Ann. Inst. Statist. Math.
Vol. 57, No. 4, 617–635 (2005)

NONLINEAR REGRESSION MODELING USING REGULARIZED LOCAL LIKELIHOOD METHOD

YOSHISUKE NONAKA* AND SADANORI KONISHI

Graduate School of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-Ku, Fukuoka 812-8581, Japan, e-mail: konishi@math.kyushu-u.ac.jp

(Received December 16, 2003; revised August 6, 2004)

Abstract. We introduce a nonlinear regression modeling strategy, using a regularized local likelihood method. The local likelihood method is effective for analyzing data with complex structure. It might be, however, pointed out that the stability of the local likelihood estimator is not necessarily guaranteed in the case that the structure of system is quite complex. In order to overcome this difficulty, we propose a regularized local likelihood method with a polynomial function which unites local likelihood and regularization. A crucial issue in constructing nonlinear regression models is the choice of a smoothing parameter, the degree of polynomial and a regularization parameter. In order to evaluate models estimated by the regularized local likelihood method, we derive a model selection criterion from an information-theoretic point of view. Real data analysis and Monte Carlo experiments are conducted to examine the performance of our modeling strategy.

Key words and phrases: Information criteria, local maximum likelihood estimates, model selection, generalized linear models, regularization.

1. Introduction

Local likelihood estimation has received considerable attention as a useful technique for analyzing data with complex structure (Tibshirani and Hastie (1987), Hjort and Jones (1996), Copas and Eguchi (1998), Eguchi and Copas (1998), Loader (1999), Eguchi and Kim (2001), Eguchi *et al.* (2003), and so on).

A local likelihood function is constructed based on first considering a parametric model for the unknown true model. It is defined as a locally weighted log-likelihood with weights determined by a kernel function and a bandwidth. When a large bandwidth is chosen, the estimator is close to the maximum likelihood estimator and tends to have a large bias. On the other hand, when a small bandwidth is chosen, the estimator depends much on the data points and tends to have a large variance. In the local modeling, bandwidth plays an important role for controling the trade-off between bias and variance of the estimator (Wand and Jones (1995), Simonoff (1996)).

Issues still remain in constructing nonlinear regression models based on the local likelihood from a finite and noisy set of data. First, the stability of local likelihood estimators is not guaranteed in the case that the structure of the system is quite complex.

*Now at Biostatistics Center, Kurume University, 67 Asahi-Machi, Kurume, Fukuoka 830-0011, Japan, e-mail: nonaka_yoshisuke@kurume-u.ac.jp.

In order to overcome this issue, we introduce a regularized local likelihood function with regularization parameter that controls the local likelihood function and the complexity of a nonlinear regression model. Second, the local likelihood procedure requires the choice of a bandwidth, the degree of the polynomial and also a regularization parameter. In order to choose these adjusted parameters, we derive model selection and evaluation criteria. These criteria are derived under model misspecification both for distributional and structural assumptions, which is usually the case in practice. Our modeling strategy can be easily applied to analyze multi-dimensional continuous data, and clear improvements are obtained for the use of the regularization parameter in the regularized local likelihood functions.

This article is organized as follows. In Section 2 we describe the proposed regularized local likelihood method in the context of generalized linear models and present an information criterion for evaluating the estimated models. In Section 3 we describe the multivariate regularized local likelihood method. Section 4 includes some applications to real data sets and numerical results, in which we use the regularized local likelihood method with Gaussian and logistic models. Some concluding remarks are given in Section 5.

2. Regularized local likelihood method

Local likelihood method has both parametric and nonparametric theoretical aspects. Eguchi and Kim (2001) bridged a gap between these theories for the local likelihood method. In this paper we call regression models based on the local likelihood method "nonlinear" regression models, since a modeling strategy is used for a curve fitting and a surface fitting. We consider a nonlinear modeling strategy based on the regularized local likelihood method in the context of generalized linear models (Nelder and Wedderburn (1972), McCullagh and Nelder (1989), Green and Silverman (1994), Fan *et al.* (1995)). We first introduce the regularized local likelihood method in the case of univariate explanatory variables.

2.1 *Model*

Suppose we have n independent observations $\{(y_i, x_i); i = 1, 2, \ldots, n\}$, where y_i are random response variables and x_i are univariate explanatory variables. It is assumed that the responses y_i are generated from an unknown true distribution $G(y \mid x)$ with density $g(y \mid x)$. Generally, a regression model consists of a random component which specifies the distribution of the response Y and systematic component which presents the structure of the conditional expectation $m(x) = E[Y \mid x]$.

It is assumed that Y has a distribution in the exponential family, taking the form

$$f(y \mid x) = \exp\left\{\frac{\theta(x)y - b(\theta(x))}{\psi(x)} + c(y, \psi(x))\right\}, \tag{2.1}$$

where $b(\cdot)$ and $c(\cdot, \cdot)$ are known functions. Here, function $\theta(\cdot)$ is called the canonical parameter and function $\psi(\cdot)$ is called the dispersion parameter. We assume that the log-likelihood $\sum_{i=1}^{n} \log f(y_i \mid x_i)$ satisfies the Bartlett (1954) identities. The mean and variance of Y can be derived easily from the well known relations

$$m(x) = E[Y \mid x] = b'(\theta(x)), \qquad \mathrm{Var}[Y \mid x] = \psi(x)b''(\theta(x)), \tag{2.2}$$

where $b'(\cdot)$ and $b''(\cdot)$ are differential of first and second order respectively. In usual generalized linear models, the unknown regression function $m(x)$ is modeled linearly via the known link function $l(\cdot)$:

$$l(m(x)) = \beta_0 + \beta_1 x. \tag{2.3}$$

The function l links the regression function to a linear space of the covariates. If $l = (b')^{-1}$, then l is called the canonical link function since in the case $l(m(x))$ is the canonical parameter in the exponential family (2.1). Expressions (2.2) and (2.3) characterize the generalized linear models.

In some situations, the use of the linear relationship (2.3) has a problem. Trial and error are required in order to search for a reasonable parametric link function. So we use the technique with local polynomial. The aim of the local modeling approach is searching for a model that describes the data well.

Here, we focus on estimating $\theta(x)$ since estimating $m(x)$ is equivalent to estimating $\theta(x) = (b')^{-1}(m(x))$. Assume that the function $\theta(x)$ is at least p-times differentiable at a point x_0. Then $\theta(x)$ can be approximated locally by a polynomial of degree p for x in a neighbourhood of x_0

$$\begin{aligned} \theta(x) &\approx \theta(x_0) + \theta^{(1)}(x_0)(x - x_0) + \cdots + \theta^{(p)}(x_0)(x - x_0)^p/p! \\ &= \boldsymbol{\beta}(x_0)^T \boldsymbol{x}(x; x_0), \end{aligned} \tag{2.4}$$

where $\beta_j(x_0) = \theta^{(j)}(x_0)/j!$ $(j = 0, 1, \ldots, p)$, $\boldsymbol{\beta}(x_0) = (\beta_0(x_0), \beta_1(x_0), \ldots, \beta_p(x_0))^T$ and $\boldsymbol{x}(x; x_0) = (1, x - x_0, \ldots, (x - x_0)^p)^T$. Then the data $\{y_1, y_2, \ldots, y_n\}$ are summarized by a model from a class of probability densities

$$\begin{aligned} &f(y_i \mid x_i; \boldsymbol{\beta}(x_0), \psi(x_0)) \\ &\quad = \exp\left\{ \frac{\boldsymbol{\beta}(x_0)^T \boldsymbol{x}(x_i; x_0) y_i - b(\boldsymbol{\beta}(x_0)^T \boldsymbol{x}(x_i; x_0))}{\psi(x_0)} + c(y_i, \psi(x_0)) \right\}, \end{aligned} \tag{2.5}$$

where $\psi(x_0)$ is the dispersion parameter at x_0.

2.2 *Estimation*

The unknown parameters $\boldsymbol{\beta}(x_0)$ and $\psi(x_0)$ in (2.5) may be estimated by maximizing the local log-likelihood function. However, when fitting a nonlinear model to data with complex structure, the local likelihood method does not yield satisfactory results. Instead of maximizing this function, we propose choosing the parameters $\boldsymbol{\beta}(x_0)$ and $\psi(x_0)$ to maximize the regularized local log-likehood function:

$$\begin{aligned} &RL(\boldsymbol{\beta}(x_0), \psi(x_0); p, h, \lambda) \\ &\quad = \sum_{i=1}^{n} w_h(x_i; x_0) \log f(y_i \mid x_i; \boldsymbol{\beta}(x_0), \psi(x_0)) - n\lambda \boldsymbol{\beta}(x_0)^T K \boldsymbol{\beta}(x_0)/2 \\ &\quad = \sum_{i=1}^{n} w_h(x_i; x_0) \left\{ \frac{\boldsymbol{\beta}(x_0)^T \boldsymbol{x}(x_i; x_0) y_i - b(\boldsymbol{\beta}(x_0)^T \boldsymbol{x}(x_i; x_0))}{\psi(x_0)} + c(y_i, \psi(x_0)) \right\} \\ &\qquad - n\lambda \boldsymbol{\beta}(x_0)^T K \boldsymbol{\beta}(x_0)/2. \end{aligned} \tag{2.6}$$

The first term in the right side of the equation (2.6) is a usual local log-likelihood function, where $w_h(x; x_0)$ is a weight function and we use the Gaussian kernel:

$$w_h(x; x_0) = (2\pi h^2)^{-1/2} \exp\left\{ -\frac{(x - x_0)^2}{2h^2} \right\}, \qquad x \in (-\infty, \infty). \tag{2.7}$$

h is a bandwidth which controls the smoothness of estimated curve. The second term in the right side of the equation (2.6) is a regularization term. Typical forms for the $(p+1) \times (p+1)$ matrix K are given in the following forms

$$\text{P1} : K = \begin{pmatrix} 0 & \mathbf{0}_p^T \\ \mathbf{0}_p & D_k^T D_k \end{pmatrix}, \qquad \text{P2} : K = \begin{pmatrix} 0 & \mathbf{0}_p^T \\ \mathbf{0}_p & I_p \end{pmatrix}, \tag{2.8}$$

where I_p is a p-dimensional identity matrix, $\mathbf{0}_p$ is a p-dimensional zero vector and D_k is a $(p-k) \times p$ matrix that represents the difference operator given by

$$D_k = \begin{pmatrix} (-1)^0{}_kC_0 & \cdots & (-1)^k{}_kC_k & 0 & \cdots & \cdots & 0 \\ 0 & (-1)^0{}_kC_0 & \cdots & (-1)^k{}_kC_k & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & (-1)^0{}_kC_0 & \cdots & (-1)^k{}_kC_k \end{pmatrix}$$

with ${}_nC_k = n!/\{k!(n-k)!\}$. The λ is a regularization parameter which controls the local log-likelihood and the complexity of the estimated model.

