Information and Posterior Probability Criteria for Model Selection in Local Likelihood Estimation

Rafael A. IRIZARRY

Local likelihood estimation has proven to be an effective method for obtaining estimates of parameters that vary with a covariate. To obtain useful estimates of such parameters, approximating models are used. In such cases it is useful to consider window based estimates. We may need to choose between competing approximating models. In this paper we propose a modification to the methods used to motivate many information and posterior probability criteria for the weighted likelihood case. We derive weighted versions for two of the most widely known criteria, namely the AIC and BIC. Via a simple modification, the criteria are also made useful for window span selection. The usefulness of the weighted version of these criteria are demonstrated through a simulation study and an application to three data sets.

KEY WORDS: Information Criteria; Posterior Probability Criteria; Model Selection; Local Likelihood.

1. INTRODUCTION

Local regression has become a popular method for smoothing scatterplots and for nonparametric regression in general. It has proven to be a useful tool in finding structure in datasets (Cleveland and Devlin 1988). Local regression estimation is a method for smoothing scatterplots $(\mathbf{x}_i, y_i)$, $i = 1, \dots, n$ in which the fitted value at $\mathbf{x}_0$ is the value of a polynomial fit to the data using weighted least squares where the weight given to $(\mathbf{x}_i, y_i)$ is related to the distance between $\mathbf{x}_i$ and $\mathbf{x}_0$. Stone (1977) shows that estimates obtained using the local regression methods have desirable theoretical properties. Recently, Fan (1993) has studied minimax properties of local linear regression.

Tibshirani and Hastie (1987) extend the ideas of local regression to a local likelihood procedure. This procedure is designed for nonparametric regression modeling in situations where weighted least squares is inappropriate as an estimation method, for example binary data. Local regression may be viewed as a special case of local likelihood estimation. Tibshirani and Hastie (1987), Staniswalis (1989), and Loader

Rafael A. Irizarry is Assistant Professor, Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205.
The author thanks David Brillinger for his mentoring, and the Associate Editor and two referees for helpful comments.

(1999) apply local likelihood estimation to several types of data where local regression is not appropriate and find it provides useful information about the data.

In local likelihood estimation, we may need to select an approximating model for the local likelihood structure. Automatic model selection criteria are desirable. In this paper we propose a way to modify criteria, based on information theory and posterior probabilities, for selecting from amongst competing models for cases where weighing observations for estimation purposes is desirable.

Section 2 gives a brief overview of local likelihood estimation. Section 3 reviews the concepts behind the information and posterior probability criteria and develops criteria for weighted estimates as those used in local likelihood estimation. As an example we develop weighted information criteria based on Akaike's (1973) AIC, Takeuchi's (1976) TIC, and Bozdogan's (1987) CAICF, and a posterior probability criterion based on Schwarz's (1978) BIC and Neath and Cavanaugh's (1997) SIC_f. In Section 4 we explore ways in which the criteria developed can be extended so as to be used for window size selection. In Section 5 we illustrate the usefulness of the weighted criteria via simulation studies. Section 6 presents examples of the criteria being applied to two binary data sets and to a signal processing problem. Section 7 gives final remarks.

2. LOCAL LIKELIHOOD ESTIMATION

Suppose we have independent observations $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ that are the realization of a response random variable Y given a $P \times 1$ covariate vector $\mathbf{x}$ which we consider to be known. Given the covariate $\mathbf{x}$, the response variable Y follows a parametric distribution $Y \sim g(y|\theta)$ where θ is a function of $\mathbf{x}$. We are interested in estimating θ using the observed data.

The log-likelihood function can be written as

$$l(\theta_1, \dots, \theta_n) = \sum_{i=1}^{n} \log g(y_i|\theta_i) \tag{1}$$

where $\theta_i = s(\mathbf{x}_i)$. A standard modeling procedure would assume a parsimonious form for the θ_is, say $\theta_i = \mathbf{x}_i'\boldsymbol{\beta}$, $\boldsymbol{\beta}$ a $P \times 1$ parameter vector. In this case the log-likelihood $l(\theta_1, \dots, \theta_n)$ would be a function of the parameter $\boldsymbol{\beta}$ that could be estimated by maximum likelihood, that is by finding the $\hat{\boldsymbol{\beta}}$ that maximizes $l(\theta_1, \dots, \theta_n)$.

The local likelihood approach is based on a more general assumption, namely that $s(\mathbf{x})$ is a "smooth" function of the covariate $\mathbf{x}$. Without more restrictive assumptions, the maximum likelihood estimate of $\boldsymbol{\theta} = \{s(\mathbf{x}_1), \dots, s(\mathbf{x}_n)\}$ is no longer useful because of over-fitting. Notice, for example, that for the case of regression with all the $\mathbf{x}_i$s distinct, the maximum likelihood estimate would simply reproduce the data.

Suppose we are interested in estimating only $\theta_0 = \theta(\mathbf{x}_0)$ for a fixed covariate value $\mathbf{x}_0$. The local likelihood estimation approach is to assume that there is some neighborhood N_0 of covariates that are "close" enough to $\mathbf{x}_0$ such that the data $\{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in N_0\}$ contain information about θ_0 through some *link function* η of the form

$$\theta_0 = s(\mathbf{x}_0) \equiv \eta(\mathbf{x}_0, \boldsymbol{\beta}) \text{ and } \theta_i = s(\mathbf{x}_i) \approx \eta(\mathbf{x}_i, \boldsymbol{\beta}), \text{ for } \mathbf{x}_i \in N_0. \tag{2}$$

Notice that we are abusing notation here since we are considering a different $\boldsymbol{\beta}$ for every $\mathbf{x}_0$. Throughout the work we will be acting as if θ_0 is the only parameter of interest and therefore not indexing variables that depend on the choice of $\mathbf{x}_0$. However, in practice we find an estimate for each $\theta_i, i = 1, \dots, n$ by repeating the procedure for $\mathbf{x}_0 = \mathbf{x}_i, i = 1, \dots, n$.

The local likelihood estimate of θ_0 is obtained by assuming that, for data in N_0, the true distribution of the data, $g(y_i|\theta_i)$ is approximated by $f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) \equiv g(y_i|\eta(\mathbf{x}_i, \boldsymbol{\beta}))$, finding the $\hat{\boldsymbol{\beta}}$ that maximizes the *weighted log-likelihood*

$$l_0(\boldsymbol{\beta}) = \sum_{\mathbf{x}_i \in N_0} w_i \log f(y_i|\mathbf{x}_i, \boldsymbol{\beta}), \tag{3}$$

and then using Equation (2) to obtain what we will call the *local likelihood estimate* $\hat{\theta}_0$. In order to obtain a useful estimate of θ_0, we need $\boldsymbol{\beta}$ to be of "small" enough dimension so that we fit a parsimonious model within N_0.

In (3) w_i is a weight coefficient related to the "distance" between $\mathbf{x}_0$ and $\mathbf{x}_i$. Throughout this paper we will assume that the weight coefficients are obtained from some function $w(s)$ satisfying Condition 1 shown in the Appendix.

Hastie and Tibshirani (1987) discuss the case where the covariate $\mathbf{x}$ is a real valued scalar and the link function, $\eta(x_i, \boldsymbol{\beta}) = \beta_0 + x_i\beta_1$, is linear. In this case the assumption being made is that the parameter function $s(x_i)$ is approximately linear within "small" neighborhoods of x_0, i.e. locally linear. Staniswalis

(1989) presents a similar approach. In this case the covariate $\mathbf{x}$ is allowed to be a vector and the link function, $\eta(\mathbf{x}_i, \beta) = \beta$, is a constant. The assumption being made here is that the parameter function $s(x_i)$ is locally constant.

If we assume a density function of the form

$$\log g(y_i|\theta_i) = C + (y_i - \theta_i)^2/\phi \tag{4}$$

where K and ϕ are constants that do not depend on the θ_is, local regression may be considered a special case of local likelihood estimation. In this case the local likelihood estimate is going to be equivalent to the estimate obtained by minimizing a sum of squares equation. The approach in Cleveland and Devlin (1988) is to consider a real valued covariate and the polynomial link function $\eta(\mathbf{x}_i, \boldsymbol{\beta}) = \sum_{j=0}^{d} x_i^j \beta_j$.

In general, the approach of local likelihood estimation, including the three above-mentioned examples, is to assume that for "small" neighborhoods around $\mathbf{x}_0$, the distribution of the data is approximated by a distribution that depends on a constant parameter $\boldsymbol{\beta}(\mathbf{x}_0)$, i.e. we have locally parsimonious models. This allows us to use the usual estimation technique of maximum likelihood. However, in the local version of maximum likelihood we often have an a priori belief that points "closer" to $\mathbf{x}_0$ contain more information about θ_0, which suggests a weighted approach.

3. MODEL SELECTION

Suppose we observe a realization of a random variable with distribution as defined in the previous section, and suppose we are interested in estimating $\theta_0 = s(\mathbf{x}_0)$, for some covariate $\mathbf{x}_0$, with a local likelihood approach. For a neighborhood N_0 around the covariate $\mathbf{x}_0$, we approximate the joint distribution of the response variable $\mathbf{Y} = \{Y_i; x_i \in N_0\}$ with

$$g_{\mathbf{Y}}(\mathbf{y}) \approx \prod_{\mathbf{x}_i \in N_0} f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) \equiv f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) \tag{5}$$

where $\mathbf{X}$ is a matrix with row entries $\mathbf{x}_i \in N_0$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_P)' \in \mathbb{R}^P$ is a $P \times 1$ parameter vector. Notice that we are suppressing the θs and using $g_{\mathbf{Y}}(\mathbf{y})$ to represent the true distribution of $\mathbf{Y}$.

