\psi^{\mathcal{M}_k}(x_i;\theta^{\mathcal{M}_k}))$ when $\psi^{\mathcal{M}_k}(x_i;\theta^{\mathcal{M}_k}) = \psi^{\mathcal{M}_k}$. In this paper, $f^{\mathcal{M}_k}(y_i|\psi^{\mathcal{M}_k}(x_i;\theta^{\mathcal{M}_k}))$ will be called the logspline model and be of our interest in model selection. The asymptotics of this family are well studied in Stone (1990, 1991) and research on other aspects are in Crain (1974, 1976a, b, 1977), Barron and Sheu (1991), Kooperberg and Stone

(1992), Kooperberg, Stone and Truong (1995), Leonard (1978), Silverman (1982), Stone (1985, 1986), and Strawderman and Tsiatis (1996). In most of these papers, the authors propose a data-driven technique to address problems in model selection, and most use AIC (Akaike 1973) or BIC (Schwartz 1978) 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.

1) The unknown underlying "true" model $\mathcal{M}_T$, which the data come from. Let θ^* and Θ^* be the parameter vector and space of $\mathcal{M}_T$.

2) "Candidate" models $\mathcal{M}_k$, which are models under consideration to fit the data. Let $\theta^{\mathcal{M}_k}$ and $\Theta^{\mathcal{M}_k}$ be the parameter vector and the parameter space of $\mathcal{M}_k$. Assume that the true model is the same as or nested in one of candidate models. When we know or assume that 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 (Stone 1990, 1991). As in many previous studies (Bozdogan 1987, Shao 1993, 1996), we consider consistent model selection between two candidate models, $\mathcal{M}_1$ and $\mathcal{M}_2$. A consistent model selection criterion chooses the better model for any sufficiently large n.

As a nonparametric candidate model, we consider the logspline model with the number of parameters $J^{\mathcal{M}_k}$ 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 pdf as the logspline model, but with a finite and fixed number of parameters for any n. Normal linear regression, Poisson regression, logistic regression and many other generalized linear models are included in this family.

3) The "encompassing" model $\mathcal{M}_\cup$, whose parameter vector $\theta^{\mathcal{M}_\cup}$ consists of all parameters in candidate models (Berger and Pericchi 1996). Let $J^{\mathcal{M}_\cup}$ be the dimension of $\theta^{\mathcal{M}_\cup}$. Note that $J^{\mathcal{M}_\cup} \geq J^{\mathcal{M}_k}$ for any k and any n. Denote parameter spaces of

$\mathcal{M}_{\cup}$ by $\Theta^{\mathcal{M}_{\cup}}$. The following example is provided for better understanding of these notations.

Example 1 *True, candidate, and encompassing models*
Suppose that there are a small number of observations of interest from the "true" model $\mathcal{M}_T$, $y_i = \theta_0^* + \theta_1^* exp(x_i) + \epsilon_i$ where ϵ_i is independently and identically distributed (iid) with $N(0,1)$, y_i is a response variable and x_i is a predictor 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 or nested in one of candidate models. An analyst may consider two "candidate" models:

- $\mathcal{M}_1 : y_i = \theta_0^{\mathcal{M}_1} + \theta_1^{\mathcal{M}_1} x_i + \theta_2^{\mathcal{M}_1} x_i^2 + \theta_3^{\mathcal{M}_1} x_i^3 + \theta_4^{\mathcal{M}_1}(x_i - 1)_+^3 + \theta_5^{\mathcal{M}_1}(x_i - 2)_+^3 + \epsilon_i$, where where $(x_i - t_j)_+ = max(0, x_i - t_j)$ and t_j's are knots in the spline (nonparametric regression spline model, Ruppert, Wand and Caroll 2003).

- $\mathcal{M}_2 : y_i = \theta_0^{\mathcal{M}_2} + \theta_1^{\mathcal{M}_2} exp(x_i) + \epsilon_i$ (parametric model).

Then, the "encompassing" model $\mathcal{M}_{\cup}$ is: $y_i = \theta_0^{\mathcal{M}_{\cup}} + \theta_1^{\mathcal{M}_{\cup}} x_i + \theta_2^{\mathcal{M}_{\cup}} x_i^2 + \theta_3^{\mathcal{M}_{\cup}} x_i^3 + \theta_4^{\mathcal{M}_{\cup}}(x_i - 1)_+^3 + \theta_5^{\mathcal{M}_{\cup}}(x_i - 2)_+^3 + \theta_6^{\mathcal{M}_{\cup}} e_i^x + \epsilon_i$. Therefore, $J^{\mathcal{M}_1} = 6$, $J^{\mathcal{M}_2} = 2$ and $J^{\mathcal{M}_{\cup}} = 7$.

In Section 2, details on the logspline model are given. In Section 3, we consider the case when a nonparametric model ($J^{\mathcal{M}_1} \to \infty$) and a parametric model ($J^{\mathcal{M}_2} < \infty$) are compared, whereas selection between parametric models ($J^{\mathcal{M}_k} < \infty$ for k=1,2) is most frequently studied in other model selection literature (Gelfand and Dey 1994, Shao 1993, *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 (Foster and George 1994), HQ (Hannan and Quinn 1979) and leave-one-out (Shao 1993) 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 pdf $f(y_i|x_i)$, 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 estimates or approximates $f(y_i|x_i)$ is defined (Stone 1991) by

$$f^{\mathcal{M}_k}(y_i|\psi^{\mathcal{M}_k}(x_i;\theta^{\mathcal{M}_k})) = \exp\left(\sum_{i=1}^{p^{\mathcal{M}_k}} \psi_i^{\mathcal{M}_k}(x_i;\theta^{\mathcal{M}_k})B_i^{\mathcal{M}_k}(y_i) - c^{\mathcal{M}_k}(\psi^{\mathcal{M}_k}(x_i;\theta^{\mathcal{M}_k}))\right) \quad (1)$$

where

$$\psi_i^{\mathcal{M}_k}(x_i;\theta^{\mathcal{M}_k}) = \sum_{j=1}^{q_i^{\mathcal{M}_k}} \theta_{ij}^{\mathcal{M}_k} A_{ij}^{\mathcal{M}_k}(x_i),$$

$$c^{\mathcal{M}_k}(\psi^{\mathcal{M}_k}(x_i;\theta^{\mathcal{M}_k})) = log\left\{\int \exp\left(\sum_{i=1}^{p^{\mathcal{M}_k}} \psi_i^{\mathcal{M}_k}(x_i;\theta^{\mathcal{M}_k})B_i^{\mathcal{M}_k}(y_i)\right) dy\right\},$$

and $A_{ij}^{\mathcal{M}_k}(x_i)$ and $B_i^{\mathcal{M}_k}(y_i)$ are spline basis functions. The total number of parameters for estimation is $J^{\mathcal{M}_k} = \sum_{i=1}^{p^{\mathcal{M}_k}} q_i^{\mathcal{M}_k}$. From now on, let $c^{\mathcal{M}_k}(\theta^{\mathcal{M}_k}) = c^{\mathcal{M}_k}(\psi^{\mathcal{M}_k}(x_i;\theta^{\mathcal{M}_k}))$ for notational simplicity. See Stone (1991) for regularity conditions for this model and Stone (1990, 1991) for asymptotic properties of this model. The following example shows that the normal (cubic) regression spline model (Ruppert, Wand and Caroll 2003) is a special case of a logspline model. This implies that the normal linear regression can be also expressed with the pdf of the logspline model.