The estimator $\hat{\boldsymbol{\beta}}(x_0)$ is obtained as a solution of $\partial RL(\boldsymbol{\beta}(x_0), \psi(x_0); p, h, \lambda)/\partial \boldsymbol{\beta} = \mathbf{0}$. This equation is generally nonlinear in $\boldsymbol{\beta}(x_0)$, so we optimize $\boldsymbol{\beta}(x_0)$ by the iterative algorithm. After the estimator $\hat{\boldsymbol{\beta}}(x_0)$ is obtained, the estimator $\hat{\psi}(x_0)$ is given as a solution of $\partial RL(\hat{\boldsymbol{\beta}}(x_0), \psi(x_0); p, h, \lambda)/\partial \psi = 0$. Replacing the unknown parameter $\boldsymbol{\beta}(x_0)$ and $\psi(x_0)$ by $\hat{\boldsymbol{\beta}}(x_0)$ and $\hat{\psi}(x_0)$ respectively and noting that parameter $\theta(x_i)$ in the exponential family (2.1) is directly given by $\theta(x_i) = \boldsymbol{\beta}(x_i)^T \boldsymbol{x}(x_i; x_i) = \beta_0(x_i)$, we obtain the estimated model as follows

$$f(y_i \mid x_i, \hat{\boldsymbol{\beta}}(x_i), \hat{\psi}(x_i)) = \exp\left\{ \frac{\hat{\beta}_0(x_i) y_i - b(\hat{\beta}_0(x_i))}{\hat{\psi}(x_i)} + c(y_i, \hat{\psi}(x_i)) \right\}. \tag{2.9}$$

2.2.1 *Gaussian model*

We consider the following model:

$$y_i = m(x_i) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2(x_i)) \qquad (i = 1, \ldots, n), \tag{2.10}$$

where $m(\cdot)$ is an unknown smooth function. Then taking $b(\theta(x_i)) = \theta(x_i)^2/2$, $\psi(x_0) = \sigma^2(x_0)$,

$$c(y_i, \psi(x_0)) = -\frac{y_i^2}{2\psi(x_0)} - \frac{1}{2}\log\{2\pi\psi(x_0)\} = -\frac{1}{2}\left\{\frac{y_i}{\sigma(x_0)}\right\}^2 - \frac{1}{2}\log\{2\pi\sigma^2(x_0)\}$$

and $l(m(x_i)) = m(x_i) = b'(\theta(x_i)) = \theta(x_i) = \boldsymbol{\beta}(x_0)^T \boldsymbol{x}(x_i; x_0)$ in the exponential family of densities (2.5), we have a nonlinear regression model with Gaussian noise which can be expressed as

$$f_N(y_i \mid x_i; \boldsymbol{\beta}(x_0), \sigma^2(x_0)) = \{2\pi\sigma^2(x_0)\}^{-1/2} \exp\left[-\frac{\{y_i - \boldsymbol{\beta}(x_0)^T \boldsymbol{x}(x_i; x_0)\}^2}{2\sigma^2(x_0)} \right]. \tag{2.11}$$

A $(p+1)$-dimensional parameter $\boldsymbol{\beta}(x_0)$ and an error variance $\sigma^2(x_0)$ in equation (2.11) are estimated by the maximization of the regularized local log-likelihood function (2.6).

Then the estimators of $\boldsymbol{\beta}(x_0)$ and $\sigma^2(x_0)$ are explicitly given by

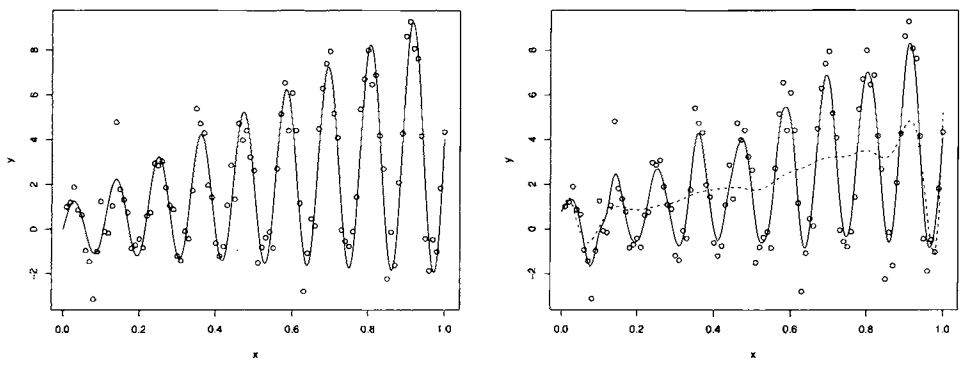
$$\hat{\boldsymbol{\beta}}(x_0) = (X^T W X + n\zeta K)^{-1} X^T W \boldsymbol{y}, \tag{2.12}$$
$$\hat{\sigma}^2(x_0) = \{\boldsymbol{y} - X\hat{\boldsymbol{\beta}}(x_0)\}^T W \{\boldsymbol{y} - X\hat{\boldsymbol{\beta}}(x_0)\} / \operatorname{tr}(W),$$

where $X = (\boldsymbol{x}(x_1; x_0), \ldots, \boldsymbol{x}(x_n; x_0))^T$, $W = \operatorname{diag}\{w_h(x_i; x_0)\}$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ and $\zeta = \lambda\sigma^2$ is another expression of the regularization parameter. Replacing the unknown parameters $\boldsymbol{\beta}(x_0)$ and $\sigma^2(x_0)$ in (2.11) by their sample estimators $\hat{\boldsymbol{\beta}}(x_i)$ and $\hat{\sigma}^2(x_i)$, we obtain the statistical model

$$f_N(y_i \mid x_i; \hat{\boldsymbol{\beta}}(x_i), \hat{\sigma}^2(x_i)) = \{2\pi\hat{\sigma}^2(x_i)\}^{-1/2} \exp\left[-\frac{\{y_i - \hat{\beta}_0(x_i)\}^2}{2\hat{\sigma}^2(x_i)}\right]. \tag{2.13}$$

The local polynomial ridge regression proposed by Seifert and Gasser (1996, 2000) may be considerd as a special case of our regularized local likelihood method. Moreover when $\lambda = 0$, the estimator $\hat{\boldsymbol{\beta}}(x_0)$ in (2.12) is equivalent to the local polynomial estimator (Stone (1977), Cleveland (1979), Fan and Gijbel (1996), Loader (1999)).

Example 2.1. We illustrate the proposed regularized local likelihood modeling by fitting curve to the simulation data. The random samples $\{(y_i, x_i); i = 1, 2, \ldots, 100\}$ were generated from the true regression model $y_i = \sin(18\pi x_i) + 5x_i \cos(18\pi x_i - \pi/2) + 4x_i + \varepsilon_i$, $\varepsilon_i \sim N(0, 0.7^2)$, where the design points x_i are uniformly distributed in $[0, 1]$. Figure 1(a) shows the true curve and the scatterplot of the data. We apply the local likelihood method and the regularized local likelihood method to the simulation data, where we fix the degree of polynomial $p = 5$ since we focus on the efficiency of regularization parameter. A broken line in Fig. 1(b) gives the smoothed curve for the bandwidth $h = 0.07$ without the help of regularization ($\lambda = \zeta = 0$). This curve is obviously oversmoothed, but it is impossible to apply the local likelihood method for a smaller h since the matrix $(X^T W X)^{-1}$ in (2.12) is not computable in practice. A solid line in Fig. 1(b) gives the smoothed curve for the bandwidth $h = 0.015$ with the help of



(a) (b)

Fig. 1. Comparison of the true curve and the smoothed curves with $p = 5$. (a) shows the true curve $y = \sin(18\pi x) + 5x\cos(18\pi x - \pi/2) + 4x$ and the scatterplot of the data. (b) shows the smoothed curve for the local likelihood methods with $h = 0.07$ (broken line) and the smoothed curve for the regularized local likelihood methods with $h = 0.015$ and $\zeta = 10^{-8}$ (solid line).

regularization, where we use the regularization parameter $\zeta = 10^{-8}$. This curve gives a good representation of the underlying function over the region $[0, 1]$. We observe that by appropriate choice of h and ζ, our nonlinear regression modeling strategy can capture the true structure generating the data for large p.

2.2.2 *Logistic model*

Let $y_1, y_2, \ldots, y_n$ be independent sequences of binary random variables taking values of 0 or 1 with conditional probabilities $\Pr(y_i = 1 \mid x_i) = \pi(x_i)$ and $\Pr(y_i = 0 \mid x_i) = 1 - \pi(x_i)$, where x_i are univariate explanatory variables. Taking $\theta(x_i) = \log[\pi(x_i)/\{1 - \pi(x_i)\}]$, $b(\theta(x_i)) = \log[1 + \exp\{\theta(x_i)\}]$, $\psi(x_0) = 1$, $c(y_i, \psi(x_0)) = 0$ and $l(m(x_i)) = \log[m(x_i)/\{1 - m(x_i)\}] = \theta(x_i) = \boldsymbol{\beta}(x_0)^T \boldsymbol{x}(x_i; x_0)$ in (2.5), we have a nonlinear logistic regression model as follows

$$f_L(y_i \mid x_i; \boldsymbol{\beta}(x_0)) = \pi(x_i; x_0)^{y_i} \{1 - \pi(x_i; x_0)\}^{1-y_i}, \tag{2.14}$$

where

$$\pi(x_i; x_0) = \frac{\exp\{\boldsymbol{\beta}(x_0)^T \boldsymbol{x}(x_i; x_0)\}}{1 + \exp\{\boldsymbol{\beta}(x_0)^T \boldsymbol{x}(x_i; x_0)\}}. \tag{2.15}$$

The unknown parameter vector $\boldsymbol{\beta}(x_0)$ is estimated by maximizing the regularized local log-likelihood (2.6), where the matrix K is given by P2 in equation (2.8). Then we can obtain the estimated model as follows

$$f_L(y_i \mid x_i, \hat{\boldsymbol{\beta}}(x_i)) = \hat{\pi}(x_i; x_i)^{y_i} \{1 - \hat{\pi}(x_i; x_i)\}^{1-y_i} \qquad (i = 1, \ldots, n), \tag{2.16}$$

where $\hat{\pi}(x_i; x_i) = \exp\{\hat{\beta}_0(x_i)\}/[1 + \exp\{\hat{\beta}_0(x_i)\}]$ is the estimated conditional probability.

