Methods for Estimating a Conditional Distribution Function

Peter HALL, Rodney C. L. WOLFF, and Qiwei YAO

Motivated by the problem of setting prediction intervals in time series analysis, we suggest two new methods for conditional distribution estimation. The first method is based on locally fitting a logistic model and is in the spirit of recent work on locally parametric techniques in density estimation. It produces distribution estimators that may be of arbitrarily high order but nevertheless always lie between 0 and 1. The second method involves an adjusted form of the Nadaraya–Watson estimator. It preserves the bias and variance properties of a class of second-order estimators introduced by Yu and Jones but has the added advantage of always being a distribution itself. Our methods also have application outside the time series setting; for example, to quantile estimation for independent data. This problem motivated the work of Yu and Jones.

KEY WORDS: Absolutely regular; Bandwidth; Biased bootstrap; Conditional distribution; Kernel methods; Local linear methods; Local logistic methods; Nadaraya–Watson estimator; Prediction; Quantile estimation; Time series analysis; Weighted bootstrap.

1. INTRODUCTION

In various statistical problems, estimating a conditional distribution function is a key aspect of inference. Consider, for example, estimating the quantile function of Y given $\mathbf{X}$, using a sample of independent data pairs $\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$. This problem was recently tackled by Yu and Jones (1998) via an ingenious application of the "double-kernel," local linear approach of Fan, Yao, and Tong (1996). A technique alternative to that of Yu and Jones would be to use the Nadaraya–Watson estimator, although this would suffer the excessive bias problems inherent in that approach (see, e.g., Chu and Marron 1991 and Fan 1993).

Another application of conditional distribution estimation, this time involving dependent data, is construction of prediction intervals for the next value in a stationary time series $\{Y_1, \ldots, Y_n\}$. If the time series is Markovian, then we may solve the prediction problem by estimating the distribution of Y_{i+1}, conditional on $Y_i = x$, and applying the result in the case where $x = Y_n$. More generally, we might wish to estimate the distribution of Y_{i+1} given a small section of the recent past, such as $(Y_i, Y_{i-1}) = (x_1, x_2)$. Both problems may be solved using methods such as those suggested by Yu and Jones (1998).

Although local linear methods of the Yu and Jones type are attractive from the viewpoint of mathematical efficiency (see, e.g., Fan 1993), they have the disadvantage of producing distribution function estimators that are not constrained either to lie between 0 and 1 or to be monotone increasing. In both these respects, Nadaraya–Watson methods are superior, despite their rather large bias. Moreover, if one passes to higher-order generalizations of the Yu and Jones approach, such as methods based on local polynomial techniques (see, e.g., Fan and Gijbels 1996), then the "nondistribution properties" from which they suffer become even more of a problem.

In this article we suggest two new techniques that largely overcome these difficulties. The first, local logistic distribution estimation, produces estimators of arbitrarily high order that always lie strictly between 0 and 1. In spirit, this approach is related to recently introduced local parametric methods for density estimation (see, e.g., Copas 1995, Hjort and Jones 1996, Loader 1996, and Simonoff 1996, sec. 3.4). Our second method is an "adjusted" version of the Nadaraya–Watson estimator. It is designed to reproduce the superior bias properties of local linear methods, while preserving the property that the Nadaraya–Watson estimator is always a distribution function. It is based on weighted, or biased, bootstrap methods (see Barbe and Bertail 1995; Hall and Presnell 1997).

The properties of positivity and monotonicity are particularly advantageous if one is going to invert a conditional distribution estimator to produce an estimator of a conditional quantile. [Early work in the latter area includes that of Bhattacharya and Gangopadhyay (1990), Jones and Hall (1990), and Sheather and Marron (1990).] Difficulties in computing quantile estimators based on distribution estimators that are not themselves distribution functions have been discussed by Yao (1995). These include existence of multiple solutions. More generally, problems inverting distribution estimators that are not monotone or constrained to lie in the interval $[0, 1]$ have been addressed by Chu and Cheng (1995).