Example 2 *Normal regression spline*

Suppose the relationship between x_i and y_i is explored with a normal regression spline,

$$y_i = \theta_0^{\mathcal{M}_k} + \theta_1^{\mathcal{M}_k}x_i + \theta_2^{\mathcal{M}_k}x_i^2 + \theta_3^{\mathcal{M}_k}x_i^3 + \sum_{j=1}^{q^{\mathcal{M}_k}-4} \theta_{j+3}^{\mathcal{M}_k}(x_i - t_j)_+^3 + \epsilon_i^{\mathcal{M}_k}$$

where $(x_i - t_j)_+ = max(0, x_i - t_j)$, t_j's are knots in the spline and $\epsilon_i^{\mathcal{M}_k} \overset{iid}{\sim} N(0, \sigma^{2\,\mathcal{M}_k})$. The number of knots increases with the sample size n.

Let $A_j^{\mathcal{M}_k}(x_i) = (1, x_i, x_i^2, x_i^3, (x_i - t_1)_+^3, \ldots, (x_i - t_{q^{\mathcal{M}_k}-4})_+^3)$ and $\theta^{\mathcal{M}_k} = (\theta_0^{\mathcal{M}_k\,T}, \ldots, \theta_{q^{\mathcal{M}_k}-1}^{\mathcal{M}_k\ \ T}, \sigma^{2\,\mathcal{M}_k})^T$. The pdf of the regression spline model is

$$f^{\mathcal{M}_k}(y_i|\psi^{\mathcal{M}_k}(x_i;\theta^{\mathcal{M}_k})) = \frac{1}{\sqrt{2\pi\sigma^{2\,\mathcal{M}_k}}} exp\left[-\frac{\left\{y_i - \sum_{j=0}^{q^{\mathcal{M}_k}-1}\theta_j^{\mathcal{M}_k}A_j^{\mathcal{M}_k}(x_i)\right\}^2}{2\sigma^{2\,\mathcal{M}_k}}\right]$$

$$= \ exp\left[\sum_{j=0}^{q^{\mathcal{M}_k}-1}\frac{\theta_j^{\mathcal{M}_k}A_j^{\mathcal{M}_k}(x_i)}{\sigma^{2\,\mathcal{M}_k}}y_i - \frac{y_i^2}{2\sigma^{2\,\mathcal{M}_k}} - \frac{\left\{\sum_{j=0}^{q^{\mathcal{M}_k}-1}\theta_j^{\mathcal{M}_k}A_j^{\mathcal{M}_k}(x_i)\right\}^2}{2\sigma^{2\,\mathcal{M}_k}} + \log\left(\frac{1}{\sqrt{2\pi\sigma^{2\,\mathcal{M}_k}}}\right)\right],$$

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

$$\psi_1^{\mathcal{M}_k}(x_i;\theta^{\mathcal{M}_k}) = \sum_{j=0}^{q^{\mathcal{M}_k}-1}\frac{\theta_j^{\mathcal{M}_k}A_j^{\mathcal{M}_k}(x_i)}{\sigma^{2\,\mathcal{M}_k}}, B_1^{\mathcal{M}_k}(y_i) = y_i, \psi_2^{\mathcal{M}_k}(x_i;\theta^{\mathcal{M}_k}) = -\frac{1}{2\sigma^{2\,\mathcal{M}_k}}, B_2^{\mathcal{M}_k}(y_i) = y_i^2,$$

and

$$c^{\mathcal{M}_k}(\theta^{\mathcal{M}_k}) = \frac{\left\{\sum_{j=0}^{q^{\mathcal{M}_k}-1}\theta_j^{\mathcal{M}_k}A_j^{\mathcal{M}_k}(x_i)\right\}^2}{2\sigma^{2\,\mathcal{M}_k}} - \log\left(\frac{1}{\sqrt{2\pi\sigma^{2\,\mathcal{M}_k}}}\right).$$

This model has $J^{\mathcal{M}_k} = q^{\mathcal{M}_k} + 1$ parameters.

3 CONDITIONS FOR CONSISTENT MODEL SELECTION CRITERIA IN LOG SPLINE MODELS

In this section, we will define a general form of information criteria, $IC^{\mathcal{M}_k}$, and find the conditions when $IC^{\mathcal{M}_k}$ is consistent.

Define the model selection criteria $IC^{\mathcal{M}_k}$ for model $\mathcal{M}_k$ with the sample size of n as

$$IC^{\mathcal{M}_k} = \sup_{\theta^{\mathcal{M}_k}\in\Theta^{\mathcal{M}_k}} \ell^{\mathcal{M}_k}(\theta^{\mathcal{M}_k}) - a(n)\,J^{\mathcal{M}_k}, \qquad (2)$$

where $\Theta^{\mathcal{M}_k}$ is the parameter space of model $\mathcal{M}_k$, $\ell^{\mathcal{M}_k}(\theta^{\mathcal{M}_k})$ is the log-likelihood, $a(n)$ is a positive non-decreasing function of n and $J^{\mathcal{M}_k}$ is the number of parameters in model $\mathcal{M}_k$. In our paper, we assume $J^{\mathcal{M}_k} = o(n^{0.5-\delta})$ for some $\delta \in (0, 0.5)$ for the convergence of the MLE (Stone 1991). As examples of (2), there are:

- $AIC^{\mathcal{M}_k} = \sup\limits_{\theta^{\mathcal{M}_k} \in \Theta^{\mathcal{M}_k}} \ell^{\mathcal{M}_k}(\theta^{\mathcal{M}_k}) - J^{\mathcal{M}_k}$, which has $a(n) = 1$ (Akaike 1973).
- $BIC^{\mathcal{M}_k} = \sup\limits_{\theta^{\mathcal{M}_k} \in \Theta^{\mathcal{M}_k}} \ell^{\mathcal{M}_k}(\theta^{\mathcal{M}_k}) - \frac{\log(n)}{2}\, J^{\mathcal{M}_k}$, which has $a(n) = \frac{\log(n)}{2}$ (Schwartz 1978).
- $RIC^{\mathcal{M}_k} = \sup\limits_{\theta^{\mathcal{M}_k} \in \Theta^{\mathcal{M}_k}} \ell^{\mathcal{M}_k}(\theta^{\mathcal{M}_k}) - \log(J^{\mathcal{M}_\cup})\, J^{\mathcal{M}_k}$, which has $a(n) = \log(J^{\mathcal{M}_\cup})$ (Foster and George 1994).
- $HQ^{\mathcal{M}_k} = \sup\limits_{\theta^{\mathcal{M}_k} \in \Theta^{\mathcal{M}_k}} \ell^{\mathcal{M}_k}(\theta^{\mathcal{M}_k}) - \log(\log(n))\, J^{\mathcal{M}_k}$, which has $a(n) = \log(\log(n))$ (Hannan and Quinn 1979).

The supremum of the likelihood, $\sup\limits_{\theta^{\mathcal{M}_k} \in \Theta^{\mathcal{M}_k}} \ell^{\mathcal{M}_k}(\theta^{\mathcal{M}_k})$, is a measure of how well the model $\mathcal{M}_k$ fits the data and $a(n)\, J^{\mathcal{M}_k}$ is a penalty to prevent choosing an overfitted model. The model that explains the data well and is parsimonious should have a high $IC^{\mathcal{M}_k}$ value. In the comparison of two models $\mathcal{M}_1$ and $\mathcal{M}_2$, we choose $\mathcal{M}_2$ over $\mathcal{M}_1$ if $IC^{\mathcal{M}_2} > IC^{\mathcal{M}_1}$.

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 1991) is satisfied so that nonparametric candidate models converge to the true model. A model selection criteria is *consistent* if

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

Equivalently, if the error rate of the model selector goes to 0, then it is called a consistent model selection criterion. Consider the following two cases to establish the consistency conditions of a model selection criterion.

- *Case* (i) when the true model $\mathcal{M}_T$ is not nested in 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 parametric model $\mathcal{M}_2$:
  For an 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 $J^{\mathcal{M}_1} > J^{\mathcal{M}_2}$, $\mathcal{M}_2$ is a better model because of parsimony.

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 appeared in many other papers to prove consistency when the number of parameters is finite (Bozdogan 1987, Shao 1993 and references contained therein). The consistency conditions of these two cases are discussed in the following theorem.

Theorem 1 *Let $y_1, \ldots, y_n$ be iid random variables from the logspline family (1). Also let $J^{\mathcal{M}_1}$ and $J^{\mathcal{M}_2}$ be the number of parameters to be estimated in a nonparametric model $\mathcal{M}_1$ and a parametric model $\mathcal{M}_2$. A model selection criterion, $IC^{\mathcal{M}_k}$, is con-*

sistent if

$$J^{\mathcal{M}_1} = o(n^{0.5-\delta}) \text{ for some } \delta \in (0, 0.5), \quad \frac{a(n)J^{\mathcal{M}_1}}{n} \to 0 \text{ and } a(n) \to \infty,$$

as $n \to \infty$.

Proof The proof is summarized as follows (See Appendix A for detailed proof). First of all, $J^{\mathcal{M}_1} = o(n^{0.5-\delta})$ for some $\delta \in (0, 0.5)$ is needed for the MLE convergence in $\mathcal{M}_1$ (Stone 1990, 1991). In *Case* (i), a model selection criterion chooses the better model $\mathcal{M}_1$ consistently if

$$\frac{a(n)J^{\mathcal{M}_1}}{n} \to 0.$$

In *Case* (ii), consistency requires

$$a(n) \to \infty.$$

■

Besides 'nonparametric vs parametric' model comparisons, which is of our interest in this paper, there are two other possible cases of model comparisons-'parametric vs parametric' and 'nonparametric vs nonparametric' model comparisons. Remark 1-3 discuss these comparisons and 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. linear vs. quadratic regression models). By setting $J^{\mathcal{M}_k} < \infty$ 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; Bozdogan 1987 and Shao 1993). 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 a parametric. Then, Theorem 1 shows that BIC, RIC and HQ can be consistent depending on $J^{\mathcal{M}_1}$, whereas AIC is inconsistent.

Remark 3 *Nonparametric vs. Nonparametric Models*

Suppose that the candidates are two nonparametric models. This case includes the selection of knots in nonparametric 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 Stone (1991) for detailed discussions on the convergence rates of the logspline models.

It is known that AIC, C_p, jackknife, bootstrap (Efron 1983) and leave-one-out cross validation are asymptotically equivalent and inconsistent when only parametric models ($J^{\mathcal{M}_k} < \infty$) are considered as candidates (Shao 1993). As another applications of Theorem 1, Corollary 1 shows inconsistency of the leave-one-out cross validation ($CV(1)$) in nonparametric vs. parametric model comparisons. $CV(1)$ is defined as

$$\hat{\Gamma}^{\mathcal{M}_k \; CV(1)} = \frac{1}{n} \sum_{i=1}^{n} \left[y_i - X_i^{\mathcal{M}_k} \hat{\theta}^{\mathcal{M}_k \; (i)} \right]^2$$

where $\hat{\theta}^{\mathcal{M}_k \; (i)}$ is the MLE of $\theta^{\mathcal{M}_k}$ without the i^{th} observation. The following Corollary can be established.

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.*

Proof: See Appendix B

4 MODEL SELECTION ERROR RATE WHEN THE SAMPLE SIZE IS FINITE

The purpose of this section is to explore error rates of model selection criteria $IC^{\mathcal{M}_k}$'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 Simulation Study with Two Candidate Models