The local likelihood approach is to estimate $\boldsymbol{\beta}$ in order to obtain an estimate of θ_0. However, suppose that before doing so we need to choose from amongst competing approximating models. For example, in local regression, when the covariate is a real valued scalar, we may need to decide if we should fit a constant, linear, or quadratic function of x.

We will consider the situation where we are choosing from amongst P competing models generated by simply restricting the general parameter space $\mathbb{R}^P$ in which $\boldsymbol{\beta}$ lies. In terms of the parameter, we represent the approximate models as

$$M_p = \left\{ f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \prod_{x_i \in N_0} f(y_i|\mathbf{x}_i, \boldsymbol{\beta}); \boldsymbol{\beta} \in \Omega_p \right\} \tag{6}$$

with Ω_p the sub-space of $\mathbb{R}^P$ defined by the following restriction: $\Omega_p = \left\{\boldsymbol{\beta} \in \mathbb{R}^P : \beta_{p+1} = \ldots = \beta_P = 0\right\}$. We will refer to p as the *number of parameters* in the approximate model.

Given that we have chosen a particular model, say M_p, we can find the $\hat{\boldsymbol{\beta}}_p$ that minimizes the weighted log-likelihood (3). Now, if instead we choose another model, say M_q, and obtain $\hat{\boldsymbol{\beta}}_q$ how do we compare these two estimates? Which one is better? As we will see in the next section, using (3) as a criterion is not practical. In this paper we develop criteria aimed at answering these questions.

3.1 Information Criteria

We wish to select the approximate model, defined by (6), that is "nearest" to the true model, defined by (5), based on the observed data $\mathbf{y}$. The principle behind information criteria is to define "nearest" using the Kullback-Leibler discrimination information (Kullback 1959)

$$K\left\{g_{\mathbf{Y}}(\mathbf{y}) : f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})\right\} = \int g_{\mathbf{Y}}(\mathbf{y}) \log \frac{g_{\mathbf{Y}}(\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})}\, d\mathbf{y}. \tag{7}$$

As done by Sawa (1978), we say that M_q is the "nearest" or best approximating model amongst the models defined by (6), if and only if $\inf_{\boldsymbol{\beta} \in \Omega_q} K\left\{g_{\mathbf{Y}}(\mathbf{y}) : f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})\right\} < \inf_{\boldsymbol{\gamma} \in \Omega_p} K\left\{g_{\mathbf{Y}}(\mathbf{y}) : f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\gamma})\right\}$ for any $1 \leq p \neq q \leq P$.

Assuming that $g_{\mathbf{Y}}(\mathbf{y})$ belongs to the set of competing models defined by (6), Akaike's (1973) Information Criterion (AIC) was developed in order to use observed data to estimate the true model. Takeuchi (1976) developed the TIC, a generalization of AIC to situations where the fitted model is not necessarily properly specified (Takeuchi (1976) is in Japanese, other references are Kitagawa (1987) and Shibata (1989)). Akaike's original work is for independent identically distributed (IID) data, however it is extended to a regression type setting in a straight-forward way (Hurvich and Tsai 1989). The approach is to choose the approximate model producing the estimate $\hat{\boldsymbol{\beta}}_p$ that minimizes $E_{\mathbf{Y}}\left[K\left\{g_{\mathbf{Y}}(\mathbf{y}) : f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X}, \hat{\boldsymbol{\beta}}_p)\right\}\right]$ with the expectation taken under the true distribution of the $\mathbf{Y}$. Since the first term on the right hand side of (7) is constant over all models, we may instead minimize the second term which can be written as

$$\mathrm{E}_{\mathbf{Y}}\left\{\sum_{\mathbf{x}_i \in N_0} - \int g(y_i|\theta_i)\, \log f(y_i|\mathbf{x}_i, \hat{\boldsymbol{\beta}}_p)\, dy_i\right\}. \tag{8}$$

Information criteria are obtained by constructing asymptotically unbiased estimates of (8).

Criteria such as AIC/TIC have been criticized for providing an estimate of (8) that is too simplistic. Furthermore, these criteria do not produce asymptotically consistent estimates of the correct or best approximating model. Several authors have proposed ways to obtain better approximations of this quantity. For example, Bozdogan (1987,1994) has developed the information-based complexity criterion (ICOMP), the inverse-Fisher information matrix criterion (IFIM), and the consistent AIC with Fisher Information (CAICF). For the case of normally distributed data, Hurvich and Tsai (1989) developed a "corrected" version of AIC and one of the referees suggested a corrected version of CAICF. These corrected criteria provide a significant improvement when the number of parameters of the true model p is close to the number of observations n.

Because we are assuming a priori knowledge that there is more information about the parameter θ_0 for data points associated with covariates "near" $\mathbf{x}_0$, it seems appropriate to consider a discrepancy measure that takes this into account. In this paper we propose the use of a weighted version of the Kullback-Leibler discrimination information (Gokhale and Kullback 1978) and use this to derive appropriate model selection information criteria.

3.2 Weighted Information Criteria

Because in local estimation we are interested in estimating only θ_0, we say, as done by Sawa (1978), that $\hat{\boldsymbol{\beta}}$ is a better estimate than $\hat{\gamma}$, if and only if

$$\mathrm{E}_{\mathbf{Y}}[K\{g(y_0|\theta_0) : f(y_0|\mathbf{x}_0, \hat{\boldsymbol{\beta}})\}] < \mathrm{E}_{\mathbf{Y}}[K\{g(y_0|\theta_0) : f(y_0|\mathbf{x}_0, \hat{\gamma})\}]. \tag{9}$$

with the expectation taken under the true distribution of the $\mathbf{Y}$.

In practice we usually have only one realization of Y_0, and thus we consider a discrepancy measure based on a weighted version of the Kullback-Leibler discrimination information that leads us to weighted maximum likelihood estimation that uses the data in N_0. Analogous to Section 2, the approach is to choose the model producing the estimate $\hat{\boldsymbol{\beta}}_p$ that minimizes

$$\mathrm{E}_{\mathbf{Y}}\left\{\sum_{\mathbf{x}_i\in N_0} -w_i \int g(y_i|\theta_i)\, \log f(y_i|\mathbf{x}_i,\hat{\boldsymbol{\beta}}_p)\, dy_i\right\}. \tag{10}$$

As in Akaike (1973), we notice that the sample version of (10), $\sum_{x_i\in N_0} -w_i \log f(y_i|\mathbf{x}_i,\hat{\boldsymbol{\beta}}_p) = -l_0(\hat{\boldsymbol{\beta}}_p)$, will underestimate (10). In general, the larger p, the more $l_0(\hat{\boldsymbol{\beta}})$ underestimates (10) making $l_0(\hat{\boldsymbol{\beta}}_p)$ a criterion that will favor larger models. We use Theorem 1 below to obtain an estimate of this bias.

Theorem 1. For any $\mathbf{x}_0$ there exists an appropriate sequence of neighborhoods $N_{0,n}$ such that under Conditions 1-4, presented in the Appendix, and with the coefficients w_i obtained from a weight function satisfying Condition 5, also in the Appendix, we have that

$$\mathrm{E}_{\mathbf{Y}}\left\{l_0(\hat{\boldsymbol{\beta}}_p) - \sum_{\mathbf{x}_i\in N_{0,n}} w_i \int g(y_i|\theta_i)\, \log f(y_i|\mathbf{x}_i,\hat{\boldsymbol{\beta}}_p)\, dy_i\right\} = \mathrm{tr}\left\{I_n(\boldsymbol{\beta}_p)J_n(\boldsymbol{\beta}_p)^{-1}\right\} + o(1)$$

with

$$I_n(\boldsymbol{\beta}) = \mathrm{E}_{\mathbf{Y}}\left[\left\{\frac{\partial}{\partial\boldsymbol{\beta}} l_0(\boldsymbol{\beta})\right\}\left\{\frac{\partial}{\partial\boldsymbol{\beta}} l_0(\boldsymbol{\beta})\right\}'\right] \text{ and } J_n(\boldsymbol{\beta}) = \mathrm{E}_{\mathbf{Y}}\left[\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} l_0(\boldsymbol{\beta})\right] \tag{11}$$

and $\boldsymbol{\beta}_p$ defined by Condition 3 in the Appendix. See the Appendix for a discussion of the theoretical justification of this theorem.

Theorem 1 provides a way to obtain a bias corrected estimate of (10). We may then choose the model M_p producing the estimate $\hat{\boldsymbol{\beta}}_p$ that minimizes the following model selection criterion:

$$\mathrm{WAIC}(p) = -2l_0(\hat{\boldsymbol{\beta}}_p) + 2\mathrm{tr}\left\{I_n(\boldsymbol{\beta}_p)J_n(\boldsymbol{\beta}_p)^{-1}\right\}. \tag{12}$$

In practice we may not be able to obtain $\mathrm{tr}\left\{I_n(\boldsymbol{\beta}_p)J_n(\boldsymbol{\beta}_p)^{-1}\right\}$. For the equally weighted case, many criteria have been developed by considering different ways of estimating this quantity (Ljung and Caines (1979), Chow (1981), Shibata (1989)). For simplicity, in this paper we propose using a sample version substituting

$$I_n(\boldsymbol{\beta}_p) \text{ with } \hat{I}_{n,p} \equiv \left\{ \frac{\partial}{\partial \boldsymbol{\beta}} l_0(\hat{\boldsymbol{\beta}}_p) \right\} \left\{ \frac{\partial}{\partial \boldsymbol{\beta}} l_0(\hat{\boldsymbol{\beta}}_p) \right\}' \text{ and } J_n(\boldsymbol{\beta}_p) \text{ with } \hat{J}_{n,p} \equiv \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} l_0(\hat{\boldsymbol{\beta}}_p). \tag{13}$$

Notice that if $w_i = 1$ for all i and the N_0 include all the data for all n, Theorem 1 reduces to the result obtained by Takeuchi (1976) and WAIC = TIC, the criterion proposed by Takeuchi. Furthermore, if $g_{\mathbf{Y}}(\mathbf{y})$ is included in one of the models defined by (6) then $I_n(\boldsymbol{\beta}_p) = J_n(\boldsymbol{\beta}_p)$ and $\mathrm{tr}\left\{I_n(\boldsymbol{\beta}_p) J_n(\boldsymbol{\beta}_p)^{-1}\right\} = p$ reducing WAIC to AIC.

