ELSEVIER

Statistics & Probability Letters 22 (1995) 7–15

STATISTICS & PROBABILITY LETTERS

On a likelihood-based approach in nonparametric smoothing and cross-validation[☆]

Probal Chaudhuri[a,*], Anup Dewanji[b]

[a] Division of Theoretical Statistics and Mathematics, Indian Statistical Institute, 203 B. T. Road, Calcutta 700035, India
[b] Applied Statistics, Surveys and Computing Division, Indian Statistical Institute, 203 B. T. Road, Calcutta 700035, India

Received August 1993; revised December 1993

Abstract

A likelihood-based generalization of usual kernel and nearest-neighbor-type smoothing techniques and a related extension of the least-squares leave-one-out cross-validation are explored in a generalized regression set up. Several attractive features of the procedure are discussed and asymptotic properties of the resulting nonparametric function estimate are derived under suitable regularity conditions. Large sample performance of likelihood-based leave-one-out cross validation is investigated by means of certain asymptotic expansions.

Keywords: Consistency; Fisher information; Generalized regression model; Maximum likelihood cross-validation; Weighted maximum likelihood

1. Introduction

Consider a set of independent observations $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$ and a generalized regression set up in which the conditional distribution of Y_i given $X_i = x_i$ has a p.d.f./p.m.f. of the form $f\{y_i \mid \theta(x_i)\}$. Here the form of f is known but θ is an unknown real-valued function that happens to be the parameter of interest. There are plenty of examples in the literature that arise in practice and fit into this structure. Specifically, usual regression with Gaussian error, logistic regression, Poisson regression, inverse Gaussian regression and gamma regression are all special examples of such a general model. In fact, all the standard examples included in "generalized linear models" (see McCullagh and Nelder, 1989) can be considered to be special cases of the preceding generalized regression set up. Besides, the conditional distribution of Y_i given $X_i = x_i$ may have a known distribution with a location structure, where $\theta(x_i)$ will be the unknown location parameter. Recently several authors have extensively explored strategies for estimating θ by constructing various types of nonparametric smoothers (see, e.g., Hastie and Tibshirani, 1986, 1990; O'Sullivan et al., 1986;

[☆] Research presented in this note was supported in part by a grant from Indian Statistical Institute.

[*] Corresponding author.

SSDI 0167-7152(94)00040-F

Stone, 1986; Staniswalis, 1989; Cox and O'Sullivan, 1990; Gu, 1990, 1992; etc.). Staniswalis (1989) (see also the "local likelihood" estimation considered by Tibshirani and Hastie, 1987; Firth et al., 1991) considered kernel smoothers that were constructed via a maximum-likelihood-type approach. The purpose of this note is to investigate certain theoretical issues that are crucial if one wants to guarantee desirable statistical properties of such likelihood-based nonparametric smoothers. We will derive some very general conditions on the model and certain weight functions (which may or may not arise from kernel functions) that ensure good asymptotic performance of the function estimates constructed using a weighted maximum likelihood approach. Also, we will try to get useful insights into the likelihood-based leave-one-out cross-validation technique by means of certain expansions that expose some key features of such a cross-validation strategy. Further, we will indicate some potential advantages in using the weighted maximum likelihood technique to construct nonparametric function estimates and point out some important related issues.

2. Estimation and cross-validation based on likelihood

From now on, we will assume that the domain of θ is a compact subset of R^d, and the support of the regressor X is contained in that set. Let x be in the domain of θ, and consider the estimate $\hat{\theta}_n(x)$ defined as

$$\hat{\theta}_n(x) = \arg\max_t \prod_{i=1}^{n} \{f(Y_i \mid t)\}^{W_{n,i}(x)}$$

assuming that a maximum exists, and it belongs to the range of θ. Here $W_{n,i}(x)$'s are some appropriately chosen nonnegative weight functions satisfying $\sum_{i=1}^{n} W_{n,i}(x) = 1$. Further, for an X_i close to x, the value of $W_{n,i}(x)$ will be large while for an X_i far away from x, the value of $W_{n,i}(x)$ will be small so that $\hat{\theta}_n(x)$ can be viewed as some kind of a local average based on data within a neighborhood of x. Examples of various types of weight functions constructed using different kernel functions can be found in Nadarya (1964), Watson (1964), Priestley and Chao (1972), Gasser and Muller (1979, 1984), Cheng and Lin (1981), Eubank (1988), etc. On the other hand, the weight functions may arise from nearest-neighbor-type local averaging also, and there a certain number of nearest neighbors of x among the data points get positive weights, and other distant neighbors are assigned zero weight.