Consider two candidate models: $\mathcal{M}_1$ (the cubic regression spline with two equally-spaced knots) and $\mathcal{M}_2$ (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 $\ell^{\mathcal{M}_1}$ and $\ell^{\mathcal{M}_2}$ as the maximum log-likelihoods for $\mathcal{M}_1$ and $\mathcal{M}_2$, and $J^{\mathcal{M}_1}$ and $J^{\mathcal{M}_2}$ 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. Suppose, because the number of parameters in $\mathcal{M}_1$ is determined with a slowly increasing function of n, $\mathcal{M}_1$ has two knots for both n=50 and 100. For example, $\mathcal{M}_1$ may be designed to have as many knots as the closest integer to $n^{1/5}$. The first true model is $\mathcal{M}_{T1}$: $y_i = 1 + sin(x_i) + 3cos(x_i) + 4log(x_i) + \epsilon_i$, 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

$$\ell^{\mathcal{M}_1} - a(n)J^{\mathcal{M}_1} < \ell^{\mathcal{M}_2} - a(n)J^{\mathcal{M}_2} \Leftrightarrow \ell^{\mathcal{M}_1} - \ell^{\mathcal{M}_2} < 3a(n),$$

the model selection error rate is

$$P(\ell^{\mathcal{M}_1} - \ell^{\mathcal{M}_2} < 3a(n)). \tag{4}$$

This error rate is "analogous" to type II error when we test "H_o: $\mathcal{M}_2$ is the true model" with the log likelihood ratio, as different $IC^{\mathcal{M}_k}$'s correspond to testing with different rejection regions. The magnitude of $a(n)$'s at each fixed n determines the rejection regions and the error rates of $IC^{\mathcal{M}_k}$'s. For example, for $n = 50$ or 100, we have

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

where $a^{AIC}(n) = 1$, $a^{HQ}(n) = \frac{3}{2}\log(\log(n))$, $a^{RIC}(n) = \log(J^{\mathcal{M}_\cup}) = \log(J^{\mathcal{M}_1})$ and $a^{BIC}(n) = \log(n)/2$. Then, the error rate also increases in the order of AIC, HQ, RIC and BIC. Different error rates of $IC^{\mathcal{M}_k}$'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.

[Figure 1 about here.]

[Table 1 about here.]

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.

Now consider the quadratic model $\mathcal{M}_{T2}$: $y_i = 1 + x_i + x_i^2 + \epsilon_i$ as the true model, from which the data are generated. 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 $IC^{\mathcal{M}_k}$'s is reversed for error rates (in the second last column of Table 1). Here, error rates are "analogous" to type I error. Error rates 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 $IC^{\mathcal{M}_k}$ 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 $IC^{\mathcal{M}_k}$ 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 Simulation Study with Three Candidate Models

In many papers (for example; Shao 1993, 1996), 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 the exactly 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 3, 4, 6 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 $IC^{\mathcal{M}_k}$'s. In this case, $IC^{\mathcal{M}_k}$ with a small $a(n)$ may perform better because a small $a(n)$ makes $IC^{\mathcal{M}_k}$ 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 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, $IC^{\mathcal{M}_k}$ 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, $IC^{\mathcal{M}_k}$ 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 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.

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 (Bozdogan 1987, Hurvich, Shumway and Tsai 1990, Shao 1993, 1996, Zheng and Lou 1995), 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.

5 SUMMARY

Many new model selection criteria have been developed in past decades, comparing 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 in other papers for parametric comparisons. It 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 a strong evidence of the best model selection criterion by itself, because it varies easily depending on the candidate models and true model.

6 REFERENCES

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.) Akademiai Kiado: Budapest*, 267-281.

Barron, A.R. and Sheu, C. (1991) Approximation of density functions by sequences of exponential families, *Annals of Statistics*, **19**, 1347-1369.

Berger, J. and Pericchi, L. (1996) The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109-122.

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

Crain, B.R. (1974) Estimation of distributions using orthogonal expansions *Annals of Statistics* **2**, 454-463.

Crain, B.R. (1976a) Exponential models, maximum likelihood estimation and the Haar conditions. *Journal of American Statistical Association*, **71**, 737-740.

Crain, B.R. (1976b) More on estimation of distributions using orthogonal expansions. *Journal of American Statistical Association*, **71**, 741-745.

Crain, B.R. (1977) An information theoretic approach to approximating a probability distribution. *SIAM Journal of Applied Mathematics*, **32**, 339-346.

Efron, B. (1983) Estimating error rate of a prediction rule: Improvement on cross validation, *Journal of American Statistical Association*, **78**, 316-331.

Foster, D. and George, E. (1994) The risk inflation criterion for multiple regression, *Annals of Statistics*, **22**, 1947-1975.

Gelfand, A.E. and Dey, D.K. (1994), Bayesian Model Choice: Asymptotics and Exact Calculations, *Journal of Royal Statistical Society. B,* **56**, 501-514.

Haughton, D. (1988) On the choice of a model fit from an exponential family, *Annals of Statistics*, **16**, 190-195.

Haughton, D. (1994) Consistency of a class of information criteria for model selection in non linear regression. *Theory Probab. Appl.*, **37**, 47-53.

Hannan, E.P., and Quinn, B.G. (1979), The determination of the order of an autoregression, *Journal of the Royal Statistical Society*, Ser. B, **41**, 190-195.

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.

Kooperberg, C. and Stone, C. (1991) A study of logspline density estimation, *Computational statistics & data analysis*, **12**, 327-347.

Kooperberg, C. and Stone, C. (1992) logspline density estimation for censored data, *Journal of Computational and Graphical Statistics* **1**, 301-328.

Kooperberg, C., Stone, C. and Truong, Y.K. (1995) Hazard regression, *Journal of American Statistical Association* **90**, 78-94.

Leonard, T. (1978) Density estimation, stochastic processes and prior information (with discussion). *Journal of Royal Statistical Society Ser. B* **40**, 113-146.

Mallow, C.L. (1973) Some comments on C_p, *Technometrics*, **15**, 661-675.

Nishii, R. (1984) Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, **12**, 758-765.

Portnoy, S. (1988) Asymptotic behavior of liklihood methods for exponential families when the number of parameters tends to infinity, *The Annals of Statistics.* **16** 1 356-366.

Potscher, B.M. (1989) Model selection under nonstationarity: autoregressive models and stochastic linear regression. *Annals of Statistics* **17**, 1257-1274.

Ruppert, D., Wand, M.P. and Caroll, R.J. (2003). *Semiparametric Regression.* Cambridge University Press, New York, 2003.

Shao, P. (1993) Linear model selection by cross-validation, *Journal of the American Statistical Association*, **88**, 486-494.

Shao, P. (1996) Bootstrap model selection, *Journal of the American Statistical Association*, **91**, 655-665.

Shibata, R. (1976) Selection of the order of an autoregressive model by Akaike's information criterion, *Biometrika* **63**, 114-126.

Shibata, R. (1981) An optimal selection of regression variables, *Bimetrika* **68**, 45-54.

Silverman, B.W. (1982) On the estimation of a probability density function by the maximum penalized likelihood method. *Annals of Statistics*, **10**, 795-810.

Strawderman, R.L. and Tsiatis, A.A. (1996) On the asymptotic properties of a flexible hazard estimator, *The Annals of Statistics* **24**, 41-63.

Stone, C. (1990) Large sample inference for log-Spline models, *The Annals of Statistics* **18**, 717-741.