The WAIC can be criticized, similarly to the AIC, for providing an approximation to the bias of $l_0(\hat{\boldsymbol{\beta}}_p)$ that is too simplistic. Furthermore, the WAIC does not produce asymptotically consistent estimates of the best approximate model. However, improvements can be easily obtained by developing the weighted version of criteria that improve AIC. For example, Bozdogan (1987) develops a modification to the AIC that provides a criteria that produces asymptotically consistent estimates of the correct model (this is under the assumption of IID data and that the correct model is one of the competing models). Corollary 1, based on the derivation of Bozdogan (1987) and presented in the Appendix, motivates the following weighted criteria

$$\mathrm{WCAICF} = -2 l_0(\hat{\boldsymbol{\beta}}_p) + 2\mathrm{tr}\left\{I_n(\boldsymbol{\beta}_p) J_n(\boldsymbol{\beta}_p)^{-1}\right\} + \log\det\left\{-J_n(\boldsymbol{\beta}_p)\right\}. \tag{14}$$

If we consider the equally weighted case then, under the assumptions made by Bozdogan (1987), we have $\mathrm{tr}\left\{I_n(\boldsymbol{\beta}_p) J_n(\boldsymbol{\beta}_p)^{-1}\right\} = p$ and $\log\det\left\{-J_n(\boldsymbol{\beta}_p)\right\} = p \log n + \log\det\{-J(\boldsymbol{\beta}_p)\}$, with $J(\boldsymbol{\beta}_p)$ the matrix of second partials of the likelihood based on one observation, and the WCAICF reduces to the CAICF. In practice, as for the WAIC, we substitute $I_n(\boldsymbol{\beta}_p)$ and $J_n(\boldsymbol{\beta}_p)$ with the sample versions given in (13).

Notice that the penalty added to the WAIC by the WCAICF is the sum of the log of the eigenvalues of the weighted information matrix. In a way, this measures how much information about the non-zero components of the parameter $\boldsymbol{\beta}_p$ is available in the data.

3.3 Posterior Probability Model Selection

A Bayesian-type approach for model selection was first suggested by Schwarz (1978). This approach is based on assigning a prior probability to each model, assigning a prior distribution to the parameter vector conditional on the model, and maximizing the posterior probabilities of the alternative models, given the observations. Schwarz considers only IID data, but the ideas are easily extended to other situations as done by and Kashyap (1982). In particular, Neath and Cavanaugh (1997) consider the case of regression.

To derive a posterior probability criterion that takes weights into account, we start by considering the approximation given by (5) to actually hold true. We act as if $f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X},\boldsymbol{\beta})$ is the distribution of $\mathbf{Y}$ within the neighborhood N_0. We then let $\Pr(M_p)$ be the prior probability of model M_p being correct, and $\mu(\boldsymbol{\beta}|M_p)$ the prior density for the parameter vector $\boldsymbol{\beta}$ conditioned on M_p being correct. Notice that $\mu(\boldsymbol{\beta}|M_p)$ is positive only if $\boldsymbol{\beta} \in \Omega_p$. By Bayes' theorem the posterior probability of M_p being the correct model is

$$\Pr(M_p|\mathbf{Y}=\mathbf{y}) = \frac{\Pr(M_p)\int_{\Omega_p} f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X},\boldsymbol{\beta})\mu(\boldsymbol{\beta}|M_p)\,d\boldsymbol{\beta}}{\sum_{q=1}^{P}\Pr(M_q)\int_{\Omega_q} f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X},\boldsymbol{\beta})\mu(\boldsymbol{\beta}|M_q)\,d\boldsymbol{\beta}}. \tag{15}$$

Since the denominator depends neither on the model nor on the data, we need only to maximize the numerator when choosing models. Furthermore, in this paper we will consider only a uniform prior for the models, so $\Pr(M_p)$ is a constant that doesn't need to be considered when maximizing the numerator.

Similar to the way we generalized information criteria to derive the WAIC, we obtain a weighted version of BIC by considering a weighted version of (15). In this case the numerator of the posterior probability that we look to maximize is

$$\int_{\Omega_p} \exp\left\{\sum_{\mathbf{x}_i\in N_0} w_i \log f(y_i|\mathbf{x}_i,\boldsymbol{\beta})\right\}\mu(\boldsymbol{\beta}|M_p)\,d\boldsymbol{\beta}.$$

The following theorem gives us an approximation of this quantity.

Theorem 2. If each model has positive prior probability $\Pr(M_p)$ and the assumption of Theorem 1 hold then

$$\log\int_{\Omega_p} \exp\left\{\sum_{i=1}^{n} w_i \log f(y_i|\mathbf{x}_i,\boldsymbol{\beta})\right\}\mu_p(\boldsymbol{\beta}|M_p)\,d\boldsymbol{\beta} = l_0(\hat{\boldsymbol{\beta}}_p) - \frac{1}{2}\log\det\left\{-\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}l_0(\hat{\boldsymbol{\beta}}_p)\right\} + O(1).$$

See the Appendix for a discussion of the theoretical considerations needed to prove this theorem.

Theorem 2 motivates using a posterior probability criterion of the form

$$\mathrm{WBIC}(p) = -2l_0(\hat{\boldsymbol{\beta}}_p) + \log\det\left\{-\hat{J}_{n,p}\right\}. \tag{16}$$

If we consider the equally weighted case then WBIC reduces to Neath and Cavanaugh's SIC_f. Furthermore, if the components of $\mathbf{Y}$ are IID then $\log\det\{-\hat{J}_{n,p}\} = p\log n + \log\det\{-J(\hat{\boldsymbol{\beta}}_p)\}$ which is asymptotically equivalent to $p\log n$ the penalty term of the BIC. Here $J(\boldsymbol{\beta}_p)$ is defined as in the previous section.

Notice that in practice the penalty terms of the WBIC and WCAICF differ by $\mathrm{tr}\left\{\hat{I}_{n,p}\hat{J}_{n,p}^{-1}\right\}$, which will be negligible compared to $\log\det\{-\hat{J}_{n,p}\}$ for large n. The fact that the WBIC and WCAICF are asymptotically equivalent is corroborated by the simulations in Section 5.

4. EXTENSION TO WINDOW SIZE SELECTION

We have assumed that the functional parameter $\boldsymbol{\theta}$ defining the distribution of $\mathbf{Y}$ depends on the regression variable $\mathbf{x}$. At this moment we are interested only in estimating a given θ_0 associated with the covariate $\mathbf{x}_0$. We have proposed using local likelihood estimation. If we were also trying to choose "optimal" window coefficients w_i in equation (3), we could define "optimal" as the coefficients that produced the best estimate $\hat{\theta}_0$ as defined by (9) or equivalently as the estimate minimizing $\mathrm{E}_{\mathbf{Y}}\{-\int g(y_0|\theta_0)\log g(y_0|\hat{\theta}_0)\,dy_0\}$. However, there may be only one observation y_0 to be used to estimate this quantity. In Section 3.2 we proposed a criterion used to compare estimates obtained by fitting approximate models, with different number of parameters, to neighborhoods of the data. The same weight coefficients were used when computing these estimates. Now we want to compare estimates obtained using different weight coefficients. The criteria developed in the previous sections are not appropriate since estimates using "heavier" weight coefficients or more non-zero coefficients would produce larger values of the criteria. This problem may be easily resolved by dividing by the total weight used in the estimation $W_0 = \sum_{i=1}^n w_i$.

Intuitively, we argue that the weighted average information (10) is an estimate of the "true" information quantity in the following way:

$$\mathrm{E}_{\mathbf{Y}}\left\{-\int g(y_0|\theta_0)\log g(y_0|\hat{\theta}_0^{(p)})\,dy_0\right\} \approx \frac{1}{W_0}E_{\mathbf{Y}}\left\{\sum_{\mathbf{x}_i\in N_0} -w_i\int \log g(y_i|\theta_i)\,f(y_i|\mathbf{x}_i,\hat{\boldsymbol{\beta}}_p)\,dy_i\right\}$$

with $\hat{\theta}_0^{(p)}$ the estimate obtained from $\hat{\boldsymbol{\beta}}_p$ using (2).

Similarly, for the Bayesian approach

$$\log \int_{\Omega_p} f(y_0|\mathbf{x},\boldsymbol{\beta})\mu(\boldsymbol{\beta}|M_p)\,d\boldsymbol{\beta} \approx \frac{1}{W_0}\log \int_{\Omega_p} \exp\left\{\sum_{i=1}^{n} w_i \log f(y_i|\mathbf{x}_i,\boldsymbol{\beta})\right\}\mu(\boldsymbol{\beta}|M_p)\,d\boldsymbol{\beta}.$$

Because we have limited knowledge of the global behavior of the function $\theta(\mathbf{x})$, we might want to consider different weight coefficients depending on how much importance we want to give to certain parts of the data. In general, it seems appropriate to weigh the central values more heavily. The question is then what portion of the data is given "significant" weight.