Although our interest in conditional distribution estimation was motivated by the problem of prediction from time series data, we introduce our methods in a more general setting that admits time series modeling as a special case. Our theoretical results also focus on the general context.

The article is organized as follows. In Section 2 we introduce our methods for estimation of a conditional distribution, and present a bootstrap scheme for choosing band-

Peter Hall is Professor, Centre for Mathematics and Its Applications, Australian National University, Canberra, ACT 0200, Australia (E-mail: *peter.hall@anu.edu.au*). Rodney C. L. Wolff is Senior Lecturer, School of Mathematics, Queensland University of Technology, Brisbane, Queensland 4001, Australia (E-mail: *r.wolff@fsc.qut.edu.au*). Qiwei Yao is Senior Lecturer in Statistics, Institute of Mathematics and Statistics, University of Kent at Canterbury, Kent CT2 7NF, UK (E-mail: *q.yao@ukc.ac.uk*). The helpful comments of the editor and a reviewer are gratefully acknowledged.

© 1999 American Statistical Association
Journal of the American Statistical Association
March 1999, Vol. 94, No. 445, Theory and Methods

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

widths. We provide numerical examples involving both simulated models and real-data applications in Section 3, including a case study with a multivariate predictor. In Section 4 we describe convergence rates, as well as asymptotic distribution properties of the proposed estimators. We relegate all technical arguments to an Appendix.

2. METHODOLOGY

We assume that data are available in the form of a strictly stationary stochastic process $\{(\mathbf{X}_i, Y_i)\}$, where Y_i is a scalar and $\mathbf{X}_i$ is a d-dimensional vector. Naturally, this includes the case where the pairs $(\mathbf{X}_i, Y_i)$ are independent and identically distributed. We wish to estimate the conditional distribution function $\pi(y|x) \equiv P(Y_i \le y|\mathbf{X}_i = \mathbf{x})$. In the time series context, $\mathbf{X}_i$ typically denotes a vector of lagged values of Y_i, in which case $\pi(\cdot|x)$ is the predictive distribution of Y_i given $\mathbf{X}_i = \mathbf{x}$. If we write $Z_i = I(Y_i \le y)$, then $E(Z_i|\mathbf{X}_i = \mathbf{x}) = \pi(y|\mathbf{x})$, and so our estimation problem may be viewed as regression of Z_i on $\mathbf{X}_i$.

To simplify discussion, here we introduce our methods and develop theory only in the case where X_i is a scalar (i.e., $d = 1$). We illustrate the multivariate case in Section 3.

2.1 Local Logistic Methods

For fixed y, write $P(x) = \pi(y|x)$ and assume that P has $r - 1$ continuous derivatives. A generalized local logistic model for $P(x)$ has the form $L(x, \boldsymbol{\theta}) \equiv A(x, \boldsymbol{\theta})/\{1 + A(x, \boldsymbol{\theta})\}$, where $A(\cdot, \boldsymbol{\theta})$ denotes a nonnegative function that depends on a vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$ that "represent" the values of $P(x), P^{(1)}(x), \ldots, P^{(r-1)}(x)$. Here, "represent" means that for each sequence $\omega_1 \in (0, 1), \omega_2, \ldots, \omega_r$ denoting potential values of $P(x), P^{(1)}(x), \ldots, P^{(r-1)}(x)$, there exist $\theta_1, \ldots, \theta_r$ such that

$$\frac{A(u, \boldsymbol{\theta})}{1 + A(u, \boldsymbol{\theta})} = \omega_1 + \omega_2(u - x) + \cdots + (r!)^{-1}\omega_r(u - x)^{r-1} + o(|u - x|^{r-1})$$