Stone, C. (1991) Asymptotics for doubly flexible logspline response models, *The Annals of Statistics* **19**, 1832-1854.

Schwarz, G. (1978) Estimating the dimension of a model, *Annals of Statistics* **6**, 461-464.

Zheng, X. and Loh, W. (1995) Consistent variable selection in linear models, *Journal of the American Statistical Association* **90**, 151-156.

APPENDIX

A Proof of Theorem 1

The following arguments are based on the assumption that MLE converges ($J^{\mathcal{M}_1} = o(n^{0.5-\delta})$ for some $\delta \in (0, 0.5)$, Stone 1990, 1991).

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.

$$
\begin{aligned}
&P[\text{selecting the better model}]\\
&= P\left[IC^{\mathcal{M}_1} > IC^{\mathcal{M}_2}\right]\\
&= P\left[\sup_{\theta^{\mathcal{M}_1}\in\Theta^{\mathcal{M}_1}} \ell^{\mathcal{M}_1}(\theta^{\mathcal{M}_1}) - a(n)J^{\mathcal{M}_1} > \sup_{\theta^{\mathcal{M}_2}\in\Theta^{\mathcal{M}_2}} \ell^{\mathcal{M}_2}(\theta^{\mathcal{M}_2}) - a(n)J^{\mathcal{M}_2}\right]\\
&= P\left[\sup_{\theta^{\mathcal{M}_1}\in\Theta^{\mathcal{M}_1}} \ell^{\mathcal{M}_1}(\theta^{\mathcal{M}_1}) - \sup_{\theta^{\mathcal{M}_2}\in\Theta^{\mathcal{M}_2}} \ell^{\mathcal{M}_2}(\theta^{\mathcal{M}_2}) > a(n)(J^{\mathcal{M}_1} - J^{\mathcal{M}_2})\right]\\
&= P\left[\sup_{\theta^{\mathcal{M}_1}\in\Theta^{\mathcal{M}_1}} \left[\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p^{\mathcal{M}_1}} \psi_j^{\mathcal{M}_1}(x_i;\theta^{\mathcal{M}_1})B_j^{\mathcal{M}_1}(y_i) - c^{\mathcal{M}_1}(\theta^{\mathcal{M}_1})\right\}\right]\right.\\
&\quad \left. - \sup_{\theta^{\mathcal{M}_2}\in\Theta^{\mathcal{M}_2}} \left[\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p^{\mathcal{M}_2}} \psi_j^{\mathcal{M}_2}(x_i;\theta^{\mathcal{M}_2})B_j^{\mathcal{M}_2}(y_i) - c^{\mathcal{M}_2}(\theta^{\mathcal{M}_2})\right\}\right] > \frac{1}{n}a(n)(J^{\mathcal{M}_1} - J^{\mathcal{M}_2})\right]
\end{aligned}
$$

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

$$
\sup_{\theta^{\mathcal{M}_1}\in\Theta^{\mathcal{M}_1}} \left[\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p^{\mathcal{M}_1}} \psi_j^{\mathcal{M}_1}(x_i;\theta^{\mathcal{M}_1})B_j^{\mathcal{M}_1}(y_i) - c^{\mathcal{M}_1}(\theta^{\mathcal{M}_1})\right\}\right] \qquad for \;\; k = 1, 2.
$$

Let $\hat{\theta}^{\mathcal{M}_1}$ be the MLE of the parameter $\theta^{\mathcal{M}_1}$ in model $\mathcal{M}_1$. By the uniqueness of the maximum likelihood estimate (Stone 1991) we have that

$$\begin{aligned}
&\sup_{\theta^{\mathcal{M}_1}\in\Theta^{\mathcal{M}_1}} && \left[\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p^{\mathcal{M}_1}}\left(\psi_j^{\mathcal{M}_1}(x_i;\theta^{\mathcal{M}_1})B_j^{\mathcal{M}_1}(y_i)-c^{\mathcal{M}_1}(\theta^{\mathcal{M}_1})\right)\right\}\right]\\
&= && \left[\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p^{\mathcal{M}_1}}\left(\psi_j^{\mathcal{M}_1}(x_i;\hat{\theta}^{\mathcal{M}_1})B_j^{\mathcal{M}_1}(y_i)-c^{\mathcal{M}_1}(\hat{\theta}^{\mathcal{M}_1})\right)\right\}\right]\\
&\to && \left[\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{p^{\mathcal{M}_1}}\left(\psi_j^{\mathcal{M}_1}(x_i;\theta^{*\,\mathcal{M}_1})B_j^{\mathcal{M}_1}(y_i)-c^{\mathcal{M}_1}(\theta^{*\,\mathcal{M}_1})\right)\right\}\right],
\end{aligned}$$

where $\psi_j^{\mathcal{M}_1}(x_i;\hat{\theta}^{\mathcal{M}_1}) \to \psi_j^{\mathcal{M}_1}(x_i;\theta^{*\,\mathcal{M}_1})$. Then, by the weak law of large numbers,

$$\begin{aligned}
&\frac{1}{n}\,\Delta\ell(n)\\
&\overset{def}{=} \quad \sup_{\theta^{\mathcal{M}_1}\in\Theta^{\mathcal{M}_1}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{p^{\mathcal{M}_1}}\psi_j^{\mathcal{M}_1}(x_i;\theta)B_j^{\mathcal{M}_1}(y_i)-c^{\mathcal{M}_1}(\theta^{\mathcal{M}_1})\right]\right\}\\
&\qquad -\sup_{\theta^{\mathcal{M}_2}\in\Theta^{\mathcal{M}_2}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{p^{\mathcal{M}_2}}\psi_j^{\mathcal{M}_2}(x_i;\theta^{\mathcal{M}_2})B_j^{\mathcal{M}_2}(y_i)-c^{\mathcal{M}_2}(\theta^{\mathcal{M}_2})\right]\right\}\\
&\to \quad E\left[\sum_{j=1}^{p^{\mathcal{M}_1}}\psi_j^{\mathcal{M}_1}(x;\theta^{*\,\mathcal{M}_1})B_j^{\mathcal{M}_1}(y)-c^{\mathcal{M}_1}(\theta^{*\,\mathcal{M}_1})\right]\\
&\qquad -E\left[\sum_{j=1}^{p^{\mathcal{M}_2}}\psi_j^{\mathcal{M}_2}(x;\theta^{*\,\mathcal{M}_2})B_j^{\mathcal{M}_2}(y)-c^{\mathcal{M}_2}(\theta^{*\,\mathcal{M}_2})\right]\\
&> \quad 0
\end{aligned}$$

