Semiparametric and Additive Model Selection Using an Improved Akaike Information Criterion

Jeffrey S. Simonoff; Chih-Ling Tsai

Journal of Computational and Graphical Statistics, Vol. 8, No. 1. (Mar., 1999), pp. 22-40.

Stable URL:
http://links.jstor.org/sici?sici=1061-8600%28199903%298%3A1%3C22%3ASAAMSU%3E2.0.CO%3B2-W

Journal of Computational and Graphical Statistics is currently published by American Statistical Association.

Semiparametric and Additive Model Selection Using an Improved Akaike Information Criterion

Jeffrey S. SIMONOFF and Chih-Ling TSAI

An improved AIC-based criterion is derived for model selection in general smoothing-based modeling, including semiparametric models and additive models. Examples are provided of applications to goodness-of-fit, smoothing parameter and variable selection in an additive model and semiparametric models, and variable selection in a model with a nonlinear function of linear terms.

Key Words: Goodness-of-fit; Kullback–Leibler discrepancy; Nonparametric regression; Smoothing spline regression estimator.

1. INTRODUCTION

Generalization of linear regression models to account for more complicated structure has been a focus of a good deal of research in recent years. The general model for regression data $\mathbf{y} = (y_1, \ldots, y_n)'$ examined here has the form

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\alpha} + \sum_{j=1}^{k} g_{j0}\left(X^{(j0)}\boldsymbol{\beta}^{(j0)}\right) + \boldsymbol{\epsilon} \\ &\equiv \mathbf{m} + \boldsymbol{\epsilon}, \end{aligned} \tag{1.1}$$

where

$$g_{j0}\left(X^{(j0)}\boldsymbol{\beta}^{(j0)}\right) = \left[g_{j0}\left(\mathbf{x}_1^{(j0)\prime}\boldsymbol{\beta}^{(j0)}\right), \ldots, g_{j0}\left(\mathbf{x}_n^{(j0)\prime}\boldsymbol{\beta}^{(j0)}\right)\right]'$$

with

$$X^{(j0)} = \left(\mathbf{x}_1^{(j0)}, \ldots, \mathbf{x}_n^{(j0)}\right)',$$

$\boldsymbol{\alpha} = \alpha\mathbf{1}$, α being a scalar and $\mathbf{1}$ being an $n \times 1$ vector, and $\mathbf{x}_i^{(j0)}$ and $\boldsymbol{\beta}^{(j0)}$ being $p_{j0} \times 1$ vectors for $i = 1, \ldots, n$. Here the functions g_{j0} are either specified, or unknown smooth

Jeffrey S. Simonoff is Professor, Department of Statistics and Operations Research, Leonard N. Stern School of Business, New York University, New York, NY 10012 (E-mail: jsimonoff@stern.nyu.edu). Chih-Ling Tsai is Professor, Graduate School of Management, University of California, Davis, CA 95616 (E-mail: cltsai@ucdavis.edu).

Journal of Computational and Graphical Statistics, Volume 8, Number 1, Pages 22–40

functions, the $X^{(j0)}$ are given $n \times p_{j0}$ matrices of predictor values ($p_{j0} < n$), and ϵ is an $n \times 1$ vector of independent errors with mean zero and variance σ_0^2. The predictor matrices are taken as fixed, or if they are random, analysis proceeds conditional on the observed values.

Many models are special cases of (1.1), including the following:

(a) linear model: $k = 1$, g_{10} the identity function;
(b) nonparametric regression model: $\boldsymbol{\alpha} = \mathbf{0}$, $k = 1$, $p_{10} = 1$, $\beta^{(10)} = 1$, g_{10} an unspecified smooth function;
(c) semiparametric model: $k = 2$, g_{10} the identity function, $p_{20} = 1$, $\beta^{(20)} = 1$, g_{20} an unspecified smooth function centered to have zero mean;
(d) additive model: $p_{j0} = 1$, $\beta^{(j0)} = 1$, and g_{j0} an unspecified smooth function centered to have zero mean, for all j; and
(e) mixed linear/nonlinear models: $\boldsymbol{\alpha}$ possibly $\mathbf{0}$, g_{j0} specified differentiable functions for $j = 1, \ldots, k$.

For discussion of models (a)–(d), including different estimation schemes, see Hastie and Tibshirani (1990) or Simonoff (1996, chap. 5). The form of the different estimators $\widehat{\mathbf{m}}$ depends on the specific model, but all estimators treated here are linear estimators. The usefulness of models (c) and (d) comes from the fact that, assuming the model is true, the $\boldsymbol{\beta}$ vectors can be estimated at optimal parametric rates in model (c), while the g functions can be estimated at optimal nonparametric rates in both models.

For any of these models, a crucial step in estimating $\boldsymbol{\beta}^{(j0)}$ and/or g_{j0} is one of model selection. If the model includes terms of the form $X^{(j0)}\boldsymbol{\beta}^{(j0)}$, this corresponds to choosing the appropriate set of predictors, balancing the desire for a simple model against the desire for closeness of fit. If the model includes unspecified smooth function(s) g_{j0}, this corresponds to choosing the smoothing parameter(s), balancing the desire for a smooth curve against the desire for closeness of fit.

Each version of the model selection problem can be executed by minimizing an approximately unbiased estimate of some measure of the discrepancy between the true regression function and the estimated regression function. This measure might be the mean average squared error (MASE)

$$\text{MASE} = \frac{1}{n} E[(\widehat{\mathbf{m}} - \mathbf{m})'(\widehat{\mathbf{m}} - \mathbf{m})],$$

often estimated using generalized cross-validation (GCV) (Craven and Wahba 1979), or the expected Kullback–Leibler discrepancy given in (2.3), often estimated using the Akaike information criterion (AIC) (Akaike 1973).

These selectors can be justified based on asymptotic arguments. Shibata (1981) showed for model (a) that if the set of candidate models does not include the true model,

$$\frac{\text{Predictive MSE of the chosen model}}{\text{Predictive MSE of the optimal model}} \xrightarrow{\text{P}} 1$$

as long as the number of candidate models does not get large too quickly (this is termed asymptotic efficiency). Härdle, Hall, and Marron (1988) demonstrated for model (b) that

the GCV- or AIC-based choice for the smoothing parameter is asymptotically optimal, in the sense that

$$\frac{\text{MISE}(\hat{h})}{\text{MISE}(\hat{h}_0)} \xrightarrow{\text{P}} 1$$

(where $\hat{h}_0$ is the smoothing parameter choice that minimizes the integrated squared error and MISE is the mean integrated squared error), as long as the level of smoothing does not get too small (similar results apply for MASE and ASE).

Despite these favorable asymptotic properties, the model selectors GCV and AIC suffer from a well-known tendency to lead to overfitting of the model in finite samples. This refers to erring on the side of closeness of fit, leading to too many predictors in a model, and an estimated function $\hat{g}_j$ that is too rough. Hurvich and Tsai (1989) showed that this effect arises in small samples due to bias, and proposed a corrected version of AIC, AIC_C, that lessens the bias (while still being asymptotically efficient). Hurvich, Simonoff, and Tsai (1998) (hereafter referred to as HST) derived a version of AIC_C for linear smoothers in the nonparametric model, and showed that its use leads to estimated regression functions that are not undersmoothed relative to those when using GCV or AIC (while still being asymptotically optimal).

In this article the selector AIC_C is generalized to model (1.1). The derivations of the criterion for three models consistent with (1.1) are given in Section 2. Section 3 provides applications of AIC_C to goodness-of-fit testing, variable and smoothing parameter selection in additive modeling and semiparametric modeling, and variable selection for a model with a nonlinear function of linear terms. Possible extensions of the proposed method are given in Section 4.