A convenient way to assign weights is by choosing a weight function $w(s)$ satisfying Condition 1 and considering different values of the span or window size h to define $w_i(h)$ using (23). Then, for each h we will have a different estimate for each model M_p, $\hat{\boldsymbol{\beta}}_p(h)$. This in turn defines estimates $\hat{\theta}_0^{(p)}(h)$ of θ_0 each with a total weight, $W_0(h) = \sum_{i=1}^n w_i(h)$, associated with it. We want to choose the best estimator according to (9). The information criteria developed for deciding on the order of the approximate model to be used may be extended to decide on what span or window size to use by basing our estimate on the average weighted log-likelihood $l_0\{\hat{\boldsymbol{\beta}}_p(h)\}/W_0(h)$. The weighted criteria are then: $\mathrm{WAIC}(p,h) = \mathrm{WAIC}(p)/W_0(h)$, $\mathrm{WCAICF}(p,h) = \mathrm{WCAICF}(p)/W_0(h)$, and $\mathrm{WBIC}(p,h) = \mathrm{WBIC}(p)/W_0(h)$ with $\mathrm{WAIC}(p)$, $\mathrm{WCAICF}(p)$ and $\mathrm{WBIC}(p)$ defined by (12), (14), and (16) respectively. Notice that these criteria penalize for both large values of p and small values of h.

5. A SIMULATION STUDY

5.1 Bozdogan's simulation

As in Bozdogan (1987) we define the n component random variable

$$Y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \beta_4 x_i^3 + \epsilon_i, i = 1, \ldots, n$$

with $\boldsymbol{\beta} = (1, 5, -1.25, 0.15)$, $x_i = i$ and the ϵ_i independent normal with mean 0 and variance σ^2. We consider $\theta_{n/2} \equiv \mathrm{E}(Y_{n/2}|x_{n/2})$ to be the parameter of interest. We compute the local likelihood estimates of $\theta_{n/2}$, using Tukey's triweight function, and consider the competing models to be polynomials of order 0 through 6, which define models with dimension 1 through 7 respectively.

Table 1. Percentage of times each dimension is chosen by each criteria in 1000 replications for varying sample size n and variance σ^2.

		Estimated Dimension									Estimated Dimension						
Exprmnt	Crit	1	2	3	**4**	5	6	7	Exprmnt	Crit	1	2	3	**4**	5	6	7
	AIC	0	0	0	67.8	16.4	10.6	5.2		AIC	0	0	0	65.9	16.9	11.2	6
	AIC_c	0	0	0	71	15	9.9	4.1		AIC_c	0	0	0	70.5	15	10	4.5
	WAIC	0	0	0	76.2	8	5.1	10.7		WAIC	0	0	0	76.7	7.4	3.6	12.3
$n = 50$	BIC	0	0	0	77	12	8	3	$n = 50$	BIC	0	0	0	75.6	12.7	8.2	3.5
$\sigma^2 = 0.25$	SIC_f	0	0	0	99	0.8	0.2	0	$\sigma^2 = 5$	SIC_f	0	0	0	99.2	0.6	0.2	0
	WBIC	0	0	0	100	0	0	0		WBIC	0	0	0	100	0	0	0
	CAICF	0	0	0	99.5	0.3	0.2	0		CAICF	0	0	0	99.7	0.3	0	0
	CAICF_c	0	0	0	99.5	0.3	0.2	0		CAICF_c	0	0	0	99.7	0.3	0	0
	WCAICF	0	0	0	100	0	0	0		WCAICF	0	0	0	100	0	0	0
	AIC	0	0	0	68.2	13.3	12.1	6.4		AIC	0	0	0	71.2	13.5	11.1	4.2
	AIC_c	0	0	0	70	13	11.1	5.9		AIC_c	0	0	0	72.4	13.4	10.4	3.8
	WAIC	0	0	0	72.2	7.8	9.8	10.2		WAIC	0	0	0	73.2	7	9.4	10.4
$n = 100$	BIC	0	0	0	80.8	9	7.2	3	$n = 100$	BIC	0	0	0	82.3	9.1	6.7	1.9
$\sigma^2 = 0.5$	SIC_f	0	0	0	99.7	0.2	0.1	0	$\sigma^2 = 5$	SIC_f	0	0	0	88.2	6.7	4.5	0.6
	WBIC	0	0	0	100	0	0	0		WBIC	0	0	0	100	0	0	0
	CAICF	0	0	0	100	0	0	0		CAICF	0	0	0	100	0	0	0
	CAICF_c	0	0	0	100	0	0	0		CAICF_c	0	0	0	100	0	0	0
	WCAICF	0	0	0	100	0	0	0		WCAICF	0	0	0	100	0	0	0
	AIC	0	0	0	68.8	14.8	10.8	5.6		AIC	0	0	0	68.4	14.3	11.5	5.8
	AIC_c	0	0	0	69.4	14.7	10.6	5.3		AIC_c	0	0	0	69	14.3	11.3	5.4
	WAIC	0	0	0	69	6.8	15.8	8.4		WAIC	0	0	0	69.2	7.2	15.7	7.9
$n = 200$	BIC	0	0	0	83.7	8.6	5.8	1.9	$n = 200$	BIC	0	0	0	83.9	7.8	6.6	1.7
$\sigma^2 = 1$	SIC_f	0	0	0	99.7	0.1	0.2	0	$\sigma^2 = 5$	SIC_f	0	0	0	99.9	0.1	0	0
	WBIC	0	0	0	100	0	0	0		WBIC	0	0	0	100	0	0	0
	CAICF	0	0	0	99.9	0.1	0	0		CAICF	0	0	0	100	0	0	0
	CAICF_c	0	0	0	99.9	0.1	0	0		CAICF_c	0	0	0	100	0	0	0
	WCAICF	0	0	0	100	0	0	0		WCAICF	0	0	0	100	0	0	0

For this example the likelihood is of the form of equation (4), therefore local likelihood is equivalent to weighted least squares and we can construct the penalty terms for the weighted criteria using

$$\operatorname{tr}\left\{I_n(\boldsymbol{\beta}_p)J_n(\boldsymbol{\beta}_p)^{-1}\right\} = \operatorname{tr}\left\{(\mathbf{X}'\mathbf{W}\mathbf{W}\mathbf{X})(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\right\} \text{ and } \log\det\left\{-J_n(\boldsymbol{\beta}_p)\right\} = \log\det(\mathbf{X}'\mathbf{W}\mathbf{X})/\sigma^2, \tag{17}$$

with $\mathbf{X}$ the regression matrix used in the local regression and $\mathbf{W}$ a diagonal matrix with entries $\mathbf{W}_{i,i} = w_i$.

Table 1 presents the percentage of times each of the competing models is chosen by various criteria in 1000 simulations. Six experiments consisting of various values of σ^2 and n are performed. The tables show that the weighted versions are choosing the correct model (dimension of 4) more frequently than their equally-weighted counterparts, i.e. WAIC outperforms AIC, WBIC outperforms SIC_f and BIC, and WCAICF barely outperforms CAICF. Notice that in all experiments the WCAICF and WBIC are the only criteria to choose the correct model 100% of the times. However, the improvement is not remarkable. This is because in this example using weighted estimates is not warranted. In fact, the least squares estimate obtained by setting $w_i = 1$ for all i has smaller MSE (for normal errors, considering (9) is equivalent to considering the MSE) than the weighted version. However, the idea of the simulation is not to show the usefulness of weighted estimates, but rather to test the criteria developed in the previous section through a simulation found in the model selection literature. We now consider 3 simulations where weighted estimates actually present an advantage.

5.2 Local Regression Simulation

Consider the case where we are given observations from the following model

$$Y_i = s(x_i) + \epsilon_i, i = 1, \dots, n$$

with ϵ_i independent normal with mean 0 and variance σ^2, $x_i \in N_0, i = 1, \dots, n$ known covariates, and $s(x)$ a "smooth" function shown in Figure 1.

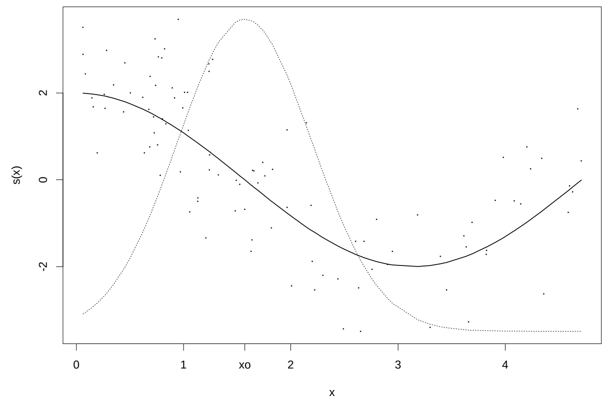

Figure 1. Mean function $s(x)$ shown with dark line, weight function shown with dotted line, and 101 simulated data points. In this example we use $s(x) = 2\cos(x)$ and $x_0 = \pi/2$.

In practice we don't know $s(x)$ and we may use local regression to obtain an estimate. For this simulation we consider $\theta_0 \equiv s(x_0)$ to be the parameter of interest with x_0 shown in Figure 1. Superimposed, in Figure 1, is the weight function to be used. In practice, for example when using the function `loess` in S-Plus, we are given the choice of locally fitting a constant, a line, or a parabola. We add the option of a cubic function for illustrative purposes. Using (9) we find that fitting a line produces the best estimate. Notice that at a visual level, it seems that a cubic polynomial would fit better than a straight line. However, since the points with $x \in (0.5, 3)$ receive "most of the weight" it turns out that fitting a line is in fact the better alternative. Order 2 is thus considered the correct choice in Table 2, which presents the same information as Table 1 but for the simulation described in this section. In this example we notice that the weighted criteria clearly outperform their equally-weighted counterparts. The WBIC and WCAICF perform best in all experiments except for $n = 50$ and $\sigma^2 = 5$, in which case WAIC performs best.