Hence,

$$P[\text{selecting the better model}] = P\left[\frac{1}{n}\,\Delta\ell(n)-\frac{a(n)(J^{\mathcal{M}_1}-J^{\mathcal{M}_2})}{n}>0\right]\to 1,$$

if

$$\frac{a(n)(J^{\mathcal{M}_1}-J^{\mathcal{M}_2})}{n}\to 0 \Leftrightarrow \frac{a(n)J^{\mathcal{M}_1}}{n}\to 0 \text{ as } n\to\infty.$$

Next, suppose that true model $\mathcal{M}_T$ is nested in parametric model $\mathcal{M}_2$ as in *Case (ii)*. Also, remind that $J^{\mathcal{M}_1} > J^{\mathcal{M}_2}$ 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 $\psi_j^{\mathcal{M}_\cup}(x_i;\hat{\theta}^{\mathcal{M}_\cup})$ converge to $\psi_j^{\mathcal{M}_\cup}(x_i;\theta^{*\,\mathcal{M}_\cup})$. Because $\Theta^{\mathcal{M}_k} \subset \Theta^{\mathcal{M}_\cup}(n)$ for k=1 and 2,

$$\sup_{\theta^{\mathcal{M}_k}\in\Theta^{\mathcal{M}_k}} \left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{p^{\mathcal{M}_k}} \psi_j^{\mathcal{M}_k}(x_i;\theta^{\mathcal{M}_k})B_j^{\mathcal{M}_k}(y_i) - c^{\mathcal{M}_k}(\theta^{\mathcal{M}_k})\right\}\right]$$
$$\leq \sup_{\theta^{\mathcal{M}_\cup}\in\Theta^{\mathcal{M}_\cup}} \left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{p^{\mathcal{M}_\cup}(n)} \psi_j^{\mathcal{M}_\cup}(x_i;\theta^{\mathcal{M}_\cup})B_j^{\mathcal{M}_\cup}(y_i) - c^{\mathcal{M}_\cup}(\theta^{\mathcal{M}_\cup})\right\}\right]$$

and

$$\sup_{\theta^{\mathcal{M}_k}\in\Theta^{\mathcal{M}_k}} \left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{p^{\mathcal{M}_k}} \psi_j^{\mathcal{M}_k}(x_i;\theta^{\mathcal{M}_k})B_j^{\mathcal{M}_k}(y_i) - c^{\mathcal{M}_k}(\theta^{\mathcal{M}_k})\right\}\right]$$
$$- \left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{p^{\mathcal{M}_\cup}(n)} \psi_j^{\mathcal{M}_\cup}(x_i;\theta^{*\,\mathcal{M}_\cup})B_j^{\mathcal{M}_\cup}(y_i) - c^{\mathcal{M}_\cup}(\theta^{*\,\mathcal{M}_\cup})\right\}\right] \geq 0.$$

Let

$$B^{\mathcal{M}_\cup}(y_i) = (B_1^{\mathcal{M}_\cup}(y_i),\ldots,B_{p^{\mathcal{M}_\cup}(n)}^{\mathcal{M}_\cup}(y_i))^T,$$

$$\psi^{\mathcal{M}_\cup}(x;\hat{\theta}^{\mathcal{M}_\cup}-\theta^{*\,\mathcal{M}_\cup}) = (\psi_1^{\mathcal{M}_\cup}(x;\hat{\theta}^{\mathcal{M}_\cup}-\theta^{*\,\mathcal{M}_\cup}),\ldots,\psi_{p^{\mathcal{M}_\cup}(n)}^{\mathcal{M}_\cup}(x;\hat{\theta}^{\mathcal{M}_\cup}-\theta^{*\,\mathcal{M}_\cup}))^T,$$

and

$$\bigtriangledown c^{\mathcal{M}_\cup}(\theta^{*\,\mathcal{M}_\cup}) = \left[\frac{dc^{\mathcal{M}_\cup}(\theta^{\mathcal{M}_\cup})}{d\theta^{\mathcal{M}_\cup}}\right]_{\theta^{\mathcal{M}_\cup}=\theta^{*\,\mathcal{M}_\cup}}.$$

Then,

$$\sup_{\theta^{\mathcal{M}_k}\in\Theta^{\mathcal{M}_k}} \left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{p^{\mathcal{M}_k}} \psi_j^{\mathcal{M}_k}(x_i;\theta^{\mathcal{M}_k})B_j^{\mathcal{M}_k}(y_i) - c^{\mathcal{M}_k}(\theta^{\mathcal{M}_k})\right\}\right]$$
$$- \left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{p^{\mathcal{M}_\cup}(n)} \psi_j^{\mathcal{M}_\cup}(x_i;\theta^{*\,\mathcal{M}_\cup})B_j^{\mathcal{M}_\cup}(y_i) - c^{\mathcal{M}_\cup}(\theta^{*\,\mathcal{M}_\cup})\right\}\right]$$
$$\leq \sup_{\theta^{\mathcal{M}_\cup}\in\Theta^{\mathcal{M}_\cup}} \left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{p^{\mathcal{M}_\cup}(n)} \psi_j^{\mathcal{M}_\cup}(x_i;\theta^{\mathcal{M}_\cup})B_j^{\mathcal{M}_\cup}(y_i) - c^{\mathcal{M}_\cup}(\theta^{\mathcal{M}_\cup})\right\}\right]$$

$$
\begin{aligned}
& - \left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{p^{\mathcal{M}_\cup}(n)} \psi_j^{\mathcal{M}_\cup}(x_i;\theta^{*\,\mathcal{M}_\cup})B_j^{\mathcal{M}_\cup}(y_i) - c^{\mathcal{M}_\cup}(\theta^{*\,\mathcal{M}_\cup})\right\}\right] \\
= & \sum_{i=1}^{n}\left[\sum_{j=1}^{p^{\mathcal{M}_\cup}(n)} \left\{(\psi_j^{\mathcal{M}_\cup}(x_i;\hat\theta^{\mathcal{M}_\cup}) - \psi_j^{\mathcal{M}_\cup}(x;\theta^{*\,\mathcal{M}_\cup}))B_j^{\mathcal{M}_\cup}(y_i)\right\} - \left\{c^{\mathcal{M}_\cup}(\hat\theta^{\mathcal{M}_\cup}) - c^{\mathcal{M}_\cup}(\theta^{*\,\mathcal{M}_\cup})\right\}\right].
\end{aligned}
$$