as $u \to x$. Arguably the simplest function A to work with is $A(u, \boldsymbol{\theta}) \equiv e^{p(u,\boldsymbol{\theta})}$, where $p(u, \boldsymbol{\theta}) = \theta_1 + \theta_2 u + \cdots + \theta_r u^{r-1}$ is a polynomial of degree $r - 1$. Fitting this model locally to indicator function data leads to an estimator $\hat{\pi}(y|x) \equiv L(0, \hat{\boldsymbol{\theta}}_{xy})$, where $\hat{\boldsymbol{\theta}}_{xy}$ denotes the minimizer of

$$R(\boldsymbol{\theta}; x, y) = \sum_{i=1}^{n} \{I(Y_i \le y) - L(X_i - x, \boldsymbol{\theta})\}^2 K_h(X_i - x), \quad (1)$$

K is a kernel function, $K_h(\cdot) = h^{-1}K(\cdot/h)$, and $h > 0$ is a bandwidth. We call this approach *local logistic distribution estimation.* Depending on bandwidth choice, this approach also furnishes consistent estimators of the derivatives $\pi^{(i)}(y|x) \equiv (\partial/\partial x)^i \pi(y|x)$, in the form $\hat{\pi}^{(i)}(y|x) = L^{(i)}(0, \hat{\boldsymbol{\theta}}_{xy})$ for $i = 1, \ldots, r - 1$, where $L^{(i)}(x, \boldsymbol{\theta}) \equiv (\partial/\partial x)^i L(x, \boldsymbol{\theta})$. In practice, $\hat{\boldsymbol{\theta}}_{xy}$ may be computed using the "downhill simplex" algorithm (see Press, Teukolsky, Vetterling, and Flannery 1992, sec. 10.2).

We expect the estimator $\hat{\pi}(y|x)$ to have bias of order h^r and variance of order $(nh)^{-1}$, under an asymptotic scheme where $h = h(n) \to 0$ and $nh \to \infty$ as $n \to \infty$. We give a more detailed account of this property in Section 4.

It is possible to fit the logistic model by matrix-weighted least squares in place of the criterion at (1), reflecting the dependence structure of the process $\{X_i\}$. However, if the process is weakly dependent (e.g., absolutely regular, as assumed in Sec. 4), then this has only a second-order effect on performance.

Our use of the logistic model here is reminiscent of methods used in connection with binary choice models. There one observes response variables Y_i that may take only the values 0 and 1, with $P(Y_i = 1|X_i = x) = G\{m(x)\}$, where G is a known link function and nonparametric or semiparametric assumptions are made about m. (See, e.g., Ahn 1995, Ahn and Manski 1993, Klein 1993, and Matzkin 1992 for work in econometrics and Chu and Cheng 1995 for discussion of binary choice from a more statistical viewpoint.) Because we do not assume that $\pi(y|x)$ is continuous in y, then, in cases where G is the identity and only smoothness assumptions are made of m, the methods here and in the next section may be viewed as a means of estimating probabilities in binary choice models, in the context of Chu and Cheng (1995).

2.2 Adjusted Nadaraya–Watson Estimator

Let $p_i = p_i(x)$, for $1 \le i \le n$, denote weights (functions of the data $X_1, \ldots, X_n$, as well as of x) with the property that each $p_i \ge 0, \sum_i p_i = 1$, and

$$\sum_{i=1}^{n} p_i(x)(X_i - x)K_h(X_i - x) = 0. \quad (2)$$

Of course, p_i's satisfying these conditions are not uniquely defined, and we specify them concisely by asking that $\prod_i p_i$ be as large as possible subject to the constraints. Define

$$\tilde{\pi}(y|x) = \frac{\sum_{i=1}^{n} I(Y_i \le y)p_i(x)K_h(X_i - x)}{\sum_{i=1}^{n} p_i(x)K_h(X_i - x)}. \quad (3)$$

Note particularly that $0 \le \tilde{\pi}(y|x) \le 1$ and $\tilde{\pi}$ is monotone in y. We show in Section 4 that $\tilde{\pi}$ is first-order equivalent to a local linear estimator, which does not enjoy either of these properties.