For a fixed value of y, the function $f(y \mid t)$ will be assumed to be differentiable with respect to t for all $t \in J$, where J is an open interval containing the range of the real-valued function θ. As a consequence of this smoothness assumption, the estimate $\hat{\theta}_n(x)$ can be computed by solving the weighted maximum likelihood equation

$$\sum_{i=1}^{n} \frac{f'\{Y_i \mid \hat{\theta}_n(x)\}}{f\{Y_i \mid \hat{\theta}_n(x)\}} W_{n,i}(x) = 0. \tag{2.1}$$

Here $f'(y \mid t)$ denotes the derivative of f with respect to t. Interestingly, for a large class of models used in practice (e.g. logistic regression model, Poisson regression model, gamma regression model, usual regression with Gaussian error, etc.), it is possible to solve (2.1) explicitly to obtain a closed-form expression for $\hat{\theta}_n(x)$. It will be appropriate to note here that this is one of the most appealing features of this approach because several other approaches considered in the literature (e.g. "penalized likelihood" as in O'Sullivan et al., 1986; Cox and O'Sullivan, 1990; Gu, 1990, 1992; or "local scoring" as in Hastie and Tibshirani, 1986, 1990) do not possess this attractive simplicity, and their implementation will typically require complex and iterative computation. Further, when the regressor is multidimensional, the "penalized likelihood" procedure becomes seriously problematic due to numerical and analytic complexities associated with the problem as well as lack of simple extension of splines in multidimension. The weighted maximum likelihood approach is completely free from such problems as the fundamental idea lying at the root of it remains unaffected whether one has to deal with univariate or multivariate regressors.

In practice, there will be a smoothing parameter intrinsically associated with the weight functions $W_{n,i}$'s $(1 \leqslant i \leqslant n)$, and its choice will influence the performance of $\hat{\theta}_n$ as an estimate of θ. To be more specific, for weight functions arising from a kernel function, the smoothing parameter is the bandwidth while in the case of nearest-neighbor-type estimation, it is the number of nearest neighbors used. Whatever the case may be, we will denote the smoothing parameter by h_n, and a brief description of an adaptive data-based procedure for choosing h_n using a likelihood-based leave-one-out cross-validation follows. Such a procedure for selecting the smoothing parameter has been used by Staniswalis (1989) and Firth et al. (1991), and their approach generalizes the earlier least-squares leave-one-out cross-validation technique considered by Stone (1974), Hardle and Marron (1985a, b), etc.

For $1 \leqslant i \leqslant n$, let $\hat{\theta}_n^{(i)}$ be an estimate of θ constructed using the weighted maximum likelihood technique applied to only $n-1$ of the data points, which are $(Y_1, X_1), \ldots, (Y_{i-1}, X_{i-1}), (Y_{i+1}, X_{i+1}), \ldots, (Y_n, X_n)$. More specifically,

$$\hat{\theta}_n^{(i)}(x) = \arg\max_t \prod_{j: 1 \leqslant j \leqslant n, j \neq i} \{f(Y_j \mid t)\}^{W_{n,j}^{(i)}(x)}$$

and the following equation holds:

$$\sum_{j: 1 \leqslant j \leqslant n, j \neq i} \frac{f'\{Y_j \mid \hat{\theta}_n^{(i)}(x)\}}{f\{Y_j \mid \hat{\theta}_n^{(i)}(x)\}} W_{n,j}^{(i)}(x) = 0.$$

Here, $W_{n,j}^{(i)}$'s $(1 \leqslant j \leqslant n, j \neq i)$ are weight functions depending on the smoothing parameter h_n, and they are based on $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$. Define a cross-validation function as

$$MLCV(h_n) = \sum_{i=1}^{n} \log[f\{Y_i \mid \hat{\theta}_n^{(i)}(X_i)\}], \tag{2.2}$$

where $MLCV$ stands for "maximum likelihood cross-validation". Then h_n will be chosen in such a way that $MLCV(h_n)$ is maximized. By suitably rescaling the range of the regressor (or equivalently the domain of θ), this maximization can be reduced to a limited numerical search if necessary.