Nonlinear regression models based on the regularized local likelihood method depend on a bandwidth h, the degree of polynomial p and a regularization parameter λ. In the next subsection we derive a model evaluation criterion for nonlinear regression models estimated by the regularized local likelihood method.

2.3 *Model selection*

Suppose that we observe a realization of a random variable with distribution as defined in Subsection 2.1. We recall that the statistical model $f(y \mid x, \hat{\boldsymbol{\beta}}(x), \hat{\psi}(x))$ defined in (2.9) is constructed within the generalized linear model framework. We want to assess the closeness of $f(y \mid x, \hat{\boldsymbol{\beta}}(x), \hat{\psi}(x))$ to the true model $g(y \mid x)$ from a predictive point of view.

Konishi and Kitagawa (1996) proposed a model selection criterion as an estimator of the Kullback-Leibler information between the true model and the estimated model. As shown in the Appendix, we can obtain an information criterion to evaluate the estimated model $f(y_i \mid x_i, \hat{\boldsymbol{\beta}}(x_i), \hat{\psi}(x_i))$ in (2.9) with $(p + 2)$-dimensional parameter estimator $\hat{\boldsymbol{\theta}}(x) = (\hat{\boldsymbol{\beta}}(x)^T, \hat{\psi}(x))^T$. The information criterion based on the regularized local likelihood method with the exponential family is given by

$$\begin{aligned} \mathrm{GIC}_{h,p,\lambda} = {} & -2 \sum_{i=1}^{n} \left\{ \frac{\hat{\beta}_0(x_i) y_i - b(\hat{\beta}_0(x_i))}{\hat{\psi}(x_i)} + c(y_i, \hat{\psi}(x_i)) \right\} \\ & + \frac{2}{n} \sum_{i=1}^{n} \mathrm{tr}\{\hat{R}(x_i)^{-1} \hat{Q}(x_i)\}, \end{aligned} \tag{2.17}$$

where we use the following notations:

$$
\begin{aligned}
&\hat{Q}(x) = \frac{1}{n\hat{\psi}(x)} \begin{bmatrix} A^{(2)}X/\hat{\psi}(x) - \lambda K\hat{\boldsymbol{\beta}}(x)\mathbf{1}_n^T\Lambda X & A^{(1)}\boldsymbol{p} - \hat{\psi}(x)\lambda K\hat{\boldsymbol{\beta}}(x)\mathbf{1}_n^T\boldsymbol{p} \\ \boldsymbol{p}^T A^{(1)T} & \hat{\psi}(x)\boldsymbol{p}^T W\boldsymbol{p} \end{bmatrix}, \\
&\hat{R}(x) = \frac{1}{n\hat{\psi}(x)} \begin{bmatrix} X^T W\Gamma X + n\hat{\psi}(x)\lambda K & A^{(1)}\mathbf{1}_n/\hat{\psi}(x) \\ \mathbf{1}_n^T A^{(1)T}/\hat{\psi}(x) & -\hat{\psi}(x)\boldsymbol{q}^T W\mathbf{1}_n \end{bmatrix}, \\
&A^{(k)} = X^T W\Lambda^k, \qquad \Lambda = \operatorname{diag}\{y_i - b'(\hat{\boldsymbol{\beta}}(x)^T\boldsymbol{x}(x_i;x))\}, \\
&\Gamma = \operatorname{diag}\{b''(\hat{\boldsymbol{\beta}}(x)^T\boldsymbol{x}(x_i;x))\}, \\
&\boldsymbol{p} = (p_1, p_2, \ldots, p_n)^T, \qquad \boldsymbol{q} = (q_1, q_2, \ldots, q_n)^T, \\
&\mathbf{1}_n = (1, 1, \ldots, 1)^T \quad (n \times 1), \\
&p_i = -\frac{\hat{\boldsymbol{\beta}}(x)^T\boldsymbol{x}(x_i;x)y_i - b(\hat{\boldsymbol{\beta}}(x)^T\boldsymbol{x}(x_i;x))}{\hat{\psi}(x)^2} + \left.\frac{\partial c(y_i,\psi(x))}{\partial\psi}\right|_{\hat{\psi}(x)}, \\
&q_i = \frac{2}{\hat{\psi}(x)^3}\{\hat{\boldsymbol{\beta}}(x)^T\boldsymbol{x}(x_i;x)y_i - b(\hat{\boldsymbol{\beta}}(x)^T\boldsymbol{x}(x_i;x))\} + \left.\frac{\partial^2 c(y_i,\psi(x))}{\partial\psi^2}\right|_{\hat{\psi}(x)}.
\end{aligned} \tag{2.18}
$$

We select a bandwidth h, the degree of polynomial p and a regularization parameter λ which minimize the information criterion (2.17).

2.3.1 *Gaussian model*

Using the equations (2.17) and (2.18), we can obtain the information criterion for the estimated Gaussian model $f_N(y_i \mid x_i; \hat{\boldsymbol{\beta}}(x_i), \hat{\sigma}^2(x_i))$ in (2.13) as follows;

$$
\begin{aligned}
\mathrm{GIC}_{h,p,\lambda} &= \sum_{i=1}^{n}\left[\log\{2\pi\hat{\sigma}^2(x_i)\} + \frac{\{y_i - \hat{\beta}_0(x_i)\}^2}{\hat{\sigma}^2(x_i)}\right] \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}\operatorname{tr}\{\hat{R}(x_i)^{-1}\hat{Q}(x_i)\}, \\
\hat{Q}(x) &= \frac{1}{n\hat{\sigma}^4(x)} \\
&\quad \times \begin{bmatrix} A_N^{(2)}X - \lambda\hat{\sigma}^2(x)K\hat{\boldsymbol{\beta}}(x)\mathbf{1}_n^T\Lambda X & \frac{1}{2\hat{\sigma}^2(x)}A_N^{(3)}\mathbf{1}_n - \frac{1}{2}A_N^{(1)}\mathbf{1}_n \\ \frac{1}{2\hat{\sigma}^2(x)}\mathbf{1}_n^T A_N^{(3)T} - \frac{1}{2}\mathbf{1}_n^T A_N^{(1)T} & \frac{1}{4\hat{\sigma}^4(x)}\mathbf{1}_n^T\Lambda_N^4 W\mathbf{1}_n - \frac{1}{4}\operatorname{tr}(W) \end{bmatrix}, \\
\hat{R}(x) &= \frac{1}{n\hat{\sigma}^2(x)} \begin{bmatrix} X^T WX + n\hat{\sigma}^2(x)\lambda K & \frac{1}{\hat{\sigma}^2(x)}A_N^{(1)}\mathbf{1}_n \\ \frac{1}{\hat{\sigma}^2(x)}\mathbf{1}_n^T A_N^{(1)T} & \frac{1}{2\hat{\sigma}^2(x)}\operatorname{tr}(W) \end{bmatrix},
\end{aligned} \tag{2.19}
$$

where $A_N^{(k)} = X^T W\Lambda_N^k$ and $\Lambda_N = \operatorname{diag}\{y_i - \hat{\boldsymbol{\beta}}(x)^T\boldsymbol{x}(x_i;x)\}$.

2.3.2 *Logistic model*

In the case of the estimated logistic model $f_L(y_i \mid x_i, \hat{\boldsymbol{\beta}}(x_i))$ in (2.16), we can obtain the information criterion as follows;

$$
\begin{aligned}
\mathrm{GIC}_{h,p,\lambda} &= -2\sum_{i=1}^{n}\log f_L(y_i \mid x_i, \hat{\boldsymbol{\beta}}(x_i)) + \frac{2}{n}\sum_{i=1}^{n}\operatorname{tr}\{\hat{R}(x_i)^{-1}\hat{Q}(x_i)\}, \\
\hat{Q}(x) &= \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{\pi}(x_i;x)\}
\end{aligned} \tag{2.20}
$$

$$\times [w_h(x_i;x)\{y_i - \hat{\pi}(x_i;x)\}\boldsymbol{x}(x_i;x) - \lambda K\hat{\boldsymbol{\beta}}(x)]\boldsymbol{x}(x_i;x)^T,$$

$$\hat{R}(x) = -\frac{1}{n}\sum_{i=1}^{n}[w_h(x_i;x)\hat{\pi}(x_i;x)\{1-\hat{\pi}(x_i;x)\}\boldsymbol{x}(x_i;x)\boldsymbol{x}(x_i;x)^T - \lambda K],$$

where $\hat{\pi}(x_i;x) = \exp\{\hat{\boldsymbol{\beta}}(x)^T\boldsymbol{x}(x_i;x)\}/[1+\exp\{\hat{\boldsymbol{\beta}}(x)^T\boldsymbol{x}(x_i;x)\}]$.

3. Multivariate regularized local likelihood method

In this section, we describe the regularized local likelihood method for the case of d explanatory variables.

3.1 *Model and estimation*

Suppose we have n independent observations $\{(y_i, \boldsymbol{x}_i); i = 1, 2, \ldots, n\}$, where y_i are ramdom response variables and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T$ are d-dimensional explanatory variable vectors. It is assumed that the conditional distribution of Y, given $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^T$, belongs to the exponential family which is given by replacing x with $\boldsymbol{x}$ in (2.1). The parameters $\theta(\cdot)$ and $\psi(\cdot)$ are related to the conditional mean and conditional variance similarly to the single covariate case.