2. DERIVATIONS OF AIC_C FOR THREE MODEL STRUCTURES

2.1 GENERAL SETTINGS OF CORRECTED AKAIKE INFORMATION CRITERIA

In Section 1, we considered five special cases of model (1.1). Hurvich and Tsai (1989) obtained the model selection criterion AIC_C for parametric linear regression models (case (a)), and showed that its bias-correction properties lead to a criterion with less of a tendency to overfit the model by choosing too many predictors. HST derived AIC_C for nonparametric regression models (case (b)), demonstrating that it led to a smoothing parameter selector that did not tend to undersmooth when applied to many different linear estimators. In this article, we focus on developing the selection criterion AIC_C for cases (c)–(e), semiparametric regression models, additive models, and known differentiable functions of linear predictors. Although it would be possible to derive the criterion in complete generality for model (1.1), giving separate derivations for these three models highlights the steps that are involved to generalize the criterion to (1.1). The derivations are similar in spirit and construction to those in Hurvich and Tsai (1989) and HST, and for this reason only sketches of the derivations are provided here.

Here, we essentially follow the notation of HST. Suppose that data $\mathbf{y}$ are generated from the true model

$$\mathbf{y} = \mathbf{m} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\epsilon \sim N(0, \sigma_0^2 I_n)$. The candidate model is

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\eta}, \tag{2.2}$$

where $\eta \sim N(0, \sigma^2 I_n)$. Let $f(\mathbf{y})$ denote the likelihood for (μ, σ^2), and let E_0 denote expectation under the true model.

A useful measure of the discrepancy between the true model (2.1) and the candidate model (2.2) is the Kullback–Leibler information (omitting terms that are not functions of the candidate model, and are hence not relevant; see Linhart and Zucchini 1986, p. 18):

$$\begin{aligned} d(\mu, \sigma^2) &= E_0[-2 \log f(\mathbf{y})] \\ &= n \log(2\pi\sigma^2) + n\frac{\sigma_0^2}{\sigma^2} + \frac{(\mathbf{m} - \boldsymbol{\mu})'(\mathbf{m} - \boldsymbol{\mu})}{\sigma^2}. \end{aligned}$$

A reasonable criterion for judging the quality of the candidate model is $\Delta = E_0[d(\widehat{\mathbf{m}}, \hat{\sigma}^2)]$, where $(\widehat{\mathbf{m}}, \hat{\sigma}^2)$ is some estimator of $(\mathbf{m}, \sigma^2)$. Ignoring the constant $n \log(2\pi)$, we have

$$\Delta = E_0\{n \log \hat{\sigma}^2\} + n\sigma_0^2 E_0\{1/\hat{\sigma}^2\} + E_0\{(\mathbf{m} - \widehat{\mathbf{m}})'(\mathbf{m} - \widehat{\mathbf{m}})/\hat{\sigma}^2\}. \tag{2.3}$$

Given a collection of competing candidate models, then, the one that minimizes Δ is preferred (see Hurvich and Tsai 1989).

Usually, Δ is not computable, since it depends on the unknown function $\mathbf{m}$. For the sake of theoretical tractability, we will make the following three assumptions:

(A.1) There exists a matrix H such that $\widehat{\mathbf{m}} = H\mathbf{y}$. That is, the estimator is a linear function of $\mathbf{y}$.

(A.2) $\widehat{\mathbf{m}}$ is an unbiased estimator of $\mathbf{m}$ (i.e., $E_0(\widehat{\mathbf{m}}) = \mathbf{m}$).

(A.3) The parametric component of candidate models (if there is one) includes its corresponding parametric component of the true model.

These three assumptions warrant further discussion. HST also focused on linear estimators, which include the usual smoothers (such as local polynomials, smoothing splines, and regression splines). Assumptions (A.2) and (A.3) are more problematic, as $\widehat{\mathbf{m}}$ is rarely unbiased, and the true model is rarely a special case of the candidate model. It must be emphasized that these assumptions arise only in the derivation of the criterion, and are not relevant to the properties of the selectors based on the criterion (and are not made when deriving those properties). As was noted earlier, under model (a) AIC_C is asymptotically efficient (a statement based on condition (A.3) not holding), and under model (b) the AIC_C-based smoothing parameter is asymptotically optimal for any linear smoother (which will not generally satisfy condition (A.2)). Chen and Shiau (1994) showed that under model (c) GCV can be used to estimate g_{20} at the optimal nonparametric convergence rate while still estimating $\beta^{(10)}$ at the usual $O(n^{-1/2})$ rate; similar results apply to AIC and AIC_C, despite potentially different small-sample properties.

Under these assumptions, Δ reduces to

$$\tilde{\Delta} = E_0\{n \log \hat{\sigma}^2\} + n^2 E_0\left\{\frac{\sigma_0^2}{\boldsymbol{\epsilon}'(I - H)'(I - H)\boldsymbol{\epsilon}}\right\} + nE_0\left\{\frac{\boldsymbol{\epsilon}'H'H\boldsymbol{\epsilon}}{\boldsymbol{\epsilon}'(I - H)'(I - H)\boldsymbol{\epsilon}}\right\}. \tag{2.4}$$

HST derived three approximations for $\tilde{\Delta}$ for the nonparametric regression context. Because the form of $\tilde{\Delta}$ for that situation (eq. (2.2) of that paper) is identical to (2.4), these three approximations are also valid for the general model (1.1). The simplest of these approximations, AIC_C, was shown to be effective for smoothing parameter selection. It takes the form

$$\text{AIC}_C = \log \hat{\sigma}^2 + \frac{1 + \text{tr}(H)/n}{1 - [\text{tr}(H) + 2]/n}. \tag{2.5}$$

Equation (2.5) is also identical to that for parametric regression models (Hurvich and Tsai 1989). In the next section, we will apply this equation to obtain selection criteria for cases (c)–(e).

2.2 CORRECTED AIC FOR THREE MODEL STRUCTURES

We provide the detailed structures of $\mathbf{m}$ and $\boldsymbol{\mu}$ for cases (c)–(e) when $k = 2$ here; discussion for $k > 2$ is given at the end of the section. The three pairs of mean functions $\mathbf{m}$ and $\boldsymbol{\mu}$ follow.

Case (c)

$$\begin{aligned} \mathbf{m} &= X^{(10)}\boldsymbol{\beta}^{(10)} + g_{20}\left(\mathbf{x}^{(20)}\right) \\ &= X^{(1)}\boldsymbol{\beta}^{(1*)} + g_{20}\left(\mathbf{x}^{(2)}\right), \end{aligned}$$

and

$$\boldsymbol{\mu} = X^{(1)}\boldsymbol{\beta}^{(1)} + g_2\left(\mathbf{x}^{(2)}\right),$$

where $\boldsymbol{\beta}^{(1)}$ is a $p_1 \times 1$ vector, $\boldsymbol{\beta}^{(1*)} = (\boldsymbol{\beta}^{(10)\prime}, \boldsymbol{\gamma}^{(10)\prime})'$, $\boldsymbol{\gamma}^{(10)}$ is a $(p_1 - p_{10}) \times 1$ vector of zeros, $X^{(1)}$ is an $n \times p_1$ matrix, $\mathbf{x}^{(2)} = \mathbf{x}^{(20)} = (x_1^{(2)}, \ldots, x_n^{(2)})'$, g_2 is an unknown smooth function, and $g_2(\mathbf{x}^{(2)}) = (g_2(x_1^{(2)}), \ldots, g_2(x_n^{(2)}))'$.