The methodology described here has been implemented by Staniswalis (1989) and Chaudhuri and Dewanji (1991) to analyze several interesting simulated as well as real data sets that include censored survival data and data arising from biological and psychological experiments giving rise to discrete and non-Gaussian continuous responses. In all the cases reported by them, this simple and convenient technique appears to work extremely well. In the following section, we explore large sample properties of the function estimate and some related asymptotic issues.

3. Some asymptotic analysis

We begin by introducing some regularity conditions on the model $f(y \mid t)$. From now on, it will be assumed that the support of $f(y \mid t)$ is the same for all $t \in J$, and for every fixed y in that support, $g(y \mid t) = \log\{f(y \mid t)\}$ is thrice continuously differentiable with respect to $t \in J$. Let Y denote the random variable with p.d.f./p.m.f. $f(y \mid t)$. Suppose that

$$E = \left[\frac{\mathrm{d}}{\mathrm{d}t} \log\{f(Y \mid t)\}\right] = E\{g'(Y \mid t)\} = 0$$

and

$$E\{g'(Y\,|\,t)\}^2 = -E\left[\frac{\mathrm{d}^2}{\mathrm{d}t^2}\log\{f(Y\,|\,t)\}\right] = -E\{g''(Y\,|\,t)\} = I(t),$$

where $I(t)$ is the usual Fisher information, which is assumed to be finite, positive and continuous for all $t \in J$. Further, for any $t \in J$, we will assume the existence of a $\delta > 0$ and a pair of nonnegative random variables $K_1(Y\,|\,t)$, $K_2(Y\,|\,t)$ satisfying $E\{K_1(Y\,|\,t)\}^2 < \infty$ and $E\{K_2(Y\,|\,t)\} < \infty$ such that

$$|g''(Y\,|\,s)| = \left|\frac{\mathrm{d}^2}{\mathrm{d}s^2}\log\{f(Y\,|\,s)\}\right| \leqslant K_1(Y\,|\,t)$$

and

$$|g'''(Y\,|\,s)| = \left|\frac{\mathrm{d}^3}{\mathrm{d}s^3}\log\{f(Y\,|\,s)\}\right| \leqslant K_2(Y\,|\,t)$$

for all $s \in (t - \delta, t + \delta) \subseteq J$. Clearly, these standard Cramér-type conditions will be satisfied for all standard models frequently used in practice including models in exponential families.

Next, we impose some conditions on the weight functions $W_{n,i}$'s that are assumed to depend only on the X_i's at this point. For any x in the domain of θ, we will assume that

$$\sum_{i=1}^{n} \{W_{n,i}(x)\}^2 \to 0 \quad \text{in probability as } n \to \infty.$$

Also, it will be assumed that there is a sequence $\{\delta_n\}$ (random or deterministic) such that $\delta_n > 0$ for all $n \geqslant 1$, δ_n tends to zero in probability as n goes to infinity, and

$$\lim_{n\to\infty} \Pr\left\{\max_{1 \leqslant i \leqslant n;\, |X_i - x| > \delta_n} W_{n,i}(x) = 0\right\} = 1.$$

Stone (1977) gave a set of sufficient conditions on weight functions for the consistency of usual nonparametric regression, where one tries to estimate the conditional mean. Our conditions are very closely related to his conditions. For weights arising from any compactly supported suitable kernel function, it is quite easy to verify that both the conditions will hold whenever the bandwidth b_n (say) satisfies $b_n \to 0$ and $nb_n^d \to \infty$ as $n \to \infty$ (here d is the dimension of x). On the other hand, for a nearest-neighbor-type approach, those two conditions on weight functions will hold provided that the number of nearest neighbors of x grows to infinity while the diameter of the set covering those neighbors tends to zero as the sample size increases. Further, it is straightforward to verify that those conditions can be made to satisfy by choosing the weight functions appropriately whenever the regressors are random with an absolutely continuous distribution having a density that remains bounded away from zero and infinity in a neighborhood of x. Alternatively, the regressors can be chosen in an appropriate deterministic way (e.g. they can be evenly distributed over a compact regressor space) so that both the conditions will hold.

3.1. Main results on the behavior of $\hat{\theta}_n$

With the assumptions on the model and the weight functions in hand, we are now ready to state our first Theorem.