Using Lemma 14 in Stone(1990), we can show

$$
c^{\mathcal{M}_\cup}(\hat\theta^{\mathcal{M}_\cup}) - c^{\mathcal{M}_\cup}(\theta^{*\,\mathcal{M}_\cup}) = \bigtriangledown c^{\mathcal{M}_\cup}(\theta^{*\,\mathcal{M}_\cup})^T \psi^{\mathcal{M}_\cup}(x;\hat\theta^{\mathcal{M}_\cup} - \theta^{*\,\mathcal{M}_\cup}) + O_p\left(\frac{J^{\mathcal{M}_\cup}(n)}{n}\right).
$$

Then

$$
\begin{aligned}
& \sum_{i=1}^{n}\left[\sum_{j=1}^{p^{\mathcal{M}_\cup}(n)} \left\{(\psi_j^{\mathcal{M}_\cup}(x_i;\hat\theta^{\mathcal{M}_\cup}) - \psi_j^{\mathcal{M}_\cup}(x;\theta^{*\,\mathcal{M}_\cup}))\, B_j^{\mathcal{M}_\cup}(y_i)\right\} - \left\{c^{\mathcal{M}_\cup}(\hat\theta^{\mathcal{M}_\cup}) - c^{\mathcal{M}_\cup}(\theta^{*\,\mathcal{M}_\cup})\right\}\right] \\
= & \sum_{i=1}^{n}\left[\left\{B^{\mathcal{M}_\cup}(y_i) - \bigtriangledown c^{\mathcal{M}_\cup}(\theta^{*\,\mathcal{M}_\cup})\right\}^T \; \psi^{\mathcal{M}_\cup}(x;\hat\theta^{\mathcal{M}_\cup} - \theta^{*\,\mathcal{M}_\cup}) + O_p\left(\frac{J^{\mathcal{M}_\cup}(n)}{n}\right)\right] \\
= & \; O_p(J^{\mathcal{M}_\cup}(n)) + O_p(J^{\mathcal{M}_\cup}(n)) \text{ (Stone 1991, Lemma 13 and (21))} \\
= & \; O_p(J^{\mathcal{M}_\cup}(n))
\end{aligned}
$$

Therefore, the difference of the *sup*'s in the following equation is bounded by $O_p(J^{\mathcal{M}_\cup}(n))$.

$$
\begin{aligned}
& P[\text{selecting the better model}] \\
= & \; P\left[\sup_{\theta^{\mathcal{M}_1}\in\Theta^{\mathcal{M}_1}} \ell^{\mathcal{M}_1}(\theta^{\mathcal{M}_1}) - \sup_{\theta^{\mathcal{M}_2}\in\Theta^{\mathcal{M}_2}} \ell^{\mathcal{M}_2}(\theta^{\mathcal{M}_2}) - a(n)(J^{\mathcal{M}_1} - J^{\mathcal{M}_2}) < 0\right] \\
= & \; P\left[\frac{\sup_{\theta^{\mathcal{M}_1}\in\Theta^{\mathcal{M}_1}} \ell^{\mathcal{M}_1}(\theta^{\mathcal{M}_1}) - \sup_{\theta^{\mathcal{M}_2}\in\Theta^{\mathcal{M}_2}} \ell^{\mathcal{M}_2}(\theta^{\mathcal{M}_2})}{J^{\mathcal{M}_\cup}(n)} - \frac{a(n)(J^{\mathcal{M}_1} - J^{\mathcal{M}_2})}{J^{\mathcal{M}_\cup}(n)} < 0\right] \\
\to & \; 1
\end{aligned}
$$

if

$$
\frac{a(n)(J^{\mathcal{M}_1} - J^{\mathcal{M}_2})}{J^{\mathcal{M}_\cup}(n)} \to \infty.
$$

Because $J^{\mathcal{M}_1}/J^{\mathcal{M}_\cup}(n) \to 1$ and $J^{\mathcal{M}_2} < \infty$, this condition is equivalent to $a(n) \to \infty$.

Therefore $IC^{\mathcal{M}_k}$ is consistent if

$$J^{\mathcal{M}_1} = o(n^{0.5-\delta}) \text{ for some } \delta \in (0, 0.5), \quad \frac{a(n) J^{\mathcal{M}_1}}{n} \to 0 \text{ and } a(n) \to \infty, \text{ as n} \to \infty.$$

■

B Proof of Corollary 1

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

$$\begin{aligned}\Gamma^{CV(1)\ \mathcal{M}_k} &= \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - X_i^{\mathcal{M}_k}\hat{\theta}^{\mathcal{M}_k\ (i)}\right\}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - X_i^{\mathcal{M}_k}\hat{\theta}^{\mathcal{M}_k}}{1-h_{ii}^{\mathcal{M}_k}}\right\}^2\end{aligned}$$

where $\hat{\theta}^{\mathcal{M}_k\ (i)}$ is the estimate of $\theta^{\mathcal{M}_k}$ without the i^{th} observation, $\hat{\theta}^{\mathcal{M}_k}$ is the estimate of $\theta^{\mathcal{M}_k}$ using all observations and $h_i^{\mathcal{M}_k}$ is the i^{th} diagonal element of the projection matrix $H^{\mathcal{M}_k} = X^{\mathcal{M}_k}(X^{\mathcal{M}_k\ T}X^{\mathcal{M}_k})^{-1}X^{\mathcal{M}_k\ T}$. 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 $\left(1-h_i^{\mathcal{M}_k}\right)^{-2} = 1 + 2h_i^{\mathcal{M}_k} + O\left\{\left(h_i^{\mathcal{M}_k}\right)^2\right\}$,

$$\begin{aligned}\Gamma^{CV(1)\ \mathcal{M}_k} &= \frac{1}{n}\sum_{i=1}^{n}\left(e_i^{\mathcal{M}_k}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(e_i^{\mathcal{M}_k}\right)^2\left[2h_i^{\mathcal{M}_k} + O\left\{\left(h_i^{\mathcal{M}_k}\right)^2\right\}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(e_i^{\mathcal{M}_k}\right)^2 + \frac{2J^{\mathcal{M}_k}\sigma^2}{n} + o_p\left(\frac{1}{n}\right),\end{aligned} \tag{6}$$

where $e_i^{\mathcal{M}_k} = y_i - X_i^{\mathcal{M}_k}\hat{\theta}^{\mathcal{M}_k}$, $e^{\mathcal{M}_k} = \left(e_1^{\mathcal{M}_k}, \ldots, e_n^{\mathcal{M}_k}\right)^T$ and I_n is the $n \times n$ identity matrix. The leave-one-out cross validation in (6) is asymptotically equivalent to AIC, which has $a(n) = 1$. Therefore, $CV(1)$ is inconsistent by Theorem 1.

■

Table 1: The rate of choosing either quadratic regression or regression spline model.