Table 2.Percentage of times each approximate model is chosen by each criteria in 1000 replications for varying sample size n and variance σ^2.

		Estimated Dimension						Estimated Dimension			
Experiment	Criterion	1	**2**	3	4	Experiment	Criterion	1	**2**	3	4
	AIC	11.6	16.4	29.8	42.2		AIC	27.2	28.8	29.1	14.9
	AIC$_c$	11.7	16.6	29.6	42.1		AIC$_c$	27.7	28.9	28.8	14.6
	WAIC	0	62.4	11.3	26.3		WAIC	3.5	70.9	13.4	12.2
$n = 50$	BIC	15.1	18.4	27.7	38.8	$n = 50$	BIC	42.8	27.7	21.7	7.8
$\sigma^2 = 1$	SIC$_f$	34.1	26.6	22	17.3	$\sigma^2 = 5$	SIC$_f$	74.1	22.4	3.5	0
	WBIC	0	99.9	0.1	0		WBIC	38.6	61.4	0	0
	CAICF	42.8	24.5	19.9	12.8		CAICF	79.7	18.8	1.5	0
	CAICF$_c$	43.1	24.6	19.7	12.6		CAICF$_c$	79.8	18.7	1.5	0
	WCAICF	0.1	99.8	0.1	0		WCAICF	45.6	54.4	0	0
	AIC	7.7	4.3	22.9	65.1		AIC	10.3	25.8	35.7	28.2
	AIC$_c$	7.8	4.4	22.9	64.9		AIC$_c$	10.3	25.9	35.8	28
	WAIC	0	50.1	7.4	42.5		WAIC	0.2	70.3	13.5	16
$n = 100$	BIC	11.9	5.5	21.9	60.7	$n = 100$	BIC	15.7	29.7	33.1	21.5
$\sigma^2 = 1$	SIC$_f$	19.6	8.1	22.4	49.9	$\sigma^2 = 5$	SIC$_f$	40	39.8	18.8	1.4
	WBIC	0	99.9	0.1	0		WBIC	6.8	93.1	0.1	0
	CAICF	21.8	8.7	21.7	47.8		CAICF	47.6	38.5	13.2	0.7
	CAICF$_c$	22	8.8	21.7	47.5		CAICF$_c$	47.7	38.5	13.1	0.7
	WCAICF	0	100	0	0		WCAICF	8.2	91.8	0	0
	AIC	13.3	1.8	9	75.9		AIC	15.8	14	28.9	41.3
	AIC$_c$	13.4	1.8	9	75.8		AIC$_c$	15.8	14	28.9	41.3
	WAIC	0	36.9	3	60.1		WAIC	0	66.4	9.7	23.9
$n = 200$	BIC	15.9	2	8.5	73.6	$n = 200$	BIC	19.6	15.1	27.6	37.7
$\sigma^2 = 1$	SIC$_f$	21.7	2.1	7.4	68.8	$\sigma^2 = 5$	SIC$_f$	31.5	22	24.6	21.9
	WBIC	0	100	0	0		WBIC	0.4	99.6	0	0
	CAICF	23.3	2.1	6.9	67.7		CAICF	36.4	22.9	22.5	18.2
	CAICF$_c$	23.3	2.1	6.9	67.7		CAICF$_c$	36.4	22.9	22.5	18.2
	WCAICF	0	100	0	0		WCAICF	0.6	99.4	0	0

5.3 Local Logistic Regression Simulation

We now perform a simulation with $Y_i, i = 1, \dots, n$ independent Bernoulli random variables with

$$\theta_i = \Pr(Y_i = 1|x_i), \text{ with } \log\left(\frac{\theta_i}{1-\theta_i}\right) = s(x_i).$$

As in the previous simulation, we consider θ_0 to be the parameter of interest with $s(x)$ and x_0 those shown in Figure 1. We use local logistic regression with the choices of locally fitting a constant, a line, a parabola, or a cubic function. According to (9), fitting a line is the correct choice, as in the previous simulation. In Table 3 we present the relevant results. We don't include the AIC$_c$ and CAICF$_c$ because they are developed for normal random variables. In this example we also see the weighted criteria outperforming their equally weighted counterparts with the WBIC and WCAICF performing the best.

Table 3. Percentage of times each approximate model is chosen, in local likelihood estimation, by each criteria in 1000 replications for varying sample size n. To avoid sparseness we consider larger n than in the previous two simulations.

	$n = 100$					$n = 200$					$n = 500$			
	Estimated Dimension					Estimated Dimension					Estimated Dimension			
Criterion	1	**2**	3	4	Criterion	1	**2**	3	4	Criterion	1	**2**	3	4
AIC	3.9	46.2	33.5	16.4	AIC	8.1	26.5	35.4	30	AIC	3.6	18.6	27.6	50.2
WAIC	0	71.6	15.3	13.1	WAIC	0	67.7	12.4	19.9	WAIC	0	59	10	31
BIC	4.8	53.5	29.5	12.2	BIC	9.1	29.6	34.8	26.5	BIC	4.5	19.5	27.9	48.1
SIC_f	14.4	70.9	13.7	1	SIC_f	15.3	42.6	30.9	11.2	SIC_f	5.7	24.3	28.6	41.4
WBIC	1	99	0	0	WBIC	0.1	99.8	0.1	0	WBIC	0	100	0	0
CAICF	18.1	70.7	10.8	0.4	CAICF	18.6	44.9	27.7	8.8	CAICF	6.5	24.6	28.4	40.5
WCAICF	3	97	0	0	WCAICF	0.1	99.8	0.1	0	WCAICF	0	100	0	0

5.4 Signal Processing Simulation

In this section we perform a simulation motivated by one presented by Ohtaki (1985), but modified to imitate a problem arising in signal processing similar to the example studied in Section 6.2. We define the n component random variable

$$Y_i = s(t_i) + \epsilon_i, i = 1, \dots, n, \text{ with } s(t) = \sum_{k=1}^{2} \rho_{k,0} \cos\{2\pi k \lambda(t) t + \psi_{k,0}\} \tag{18}$$

with $t_i = i/44100$ and ϵ_i independent normal with mean 0 and variance 1. We consider $t \in (0, 0.040)$, i.e. n=1764, and let $\lambda(t)$ be constant and equal to 660 Hz for $t \in (0, .025]$ and gradually changing to 832 Hz for $t \in (0.025, 0.040)$. We can think of Y as a digital sample (sampled at 44100 Hz) of a 40 millisecond segment of a sound signal produced by an instrument, with timbre defined by the ρs and ψs, playing an F note for 25 milliseconds and then *bending* to $F\sharp$. Figure 2 shows the simulated signal.

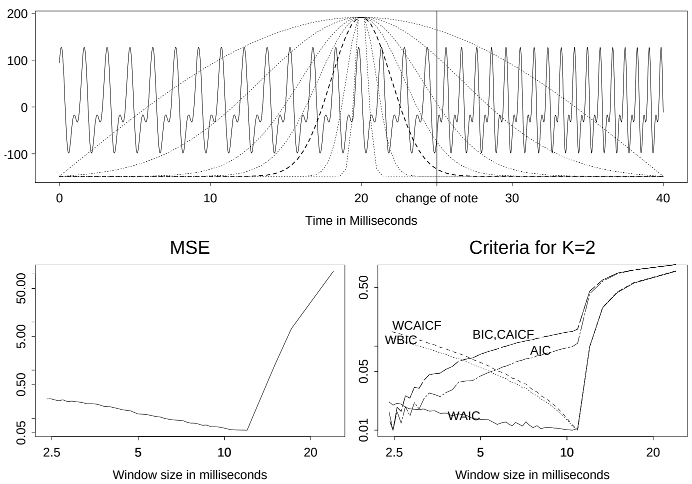

Figure 2. Simulated sound signal with w(t;h) superimposed for various values of h. The weight function that is slightly darker has the "optimal" window size. MSE of estimates obtained for $K = 2$ plotted against window size. Average values of model selection criteria against window size.

We are interested in estimating $s(t_0)$, or equivalently the $\rho_{k,0}$s, $\psi_{k,0}$s and $\lambda(t_0)$, with $t_0 = 0.02$. As described in Section 6.2, most estimation techniques used in the signal processing literature are equivalent to local likelihood estimation with (2) defined by

$$\theta_i = s(t_i) \approx \eta(t_i, \boldsymbol{\beta}) = \sum_{k=1}^{K} \rho_k \cos(k\lambda t_i + \psi_k) \text{ for } t_i \in N_0 \tag{19}$$

with $\boldsymbol{\beta} = (\rho_1, \psi_1, \ldots, \rho_K, \psi_K, \lambda)$ and N_0 an appropriately sized segment of the signal. Since λ is constant, (19) defines approximate models for (18). A model selection problem is to choose a K and a window size.

In this simulation, we consider various weight functions by defining $w(t;h) = \phi\{3 \times (t - 0.02)/h\}$ with ϕ the standard normal density. Since $\phi(s) \approx 0$ for $s > 3$ we call h the window size. In Figure 2 we superimpose weight functions obtained for various values of h. Using (9) we find that the best choice of window size is about $h = 12$ milliseconds with $K = 2$, which agrees with our intuition since the note change occurs 5 milliseconds after t_0 and $K = 2$ is the actual number of sinusoidal components in (18). In Figure 2 we see a plot of MSE against h of the estimates obtained using $K = 2$. The worst estimates occur for window sizes larger than 12. The MSE slowly increases as the window sizes become smaller than 12.