Theorem 3.1. *Suppose that the regularity conditions assumed on $f(y\,|\,t)$ and the conditions imposed on $W_{n,i}$'s ($1 \leqslant i \leqslant n$) at the beginning of the section hold. Further, assume that $\theta(x)$ is continuous in x. Then there exists a root $\hat{\theta}_n(x)$ of the estimating equation* (2.1) (*see Section* 2), *which will be a maximizer of our weighted likelihood and a consistent estimate for $\theta(x)$.*

Proof. First note that Eq. (2.1) can be restated as

$$\sum_{i=1}^{n} g'(Y_i \mid t) W_{n,i}(x) = 0.$$

For any fixed $\varepsilon > 0$ that is sufficiently small, we have the Taylor expansion

$$\sum_{i=1}^{n} g'\{Y_i \mid \theta(x) + \varepsilon\} W_{n,i}(x)$$

$$= \sum_{i=1}^{n} g'\{Y_i \mid \theta(X_i)\} W_{n,i}(x) + \sum_{i=1}^{n} \{\theta(x) + \varepsilon - \theta(X_i)\} g''\{Y_i \mid \xi_i(x)\} W_{n,i}(x),$$

where $\xi_i(x)$ lies between $\theta(x) + \varepsilon$ and $\theta(X_i)$. In view of the conditions imposed on $f(y \mid t)$ and the weight functions, the first term in the preceding expansion has zero conditional mean given all of the X_i's ($1 \leqslant i \leqslant n$), and its conditional variance tends to zero as n tends to infinity. The continuity of θ and the conditions imposed on the weight functions and g'' imply that

$$\left| \sum_{i=1}^{n} \{\theta(x) - \theta(X_i)\} g''\{Y_i \mid \xi_i(x)\} W_{n,i}(x) \right|$$

$$\leqslant \left\{ \max_{1 \leqslant i \leqslant n;\, |X_i - x| \leqslant \delta_n} |\theta(x) - \theta(X_i)| \right\} \sum_{i=1}^{n} |g''\{Y_i \mid \xi_i(x)\}| W_{n,i}(x) \to 0 \quad \text{in probability as } n \to \infty.$$

On the other hand, we can write

$$\sum_{i=1}^{n} \varepsilon g''\{Y_i \mid \xi_i(x)\} W_{n,i}(x) = \sum_{i=1}^{n} \varepsilon [g''\{Y_i \mid \theta(X_i)\} + I\{\theta(X_i)\}] W_{n,i}(x)$$

$$- \sum_{i=1}^{n} I\{\theta(X_i)\} W_{n,i}(x) + \sum_{i=1}^{n} \varepsilon \{\xi_i(x) - \theta(X_i)\} g'''\{Y_i \mid \psi_i(x)\} W_{n,i}(x),$$

where $\psi_i(x)$ lies between $\theta(X_i)$ and $\xi_i(x)$. It is straightforward to verify using the conditions imposed on g'' and the weight functions that the first term on the right-hand side of the above equation tends to zero in probability as n tends to infinity. Also, since I has been assumed to be a continuous and positive function, the sum $\sum_{i=1}^{n} I\{\theta(X_i)\} W_{n,i}(x)$ must remain positive and bounded away from zero in probability as n tends to infinity. Finally, the assumptions made on g''' imply that

$$\left| \sum_{i=1}^{n} \varepsilon \{\xi_i(x) - \theta(X_i)\} g'''\{Y_i \mid \psi_i(x)\} W_{n,i}(x) \right|$$

$$\leqslant \left\{ \max_{1 \leqslant i \leqslant n;\, |X_i - x| \leqslant \delta_n} |\xi_i(x) - \theta(X_i)| \right\} \sum_{i=1}^{n} |g'''\{Y_i \mid \psi_i(x)\}| W_{n,i}(x) \to 0 \quad \text{in probability as } n \to \infty.$$

Combining all of these observations, we now have

$$\lim_{n \to \infty} \Pr\left[\sum_{i=1}^{n} g'\{Y_i \mid \theta(x) + \varepsilon\} W_{n,i}(x) < 0 \right] = 1.$$