Assume that the function $\theta(\boldsymbol{x})$ is differentiable at a point $\boldsymbol{x}_0$. Then $\theta(\boldsymbol{x})$ can be approximated locally by a polynomial of degree 1:

$$\theta(\boldsymbol{x}) \approx \theta(\boldsymbol{x}_0) + \frac{\partial\theta(\boldsymbol{x}_0)}{\partial\boldsymbol{x}^T}(\boldsymbol{x}-\boldsymbol{x}_0) = \boldsymbol{\beta}^*(\boldsymbol{x}_0)^T\boldsymbol{x}^*(\boldsymbol{x};\boldsymbol{x}_0), \tag{3.1}$$

where $\boldsymbol{\beta}^*(\boldsymbol{x}_0) = (\beta_0^*(\boldsymbol{x}_0), \beta_1^*(\boldsymbol{x}_0), \ldots, \beta_d^*(\boldsymbol{x}_0))^T$, $\beta_0^*(\boldsymbol{x}_0) = \theta(\boldsymbol{x}_0)$, $\beta_j^*(\boldsymbol{x}_0) = \partial\theta(\boldsymbol{x}_0)/\partial x_j$ $(j = 1, 2, \ldots, d)$ and $\boldsymbol{x}^*(\boldsymbol{x};\boldsymbol{x}_0) = (1, (\boldsymbol{x}-\boldsymbol{x}_0)^T)^T$. More generally, approximation (3.1) can be expanded to the higher order term (see Ruppert and Wand (1994)). But we use approximation (3.1) in order not to increase the number of unknown parameters.

Then the data $\{y_1, y_2, \ldots, y_n\}$ were summarized by a model from a class of probability densities $f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}^*(\boldsymbol{x}_0), \psi(\boldsymbol{x}_0))$ which are expressed by (2.5). We propose choosing unknown parameters $\boldsymbol{\beta}^*(\boldsymbol{x}_0)$ and $\psi(\boldsymbol{x}_0)$ to maximize the regularized local log-likehood function

$$\begin{aligned} RL(\boldsymbol{\beta}^*(\boldsymbol{x}_0), \psi(\boldsymbol{x}_0); H, \lambda) &= \sum_{i=1}^{n} w_H(\boldsymbol{x}_i;\boldsymbol{x}_0)\left\{\frac{\boldsymbol{\beta}^*(\boldsymbol{x}_0)^T\boldsymbol{x}^*(\boldsymbol{x}_i;\boldsymbol{x}_0)y_i - b(\boldsymbol{\beta}^*(\boldsymbol{x}_0)^T\boldsymbol{x}^*(\boldsymbol{x}_i;\boldsymbol{x}_0))}{\psi(\boldsymbol{x}_0)}\right. \\ &\qquad \left. + c(y_i, \psi(\boldsymbol{x}_0))\right\} \\ &\quad - n\lambda\boldsymbol{\beta}^*(\boldsymbol{x}_0)^T K\boldsymbol{\beta}^*(\boldsymbol{x}_0)/2. \end{aligned} \tag{3.2}$$

The weight function $w_H(\boldsymbol{x}_i;\boldsymbol{x}_0)$ is the Gaussian product kernel:

$$w_H(\boldsymbol{x};\boldsymbol{x}_0) = \frac{K_d(H^{-1}(\boldsymbol{x}-\boldsymbol{x}_0))}{|\det(H)|}, \qquad K_d(\boldsymbol{u}) = (2\pi)^{-d/2}\prod_{j=1}^{d}\exp\left(-\frac{1}{2}u_j^2\right), \tag{3.3}$$

where a matrix H is a $p \times p$ nonsingular bandwidth matrix. We use $H = hI_d$, where I_d is a d-dimensional identity matrix and h is a bandwidth. The regularization term is

given by equation:

$$\boldsymbol{\beta}^*(\boldsymbol{x}_0)^T K \boldsymbol{\beta}^*(\boldsymbol{x}_0) = \boldsymbol{\beta}^*(\boldsymbol{x}_0)^T \begin{bmatrix} 0 & \mathbf{0}_d^T \\ \mathbf{0}_d & I_d \end{bmatrix} \boldsymbol{\beta}^*(\boldsymbol{x}_0), \tag{3.4}$$

where $\mathbf{0}_d$ is a d-dimensional zero vector.

The estimators $\hat{\boldsymbol{\beta}}^*(\boldsymbol{x}_0)$ and $\hat{\psi}(\boldsymbol{x}_0)$ are obtained by the iterative algorithm. Replacing the unknown parameters $\boldsymbol{\beta}^*(\boldsymbol{x}_0)$ and $\psi(\boldsymbol{x}_0)$ by $\hat{\boldsymbol{\beta}}^*(\boldsymbol{x}_0)$ and $\hat{\psi}(\boldsymbol{x}_0)$ respectively and noting that the parameter $\theta(\boldsymbol{x}_i)$ in the exponential family $f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}^*(\boldsymbol{x}_0), \psi(\boldsymbol{x}_0))$ is directly given by $\theta(\boldsymbol{x}_i) = \boldsymbol{\beta}^*(\boldsymbol{x}_i)^T \boldsymbol{x}^*(\boldsymbol{x}_i; \boldsymbol{x}_i) = \beta_0^*(\boldsymbol{x}_i)$, we obtain the estimated model:

$$f(y_i \mid \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}^*(\boldsymbol{x}_i), \hat{\psi}(\boldsymbol{x}_i)) = \exp\left\{ \frac{\hat{\beta}_0^*(\boldsymbol{x}_i) y_i - b(\hat{\beta}_0^*(\boldsymbol{x}_i))}{\hat{\psi}(\boldsymbol{x}_i)} + c(y_i, \hat{\psi}(\boldsymbol{x}_i)) \right\}. \tag{3.5}$$

Example 3.1. We illustrate the proposed regularized local likelihood modeling by fitting surface to the simulation data. The random samples $\{(y_i, \boldsymbol{x}_i); i = 1, 2, \ldots, 300\}$ are generated from the true model $y_i = \sin(5\pi x_{i1}) + \cos(2\pi x_{i2}) + \varepsilon_i$, $\varepsilon_i \sim N(0, 0.1^2)$, where the design points x_{i1} and x_{i2} in $\boldsymbol{x}_i = (x_{i1}, x_{i2})^T$ are uniformly distributed in $[-1, 1]$. Figure 2 shows the true surface, where x_1, x_2 in $\boldsymbol{x} = (x_1, x_2)^T$ and y are expressed as X-, Y- and Z-axis respectively. We apply the local likelihood method and the regularized local likelihood method to the simulation data respectively, to examine the effectiveness of the regularization parameter $\zeta = \lambda\sigma^2$, where we use the Gaussian model. Figure 3(a) shows the estimated surface for the bandwidth matrix $H = h \times I_2$ ($h = 0.07$) without the help of regularization ($\lambda = \zeta = 0$). This surface is obviously undersmoothed in the boundary region. We use a large bandwidth $h = 0.1$ and the corresponding estimated surface is given in Fig. 3(b), but the fitting in the boundary region cannot be improved. This result implies that the estimated surface is unstable in the boundary region. Figure 3(c) shows the estimated surface for the bandwidth $h = 0.07$ with the help of regularization ($\zeta = 10^{-3}$). This surface gives a good representation of the underlying function over the boundary region. Figure 3(d) shows the four curves for the true and estimated surfaces in the boundary region ($x_2 = -1$), where the solid line in Fig. 3(d) is the true model, the broken line in Fig. 3(d) is the local likelihood model ($h = 0.07$), the dotted line in Fig. 3(d) is the local likelihood model ($h = 0.1$) and the

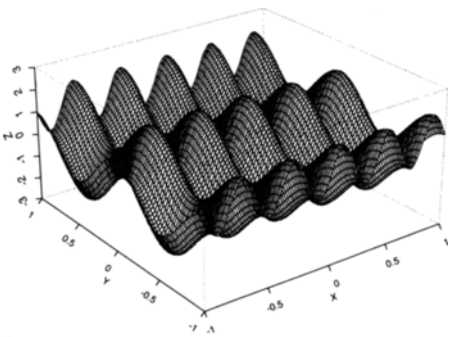


Fig. 2. True surface $y = \sin(5\pi x_1) + \cos(2\pi x_2)$.

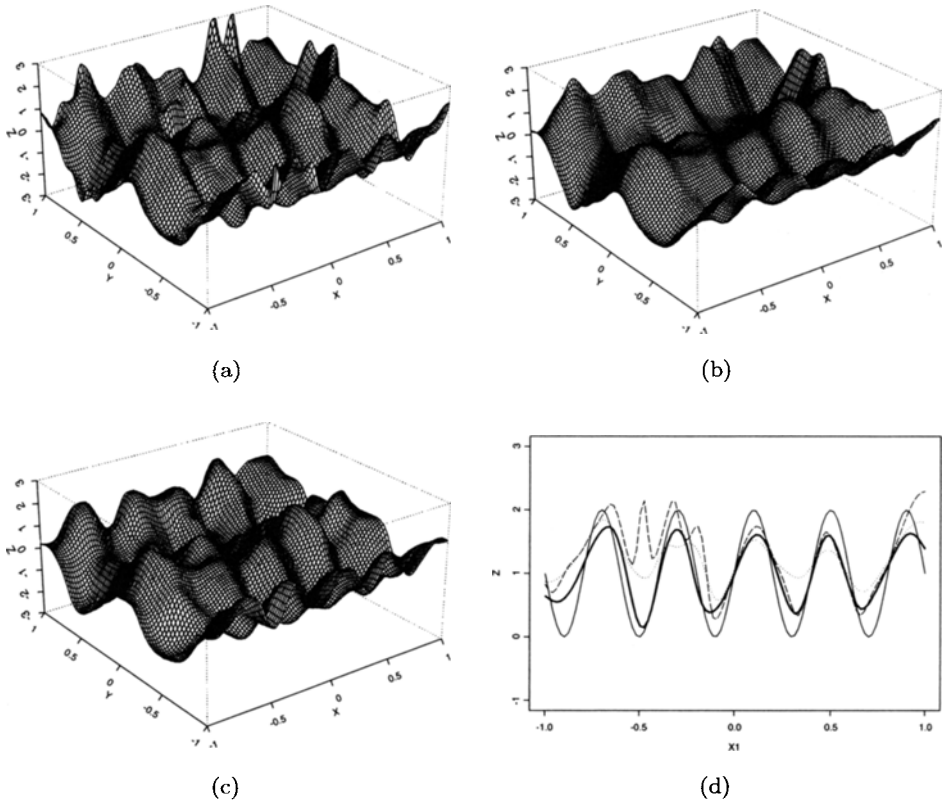


Fig. 3. Comparison of the estimated surfaces. (a) and (b) show the estimated surfaces for the local likelihood method with $h = 0.07$ and $h = 0.1$ respectively. (c) shows the estimated surface for the regularized local likelihood method with $h = 0.07$ and $\zeta = 10^{-3}$. (d) shows the four curves for the true and estimated surfaces in the boundary region ($x_2 = -1$).

thick line in Fig. 3(d) is the regularized local likelihood model ($h = 0.07, \zeta = 10^{-3}$). In this figure, the thick line gives the most appropriate estimate of the true model since this gives a stable sine curve.