We create 5000 simulated signals and minimize WAIC, WBIC, WCAICF, and their unweighted counterparts to choose between $K = 1, 2$, or 3 and amongst various values of h. Because n is relatively large we do not include AIC$_c$ and WCAICF$_c$. Table 3 shows the percent of times each criteria chooses values of each K. We subdivide these totals into percentage of times the criteria choose $12 < h$, $6 < h \leq 12$, and $h \leq 6$.

Table 4.Percentage of times each approximate model is chosen subdivided into 3 different regions of chosen window sizes in 5000 simulations.

Dimension	$K=1$				$\mathbf{K=2}$				K=3			
Criterion	$12 < h$	$6 < h \leq 12$	$h \leq 6$	Tot	$12 < h$	$\mathbf{6 < h \leq 12}$	$h \leq 6$	Tot	$12 < h$	$6 < h \leq 12$	$h \leq 6$	Tot
AIC	0	0	0	0	0	0	30	30	0	0	70	70
WAIC	0	0	0	0	0	8	45	53	0	7	40	47
SIC$_f$	0	0	0	0	0	0	28	28	0	0	72	72
WBIC	0	0	0	0	0	79	18	97	0	2	1	3
CAICF	0	0	0	0	0	0	0	0	0	0	100	100
WCAICF	0	0	0	0	0	87	12	99	0	1	0	1

Based on the MSE, we characterize these choices as window size that are: too large, appropriate, and too small respectively. Notice that the equally weighted criteria consistently choose h that are too small. The WCAICF and WBIC choose the correct K more than 97% of the time. They also perform quite well at choosing appropriate window sizes. The WAIC outperforms the equally weighted criteria, but is inferior to the other two weighted criteria. Figure 2 also shows the average values (scaled to fit in same plot) of the different criteria obtained for $K = 2$, plotted against h. We see that all the weighted criteria have a minimum near the optimal window size of 12 and that the reason the WCAICF and WBIC outperform the WAIC is that the latter's penalty for small windows is not "strong" enough.

6. EXAMPLES

6.1 Weighted Logistic Regression

An advantage of local likelihood over local regression is that it is applicable in situations with non-continuous data. Tibshirani and Hastie (1987) and Loader (1999) present examples of local likelihood estimation applied to binary data. The model they consider assumes the binary observations are the outcomes of independent Bernoulli random variables $Y_i, i = 1, \dots, n$ with probability of success p_i. Both consider age as one of the covariates of interest and use local likelihood estimation to get an estimate of $p_i, i = 1, \dots, n$. For a given age x_0 an appropriate neighborhood N_0 is chosen and a linear function of age is fitted using (2) as follows

$$\theta_i \equiv l(p_i) \approx \eta(x_i, \boldsymbol{\beta}) = \beta_0 + x_i\beta_1, \text{ for } x_i \in N_0 \tag{20}$$

with $l(p_i) = \log\{p_i/(1-p_i)\}$. The $\boldsymbol{\beta}$ that minimizes the weighted log-likelihood

$$l_0(\boldsymbol{\beta}) = \sum_{x_i \in N_0} w_i \left[y_i(\beta_0 + x_i\beta_1) - \log\left\{1 + \exp(\beta_0 + x_i\beta_1)\right\}\right] \tag{21}$$

is used to obtain an estimate of $p_0 = l^{-1}(\theta_0)$.

In (20) instead of a linear function of age, we may fit a constant $\eta(x, \boldsymbol{\beta}) = \beta_0$ or a quadratic function $\eta(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \beta_2 x^2$. Weighted criteria may be used to help us make this decision. Furthermore, we may use the criteria to decide on a span or size of the neighborhood N_0.

In this case, the penalties needed to construct the criteria are

$$\operatorname{tr}\left\{I_n(\boldsymbol{\beta}_p)J_n(\boldsymbol{\beta}_p)^{-1}\right\} \approx \operatorname{tr}\left\{(\mathbf{X}'\mathbf{WVWX})(\mathbf{X}'\mathbf{WVX})^{-1})\right\} \text{ and } \log\det\left\{-J_n(\boldsymbol{\beta}_p)\right\} = \log\det(\mathbf{X}'\mathbf{WVX})$$

where $\mathbf{X}$ is the design matrix used in the local regression, $\mathbf{W}$ is a diagonal matrix with entries $\mathbf{W}_{i,i} = w_i$, and $\mathbf{V}$ the diagonal variance matrix with entries $\mathbf{V}_{i,i} = l^{-1}\{\eta(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_p)\}[1 - l^{-1}\{\eta(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_p)\}]$.

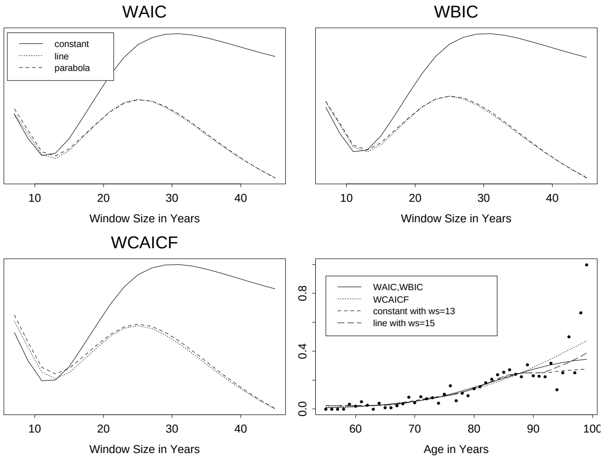

Figure 3. Values obtained for WAIC, WCAICF, and WBIC when using different window sizes and approximate models and final local likelihood estimate for a mortality data set.

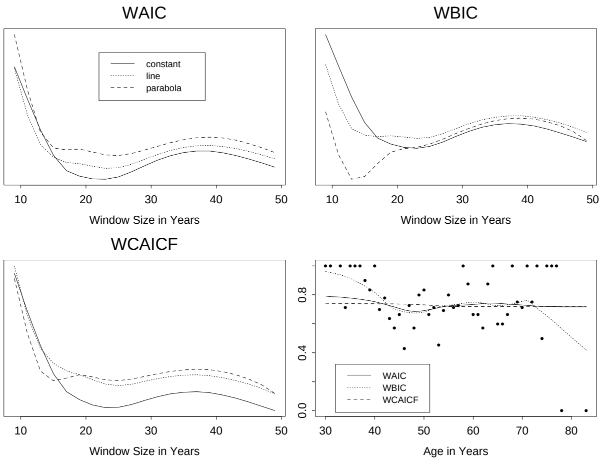

Figure 4. Values obtained for WAIC, WBIC, and WCAICF when using different window sizes and approximate models and final local likelihood estimate for a clinical trial data set.

For the mortality data presented in Loader (1999), we fix the covariate of interest at $x_0 = 77$ and find the estimate of θ_0 using various window sizes when fitting a constant, linear, and quadratic function. In Figure 3 we see the resulting WAIC, WBIC, and WCAICF. The equally weighted counterparts are not shown because they all choose the smallest window size considered, which is not practical since estimates would reproduce the data. The minimum of WAIC and WBIC is obtained by fitting a quadratic function with a window size of 45 years. WCAICF chooses a linear function with the same window size. All criteria have a local minimum near a window size of 13 years when fitting a constant function and near a window size of 15 years when fitting a linear function. It is encouraging to see the criteria automatically choosing smaller models for smaller window sizes. In Figure 3 we also see the final local likelihood estimates of $p_i, i = 1, \ldots, n$ obtained when using the models and window sizes chosen by the criteria. Except for slightly different predictions for older ages, the resulting estimates are quite similar.

A similar analysis is performed for the clinical trial data presented in Tibshirani and Hastie (1987). In Figure 4 we see the values of the weighted criteria obtained for the three competing models at the different window sizes. The minimum for the WAIC occurs at window size 23 when fitting a constant function, the minimum for the WBIC is at window size 12 when fitting a constant, and the WCAICF chooses the largest window size of 49 when fitting a constant. In Figure 4 we also see the final local likelihood estimates of $p_i, i = 1, \ldots, n$ obtained when using these choices.

6.2 Application in Signal Processing

The study of musical sound has become a popular research field within signal processing. Stochastic harmonic regression models, $y_i = s(t_i) + \epsilon(t_i)$, have been used to analyze musical sound waves. Harmonic parameters in sound analysis models are considered to be time-varying; it is thus useful to consider window based estimates when performing estimation. See Rodet (1997) for details.

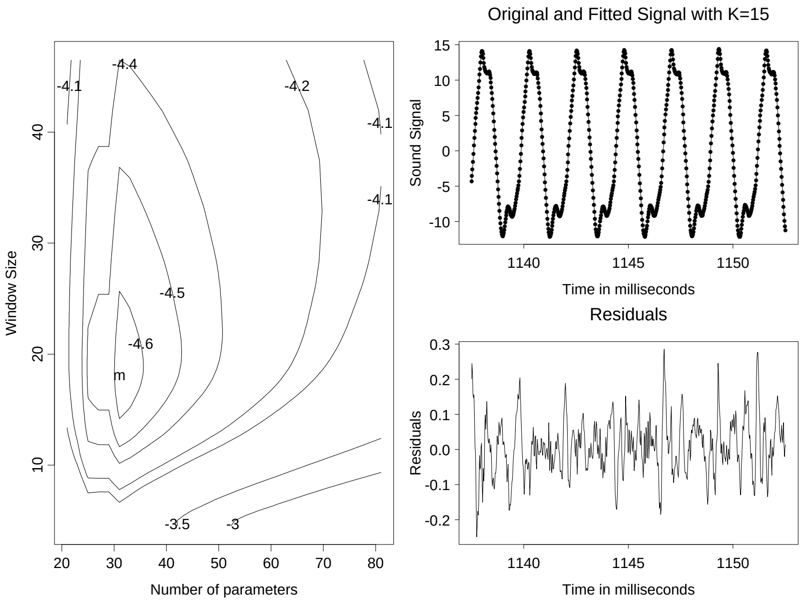

Figure 5. WAIC and WBIC plotted against pairs of number of parameters and window sizes.