Arguing along the same line via Taylor expansion of $\sum_{i=1}^{n} g'\{Y_i \mid \theta(x) - \varepsilon\} W_{n,i}(x)$, one can show that

$$\lim_{n\to\infty} \Pr\left[\sum_{i=1}^{n} g'\{Y_i \mid \theta(x) - \varepsilon\} W_{n,i}(x) > 0 \right] = 1.$$

Therefore, as n tends to infinity, Eq. (2.1) will have a root lying between $\theta(x) - \varepsilon$ and $\theta(x) + \varepsilon$ with probability tending to one as the sample size grows to infinity. Since this is true for any given $\varepsilon > 0$, the Theorem is now established. □

Our preceding Theorem guarantees the existence of at least one solution of (2.1) that is consistent. In some situations, Eq. (2.1) may have multiple roots (e.g. when our weighted likelihood has multiple maxima). However, for models that belong to exponential families, the log-concavity of the weighted likelihood in large samples guarantees unique solution of (2.1). From now on, we will assume that $\hat{\theta}_n(x)$ is a consistent solution of (2.1). Then we have the following simple Taylor expansion:

$$\sum_{i=1}^{n} g'\{Y_i \mid \theta(X_i)\} W_{n,i}(x) = \sum_{i=1}^{n} \{\theta(X_i) - \hat{\theta}_n(x)\} g''\{Y_i \mid \eta_i(x)\} W_{n,i}(x),$$

where $\eta_i(x)$ lies between $\theta(X_i)$ and $\hat{\theta}_n(x)$. Assuming that (see (a) in the proof of Theorem 3.2 that follows) $\sum_{i=1}^{n} g''\{Y_i \mid \eta_i(x)\} W_{n,i}(x) \neq 0$, we can rewrite the preceding equation as

$$\hat{\theta}_n(x) - \theta(x) = \frac{\sum_{i=1}^{n} \{\theta(X_i) - \theta(x)\} g''\{Y_i \mid \eta_i(x)\} W_{n,i}(x)}{\sum_{i=1}^{n} g''\{Y_i \mid \eta_i(x)\} W_{n,i}(x)} - \frac{\sum_{i=1}^{n} g'\{Y_i \mid \theta(X_i)\} W_{n,i}(x)}{\sum_{i=1}^{n} g''\{Y_i \mid \eta_i(x)\} W_{n,i}(x)}.$$

Let us denote the first term in the above decomposition by $B_n(x)$ and the second term by $V_n(x)$. In view of the arguments used in the proof of Theorem 3.1, it is now obvious that $B_n(x)$ converges to zero in probability as n tends to infinity whenever our previous conditions on the model and the weight functions hold. In fact, the asymptotic behavior of $B_n(x)$ depends mainly on the behavior of θ in a neighborhood of x, and we have assumed θ to be a continuous function in the statement of Theorem 3.1. On the other hand, we have the following Theorem that describes the limiting behavior of $V_n(x)$.

Theorem 3.2. *Suppose that all the conditions assumed in Theorem* 3.1 *hold, and we have*

$$\frac{\max_{1\leqslant i\leqslant n} W_{n,i}(x)}{[\sum_{i=1}^{n} \{W_{n,i}(x)\}^2]^{1/2}} \to 0 \quad \textit{in probability as } n \to \infty.$$

Assume further that there is a $\rho > 0$ *such that* $\sup_{t\in J} E\{g'(Y \mid t)\}^{2+\rho} < \infty$, *where* Y *is a random variable having* $f(Y \mid t)$ *as the p.d.f./p.m.f. as before. Define* $\{\sigma_n(x)\}^2 = [I\{\theta(x)\}]^{-1} \sum_{i=1}^{n} \{W_{n,i}(x)\}^2$, *where recall that* $I\{\theta(x)\}$ *is the Fisher information associated with the model* $f\{y \mid \theta(x)\}$. *Then* $\{\sigma_n(x)\}^{-1} V_n(x)$ *converges weakly to a standard normal random variable as n tends to infinity.*

Proof. It is easy to see that the conditions assumed in Theorem 3.1 yield the following:

(a) The sum $\sum_{i=1}^{n} g''\{Y_i \mid \eta_i(x)\} W_{n,i}(x)$ converges to $-I\{\theta(x)\}$ in probability as n tends to infinity in view of the continuity of θ and I, the conditions imposed on g'' and the weight functions, and some of the arguments used in the proof of Theorem 3.1.