Case (d)

$$\begin{aligned} \mathbf{m} &= g_{10}\left(\mathbf{x}^{(10)}\right) + g_{20}\left(\mathbf{x}^{(20)}\right) \\ &= g_{10}\left(\mathbf{x}^{(1)}\right) + g_{20}\left(\mathbf{x}^{(2)}\right), \end{aligned}$$

and

$$\boldsymbol{\mu} = g_1\left(\mathbf{x}^{(1)}\right) + g_2\left(\mathbf{x}^{(2)}\right),$$

where g_1 and g_2 are both unknown smooth functions, $g_2\left(\mathbf{x}^{(2)}\right)$ is defined as in case (c) above, and $g_1\left(\mathbf{x}^{(1)}\right) = (g_1(x_1^{(1)}), \ldots, g_1(x_n^{(1)}))'$ and $\mathbf{x}^{(1)} = (x_1^{(1)}, \ldots, x_n^{(1)})'$.

Case (e)

$$\begin{aligned} \mathbf{m} &= g_{10}\left(X^{(10)}\boldsymbol{\beta}^{(10)}\right) + g_{20}\left(X^{(20)}\boldsymbol{\beta}^{(20)}\right) \\ &= g_{10}\left(X^{(1)}\boldsymbol{\beta}^{(1*)}\right) + g_{20}\left(X^{(2)}\boldsymbol{\beta}^{(2*)}\right) \end{aligned}$$

and

$$\boldsymbol{\mu} = g_1\left(X^{(1)}\boldsymbol{\beta}^{(1)}\right) + g_2\left(X^{(2)}\boldsymbol{\beta}^{(2)}\right),$$

where $X^{(1)}$ and $\boldsymbol{\beta}^{(1*)}$ are defined as in case (c) above, $\boldsymbol{\beta}^{(2)}$ is a $p_2 \times 1$ vector, $\boldsymbol{\beta}^{(2*)} = (\boldsymbol{\beta}^{(20)\prime}, \boldsymbol{\gamma}^{(20)\prime})'$, $\boldsymbol{\gamma}^{(20)}$ is a $(p_2 - p_{20}) \times 1$ vector of zeros, both g_1 and g_2 are known differentiable functions, $g_j(X^{(j)}\boldsymbol{\beta}^{(j)}) = (g_j(\mathbf{x}_1^{(j)\prime}\boldsymbol{\beta}^{(j)}), \ldots, g_j(\mathbf{x}_n^{(j)\prime}\boldsymbol{\beta}^{(j)}))'$, $X^{(j)} = (\mathbf{x}_1^{(j)}, \ldots, \mathbf{x}_n^{(j)})'$ and $\mathbf{x}_i^{(j)}$ are $p_j \times 1$ vectors for $j = 1, 2$ and $i = 1, \ldots, n$.

For each of these cases, the candidate model is fit to the data. The resulting H matrices for estimating $\widehat{\mathbf{m}}$ in cases (c)–(e) are

$$H_c = H^* + S_2,$$

or

$$H_c = \tilde{H} + S_2$$

for case (c) (see Speckman [1988, eqns. (5.2a) and (5.2b)], although Chen and Shiau [1991] proposed a different estimation scheme);

$$H_d = I - (I - S_2)(I - S_1S_2)^{-1}(I - S_1)$$

for case (d) (see Hastie and Tibshirani 1990, p. 120); and

$$H_e = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}',$$

for case (e), respectively, where S_1 and S_2 are $n \times n$ smoother matrices that depend on $X^{(1)}$ and $X^{(2)}$, respectively, $H^* = (I - S_2)X^{(1)}\left(X^{(1)\prime}(I - S_2)X^{(1)}\right)^{-1}X^{(1)\prime}(I - S_2)$, $\tilde{H} = (I - S_2)X^{(1)}\left(X^{(1)\prime}(I - S_2)'(I - S_2)X^{(1)}\right)^{-1}X^{(1)\prime}(I - S_2)'(I - S_2)$, and

$$\begin{aligned}\tilde{X} &= \operatorname{diag}\left[\frac{\partial g_1\left(X^{(1)}\boldsymbol{\beta}^{(1)}\right)}{\partial\boldsymbol{\beta}^{(1)}}, \frac{\partial g_2\left(X^{(2)}\boldsymbol{\beta}^{(2)}\right)}{\partial\boldsymbol{\beta}^{(2)}}\right] \\ &= \operatorname{diag}\left[\frac{\partial g_1\left(X^{(1)}\boldsymbol{\beta}^{(1)}\right)}{\partial\left(X^{(1)}\boldsymbol{\beta}^{(1)}\right)}X^{(1)}, \frac{\partial g_2\left(X^{(2)}\boldsymbol{\beta}^{(2)}\right)}{\partial\left(X^{(2)}\boldsymbol{\beta}^{(2)}\right)}X^{(2)}\right]\end{aligned}$$

The scale parameter estimates for cases (c)–(e) are $\hat{\sigma}_c^2 = \mathbf{y}'(I - H_c)'(I - H_c)\mathbf{y}/n$, $\hat{\sigma}_d^2 = \mathbf{y}'(I - H_d)'(I - H_d)\mathbf{y}/n$ and $\hat{\sigma}_e^2 = \mathbf{y}'(I - H_e)'(I - H_e)\mathbf{y}/n$, respectively. Replacing H by H_c, H_d or H_e and $\hat{\sigma}^2$ by $\hat{\sigma}_c^2$, $\hat{\sigma}_d^2$ or $\hat{\sigma}_e^2$, respectively, into (2.5) then gives the corrected Akaike information criteria for semiparametric regression models, additive models, and models with known differentiable functions of unknown linear predictors.

Generalization of the functions $\mathbf{m}$ and $\boldsymbol{\mu}$ to $k > 2$ is straightforward. Unfortunately, the form of the smoother matrix H is complicated in this circumstance. Hastie and Tibshirani (1990, pp. 121–123) discussed fitting models (c) and (d) when there are two

smooth terms. They showed that if smoothers such as smoothing or regression splines are used, there exists a solution to the associated penalized least squares problem. If cubic smoothing splines are used (as is the case in the examples in the next section), that solution is unique as long as the predictor variables are not perfectly collinear, and the so-called backfitting algorithm (which is the basis of fitting in the examples in the next section) converges to that solution, yielding as a by-product the required value of $\text{tr}(H)$.

3. APPLICATIONS OF AIC_C

In this section we give several applications of AIC_C to problems of goodness-of-fit, variable selection, and smoothing parameter selection.

3.1 GOODNESS-OF-FIT TESTING

Classical goodness-of-fit tests are generally either parametric (such as likelihood ratio tests) or nonparametric (such as χ^2 tests). The latter tests often have low power against unspecified alternatives, while the former are only valid if the parametric assumptions are appropriate. Many authors have noted the possibility of using nonparametric regression methods to create goodness-of-fit tests based on the idea of comparing a smooth estimate of a functional (a density function, probability vector or regression function) to the null parametric form. Bickel and Rosenblatt (1973) and Simonoff (1985) gave examples of this idea for continuous density and categorical probability functions, respectively, while Hart (1997) described many approaches in the regression context.

The AIC_C criterion can be used as a goodness-of-fit criterion because of its role as an estimate of the expected Kullback–Leibler discrepancy (2.3). The idea is to compare the estimated Kullback–Leibler discrepancy between the true regression function and the parametric null model to the estimated Kullback–Leibler discrepancy between the true regression function and an estimated smooth regression function; that is, the test statistic is

$$A = \text{AIC}_C(\hat{m}_p) - \text{AIC}_C(\hat{m}_s),$$

where $\hat{m}_p$ is a parametric fit and $\hat{m}_s$ is the smooth fit obtained when choosing the smoothing parameter to minimize the AIC_C criterion. If the null model is not appropriate, AIC_C will be much lower for $\hat{m}_s$ than for $\hat{m}_p$, so large values of A lead to rejection of the null hypothesis.