True model			$\mathcal{M}_{T1}^{\dagger}$				$\mathcal{M}_{T2}^{\ddagger}$	
			$x_i \in [1,3]$		$x_i \in [3,5]$		$x_i \in [3,5]$	
	n	a(n)	Spline$^{\sharp}$	Quad*	Spline	Quad	Spline	Quad
AIC	50	1.00	0.999	0.001	0.489	0.511	0.149	0.851
BIC	50	1.96	0.976	0.024	0.137	0.863	0.014	0.986
HQ	50	1.36	0.997	0.003	0.317	0.683	0.064	0.936
RIC	50	1.79	0.986	0.014	0.175	0.825	0.021	0.979
AIC	100	1.00	1.000	0.000	0.684	0.316	0.128	0.872
BIC	100	2.30	1.000	0.000	0.192	0.808	0.006	0.994
HQ	100	1.53	1.000	0.000	0.453	0.547	0.033	0.967
RIC	100	1.79	1.000	0.000	0.349	0.651	0.019	0.981

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

‡ $\mathcal{M}_{T2} : y_i = 1 + x_i^2 + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0, 0.15^2)$

$^{\sharp}$ $\mathcal{M}_1 : y_i = \beta_0^{\mathcal{M}_1} + \beta_1^{\mathcal{M}_1} x_i + \beta_2^{\mathcal{M}_1} x_i^2 + \beta_3^{\mathcal{M}_1} x_i^3 + \beta_4^{\mathcal{M}_1} (x_i - t_1)_+^3 + \beta_5^{\mathcal{M}_1} (x_i - t_2)_+^3 + \epsilon_i^{\mathcal{M}_1}$
where t_1 and t_2 are equally spaced knots within the range of x_i.

* $\mathcal{M}_2 : y_i = \beta_0^{\mathcal{M}_2} + \beta_1^{\mathcal{M}_2} x_i + \beta_2^{\mathcal{M}_2} x_i^2 + \epsilon_i^{\mathcal{M}_2}$

Table 2: The rate of choosing either quadratic regression, exact or regression spline model: True model is $y_i = 1+\sin(x_i)+3\cos(x_i)+4\log(x_i)+\epsilon_i$ where $\epsilon_i \overset{iid}{\sim} N(0, 0.15^2)$.

			$x_i \in [1,3]$			$x_i \in [3,5]$		
Model selector	n	a(n)	Spline[♯]	Quad[⁎]	Exact[†]	Spline	Quad	Exact
AIC	50	1.00	0.172	0.000	0.828	0.147	0.279	0.574
BIC	50	1.96	0.031	0.001	0.969	0.023	0.531	0.446
HQ	50	1.36	0.091	0.000	0.909	0.071	0.387	0.542
RIC	50	1.79	0.041	0.000	0.959	0.032	0.496	0.471
AIC	100	1.00	0.147	0.000	0.854	0.146	0.115	0.739
BIC	100	2.30	0.012	0.000	0.988	0.011	0.350	0.640
HQ	100	1.53	0.053	0.000	0.947	0.050	0.207	0.740
RIC	100	1.79	0.033	0.000	0.967	0.028	0.256	0.716

[♯] $\mathcal{M}_1$: $y_i = \beta_0^{\mathcal{M}_1} + \beta_1^{\mathcal{M}_1} x_i + \beta_2^{\mathcal{M}_1} x_i^2 + \beta_3^{\mathcal{M}_1} x_i^3 + \beta_4^{\mathcal{M}_1} (x_i - t_1)_+^3 + \beta_5^{\mathcal{M}_1} (x_i - t_2)_+^3 + \epsilon_i^{\mathcal{M}_1}$ where t_1 and t_2 are equally spaced knots within the range of x_i.

[⁎] $\mathcal{M}_2$: $y_i = \beta_0^{\mathcal{M}_2} + \beta_1^{\mathcal{M}_2} x_i + \beta_2^{\mathcal{M}_2} x_i^2 + \epsilon_i^{\mathcal{M}_2}$,

[†] $\mathcal{M}_3$: $y_i = \beta_0^{\mathcal{M}_3} + \beta_1^{\mathcal{M}_3} \sin(x_i) + \beta_2^{\mathcal{M}_3} \cos(x_i) + \beta_3^{\mathcal{M}_3} \log(x_i) + \epsilon_i^{\mathcal{M}_3}$,

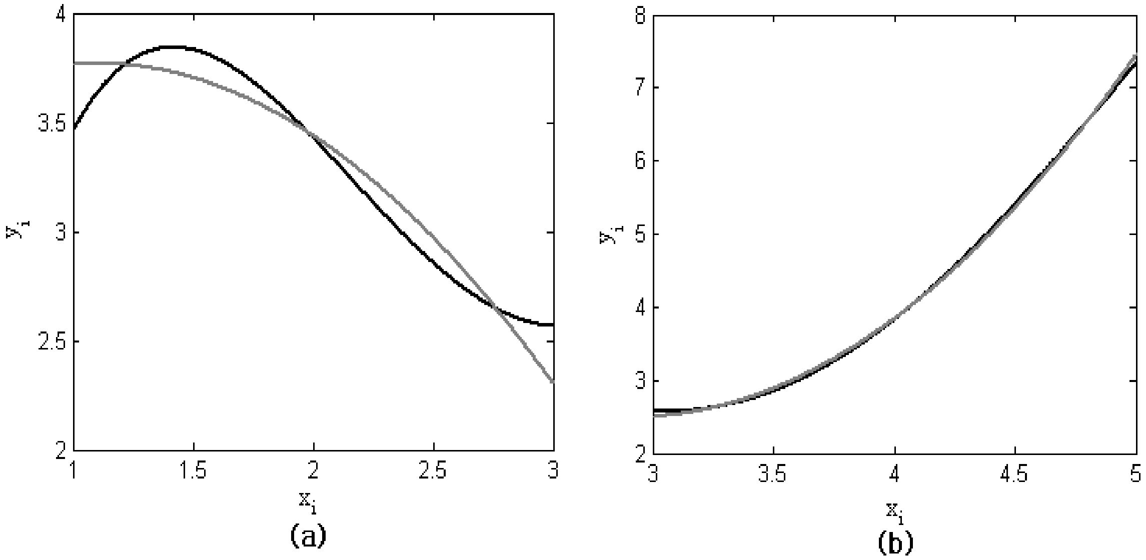

Figure 1: The true mean function $y_i = 1 + sin(x_i) + 3cos(x_i) + 4log(x_i)$ (dark-colored line) and it's closest quadratic function (light-colored line) when (a) $x_i \in [1, 3]$ and (b) $x_i \in [3, 5]$.