(b) Let $\alpha_n(x)$ be the ratio defined as

$$\alpha_n(x) = \frac{\sum_{i=1}^n I\{\theta(X_i)\}\{W_{n,i}(x)\}^2}{I\{\theta(x)\}\sum_{i=1}^n \{W_{n,i}(x)\}^2}.$$

Then the continuity of θ and I together with one of the conditions assumed on the weight functions will imply that $\alpha_n(x)$ tends to one in probability as n tends to infinity.

(c) Given all of the X_i's $(1 \leqslant i \leqslant n)$, the conditional mean of the sum of independent random variables $\sum_{i=1}^n g'\{Y_i | \theta(X_i)\} W_{n,i}(x)$ is zero, and its conditional variance is $\sum_{i=1}^n I\{\theta(X_i)\}\{W_{n,i}(x)\}^2$.

The proof of the Theorem is now complete using the observations made in (a)–(c) and an application of Lindeberg's central limit theorem exploiting the condition on weight functions and the moment condition on $g'(Y|t)$ assumed in the statement of the Theorem. □

As already mentioned, Staniswalis (1989) investigated a kernel-based approach to estimate a function parameter nonparametrically using the likelihood and briefly (somewhat casually) discussed the asymptotic properties of constructed estimates. Such a kernel smoothing technique is a special case of the general weighted maximum likelihood approach considered here. However, though the approach here is very general, the conditions imposed to derive the asymptotic results are neither very strong nor un-natural, and we have tried to state the conditions in a way so that they become quite easy to comprehend, verify and implement in specific situations.

3.2. Likelihood-based cross-validation: some heuristics

So far we have investigated the asymptotic behavior of $\hat{\theta}_n$ by imposing conditions on the weight functions, which were assumed to be functions of the X_i's only without considering a data-based adaptive selection of the smoothing parameter. However, the practical implementation of the procedure will involve selection of the smoothing parameter by maximizing the cross-validation function described in (2.2) (see Section 2), and it is quite relevant to explore the asymptotic properties of this likelihood-based cross-validation criterion. Using the regularity conditions assumed on the model $f(y|t)$ and a second-order Taylor expansion ignoring the remainder term, we can write

$$\begin{aligned} MLCV(h_n) &= \sum_{i=1}^n \log[f\{Y_i | \hat{\theta}_n^{(i)}(X_i)\}] = \sum_{i=1}^n g\{Y_i | \hat{\theta}_n^{(i)}(X_i)\} \\ &\approx \sum_{i=1}^n g\{Y_i | \theta(X_i)\} + \sum_{i=1}^n \{\hat{\theta}_n^{(i)}(X_i) - \theta(X_i)\} g'\{Y_i | \theta(X_i)\} \\ &\quad + \sum_{i=1}^n \{\hat{\theta}_n^{(i)}(X_i) - \theta(X_i)\}^2 g''\{Y_i | \theta(X_i)\}. \end{aligned}$$

Clearly, approximating $MLCV(h_n)$ by such an asymptotic expansion is meaningful provided that the estimate $\hat{\theta}_n^{(i)}(X_i)$ is close to $\theta(X_i)$ for each i. The first term in this approximating expansion is completely free from h_n. Also, since $\hat{\theta}_n^{(i)}(X_i)$ is the leave-one-out estimate of $\theta(X_i)$ based on $(Y_1, X_1), \ldots, (Y_{i-1}, X_{i-1}), (Y_{i+1}, X_{i+1}), \ldots, (Y_n, X_n)$, the second term in the expansion has zero expectation, and the third term has expectation

$$-E\left[\sum_{i=1}^n \{\hat{\theta}_n^{(i)}(X_i) - \theta(X_i)\}^2 I\{\theta(X_i)\}\right]$$

assuming that all the expectations exist finitely. This indicates that the strategy of choosing h_n by maximizing $MLCV(h_n)$ will asymptotically yield a value of h_n, which will be an approximate minimizer of the weighted sum of squares

$$\sum_{i=1}^{n} \{\hat{\theta}_n^{(i)}(X_i) - \theta(X_i)\}^2 I\{\theta(X_i)\}.$$

The appearance of the Fisher information as the weight function in the above weighted sum of squares is a very desirable and noteworthy feature in view of Theorem 3.2.