Consider a situation with one predictor variable and null model being a simple linear model,

$$H_0 : y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n.$$

If the smoothing method used has as a special case the least squares regression line, the statistic A is just a comparison between the AIC_C value for that special case and the minimized value; thus, $A \geq 0$. Examples of such smoothing methods include local linear estimators and cubic smoothing spline estimators. The tail probability for an observed value of the test, A_0, can be determined using Monte Carlo, as follows:

- Fit the least squares regression line to the data, obtaining fitted values $\hat{\mathbf{y}}$ and residuals $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$.
- Resample with replacement from $\mathbf{e}$ to obtain $\mathbf{e}^*$, and form a replicated target variable as $\mathbf{y}^* = \hat{\mathbf{y}} + \mathbf{e}^*$. Determine the value of A for this data set (call it A^*).
- Repeat step (2) many times (e.g., 500–1,000). The estimated tail probability is the proportion of A^* values that are greater than or equal to A_0.

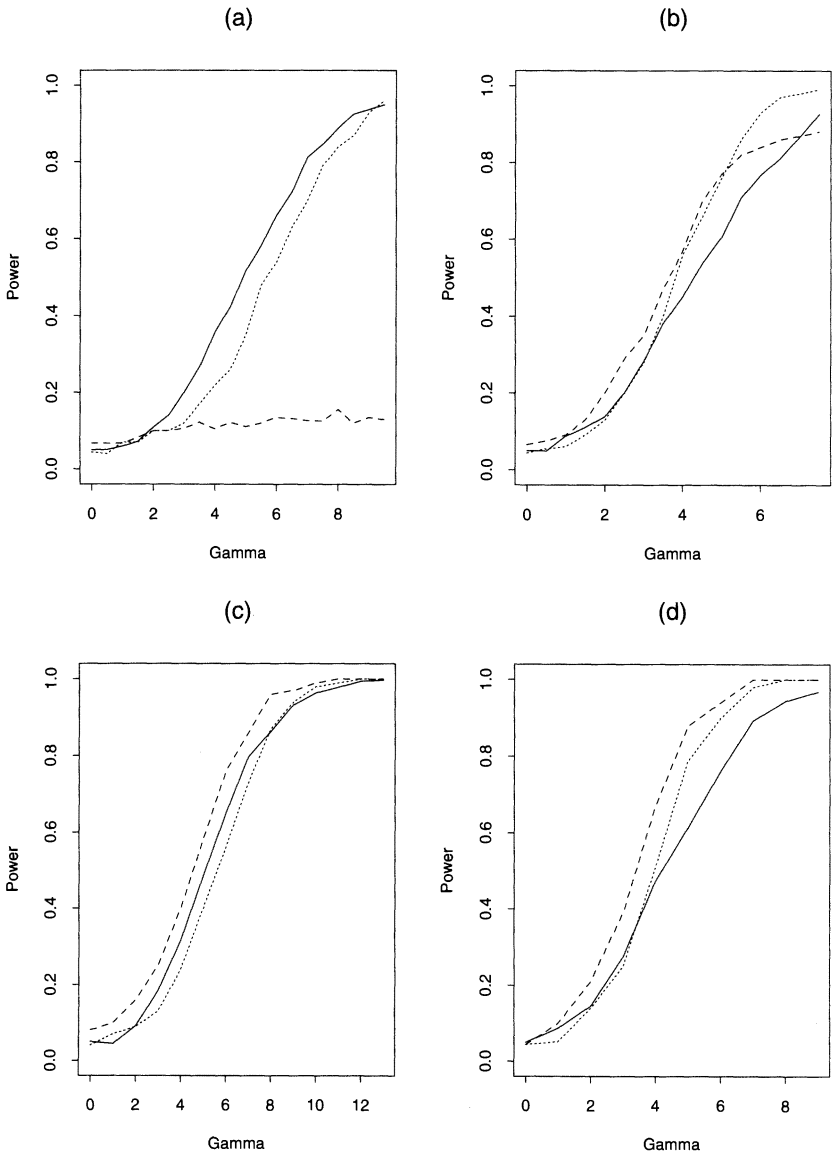

Figure 1. Estimated power functions for A (solid line), test of Eubank, Li and Wang (dashed line), and test of Eubank and Hart (dotted line). (a) $\beta_1 = 5$, n=50; (b) $\beta_1 = 5$, n=100; (c) $\beta_1 = 10$, n=50; (d) $\beta_1 = 10$, n=100.

A small Monte Carlo simulation was used to assess the power of this test. The Monte Carlo has the same structure as that used in Eubank, Li, and Wang (1997). The predictor values are taken to be $x_i = t_i - .5$, where

$$t_i = \frac{2i-1}{2n}, \quad i = 1, \ldots n,$$

with $n = 50$ or 100, respectively. The errors are distributed as standard normals. The alternative hypothesis has the form

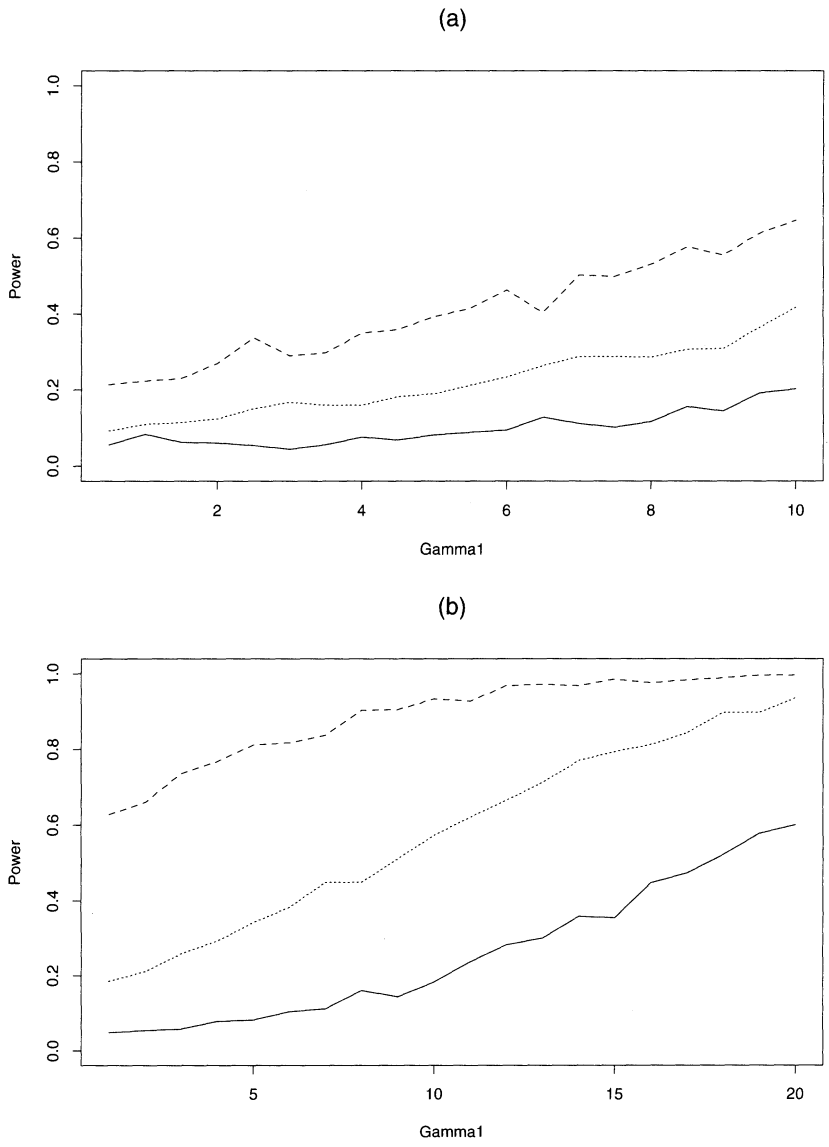

Figure 2. Estimated power functions for A (two predictors, n=50) (a) $\beta_1 = \beta_2 = 5$; (b) $\beta_1 = \beta_2 = 10$. In each plot estimated power functions are given for $\gamma_2 = 0$ (solid line), $\gamma_2 = \beta_2$ (dotted line), and $\gamma_2 = 2\beta_2$ (dashed line).