In Section 2, the local likelihood method may cause the instability of estimator for large degrees of polynomial p. In multivariate case, this issue may be occurred for large number of explanatory variables d.

3.2 *Model selection*

The crucial issue is how to choose a bandwidth matrix H and a regularization parameter λ. We derive an information criterion to evaluate the estimated model in (3.5). Similarly to the way in Subsection 2.3, we obtain the information criterion based on the multivariate regularized local likelihood method by

$$\text{GIC}_{H,\lambda} = -2\sum_{i=1}^{n}\left\{\frac{\hat{\beta}_0^*(\boldsymbol{x}_i)y_i - b(\hat{\beta}_0^*(\boldsymbol{x}_i))}{\hat{\psi}(\boldsymbol{x}_i)} + c(y_i, \hat{\psi}(\boldsymbol{x}_i))\right\} \tag{3.6}$$

$$+\frac{2}{n}\sum_{i=1}^{n}\mathrm{tr}\{\hat{R}(\boldsymbol{x}_i)^{-1}\hat{Q}(\boldsymbol{x}_i)\},$$

where

$$\hat{Q}(\boldsymbol{x})=\frac{1}{n\hat{\psi}(\boldsymbol{x})}\begin{bmatrix}A^{(2)}\boldsymbol{X}/\hat{\psi}(\boldsymbol{x})-\lambda K\hat{\boldsymbol{\beta}}^{*}(\boldsymbol{x})\mathbf{1}_n^T\Lambda\boldsymbol{X} & A^{(1)}\boldsymbol{p}-\hat{\psi}(\boldsymbol{x})\lambda K\hat{\boldsymbol{\beta}}^{*}(\boldsymbol{x})\mathbf{1}_n^T\boldsymbol{p}\\ \boldsymbol{p}^TA^{(1)T} & \hat{\psi}(\boldsymbol{x})\boldsymbol{p}^T\boldsymbol{W}\boldsymbol{p}\end{bmatrix},$$

$$\hat{R}(\boldsymbol{x})=\frac{1}{n\hat{\psi}(\boldsymbol{x})}\begin{bmatrix}\boldsymbol{X}^T\boldsymbol{W}\Gamma\boldsymbol{X}+n\hat{\psi}(\boldsymbol{x})\lambda K & A^{(1)}\mathbf{1}_n/\hat{\psi}(\boldsymbol{x})\\ \mathbf{1}_n^TA^{(1)T}/\hat{\psi}(\boldsymbol{x}) & -\hat{\psi}(\boldsymbol{x})\boldsymbol{q}^T\boldsymbol{W}\mathbf{1}_n\end{bmatrix},$$

$$A^{(k)}=\boldsymbol{X}^T\boldsymbol{W}\Lambda^k,\qquad \boldsymbol{X}=(\boldsymbol{x}^*(\boldsymbol{x}_1;\boldsymbol{x}),\ldots,\boldsymbol{x}^*(\boldsymbol{x}_n;\boldsymbol{x}))^T,\qquad \boldsymbol{W}=\mathrm{diag}\{w_H(\boldsymbol{x}_i;\boldsymbol{x})\}.$$

The notations Λ, Γ, $\boldsymbol{p}$ and $\boldsymbol{q}$ are given by (2.18), where we use $\hat{\boldsymbol{\beta}}^{*}(\boldsymbol{x})^T\boldsymbol{x}^*(\boldsymbol{x}_i;\boldsymbol{x})$ and $\hat{\psi}(\boldsymbol{x})$ instead of $\hat{\boldsymbol{\beta}}(x)^T\boldsymbol{x}(x_i;x)$ and $\hat{\psi}(x)$ respectively. We select a bandwidth matrix H and a regularization parameter λ which mininize the information criterion (3.6).

We give the information criteria in the case of Gaussian and logistic models.

3.2.1 *Gaussian model*

We use the Gaussian regression model $f_N(y_i\mid\boldsymbol{x}_i;\hat{\boldsymbol{\beta}}^{*}(\boldsymbol{x}_i),\hat{\sigma}^2(\boldsymbol{x}_i))$ in (2.13), where the estimators $\hat{\boldsymbol{\beta}}^{*}(\boldsymbol{x}_0)$ and $\hat{\sigma}^2(\boldsymbol{x}_0)$ are given by

$$\hat{\boldsymbol{\beta}}^{*}(\boldsymbol{x}_0)=(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}+n\zeta Q)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y},\tag{3.7}$$

$$\hat{\sigma}^2(\boldsymbol{x}_0)=\{\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^{*}(\boldsymbol{x}_0)\}^T\boldsymbol{W}\{\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^{*}(\boldsymbol{x}_0)\}/\mathrm{tr}(\boldsymbol{W}).$$

Then we obtain the information criterion for the estimated Gaussian model as follows;

$$\mathrm{GIC}_{H,\lambda}=\sum_{i=1}^{n}\left[\log\{2\pi\hat{\sigma}^2(\boldsymbol{x}_i)\}+\frac{\{y_i-\hat{\beta}_0^{*}(\boldsymbol{x}_i)\}^2}{\hat{\sigma}^2(\boldsymbol{x}_i)}\right]+\frac{2}{n}\sum_{i=1}^{n}\mathrm{tr}\{\hat{R}(\boldsymbol{x}_i)^{-1}\hat{Q}(\boldsymbol{x}_i)\},\tag{3.8}$$

$$\hat{Q}(\boldsymbol{x})=\frac{1}{n\hat{\sigma}^4(\boldsymbol{x})}\times\begin{bmatrix}\frac{1}{\hat{\sigma}^2(\boldsymbol{x})}A_N^{(2)}\boldsymbol{X}-\lambda\hat{\sigma}^2(\boldsymbol{x})K\hat{\boldsymbol{\beta}}^{*}(\boldsymbol{x})\mathbf{1}_n^T\Lambda_N\boldsymbol{X} & \frac{1}{2\hat{\sigma}^2(\boldsymbol{x})}A_N^{(3)}\mathbf{1}_n-\frac{1}{2}A_N^{(1)}\mathbf{1}_n\\ \frac{1}{2\hat{\sigma}^2(\boldsymbol{x})}\mathbf{1}_n^TA_N^{(3)T}-\frac{1}{2}\mathbf{1}_n^TA_N^{(1)T} & \frac{1}{4\hat{\sigma}^4(\boldsymbol{x})}\mathbf{1}_n^T\Lambda_N^4\boldsymbol{W}\mathbf{1}_n-\frac{1}{4}\mathrm{tr}(\boldsymbol{W})\end{bmatrix},$$

$$\hat{R}(\boldsymbol{x})=\frac{1}{n\hat{\sigma}^2(\boldsymbol{x})}\begin{bmatrix}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}+n\hat{\sigma}^2(\boldsymbol{x})\lambda K & \frac{1}{\hat{\sigma}^2(\boldsymbol{x})}A_N^{(1)}\mathbf{1}_n\\ \frac{1}{\hat{\sigma}^2(\boldsymbol{x})}\mathbf{1}_n^TA_N^{(1)T} & \frac{1}{2\hat{\sigma}^2(\boldsymbol{x})}\mathrm{tr}(\boldsymbol{W})\end{bmatrix},$$

where $A_N^{(k)}=\boldsymbol{X}^T\boldsymbol{W}\Lambda_N^k$ and $\Lambda_N=\mathrm{diag}\{y_i-\hat{\boldsymbol{\beta}}^{*}(\boldsymbol{x})^T\boldsymbol{x}^*(\boldsymbol{x}_i;\boldsymbol{x})\}$.