It is possible for the estimator at (3) to assume the form 0/0, due to no datum X_i lying in the local window. More generally, sparsity of design points can sometimes be a difficulty. However, this is generally somewhat less of a problem for the modified Nadaraya–Watson estimator than for local linear estimators in the same setting, in that sparsity tends to produce small cusps in the estimator $\tilde{\pi}$, rather than the large fluctuations associated with design sparsity for local polynomials. Problems with data sparseness for either $\hat{\pi}$ or $\tilde{\pi}$ may be overcome by locally increasing the bandwidth or by imputing new design points via interpolation. [See Hall and Turlach (1997) and Seifert and Gasser (1996a,b) for discussion of these approaches in more conventional problems involving nonparametric regression.]

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Another way to view the biased bootstrap estimator $\tilde{\pi}$ is as a local linear estimator in which the weights for the least squares step are taken to be $p_i(x)K_h(X_i - x)$, rather than simply $K_h(X_i - x)$, for $1 \le i \le n$. To appreciate why this is so, we refer to the definition of general local linear estimators given by Fan and Gijbels (1996, p. 20) and note that in view of (2), with the suggested change of weights, their estimator $\hat{m}_0$ reduces to

$$\hat{m}_0(x) = \left\{\sum_{i=1}^{n} w_i(x)I(Y_i \le y)\right\} \Big/ \left\{\sum_{i=1}^{n} w_i(x)\right\},$$

where

$$w_i(x) = p_i(x)K_h(X_i - x)\sum_{j=1}^{n} p_j(x)(x - X_j)^2 K_h(x - X_j).$$

Therefore, $\hat{m}_0 = \tilde{\pi}(y|x)$.

Computation of the p_i's is simplified by the fact that

$$p_i(x) = n^{-1}\{1 + \lambda(x - X_i)K_h(X_i - x)\}^{-1},$$

where λ (a function of the data and of x) is uniquely defined by (2). It is easily computed using a Newton–Raphson argument. Condition (2) ensures that $\sum_i p_i = 1$.

2.3 Bandwidth Choice

Particularly in the time series case, deriving asymptotically optimal bandwidths for either the local logistic or the biased bootstrap methods is a tedious matter. Using plug-in methods requires explicit estimation of complex functions using dependent data; using the bootstrap calls for selection of subsidiary smoothing parameters and resampling of time series data; and using cross-validation demands selection of the amount of data that are left out. Such complexity is arguably not justified, not in the least because the target function $P(x) = \pi(y|x)$ is often approximately monotone and so has only limited opportunity for complex behavior. For example, P is exactly monotone if the joint distribution of (X_i, Y_i) is normal.

Instead, we suggest an approximate parametric method, as follows. We fit a parametric model

$$Y_i = a_0 + a_1 X_i + \cdots + a_k X_i^k + \sigma\varepsilon_i,$$

where ε_i is standard normal, $a_0, \ldots, a_k, \sigma$ are estimated from the data, and k is determined by the Akaike information criterion (AIC). We form a parametric estimator $\hat{\pi}(y|x)$ based on the model. By Monte Carlo simulation from the model, we compute a bootstrap version of $\{Y_1^*, \ldots, Y_n^*\}$ based on given observations $\{X_1, \ldots, X_n\}$, and thence a bootstrap version $\hat{\pi}_h^*(y|x) = \hat{\pi}^*(y|x)$ of $\hat{\pi}(y|x)$, derived from (1) with $\{(X_i, Y_i)\}$ replaced by $\{(X_i, Y_i^*)\}$. The bootstrap estimator of the absolute deviation error of $\hat{\pi}(y|x)$ is

$$M(h; x, y) = E[|\hat{\pi}_h^*(y|x) - \hat{\pi}(y|x)| | \{(X_i, Y_i)\}].$$

Choose $h = \hat{h}(x, y)$ to minimize $M(h; x, y)$. Sometimes we use the x-dependent bandwidth $\hat{h}(x)$, which minimizes

$$M(h; x) = \int M(h; x, y)\hat{\pi}(y|x)\,dy. \tag{4}$$

The foregoing approach can also be applied to choosing h for estimating $\tilde{\pi}(y|x)$.