Table 1. AIC_C values for additive models based on 1993 automobile data, along with associated degrees of freedom for the spline estimates for the predicting variables

Predicting variables	*AIC_C value*	*Degrees of freedom*
Engine size	3.6112	8.22
Horsepower	3.6707	4.65
Weight	3.2683	3.69
Engine size, Horsepower	3.5656	(2.62, 4.10)
Engine size, Weight	3.2738	(1, 2.45)
Horsepower, Weight	3.2915	(1, 2.70)
Engine size, Horsepower, Weight	3.2916	(1, 1, 2.43)

$$H_a : y_i = \beta_0 + \beta_1 x_i + \gamma[t_i^2 - d_1 x_i - d_0] + \epsilon_i,$$

where $d_0 = \sum t_i^2/n$ and $d_1 = [\sum t_i^2 x_i]/[\sum x_i^2]^{1/2}$, and $\gamma = j\beta_1/10, j = 1, \ldots, 20$ with $\beta_1 = 5$ or 10. There were 1,000 Monte Carlo replications for each simulation setting. Other choices of n and β_1 were also examined, with broadly similar results (as would be expected, power increases with larger sample size and larger γ).

Figure 1 gives the estimated power functions for three smoothing-based tests: A (based on a cubic smoothing spline estimator); a test based on a Fourier series estimator proposed by Eubank, Li, and Wang (1997); and a test based on the estimated order of a Fourier series estimator proposed by Eubank and Hart (1992). It can be seen that the power of A is quite competitive with those of the other tests, being broadly similar to that of the test of Eubank and Hart (1992). (Eubank et al. [1997] pointed out that the reason for the poor power of their test for alternative (a) is that the optimal order of the Fourier series estimator in this case is one, and in that case the associated cosine function is nearly orthogonal to t^2.) Since Eubank and Hart (1992) showed that their test can be much more powerful than classical tests—such as the Cramér–von Mises test for high frequency alternatives—these results imply that the test A can be much more powerful than classical tests. Note also that the test A can be based on any linear smoother (not just cubic smoothing splines, as was done here), leaving open the possibility for it to be constructed to have better performance in specific circumstances.

Figure 2 gives estimated power functions for A for a model with two predictors (here $\hat{m}_s$ is based on an additive model of the two predictors). The alternative hypothesis has the form

$$H_a : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \gamma_1 \left[t_i^2 - d_1 x_{1i} - d_0\right] + \gamma_2 \left[t_i^2 - d_1 x_{2i} - d_0\right] + \epsilon_i,$$

where $x_{1i} = t_i - .5$, $x_{2i} = |x_{1i}|$, $\gamma_1 = j\beta_1/10, j = 1, \ldots, 20$ with $\beta_1 = \beta_2 = 5$ or 10. Results for $n = 50$ and $\gamma_2 = 0$, β_2, and $2\beta_2$ are given, based on 500 Monte Carlo replications for each simulation setting. Not surprisingly, the power of the test is not as high when there are two predictors as when there is only one predictor, but the test still is effective in identifying deviations from linearity in one (or especially) both predictors.

3.2 ADDITIVE MODELS

Additive models generalize linear models to allow arbitrary smooth functions of each predictor to enter the model in an additive fashion. Fitting such models requires both smoothing parameter selection (here joint selection for all of the variables in the model) and variable selection (to determine which smooth functions should be included in the model). Consider the following example. The data examined are properties of 93, 1993 model, U.S. automobiles (Lock 1993). The goal was to model the highway gasoline mileage in miles per gallon as an additive function of smooth representations for engine size, horsepower, and weight. Table 1 summarizes the use of AIC_C in the fitting of the additive model using cubic smoothing splines. The best model is one based on only weight, with $AIC_C = 3.2683$. Figure 3 illustrates how this model fitting works. Although horsepower is strongly related to highway mileage marginally (Figure 3(a)), when a smooth function of weight is also included, the partial residuals given weight show that the relationship with horsepower has been removed (Figure 3(b)). Figure 3(c) gives the final model of choice of mileage on weight alone, showing that mileage of the auto is monotonically inversely related to the weight in a nonlinear fashion.

3.3 SEMIPARAMETRIC MODELS

Fitting semiparametric models involves both variable selection of the parametric part of the model (and variable selection of the nonparametric part if it includes more than one predictor) and smoothing parameter selection for the nonparametric part. In this section two examples are given where AIC_C is used for both of these purposes.

In recent years, ratings of universities, colleges, graduate schools, and professional schools by different magazines have become very common. These ratings are used by prospective students and their parents to help choose the college that they will attend, but how are they related to objective quality measures? The data analyzed here come from the 48 U.S. research universities rated in the top 50, with complete data given in *U.S. News and World Report* (1996). The magazine provides data on the percentage of classes with size greater than 50 (Size $>$ 50); the percentage of freshman students entering in 1991–1994 who returned for their sophomore years (Retention); the percentage of students entering in 1986–1989 who graduated within six years (Graduation); and the logged education expenditure per student (Expenditure). The ranking of the university by the magazine is the target variable in a semiparametric model, with these variables being the candidate predictor variables, and the academic reputation of the university (Reputation, as determined by a survey of university presidents, provosts, and deans) being controlled for as a smooth control variable. That is, since a university's academic reputation is presumably slow to change, what factors within more direct control of the university are related to its rating, given that reputation? The control variable is allowed to enter the model nonlinearly (as an unspecified smooth function) since its marginal relationship with university rank is nonlinear; see Figure 4.

Table 2 summarizes the use of AIC_C in the fitting of the semiparametric model. For each possible set of predicting variables, the smoothing parameter for academic reputation