3.2.2 *Logistic model*

In the case of logistic model, we can obtain the information criterion as follows

$$\mathrm{GIC}_{H,\lambda}=-2\sum_{i=1}^{n}\log f_L(y_i\mid\boldsymbol{x}_i,\hat{\boldsymbol{\beta}}^{*}(\boldsymbol{x}_i))+\frac{2}{n}\sum_{i=1}^{n}\mathrm{tr}\{\hat{R}(\boldsymbol{x}_i)^{-1}\hat{Q}(\boldsymbol{x}_i)\},\tag{3.9}$$

$$\hat{Q}(\boldsymbol{x})=\frac{1}{n}\sum_{i=1}^{n}\{y_i-\hat{\pi}(\boldsymbol{x}_i;\boldsymbol{x})\}\times[w_H(\boldsymbol{x}_i;\boldsymbol{x})\{y_i-\hat{\pi}(\boldsymbol{x}_i;\boldsymbol{x})\}\boldsymbol{x}^*(\boldsymbol{x}_i;\boldsymbol{x})-\lambda K\hat{\boldsymbol{\beta}}^{*}(\boldsymbol{x})]\boldsymbol{x}^*(\boldsymbol{x}_i;\boldsymbol{x})^T,$$

$$\hat{R}(\boldsymbol{x})=-\frac{1}{n}\sum_{i=1}^{n}[w_H(\boldsymbol{x}_i;\boldsymbol{x})\hat{\pi}(\boldsymbol{x}_i;\boldsymbol{x})\{1-\hat{\pi}(\boldsymbol{x}_i;\boldsymbol{x})\}\boldsymbol{x}^*(\boldsymbol{x}_i;\boldsymbol{x})\boldsymbol{x}^*(\boldsymbol{x}_i;\boldsymbol{x})^T-\lambda K],$$

for the estimator $\hat{\boldsymbol{\beta}}^*(\boldsymbol{x})$ which maximizes the regularized local likelihood function (3.2), where $\hat{\pi}(\boldsymbol{x}_i;\boldsymbol{x}) = \exp\{\hat{\boldsymbol{\beta}}^*(\boldsymbol{x})^T\boldsymbol{x}^*(\boldsymbol{x}_i;\boldsymbol{x})\}/[1+\exp\{\hat{\boldsymbol{\beta}}^*(\boldsymbol{x})^T\boldsymbol{x}^*(\boldsymbol{x}_i;\boldsymbol{x})\}]$ and $f_L(y_i \mid \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}^*(\boldsymbol{x}_i)) = \hat{\pi}(\boldsymbol{x}_i;\boldsymbol{x}_i)^{y_i}\{1-\hat{\pi}(\boldsymbol{x}_i;\boldsymbol{x}_i)\}^{1-y_i}$.

4. Real examples and numerical results

In this section we use a real data example and Monte Carlo simulations to investigate the performance of the regularized local likelihood modeling.

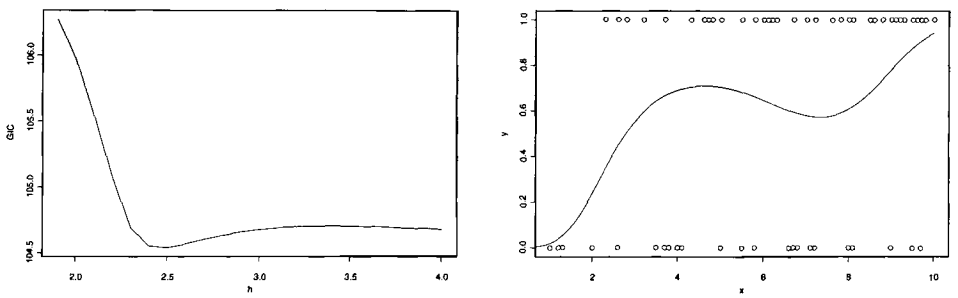
4.1 *Summer rainfall data*

We apply the proposed modeling procedure to the rainfall data in Kagoshima. In general, investigators in national meteorological observatories observe the sky at fixed times. For example, when they observe the clouds, they look out over the sky and observe the rate of the clouds which is called "cloud amount". They classify the weather condition as "Blue Sky" or "Cloudy". However, a weather condition such as "Rain" does not relate to cloud amount directly. Therefore, we investigate the relationship between cloud amount and "Rain" using our nonlinear modeling procedure.

We use the data $\{(y_i, x_i); i = 1, 2, \ldots, 92\}$, where y_i are the binary response variables having the value 1 (if the wheather is "Rain") or 0 (otherwise) and $x_i \in [0, 10]$ are the average daily cloud amounts. These variables were observed in Kagoshima from June, 2001 to August, 2001. We apply our nonlinear logistic regression model in (2.16) to this data and select the bandwidth h, the degree of polynomial p and the regularization parameter λ by minimizing GIC in (2.20).

Figure 4(a) shows the minimum GIC with respect to the bandwidth h, where the degree of polynomial $p = 3$ and the regularization parameter $\lambda = 1.00 \times 10^{-7}$. We observe that the optimal bandwidth is $\hat{h} = 2.5$, and show the estimated curve in Fig. 4(b). In general, we tend to predict the probability of precipitation that is higher as the cloud amount increases. The estimated curve is not monotone increasing, since the probability of precipitation is influenced by cloud types in addition to the cloud amount.

We examine the relationship among cloud amount, humidity and "Rain" using our nonlinear regression modeling. We use the data $\{(y_i, \boldsymbol{x}_i); i = 1, 2, \ldots, 92\}$, where $\boldsymbol{x}_i = (x_{i1}, x_{i2})$, x_{i1} are the average daily cloud amounts and x_{i2} are the average daily values



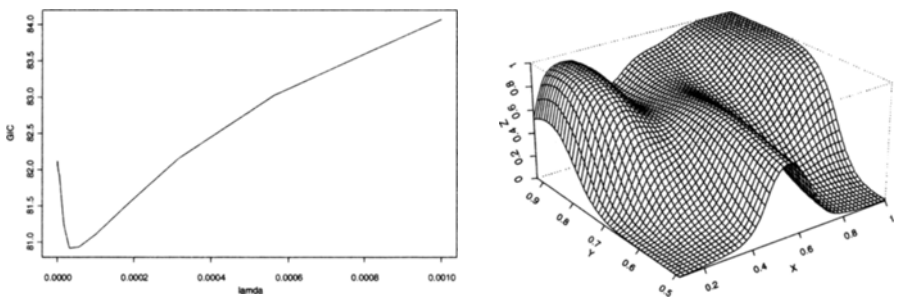
(a) (b)

Fig. 4. (a) The relationship between the bandwidth h and GIC with the number of polynomial $p = 3$ and the regularization parameter $\lambda = 1.00 \times 10^{-7}$. (b) The smoothed curve based on the regularized local likelihood model and GIC ($\hat{h} = 2.5$, $\hat{\lambda} = 1.00 \times 10^{-7}$).

of humidity. The data were observed in Kagoshima from June, 2001 to August, 2001, respectively. We apply our nonlinear logistic regression model to this data and select the bandwidth matrix H and the regularization parameter λ by minimizing GIC in (3.9). We use $H = hI_2$ (I_2 is 2×2 identity matrix) as the bandwidth matrix.

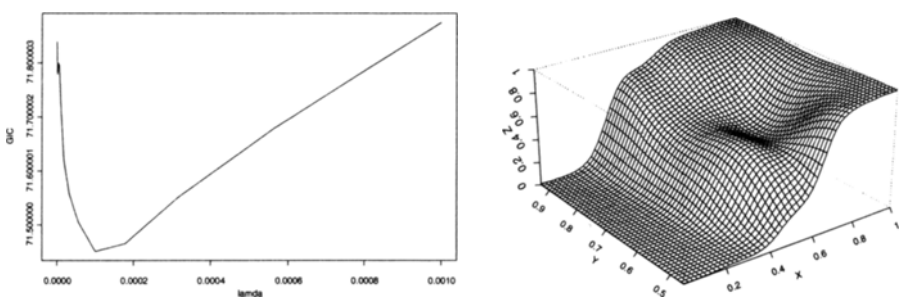
Figure 5(a) shows the minimum GIC with respect to the regularization parameter λ, with the optimal bandwidth. We select the optimal bandwidth $\hat{h} = 0.11$ and regularization parameter $\hat{\lambda} = 3.16 \times 10^{-5}$, and show the corresponding estimated surface in Fig. 5(b). We observe that the probability of precipitation is influenced by the humidity rather than the cloud amount, however the relationship between the humidity and the probability of precipitation is not monotonic. We conclude that the other factors, such as the atmospheric pressure and the sort of clouds, may influence the probability of precipitation.

Figure 6 shows the result of the nonlinear logistic regression model for the data $\{(y_i, \boldsymbol{x}_i); i = 1, 2, \ldots, 92\}$ observed in Niigata from June, 2001 to August, 2001. Figure 6(a) shows the minimum GIC with respect to the regularization parameter λ for the optimal bandwidth, and (b) shows the estimated surface ($\hat{h} = 0.1$, $\hat{\lambda} = 1.00 \times 10^{-4}$)



(a) (b)

Fig. 5. (a) The relationship between the regularization parameter λ and GIC with the optimal bandwidth h. (b) The smoothed surface based on the regularized local likelihood model and GIC ($\hat{h} = 0.11$, $\hat{\lambda} = 3.16 \times 10^{-5}$).



(a) (b)

Fig. 6. (a) The relationship between the regularization parameter λ and GIC with the optimal bandwidth h (in Niigata). (b) The smoothed surface based on the regularized local likelihood model and GIC ($\hat{h} = 0.1$, $\hat{\lambda} = 1.00 \times 10^{-4}$).

based on the regularized local likelihood method and GIC. It is clear that the estimated surface varies according to the locality where the data were observed, by comparing Fig. 5(b) with Fig. 6(b). This difference of the estimated surfaces may be caused by the property of summer climate in each place. In this example, Kagoshima is located in the Pacific coast, while Niigata is located in the coast of the Sea of Japan. The probability of precipitation in Kagoshima is related to the humidity because of the large influence of rain shower and typhoon. Conversely, the probability of precipitation in Niigata is related to the cloud amount because of the small influence of rain shower and typhoon.

4.2 *Monte Calro simulations*

Monte Carlo simulations were conducted to investigate the performance of the nonlinear modeling strategy based on the regularized local likelihood method. For the simulation study, repeated random samples $\{(y_i, \boldsymbol{x}_i); i = 1, 2, \ldots, n\}$ were generated from the true regression model $y_i = m(\boldsymbol{x}_i) + \varepsilon_i$, where $\boldsymbol{x}_i = (x_{i1}, x_{i2})^T$ and the design points x_{i1}, x_{i2} are uniformly distributed in $[-1, 1]$. The errors ε_i are assumed to be independently distributed according to a mixture of normal distributions $rN(0, (0.5R_y)^2) + (1 - r)N(0, (0.1R_y)^2)$, where R_y is the range of $m(\boldsymbol{x})$ over $\boldsymbol{x}$ and $r = 0.1, 0.5$. For the true curve $m(\boldsymbol{x})$ we consider the following function: $m(\boldsymbol{x}_i) = \sin(2\pi x_{i1})/2 + \cos(4\pi x_{i2})/2$.