The estimation procedures presented in current sound analysis research are based on the assumption that short segments of the acoustic signal, called time-frames, it may be considered to follow a deterministic harmonic signal plus stochastic noise model. Estimates for the deterministic part of the signal at a given time t_0 are found using weighted least squares estimation, which implies that we may view the procedure as a case of local likelihood estimation with (2) now defined as in (19). As mentioned in section 5.4, two model selection problems are choosing the *number of partials* K to include in the approximate model and the window size h of the time-frames considered for estimation purposes.

We apply our procedure to the sound signal of a clarinet. Previous estimation procedures usually fit approximate models with many parameters ($K > 50$) to short time-frames (around $h = 2$ milliseconds). Arbitrarily fitting too many parameters to small amounts of data may result in estimates that are hard to interpret. The weighted criteria developed in this paper provide a data driven procedure for choosing from amongst the different possible estimates obtained using different window sizes and values of K. In Figure 5 we see a contour plot of the values of the WCAICF when considering windows sizes between 9 and 50 milliseconds and fitting models with $K = 1, \dots, 48$. The pairs chosen are $K = 15$ (31 parameters) and $h = 18$ milliseconds. The estimates obtained by using the model and window size chosen by the WCAICF are also shown in Figure 5. Notice how well the estimates fit the data (the original signal has a range of about 25, the residual's range is about 0.5). The WBIC produces similar results. The WAIC chooses $K = 16$ and window size of 11 milliseconds. The equally weighted criteria all choose the smallest window size considered.

For this particular case we can show that the penalty in (12) and (16) can be well approximated with

$$2(K+1)\frac{\int w(s)^2\,ds}{\int w(s)\,ds} \text{ and } 2K\log\int w(s)\,ds$$

respectively, with $w(s)$ the weight function, making the procedure computationally fast.

7. CONCLUSIONS AND EXTENSIONS

We have presented model selection criteria to be used in local likelihood estimation. They were developed as weighted versions of well known criteria, namely AIC, BIC, and CAICF. However, we believe that most of the information and posterior probability criteria presented in the model selection literature can be extended to weighted versions, intended for use in local likelihood settings, in similar ways to the ones presented in this paper.

The theoretical results used to justify the criteria are based on an asymptotic approximation. However, simulations show that the WBIC works well in many situations dealing with finite samples. The WAIC does not seem to work as well, but a rather simple modification leads us to the WCAICF which, simulations demonstrate, works as well or better than the WBIC. This is not surprising since, in practice, the two criteria are essentially equivalent for large n.

In the simulations, the WAIC did outperform the WBIC and WCAICF in one particular instance: a small sample with large variance. The main difference between the WAIC and the other two is the $\det \log -\hat{J}_{n,p}$ term defined by (13). This can be become the dominant term of the criteria in some cases with small n. Consider the case of simple linear regression with normal errors. The $\det \log -\hat{J}_{n,p}$ term grows with the range of the covariates. With large enough variance the difference between the $l_0(\hat{\boldsymbol{\beta}})$s obtained for different models can be small enough to make $\det \log -\hat{J}_{n,p}$ the dominant term resulting in under-fitting. Given a particular application, one should examine the behavior of this term to avoid under-fitting. It is also important to consider that the appropriateness of $\det \log -\hat{J}_{n,p}$ as an estimate of $J_n(\boldsymbol{\beta}_p)$ will depend on how close the approximate models are to the generating model. These and other factors, such as the value of parameters and the size of the largest approximate model, can affect the performance of model selection criteria. See Soofi (1997) for a thorough discussion with illustrative examples. Given the broad spectrum of possible applications it is beyond the scope of this paper to give a complete set of guidelines for when each criteria should be used. We can, however, discuss some of our experiences with the application of these criteria.

In the applications presented in Section 6.1 we consider the largest window size to be one that would include all of the covariates. However, Figures 3 and 4 seem to suggest that the criteria will choose larger window sizes if these are permitted. Moreover, if the largest window is restricted, say to 30 or less, the criteria would choose smaller window sizes. In practice we can never consider all possible window sizes and all possible approximate models. Scientific knowledge of the problem should be considered in conjunction with the data-driven criteria. As suggested in part by the simulations, this is especially important when n is relatively small, which is the case in these two examples. Yet in many situations we can argue, from a practical point of view, that we are only interested in certain window sizes. For example, for the data in Tibshirani and Hastie (1987) a physiologist may know (or suspect) that the risk of disease changes at least every 25 years. This would lead us to consider window sizes smaller than 25 years and to an estimate such as the one chosen by WBIC (shown in Figure 4). This estimate purveys that the chance of survival is larger for young patients and smaller for the older, which agrees with our intuition. Since we used a model selection criteria, our choice of window size and approximate model is, in part, data-driven. Under the mentioned restriction, all the criteria choose roughly the same model. The simulation results shown in Table 3 suggest that for this size n all three criteria perform similarly.

When computing the local likelihood estimates shown in Figures 3 and 4, a separate estimate was obtained for each $x_i = 1, \dots, n$. However, we used the window size and approximate model chosen for $x_{n/2}$. A more appropriate procedure would be to choose a window size and approximate model for each i. This of course is computationally more time consuming and, in this particular case, we don't expect the results to change much. However, in the signal processing example a procedure such as this is quite useful.

The sound signal studied in Section 6.2 was a segment of a more complicated 3 second duration note. In most situations these types of signals have parts that are "stable" and others that are less so. This suggests that different window sizes should be used in different parts of the signal. Considering a procedure that performs model selection for each t_i seems appropriate. The final result is an estimate of the deterministic harmonic signal $s(t)$ and of its time-varying harmonic parameters. We can perform residual analysis by listening to the fitted and residual signals. The WCAICF and WBIC estimates are indistinguishable to the ear and the sound of the residuals are as we would expect. For example, the clarinet residuals sound like air and spit going through a mouth-piece. The results obtained by this procedure greatly improve those obtained by the fixed predetermined window size procedures suggested by the sound signal literature. The WAIC estimates, however, sound too "noisy", thus suggesting that the chosen window sizes are too small. Equally weighted criteria are not useful because they tend to choose the smallest window considered; in fact it was for this example that the weighted criteria presented in this paper were first developed. The results of this *residual analysis by ear* are in agreement with the simulation results. Examples of the results and sounds mentioned here can be found in the Demo Section of the author's web page `http://biosun01.biostat.jhsph.edu/~ririzarr`. The S-Plus code used for the simulations and examples can be found in the Software Section.

APPENDIX

Discussion of Theorems 1 and 2.

One of the assumptions stated in Theorem 1 is that we have an "appropriate sequence of neighborhoods $N_{0,n}$". Notice that we have not defined what is meant by "appropriate sequence of neighborhood". A theoretically rigorous description requires assumptions on the behavior of the function $\theta(\mathbf{x})$ and the density families $g_{\mathbf{Y}}(\mathbf{y})$ within arbitrarily small neighborhoods of $\mathbf{x}$. Since the purpose of this paper is to present and motivate new model selection criteria and to show how they can be used in practice, these theoretical details are left as future work. We will only discuss what is meant by this in a heuristic fashion. The asymptotic theory presented in, for example, Staniswalis (1989) and Loader (1996), is developed under the assumption that as the size (or radius) of the neighborhood of the covariate of interest $\mathbf{x}_0$ tends to 0, the difference between the true and the approximating distributions within such neighborhood becomes negligible. Furthermore, we assume that despite the fact that the neighborhoods become arbitrarily small, the number of data points in the neighborhood somehow tends to ∞. The idea is that, asymptotically, the behavior of the data within a given neighborhood is like the one assumed in classical asymptotic theory for non-IID data: the small window size assures that the difference between the true and approximating models is negligible and the large number of independent observations is available to estimate a parameter of fixed dimension that completely specifies the joint distribution. This concept motivates us to prove Theorems 1 and 2 for the special case where for all n the neighborhood $N_{0,n}$ actually contains all the covariates $\mathbf{x}_1, \dots, \mathbf{x}_n$. Notice that these conditions, together with Condition 5, suggest that for these results to be useful in practice, we need to have the number of data points given significant weight to be large. The results don't fit well in situations where n is large but the window size is small enough so that the data points receiving considerable weight are considerably smaller than n. In the case of the signal processing example these approximations seem more than adequate. For the other two examples they seem adequate enough.

We assume that that for any of the models defined by (6) the following 4 conditions hold:

Condition 1. For any $\beta \in \Omega_p$ both the gradient vector $\frac{\partial}{\partial\beta} l_0(\beta)$ and the Hessian matrix $\frac{\partial}{\partial\beta\partial\beta'} l_0(\beta)$ are well defined with probability 1.