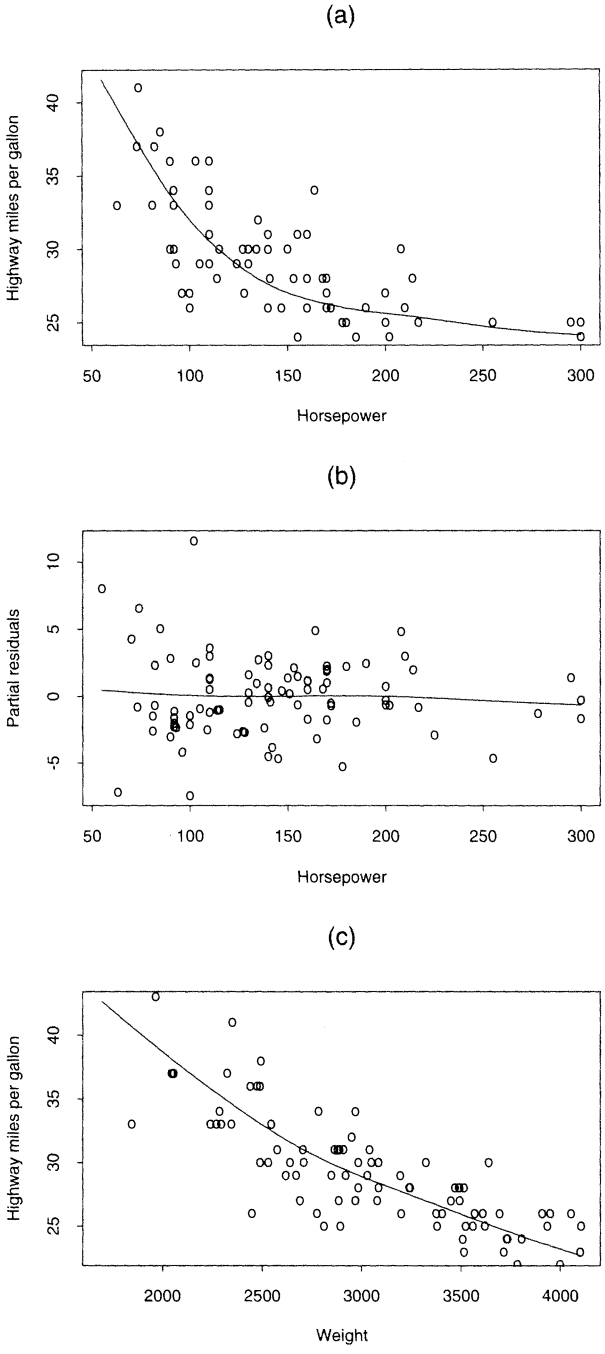

Figure 3. Additive model fits for 1993 automobile mileage data. (a) Highway mileage versus horsepower, with AIC_C-based smoothing spline estimate superimposed. (b) Partial residuals for AIC_C-based additive model of highway mileage on horsepower, given weight, with smoothing spline estimate superimposed. (c) Highway mileage versus weight, with AIC_C-based smoothing spline superimposed.

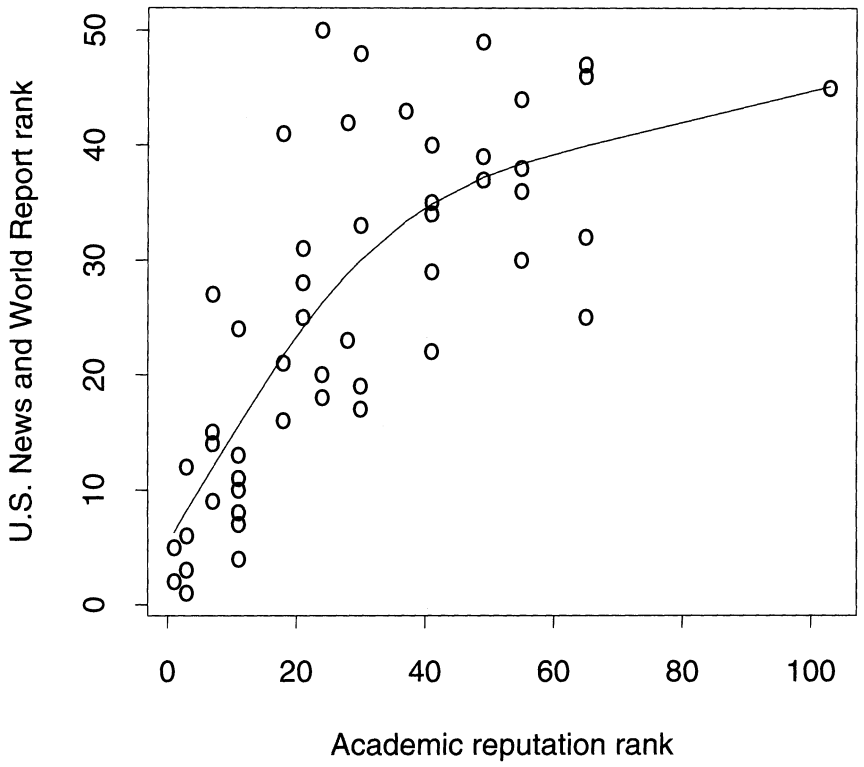

Figure 4. Scatterplot of U.S. News and World Report *rank versus academic reputation rank for top 50 universities, with AIC_C-based cubic smoothing spline estimate superimposed.*

is chosen to minimize AIC_C, with the final model chosen to be the one with smallest AIC_C value. It can be seen that the model chosen is based on only graduation rate and logged educational expenditure, with percentage of classes greater than 50 and freshman retention rate not adding to the fit. The degrees of freedom for the spline estimate in the semiparametric model equals 1, which corresponds to a linear fit; that is, the final model chosen is a multiple linear regression model,

$$\begin{aligned} \textit{U.S. News} \text{ rank} \;=\; & 203.973 - .802 \times \text{Graduation rate} \\ & -27.005 \times \text{Logged expenditure per student} \\ & +.176 \times \text{Academic reputation}, \end{aligned}$$

with $R^2 = .924$ and $F = 177.35$ on $(3, 44)$ degrees of freedom, which is highly significant (the individual t statistics for each variable are -9.79, -8.16, and 5.33, respectively). Thus, using AIC_C allows model selection of the parametric part of the semiparametric model, as well as recognition that a fully parametric model (without a smooth part) accounts for the structure in the data. The coefficients are all easily understandable, implying that a higher graduation rate, higher expenditures, and a better academic reputation are associated with a better *U.S. News and World Report* ranking (each given the other two).

Semiparametric models also can be generalized to include more than one smooth

Table 2. AIC_C values for semiparametric models based on *U.S. News* university rank data, along with associated degrees of freedom for the spline estimate for academic reputation

Predicting variables	*AIC_C value*	*Degrees of freedom*
None	5.4656	3.54
Size > 50	5.3501	2.58
Retention	5.0984	2.50
Graduation	4.6880	3.08
Expenditure	5.0515	1.68
Size > 50, Retention	4.7886	2.69
Size > 50, Graduation	4.6804	3.33
Size > 50, Expenditure	5.0442	1.55
Retention, Graduation	4.7326	2.84
Retention, Expenditure	4.5334	1.38
Graduation, Expenditure	3.9595	1.00
Size > 50, Retention, Graduation	4.6665	2.87
Size > 50, Retention, Expenditure	4.3663	1.00
Size > 50, Graduation, Expenditure	4.0141	1.00
Retention, Graduation, Expenditure	3.9699	1.00
Size > 50, Retention, Graduation, Expenditure	4.0128	1.00

term. Consider again the 1993 automobile data analyzed in Section 3.2. Table 3 summarizes a model fit to that data where, in addition to the three predictors that potentially enter the model through smooth functions (engine size, horsepower, and weight), three additional indicator variables potentially enter the model linearly. These three predictors correspond to whether or not the automobile can be bought with manual transmission, is of domestic origin, and is a van, respectively. Table 3 gives the linear predictors that minimize AIC_C given the smooth terms that are included in the model. It is apparent that including whether or not the auto is a van is useful in the model (and better than not including it at all; the AIC_C values for each set of smooth terms is smaller in Table 3 than that in Table 1, where no linear terms were considered); for the model including a smooth horsepower function, whether or not the auto is of domestic origin is useful also. The availability of manual transmission is never included among the useful linear predictors based on the AIC_C criterion. The model with minimized AIC_C includes Van as a linear predictor, and horsepower and weight as smooth terms, but the degrees of freedom for the horsepower term equals one, implying a semiparametric model with two linear predictors and one smooth term,

$$\text{Highway MPG} = 48.18 - .0172 \times \text{Horsepower} - 4.355 \times \text{Van} + g(\text{Weight}).$$

This model uses only 1.01 more effective degrees of freedom than the AIC_C-based smooth model on weight alone (since there are fewer degrees of freedom associated with the smooth term), while having residual sum of squares more than 10% smaller. The coefficients of the linear predictors are intuitive, with higher horsepower associated with worse mileage (given auto type and weight), and a van having on average more than four miles per gallon worse highway mileage than a regular automobile, given horsepower and weight.