Brillinger (1977, 1986) mentioned about "conditional M-estimates" and Stone (1977) briefly discussed them in a very general and abstract set up. It is not difficult to observe that our weighted maximum likelihood estimates can be viewed as special cases of these "conditional M-estimates". However, neither Brillinger (1977, 1986) nor Stone (1977) indicated how to determine the appropriate degree of smoothing for such estimates and what kind of cross-validation can possibly be used. Staniswalis (1989) and Firth et al. (1991) used likelihood-based leave-one-out cross-validation to select the smoothing parameter associated with their kernel smoothers. But none of them provided any theoretical justification for using the likelihood-based leave-one-out cross-validation. While we have not undertaken formal analytic investigations into such cross-validation in this note, the observations and heuristics presented in this section are quite promising and provide valuable insights.

Acknowledgement

The authors are thankful to an anonymous referee for a prompt and careful reading of an earlier draft of the manuscript. Comments from the referee were helpful in preparing the revision.

References

Brillinger, D.R. (1977), Comment on paper by C.J. Stone, *Ann. Statist.* **5**, 622–623.

Brillinger, D.R. (1986), Comment on paper by T. Hastie and R. Tibshirani, *Statist. Sci.* **1**, 310–312.

Chaudhuri, P. and A. Dewanji (1991), Likelihood based nonparametrics: kernel smoothing and cross-validation in generalized regression, Technical Report No. 22/91, Division of Theoretical Statistics & Mathematics, Indian Statistical Institute, Calcutta.

Cheng, K.F. and P.E. Lin (1981), Nonparametric estimation of a regression function, *Z. Wahrsch. verw. Gebiete* **57**, 223–233.

Cox, D.D. and F. O'Sullivan (1990), Asymptotic analysis of penalized likelihood and related estimators, *Ann. Statist.* **18**, 1676–1695.

Eubank, R.L. (1988), *Spline Smoothing and Nonparametric Regression* (Marcel Dekker, New York).

Firth, D., J. Glosup and D.V. Hinkley (1991), Model checking with nonparametric curves, *Biometrika* **78**, 245–252.

Gasser, T. and H.G. Muller (1979), Kernel estimation of regression functions, in: T. Gasser and M. Rosenblatt, eds., *Smoothing Techniques for Curves Estimation* (Springer, Heidelberg), pp. 23–68.

Gasser, T. and H.G. Muller (1984), Estimating regression functions and their derivatives by the kernel method, *Scand. J. Statist.* **11**, 171–185.

Gu, C. (1990), Adaptive spline smoothing in non-Gaussian regression models, *J. Amer. Statist. Assoc.* **85**, 801–807.

Gu, C. (1992), Cross-validating non-Gaussian data, *J. Comput. Graphical Statist.* **1**, 169–179.

Hardle, W. and J.S. Marron (1985a), Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression, *Biometrika* **72**, 481–484.

Hardle, W. and J.S. Marron (1985b), Optimal bandwidth selection in nonparametric regression function estimation, *Ann. Statist.* **13**, 1465–1482.

Hastie, T. and R. Tibshirani (1986), Generalized additive models (with discussion) *Statist. Sci.* **1**, 297–318.

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

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

Nadaraya, E.A. (1964), On estimating regression, *Theory Probab. Appl.* **9**, 141–142.

O'Sullivan, F., B.S. Yandell and W.J. Raynor (1986), Automatic smoothing of regression functions in generalized linear models, *J. Amer. Statist. Assoc.* **81**, 96–103.

Priestley, M.B. and M.T. Chao (1972), Nonparametric function fitting, *J. Roy. Statist. Soc.* Ser. B **34**, 384–392.

Staniswalis, J.G. (1989), The kernel estimate of a regression function in likelihood based models, *J. Amer. Statist. Assoc.* **84**, 276–283.

Stone, C.J. (1977), Consistent nonparametric regression (with discussion), *Ann. Statist.* **5**, 595–645.

Stone, C.J. (1986), The dimensionality reduction principle for generalized additive models, *Ann. Statist.* **14**, 590–606.

Stone, M. (1974), Crossvalidatory choice and assessment of statistical predictions, *J. Roy. Statist. Soc.* Ser. B **36**, 111–123.

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

Watson, G.S. (1964), Smooth regression analysis, *Sankhyā* Ser. A **26**, 359–372.