We fit the nonlinear regression model with Gaussian noise defined by (2.13) to the simulated data. The model is estimated by the maximization of the local log-likelihood function and the regularized local log-likelihood function (2.6). The values of the bandwidth matrix $H = hI_2$ and the regularization parameter λ are chosen as the minimizers of the criteria GIC and Modified AIC given by

$$\mathrm{MAIC}_{H,\lambda} = 2\sum_{i=1}^{n}\left[\log(2\pi\hat{\sigma}^2(\boldsymbol{x}_i)) + \frac{\{y_i - \hat{\boldsymbol{\beta}}^*(\boldsymbol{x}_i)^T\boldsymbol{x}^*(\boldsymbol{x}_i; \boldsymbol{x}_i)\}^2}{\hat{\sigma}^2(\boldsymbol{x}_i)}\right] + 2\,\mathrm{tr}(S),$$

where S is the $n \times n$ smoother matrix (Hastie and Tibshirani (1990)) given by $(\boldsymbol{s}(\boldsymbol{x}_1), \ldots, \boldsymbol{s}(\boldsymbol{x}_n))^T$ with $\boldsymbol{s}(\boldsymbol{x}_i) = [\boldsymbol{e}_i^T\boldsymbol{X}\{\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + n\zeta Q\}^{-1}\boldsymbol{X}^T\boldsymbol{W}]^T$ and $\boldsymbol{e}_i$ are the n-dimensional vectors having the value 1 in the i-th entry and zero elsewhere. The cross-validation (CV) and the generalized cross-validation (GCV) were also used for the choice of the adjusted parameters. These criteria were examined by comparing the mean squared error:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{\boldsymbol{\beta}}^*(\boldsymbol{x}_i)^T\boldsymbol{x}^*(\boldsymbol{x}_i; \boldsymbol{x}_i) - m(\boldsymbol{x}_i)\}^2$$

and the standard deviations (SD) of the selected h and λ.

Tables 1 and 2 compare the MSE between the true and estimated functions, in which the fitted functions are obtained by averaging over 100 repeated Monte Carlo trials so we set this as AMSE; h and λ in each table are the averages of optimal h and λ, and (SD) is the standard deviation of h and λ. In the case where $n = 100$ and $r = 0.1$ in Table 1, RLLM & GIC and RLLM & GCV are superior to other methods in the sense of decreasing MSE. In the other cases where $r = 0.5$ in Table 2, LLM & GIC and RLLM & GCV are superior to other methods in that sense. Moreover, the selected models based on GIC are more stable since the optimal h has smaller variance than the use of CV and GCV. This implies that the proposed modeling procedure gives the fitted functions that capture the true structure in practical applications.

Table 1. The result of the surface when $r = 0.1$.

LLM	GIC	MAIC	CV	GCV
h	0.119	0.063	0.126	0.096
(SD)	0.015	0.008	0.034	0.014
AMSE$\times 10^4$	395	331	454	314
(SD)	1.03×10^{-2}	5.72×10^{-3}	2.28×10^{-2}	5.95×10^{-3}
RLLM	GIC	MAIC	CV	GCV
h	0.076	0.054	0.088	0.080
(SD)	0.009	0.002	0.024	0.014
λ	4.58×10^{-1}	8.40×10^{-3}	3.66×10^{-2}	8.56×10^{-2}
(SD)	5.05×10^{-2}	2.51×10^{-2}	7.46×10^{-2}	1.08×10^{-1}
AMSE$\times 10^4$	291	355	299	292
(SD)	4.02×10^{-3}	5.36×10^{-3}	5.57×10^{-3}	4.98×10^{-3}

Table 2. The result of the surface when $r = 0.5$.

LLM	GIC	MAIC	CV	GCV
h	0.123	0.064	0.205	0.170
(SD)	0.017	0.007	0.065	0.055
AMSE$\times 10^4$	1183	1971	1348	1283
(SD)	2.21×10^{-2}	3.34×10^{-2}	2.51×10^{-2}	2.84×10^{-2}
RLLM	GIC	MAIC	CV	GCV
h	0.084	0.051	0.154	0.134
(SD)	0.014	0.002	0.062	0.042
λ	3.53×10^{-2}	7.22×10^{-5}	1.89×10^{-2}	2.20×10^{-2}
(SD)	7.65×10^{-3}	1.50×10^{-4}	1.59×10^{-2}	1.49×10^{-2}
AMSE$\times 10^4$	1232	2185	1241	1186
(SD)	2.34×10^{-2}	3.45×10^{-2}	3.03×10^{-2}	3.24×10^{-2}

Figure 7 shows the comparison between GIC and MAIC, where we use the regularization parameter $\zeta = 10^{-5}$. Figure 7(a) indicates the relationship between the bandwidth h and the log-likelihood term given by

$$2\sum_{i=1}^{n}\left[\log(2\pi\hat{\sigma}^2(\boldsymbol{x}_i)) + \frac{\{y_i - \hat{\boldsymbol{\beta}}^*(\boldsymbol{x}_i)^T\boldsymbol{x}^*(\boldsymbol{x}_i;\boldsymbol{x}_i)\}^2}{\hat{\sigma}^2(\boldsymbol{x}_i)}\right],$$

which is a monotone increasing function with respect to the bandwidth h. Figure 7(b) shows the comparison between GIC and MAIC with respect to the corrected bias terms, where the dotted line indicates the bias of MAIC and the solid line the bias of GIC. The biases present a monotone decrease function with respect to the bandwidth h. Figure 7(c) shows the results for GIC and MAIC. In the case of MAIC, the optimal bandwidth tends to be too small to select as shown in Fig. 7(c), since the estimator may be incomputable for smaller values of h. On the contrary, GIC always selects the optimal bandwidth as in Fig. 7(c), and has smaller variance than the cases of CV and GCV.

Nonaka *et al.* (2003) examined the efficiency of our proposed method through a Monte Carlo simulation for the one-dimensional explanatory variable x.

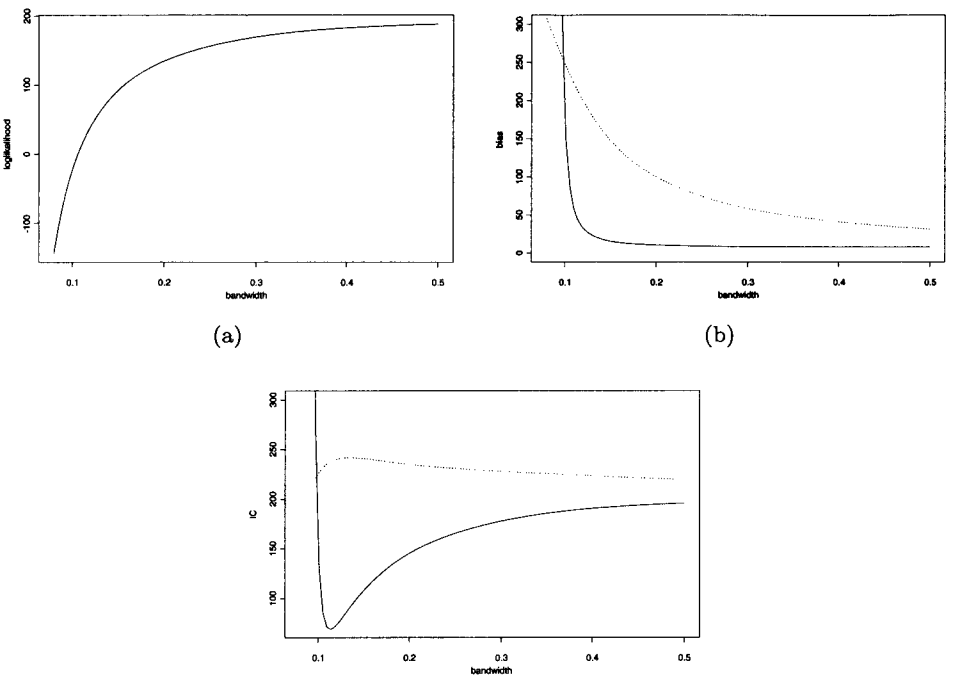


Fig. 7. (a) The relationship between the bandwidth h and the log-likelihood term with the regularization parameter $\zeta = 10^{-5}$. (b) The relationship between the bandwidth h and the corrected bias term. The dotted line expresses the bias of MAIC and the solid line expresses the bias of GIC. (c) The relationship between the bandwidth h and the information criterion (IC). The dotted line expresses the value of MAIC and the solid line expresses the value of GIC.

5. Concluding remarks

The main aim of this paper was to introduce the regularized local likelihood method in constructing a nonlinear regression model; determining a set of kernel function with bandwidth, estimating the unknown parameters by regularization and then evaluating the constructed model to select a suitable one. The estimated curve based on the local likelihood method tends to be unstable for a small bandwidth h and a higher degree of polynomial p. In particular, the estimated surface can not be calculated for a small $H = hI_d$ in the multivariate case. We proposed the nonlinear regression modeling procedure based on the regularized local likelihood method in order to obtain a stable estimator, and derived a model selection criterion for evaluating constructed models from an information-theoretic point of view.

We applied the regularized local likelihood method to summer rainfall data and simulated data. We observed that our method is effective in constructing nonlinear regression models for the multivariate data, and that the proposed strategy using the information criterion GIC yields stable parameter estimates. It may be applied to construct Gaussian, logistic and Poisson nonlinear regression models, and provides a tool to draw information about the system under consideration from a finite and noisy data

set. We would recommend implementing nonlinear regression modeling based on the regularized local likelihood method, using the information criterion GIC.

Acknowledgements

The authors would like to thank the referees and the Editor for their helpful comments and suggestions.

Appendix: Derivation of the information criterion

We derive an information criterion to evaluate models estimated by the regularized local likelihood method.