Table 3. AIC_C values for semiparametric models based on 1993 automobile data, along with associated degrees of freedom for the spline estimates for the predicting variables for the smooth terms. The set of linear predictors that yields the minimized AIC_C value given the smooth terms included in the model is given for each set of smooth terms

Predicting variables		AIC_C	*Degrees*
Linear predictors	*Smooth terms*	*value*	*of freedom*
Van	Engine size	3.4610	4.75
Van, Domestic	Horsepower	3.3546	4.72
Van	Weight	3.2015	2.65
Van	Engine size, Horsepower	3.3054	(2.67, 5.02)
Van	Engine size, Weight	3.2227	(1, 2.66)
Van	Horsepower, Weight	3.1880	(1, 2.70)
Van	Engine size, Horsepower, Weight	3.2134	(1, 1, 2.68)

3.4 MODELS WITH NONLINEAR FUNCTIONS OF LINEAR TERMS

Atkinson (1985, pp. 48–50, 122–123) analyzed data given originally in Ruppert and Carroll (1980) examining the relationship between the salinity of water in Pamlico Sound, NC, and the salinity lagged two weeks, a seasonal effect, and river discharge. After correcting an apparently erroneous data value, Atkinson explored fitting a linear model for both salinity and logged salinity on all three predictors. Table 4 summarizes variable selection for these two model fits for these data. As can be seen, in both cases lagged salinity and river discharge are deemed important predictors, while the seasonal effect is not.

An alternative approach to fitting a linear model with logged salinity as the target is to fit an exponential model with salinity as the target,

$$\text{Salinity} = \exp(\beta_0 + \beta_1 \text{Lagged salinity} + \beta_2 \text{Discharge} + \beta_3 \text{Season}) + \epsilon.$$

Fitting this model, which is an example of model (e) from Section 1, has the advantage of addressing goodness-of-fit in the original units of the target variable, rather than logged units, and is appropriate if the errors are additive in the original scale, rather than multiplicative. Table 4 summarizes variable selection for this exponential model, and shows that AIC_C chooses the model that uses all three predictors.

The AIC_C value for this exponential model is smaller than that for the best linear model, suggesting that the nonlinear model is more appropriate. This is consistent with Atkinson's results, and is supported by an application of the goodness-of-fit test A of Section 3.1, which equals .1124, with (Monte Carlo-based) tail probability .09. The AIC_C values for the exponential model and the linear model with logged salinity as the target are not comparable, but Figure 5 highlights an advantage of the exponential model. The figure gives density estimates of the residuals from the exponential model fit, and from a back-transformed version of the log-linear model fit on the predictors (i.e., the residuals after exponentiating the fitted values from the model using logged salinity as the target). The density estimates are penalized likelihood estimates, with smoothing parameters chosen using AIC_C (see Simonoff 1998). The density estimates show that the residuals from the exponential model are better centered around zero (mean $-.003$ vs. .04, median .07 vs. .10, and mode .09 vs. .11, respectively) and less variable (standard deviation .91 versus standard deviation 1.00). Thus, working in the original scale using the nonlinear

Table 4. AIC_C values for models based on salinity data

Predicting variables	AIC_C *value*
Linear model on salinity	
Lagged salinity	1.9925
Season	3.4027
Discharge	2.8768
Lagged salinity, Season	2.0406
Lagged salinity, Discharge	1.3379
Season, Discharge	2.9016
Lagged salinity, Season, Discharge	1.3896
Linear model on logged salinity	
Lagged salinity	−2.4236
Season	−1.0015
Discharge	−1.5769
Lagged salinity, Season	−2.3892
Lagged salinity, Discharge	−3.2681
Season, Discharge	−1.5545
Lagged salinity, Season, Discharge	−3.2275
Exponential model on salinity	
Lagged salinity	2.1174
Season	3.4020
Discharge	2.8345
Lagged salinity, Season	2.1354
Lagged salinity, Discharge	1.3017
Season, Discharge	2.8492
Lagged salinity, Season, Discharge .	1.2361

(exponential) model apparently results in a better fit for salinity than working in the logged scale using a linear model.

4. CONCLUSION

In this article we have derived and illustrated the AIC_C criterion for general regression models, including semiparametric and additive models. The results given here can be generalized in several ways. The selection criteria can be easily generalized to models of the form (1.1) by modifying the criterion in the ways derived in cases (c)–(e) of Section 2. Linear estimators that are piecewise smooth, but allow for possible discontinuities in the regression function, could also be fit using AIC_C; for an example of such an estimator, see Koo (1997).

Models based on non-Gaussian distributions can be accommodated using quasi-likelihood and generalized linear models. Hurvich and Tsai (1995) obtained a version of AIC_C for the quasi-likelihood model with a parametric linear predictor using a linear approximation to the Kullback–Leibler distance. This approach also can be used to obtain AIC_C for the quasi-likelihood model with a nonparametric smooth function, as follows. Suppose data $\mathbf{y}$ are generated from the true extended quasi–likelihood model (Nelder and Pregibon 1987; McCullagh and Nelder 1989, p. 350)

$$Q^+(\mathbf{y};\mathbf{m},\sigma_0^2) = -\frac{n}{2}\log(\sigma_0^2) - \frac{Q(\mathbf{y};\mathbf{y}) - Q(\mathbf{y};\mathbf{m})}{\sigma_0^2}, \tag{4.1}$$

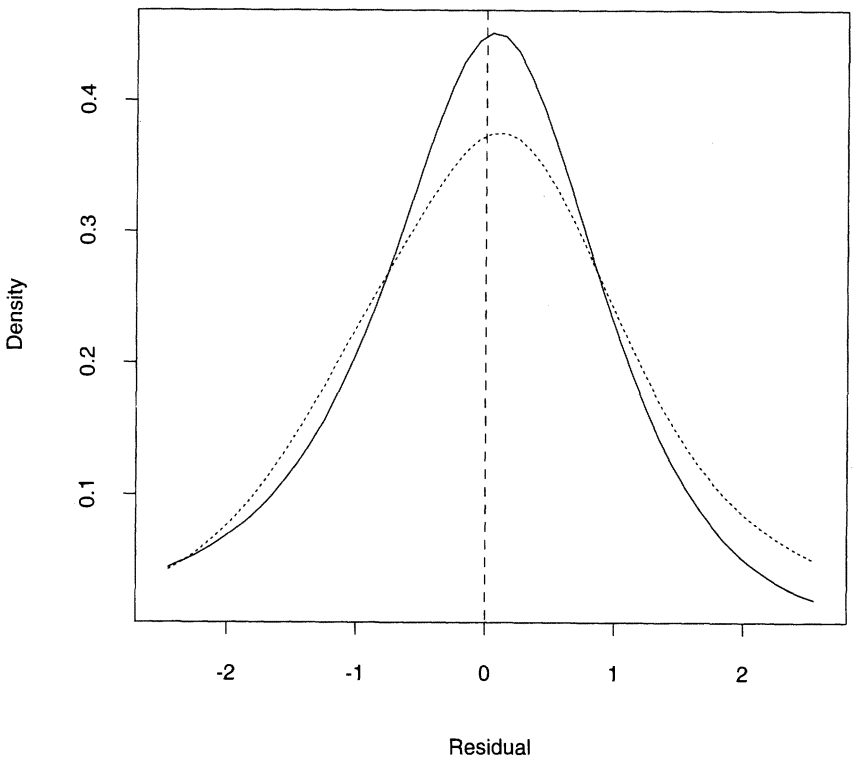

Figure 5. Penalized likelihood density estimates of residuals from exponential fit (solid line) and back-transformed log-linear fit (dotted line) for salinity data.