Condition 2. For any $\beta \in \Omega_p$ we have $\mathrm{E}_{\mathbf{Y}} \left| \frac{\partial}{\partial\beta} l_0(\beta) \right| < \infty$ and $\mathrm{E}_{\mathbf{Y}} \left| \frac{\partial}{\partial\beta\partial\beta'} l_0(\beta) \right| < \infty$. Here the comparisons are taken component-wise.

Condition 3. For the neighborhood $N_{0,n}$ under consideration, there exists a unique $\beta_p \in \Omega_p$ such that

$$\mathrm{E}_{\mathbf{Y}} \left\{ \frac{\partial}{\partial\beta} l_0(\beta_p) \right\} = 0. \tag{22}$$

For any $\epsilon > 0$,

$$\sup_{||\beta - \beta_p|| > \epsilon} \quad l_0(\beta) - l_0(\beta_p)$$

diverges to $-\infty$ for $\beta_p \in \Omega_p$.

Condition 4. For any $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\sup_{||\beta-\beta_p||<\delta} \left| \mathrm{E}_{\mathbf{Y}}(\hat{\beta}_p - \beta_p)' J_n(\beta_p)(\hat{\beta}_p - \beta_p) - \mathrm{tr}\left\{ I_n(\beta_p) J_n(\beta_p)^{-1} \right\} \right|$$

with $I_n(\beta_p)$ and $J_n(\beta_p)$ defined as in (11) and $\hat{\beta}_p$ the local likelihood estimate obtained under model M_p.

We also assume that the weight function satisfies the following condition.

Condition 5. The function $w(s)$ is non-negative, bounded, of bounded variation, and has support $[0,1]$ with $\int_0^1 w(s)\,ds > 0$.

Weight coefficients can then be defined via

$$w_i = w\{|\mathbf{x}_i - \mathbf{x}_0|/2h + 1/2\} \tag{23}$$

with $|\mathbf{x}_i - \mathbf{x}_0|$ some distance and $h > 0$ a span of the same order as the size of $N_{0,n}$.

Under these conditions the proof of Theorem 1 follows in the same way as in Shibata (1989).

Proof. Expand $l_0(\hat{\beta}_p)$ around β_p to obtain

$$l_0(\hat{\beta}_p) = l_0(\beta_p) + (\hat{\beta}_p - \beta_p)' \frac{\partial}{\partial \beta} l_0(\beta_p) + \frac{1}{2}(\hat{\beta}_p - \beta_p)' \frac{\partial^2}{\partial \beta \partial \beta'} l_0(\tilde{\beta}_p)(\hat{\beta}_p - \beta_p)$$

with $\tilde{\beta}_p$ between $\hat{\beta}_p$ and β_p.

Condition 3 implies we can expand $\sum_{i=1}^n w_i \int g(y_i|\theta_i) \log f(y_i|\mathbf{x}_i, \hat{\beta})\,dy_i$ as

$$\sum_{i=1}^n w_i \int g(y_i|\theta_i) \log f(y_i|\mathbf{x}_i, \beta_p)\,dy_i + \frac{1}{2}(\hat{\beta}_p - \beta_p)' \left\{ \sum_{i=1}^n w_i \int g(y_i|\theta_i) \frac{\partial^2}{\partial \beta \partial \beta'} \log f(y_i|\mathbf{x}_i, \tilde{\beta}_p)\,dy_i \right\} (\hat{\beta}_p - \beta_p)$$

with $\tilde{\beta}_p$ between $\hat{\beta}_p$ and β_p.

Now using Condition 4 the expectation

$$\mathrm{E}_{\mathbf{Y}} \left\{ \sum_{i=1}^n w_i \int g(y_i|\theta_i) \log f(y_i|\mathbf{x}_i, \hat{\beta}_p)\,dy_i \right\} = \mathrm{E}_{\mathbf{Y}}\left\{ l_0(\beta_p) \right\} - \frac{1}{2}\mathrm{tr}\left\{ I_n(\beta_p) J_n(\beta_p)^{-1} \right\} + o(1).$$

We expand $l_0(\beta_p)$ around $\hat{\beta}_p$. Since $\partial l_0(\hat{\beta})/\partial \beta = 0$ we have

$$l_0(\beta_p) = l_0(\hat{\beta}_p) + \frac{1}{2}(\beta_p - \hat{\beta}_p)' \frac{\partial^2}{\partial \beta \partial \beta'} l_0(\tilde{\beta}_p)(\beta_p - \hat{\beta}_p)$$

and we can obtain the result from the theorem.

The following corollary is obtained by extending Theorem 1 in the same way Bozdogan (1987) extends the result of Akaike (1973).

Corollary 1. Under the same conditions of Theorem 1 we have that

$$\mathrm{E}_{\mathbf{Y}}\left\{ l_0(\hat\beta_p) - \sum_{\mathbf{x}_i \in N_{0,n}} w_i \int g(y_i|\theta_i) \log f(y_i|\mathbf{x}_i, \hat\beta_p)\, dy_i \right\} = \mathrm{tr}\left\{ I_n(\beta_p) J_n(\beta_p)^{-1} \right\} + \log\det\left\{ -J_n(\beta_p) \right\} + o(1).$$

The proof of Theorem 2 follows in the same way as Chow (1981) or Neath and Cavanaugh (1997).

Proof. Following the proof of a well known theorem of Jeffreys (1961), we can show that the conditional density of β given the data and that model M_p is true can be approximated with

$$(2\pi)^{\frac{p}{2}} \det\left\{ -\frac{\partial^2}{\partial\beta\partial\beta'} \log f(y_i|\mathbf{x}_i, \hat\beta_p) \right\}^{\frac{1}{2}} \exp\left\{ \frac{1}{2}(\beta - \hat\beta_p)' \frac{\partial^2}{\partial\beta\partial\beta'} l_0(\beta_p)(\beta - \hat\beta_p) \right\} \{1 + O(n^{1/2}).\}$$

Evaluating this equation and taking natural logarithms we arrive at the result of the theorem. Notice that $p/2\log(2\pi)$ is of order $O(1)$ and, for simplicity, we leave this term out in the result of Theorem 2.

REFERENCES

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," in Petrov, B. and Csaki, B., editors, *Second International Symposium on Information Theory*, pp. 267–281, Budapest: Academiai Kiado.

Bozdogan, H. (1987), "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions," *Psychometrica*, 52, 345–370.

——— (1994), "Mixture-model cluster analysis using a new informational complexity and model selection criteria," in Bozdogan, H., editor, *Multivariate Statistical Modeling*, volume 2, pp. 69–113, The Netherlands: Dordrecht.

Chow, G. C. (1981), "A comparison of the information and posterior probability criteria for model selection," *Journal of Econometrics*, 16, 21–33.

Cleveland, W. S. and Devlin, S. J. (1988), "Locally weighted regression: An approach to regression analysis by local fitting," *Journal of the American Statistical Association*, 83, 596–610.

Fan, J. (1993), "Local linear regression smoothers and their minimax efficiencies," *Annals of Statistics*, 21, 196–216.

Gokhale, D. V. and Kullback, S. (1978), *The Information in Contingency Tables.* Marcel Dekker.

Hastie, T. and Tibshirani, R. (1987), "Generalized additive models: Some applications," *Journal of the American Statistical Association*, 82, 371–386.

Hurvich, C. M. and Tsai, C.-L. (1989), "Regression and time series model selection in small samples," *Biometrika*, 76, 297–307.

Kashyap, R. L. (1982), "Optimal choice of AR and MA parts in autoregressive moving average models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-4, 99–104.

Kitagawa, G. (1987), Reply to comments on "Non-gaussian state-space modeling of nonstationary time series," *Journal of the American Statistical Association*, 82, 1060–1063.

Kullback, S. (1959), *Information Theory and Statistics*, New York: John Wiley & Sons.

Ljung, L. and Caines, P. E. (1979), "Asymptotic normality of prediction error estimators for approximate system models," *Stochastics*, 3, 29–46.

Loader, C. R. (1996), "Local likelihood density estimation," *The Annals of Statistics*, 24, 1602–1618.

——— (1999), *Local Regression and Likelihood*, New York: Springer.

Neath, A. A. and Cavanaugh, J. E. (1997), "Regression and time series model selection using variants of the Schwarz information criteria," *Communications in Statistics – Theory and Methods*, 26, 559–580.

Ohtaki, M. (1985), "On the application of linear models to local regions in regression," in *Statistical Theory and Data Analysis: Proceedings of the Pacific Area Statistical Conference*, pp. 529–545.

Rodet, X. (1997), "Musical sound signals analysis/synthesis: Sinusoidal+residual and elementary waveform models," in *Proceedings of the IEEE Time-Frequency and Time-Scale Workshop*.

Sawa, T. (1978), "Information criteria for discriminating among alternative regression models," *Econometrica*, 46, 1273–1291.

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

Shibata, R. (1989), "Statistical aspects of model selection," in Williems, J. C., editor, *From Data to Model*, pp. 215–240, New York: Springer-Verlag.

Soofi, E. S. (1997), "Information theoretic regression methods," in Fomby, T. B. and Hill, R. C., editors, *Advances in Econometrics: Applying Maximum Entropy to Econometric Problems*, volume 12, pp. 52–83, JAI Press Inc.

Staniswalis, J. G. (1989), "The kernel estimate of a regression function in likelihood-based models," *Journal of the American Statistical Association*, 84, 276–283.

Stone, C. J. (1977), "Consistent nonparametric regression," *The Annals of Statistics*, 5, 595–620.

Takeuchi, K. (1976), "Distribution of information statistics and a criterion of model fitting," *Suri-Kagaku*, 153, 12–18. (In Japanese).

Tibshirani, R. and Hastie, T. (1987), "Local likelihood estimation," *Journal of the American Statistical Association*, 82, 559–567.