Suppose that $z_1, z_2, \ldots, z_n$ are future observations for the response variable Y drawn from $g(y \mid x)$. Let $f(\boldsymbol{z} \mid X; \hat{\boldsymbol{\theta}}(X)) = \prod_{i=1}^n f(z_i \mid x_i; \hat{\boldsymbol{\beta}}(x_i), \hat{\psi}(x_i))$ and $g(\boldsymbol{z} \mid X) = \prod_{i=1}^n g(z_i \mid x_i)$. An information criterion may be derived as an estimator of the Kullback-Leibler information (Kullback and Leibler (1951))

$$KL\{g, f\} = E_{G(\boldsymbol{z}|X)}[\log g(\boldsymbol{z} \mid X)] - E_{G(\boldsymbol{z}|X)}[\log f(\boldsymbol{z} \mid X; \hat{\boldsymbol{\theta}}(X))] \tag{A.1}$$

conditional on $\hat{\boldsymbol{\theta}}(X) = (\hat{\boldsymbol{\beta}}(X)^T, \hat{\psi}(X))^T$.

The first term in the right-hand side of equation (A.1) is constant over all models and only the second term

$$E_{G(\boldsymbol{z}|X)}[\log f(\boldsymbol{z} \mid X; \hat{\boldsymbol{\theta}}(x_0))] = \int \log f(z \mid x; \hat{\boldsymbol{\theta}}(x_0)) dG(z \mid x), \tag{A.2}$$

is relevant. Hence, instead of minimizing the Kullback-Leibler information (A.1), we maximize the expected log-likelihood (A.2) that depends on the unknown true distribution $G(\boldsymbol{z} \mid X)$. An estimate of the expected log-likelihood is the log-likelihood

$$\sum_{i=1}^n \log f(y_i \mid x_i; \hat{\boldsymbol{\theta}}(x_0)), \tag{A.3}$$

obtained by replacing the unknown distribution $G(\boldsymbol{z} \mid X)$ by the empirical distribution. Then the bias of the log-likelihood in estimating the expected log-likelihood is given by

$$\mathrm{b}(G) = E_{G(\boldsymbol{y}|X)}[\log f(\boldsymbol{y} \mid X; \hat{\boldsymbol{\theta}}(x_0)) - E_{G(\boldsymbol{z}|X)}[\log f(\boldsymbol{z} \mid X; \hat{\boldsymbol{\theta}}(x_0))]].$$

Konishi and Kitagawa (1996) considered an asymptotic bias for a statistical model with functional estimators and gave the bias by a function of the empirical influence function of estimators and the score function of a specified parametric model. It may be seen that the regularized local likelihood estimator $\hat{\boldsymbol{\theta}}(x_0) = (\hat{\boldsymbol{\beta}}(x_0)^T, \hat{\psi}(x_0))^T$ can be expressed as $\hat{\boldsymbol{\theta}}(x_0) = T(\hat{G})$ for the functional $T(\cdot)$ defined by

$$\int \frac{\partial}{\partial \boldsymbol{\theta}} \{w_h(x; x_0) \log f(z \mid x; \boldsymbol{\theta}(x_0)) - \lambda \boldsymbol{\beta}(x_0)^T K \boldsymbol{\beta}(x_0)/2\} \Big|_{T(G)} dG(z) = \mathbf{0}, \tag{A.4}$$

where G and $\hat{G}$ are respectively the joint distribution of (x, y) and the empirical distribution function based on the observed data. Replacing G in (A.4) by $G_\varepsilon = (1-\varepsilon)G + \varepsilon \delta_{(y,x)}$

with $\delta_{(y,x)}$ being a point of mass at (y, x) and differentiating with respect to ε yield the influence function of the regularized estimator $\hat{\boldsymbol{\theta}}(x_0) = T(\hat{G})$ in the form

$$\text{(A.5)}\quad T^{(1)}(z \mid x; G) = R(G)^{-1}\frac{\partial}{\partial \boldsymbol{\theta}}\{w_h(x; x_0)\log f(z \mid x; \boldsymbol{\theta}(x_0)) - \lambda \boldsymbol{\beta}(x_0)^T K \boldsymbol{\beta}(x_0)/2\}\Big|_{T(G)},$$

where

$$\text{(A.6)}\quad R(G) = -\int \left\{\frac{\partial^2 A(z \mid x; \boldsymbol{\theta}(x_0))}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right\}\Big|_{T(G)} dG(z),$$

with $A(z \mid x; \boldsymbol{\theta}(x_0)) = w_h(x; x_0)\log f(z \mid x; \boldsymbol{\theta}(x_0)) - \lambda \boldsymbol{\beta}(x_0)^T K \boldsymbol{\beta}(x_0)/2$.

It follows from Theorem 2.1 in Konishi and Kitagawa (1996) (see also Konishi (1999)) that the bias is asymptotically given by $\mathrm{b}(G) = \mathrm{tr}\{R(G)^{-1}Q(G)\} + o(1/n)$, where

$$Q(G) = \int \left\{\frac{\partial A(z \mid x; \boldsymbol{\theta}(x_0))}{\partial \boldsymbol{\theta}}\frac{\partial \log f(z \mid x; \boldsymbol{\theta}(x_0))}{\partial \boldsymbol{\theta}^T}\right\}\Big|_{T(G)} dG(z).$$

By replacing the unknown distribution G by the empirical distribution $\hat{G}$, we have an information criterion

$$\text{(A.7)}\quad \mathrm{GIC}_{h,p,\lambda}(x_0) = -2\sum_{i=1}^{n}\log f(y_i \mid x_i; \hat{\boldsymbol{\theta}}(x_0)) + 2\,\mathrm{tr}\{R(\hat{G})^{-1}Q(\hat{G})\}.$$

where

$$Q(\hat{G}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial A(y_i \mid x_i; \boldsymbol{\theta}(x_0))}{\partial \boldsymbol{\theta}}\frac{\partial \log f(y_i \mid x_i; \boldsymbol{\theta}(x_0))}{\partial \boldsymbol{\theta}^T}\right\}\Big|_{\hat{\boldsymbol{\theta}}(x_0)},$$

$$R(\hat{G}) = -\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial^2 A(y_i \mid x_i; \boldsymbol{\theta}(x_0))}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right\}\Big|_{\hat{\boldsymbol{\theta}}(x_0)}.$$

It might be noticed here that the information criterion based on the local method has two problems: the information criterion $\mathrm{GIC}_{h,p,\lambda}(x_0)$ in (A.7) depends on the point x_0, and this method assesses the closeness of $f(y \mid x; \hat{\boldsymbol{\theta}}(x_0))$ to the model $g(y \mid x)$ for a fixed point x_0. In order to assess the closeness of $f(y \mid x; \hat{\boldsymbol{\theta}}(x))$ to the model $g(y \mid x)$, we modify the information criterion (A.7) in the following:

$$\text{(A.8)}\quad \mathrm{GIC}_{h,p,\lambda} = -2\sum_{i=1}^{n}\log f(y_i \mid x_i; \hat{\boldsymbol{\theta}}(x_i)) + \frac{2}{n}\sum_{i=1}^{n}\mathrm{tr}\{\hat{R}(x_i)^{-1}\hat{Q}(x_i)\},$$

where we replace $Q(\hat{G})$ and $R(\hat{G})$ with $\hat{Q}(x_0)$ and $\hat{R}(x_0)$, respectively. For the problem of choosing among different models, we select the model for which the value of the information criterion $\mathrm{GIC}_{h,p,\lambda}$ is smallest.

Irizarry (2001) proposed the use of a weighted version of the Kullback-Leibler information and derived the model selection criterion WAIC. That assesses the closeness of $f(y \mid x; \hat{\boldsymbol{\theta}}(x_0))$ to the model $g(y \mid x)$ for a fixed point x_0.

REFERENCES

Bartlett, M. S. (1954). A note on some multiplying factors for various χ approximations, *Journal of the Royal Statistical Society, Series, B*, **16**, 296–298.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association*, **74**, 829–836.

Copas, B. J. and Eguchi, S. (1998). Sensitivity approximations for selectivity bias in observational data analysis, Research Memo., No. 660, The Institute of Statistical Mathematics, Tokyo.

Eguchi, S. and Copas, J. (1998). A class of local likelihood methods and near-parametric asymptotics, *Journal of the Royal Statistical Society, Series, B*, **60**, 709–724.

Eguchi, S. and Kim, T. Y. (2001). Local likelihood method and theory for a bridge of parametric and nonparametric regression, Research Memo., No. 806, The Institute of Statistical Mathematics, Tokyo.

Eguchi, S., Kim, T. Y. and Park, B. U. (2003). Local likelihood method: A bridge over parametric and nonparametric regression, *Journal of Nonparametric Statistics*, **15**, 665–683.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman & Hall, London.

Fan, J., Heckman, N. E. and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions, *Journal of the American Statistical Association*, **90**, 141–150.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman & Hall, London.

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

Hjort, N. L. and Jones, M. L. (1996). Locally parametric nonparametric density estimation, *Annals of Statistics*, **24**, 1667–1678.

Irizarry, R. A. (2001). Information and posterior probability criteria for model selection in local likelihood estimation, *Journal of the American Statistical Association*, **96**, 303–316.

Konishi, S. (1999). Statistical model evaluation and information criteria, *Multivariate Analysis, Design of Experiments and Survey Sampling* (ed. S. Ghosh), 369–399, Marcel Dekker, New York.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection, *Biometrika*, **83**, 875–890.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *Annals of Mathematical Statistics*, **22**, 79–86.

Loader, C. R. (1999). *Local Regression and Likelihood*, Springer, New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Chapman & Hall, London.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series, A*, **135**, 370–384.

Nonaka, Y., Ando, T. and Konishi, S. (2003). Nonlinear regression modeling using regularized local likelihood (in Japanese), *Journal of the Japanese Society of Computational Statistics*, **16**, 43–57.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression, *Annals of Statistics*, **22**, 1346–1370.

Seifert, B. and Gasser, T. (1996). Variance properties of local polynomials and ensuing modifications, *Statistical Theory and Computational Aspects of Smoothing* (eds. W. Härdle and M. G. Schimek), 50–127, Physica, Heidelberg.

Seifert, B. and Gasser, T. (2000). Data adaptive ridging in local polynomial regression, *Journal of the Computational and Graphical Statistics*, **9**, 338–360.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*, Springer, New York.

Stone, C. J. (1977). Consistent nonparametric regression (with discussion), *Annals of Statistics*, **5**, 595–645.

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

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, Chapman & Hall, London.