where $Q(\mathbf{y};\mathbf{m}) = \mathbf{y}'\boldsymbol{\theta}_0 - b(\boldsymbol{\theta}_0) + c(\mathbf{y})$ (see McCullagh and Nelder 1989, p. 336), $E_0(\mathbf{y}) = \mathbf{m}$, $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{10},\ldots,\boldsymbol{\theta}_{n0})'$, $b(\cdot)$ and $c(\cdot)$ are suitably chosen functions, and the relationship between the mean of $\mathbf{y}$ and covariate $\mathbf{x}$ (an $n \times 1$ vector) is linked by the function $k(\mathbf{m}) = \boldsymbol{\theta}_0 = h(\mathbf{x})$. The candidate model is the same as Equation (4.1), replacing $\mathbf{m}$, $\boldsymbol{\theta}_0$, σ_0^2, and E_0 by $\boldsymbol{\mu}$, $\boldsymbol{\theta}$, σ^2, and E, respectively, with E denoting expectation under the candidate model. Applying the same techniques as were used in Section 2, we obtain

$$\begin{aligned}\Delta &= E_0[d(\hat{\mathbf{m}},\hat{\sigma}^2)] \\ &= E_0[n\log(\hat{\sigma}^2)] + E_0[2\{\mathbf{y}'\mathbf{y} - b(\mathbf{y}) - \mathbf{m}'\hat{\boldsymbol{\theta}} + b(\hat{\boldsymbol{\theta}})\}/\hat{\sigma}^2],\end{aligned}$$

where $\hat{\boldsymbol{\theta}} = H\mathbf{y}$, $\hat{\sigma}^2 = (\mathbf{y}-\hat{\mathbf{m}})'\hat{V}(\mathbf{y}-\hat{\mathbf{m}})/n$, $\hat{\mathbf{m}} = k^{-1}(\hat{\boldsymbol{\theta}})$, $\hat{V}$ is

$$V = \frac{\partial^2 b(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$$

evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and H can be obtained from O'Sullivan, Yandell, and Raynor (1986) or Green and Silverman (1994, chap. 5). By making the assumptions (A.2) and (A.3) and by using appropriate approximations, Δ is approximately

$$\tilde{\Delta} = E_0[n\log(\hat{\sigma}^2)] + E_0\left[\frac{n^2\sigma_0^2}{\boldsymbol{\epsilon}^{*\prime}(I-H^*)'(I-H^*)\boldsymbol{\epsilon}^*}\right] + E_0\left[\frac{n\boldsymbol{\epsilon}^{*\prime}H^{*\prime}H^*\boldsymbol{\epsilon}^*}{\boldsymbol{\epsilon}^{*\prime}(I-H^*)'(I-H^*)\boldsymbol{\epsilon}^*}\right],$$

where $\sigma_0^2 = 2E_0[\mathbf{y}'\mathbf{y} - b(\mathbf{y}) - \mathbf{m}'\boldsymbol{\theta} + b(\boldsymbol{\theta})]/n$, $\boldsymbol{\epsilon}^* = V_0^{-1/2}(\mathbf{y} - \mathbf{m})$, V_0 is V evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, and $H^* = V_0^{1/2} H V_0^{1/2\prime}$. Then, AIC_C has the same form as (2.5), replacing H with $\hat{H}^* = \hat{V}^{1/2} H \hat{V}^{1/2\prime}$.

The AIC_C criterion can be obtained in an analogous way for the nonparametric generalized linear model (Green and Silverman 1994, p. 98), generalized additive model (Hastie and Tibshirani 1990), generalized partially linear single index model (Carroll, Fan, Gijbels, and Wand 1997), generalized semilinear model (Emond and Self 1997) and smoothing methods for categorical data (Simonoff 1996, chap. 6). The close connection between categorical data smoothing and local likelihood density estimation and penalized likelihood density estimation means that AIC_C also can be used as a smoothing parameter selector in the density estimation context, as was done in Figure 5 (see Simonoff 1998).

The S-Plus functions and data sets used for the examples in Section 3 can be obtained via the World Wide Web at the address http://www.stern.nyu.edu/~jsimonof/aiccsemi.dmp.

ACKNOWLEDGMENTS

The authors thank Cliff Hurvich for many helpful discussions of this material. The comments of the Associate Editor and two anonymous referees helped to improve the presentation here. Chih-Ling Tsai's research was supported in part by National Science Foundation grant DMS–95–10511.

[Received October 1997. Revised February 1998.]

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *2nd International Symposium on Information Theory*, eds. B.N. Petrov and F. Csaki, Budapest: Akademia Kiado, pp. 267–281.

Atkinson, A. C. (1985), *Plots, Transformations, and Regression*, Oxford: Clarendon Press.

Bickel, P. J., and Rosenblatt, M. (1973), "On Some Global Measures of the Deviation of Density Function Estimates," *The Annals of Statistics*, 1, 1071–1095.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 92, 477–489.

Chen, H., and Shiau, J.-J. H. (1991), "A Two–Stage Spline Smoothing Method for Partially Linear Models," *Journal of Statistical Planning and Inference*, 27, 187–201.

——— (1994), "Data–Driven Efficient Estimators for a Partially Linear Model," *The Annals of Statistics*, 22, 211–237.

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data with Spline Functions," *Numerische Mathematik*, 31, 377–403.

Emond, M. J., and Self, S. G. (1997), "An Efficient Estimator for the Generalized Semilinear Model," *Journal of the American Statistical Association*, 92, 1033–1040.

Eubank, R. L., and Hart, J. D. (1992), "Testing Goodness-of-Fit in Regression via Order Selection Criteria," *The Annals of Statistics*, 20, 1412–1425.

Eubank, R. L., Li, C.-S., and Wang, S. (1997), "Testing Lack-of-Fit of Parametric Regression Models Using Nonparametric Regression Techniques," in *Proceedings of the New York University Symposium on Recent Developments in Smoothing Methods*, May 30, 1997, pp. 103–131.

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, London: Chapman and Hall.

Härdle, W., Hall, P., and Marron, J. S. (1988), "How Far Are Automatically Chosen Regression Smoothing Parameters From Their Optimum?," *Journal of the American Statistical Association*, 83, 86–101.

Hart, J. D. (1997), *Nonparametric Smoothing and Lack-of-Fit Tests*, New York: Springer–Verlag.

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Hurvich, C.M., and Tsai, C.-L. (1989), "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297–307.

——— (1995), "Model Selection for Extended Quasi-Likelihood Models in Small Samples," *Biometrics*, 51, 1077–1084.

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1998), "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion," *Journal of the Royal Statistical Society*, Ser. B, 60, 271–293.

Koo, J.-Y. (1997), "Spline Estimation of Discontinuous Regression Functions," *Journal of Computational and Graphical Statistics*, 6, 266–284.

Linhart, H., and Zucchini, W. (1986), *Model Selection*, New York: Wiley.

Lock, R. H. (1993), "1993 New Cars Data," *Journal of Statistics Education*, 1, No. 1.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, (2nd ed.), London: Chapman and Hall.

Nelder, J. A., and Pregibon, D. (1987), "An Extended Quasi-Likelihood Function," *Biometrika*, 74, 221–232.

O'Sullivan, F., Yandell, B. S., and Raynor, W. J., Jr. (1986), "Automatic Smoothing of Regression Functions in Generalized Linear Models," *Journal of the American Statistical Association*, 81, 96–103.

Ruppert, D., and Carroll, R. J. (1980), "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association*, 75, 828–838.

Shibata, R. (1981), "An Optimal Selection of Regression Variables," *Biometrika*, 68, 45–54.

Simonoff, J. S. (1985), "An Improved Goodness-of-Fit Statistic for Sparse Multinomials," *Journal of the American Statistical Association*, 80, 671–677.

——— (1996), *Smoothing Methods in Statistics*, New York: Springer–Verlag.

——— (1998), "Three Sides of Smoothing: Categorical Data Smoothing, Nonparametric Regression, and Density Estimation," *International Statistical Review*, 66, 137–156.

Speckman, P. L. (1988), "Kernel Smoothing in Partial Linear Models," *Journal of the Royal Statistical Society*, Ser. B, 50, 413–436.

U.S. News and World Report (1996), *1997 America's Best Colleges*, Washington, D.C.: author.