In the event that we are working with time series data (e.g., $X_i = Y_{i-m}$ for some $m \ge 1$), we propose an alternative resampling scheme as follows. Assume that the data $\{Y_{-m+1}, \ldots, Y_n\}$ represent a segment of a Gaussian autoregression. Estimate its parameters, and resample the segment $\{Y_{-m+1}^*, \ldots, Y_n^*\}$ from the parametric model. The bootstrap estimator $\pi_h^*(y|x)$ is calculated using this segment and then substituted into the foregoing formula for $M(h; x, y)$.

3. NUMERICAL PROPERTIES

3.1 Simulation Studies

We compared various estimators of the conditional distribution function $\pi(\cdot|\cdot)$ through two simulated models, one with independent observations and one with nonlinear time series. These estimators are the Nadaraya–Watson estimator (NW), the local linear regression estimator (LL), the adjusted Nadaraya–Watson estimator (ANW), and the local logistic estimators with $r = 2$ (LG-2) and $r = 3$ (LG-3). For each simulated sample, the performance of the estimator was evaluated in terms of mean absolute deviation error (MADE):

$$\text{MADE} = \frac{\sum_i |\pi_e(y_i|x_i) - \pi(y_i|x_i)|I(.001 \le \pi(y_i|x_i) \le .999)}{\sum_i I(.001 \le \pi(y_i|x_i) \le .999)},$$

where $\pi_e(\cdot|\cdot)$ denotes an estimator of $\pi(\cdot|\cdot)$, and $\{(x_i, y_i)\}$ are grid points that are specified later. We conducted the simulation in two stages. First, we calculated MADEs for the various estimators over grid points evenly distributed across the entire sample space. For each estimator, we used the optimal bandwidth defined by

$$h_{\text{op}}(x) = \int h_{\text{op}}(x, y)\pi(y|x)\,dy,$$

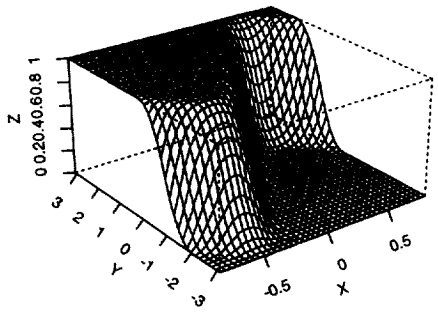
where $h_{\text{op}}(x, y)$ is the minimizer of the asymptotic mean squared error (up to first order) of the estimator. This guarantees a fair comparison among different methods. Second, we demonstrated the usefulness of the bootstrap scheme for choosing bandwidths proposed in Section 2.3 by evaluating MADEs for some fixed values of x. We used the x-dependent bandwidth $\hat{h}(x)$, which minimizes (4). Throughout this section we used the Gaussian kernel.

Example 1. We consider the simple model

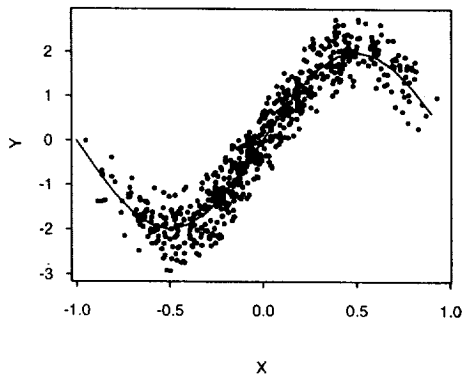
$$Y_i = 2\sin(3.1416X_i) + \varepsilon_i,$$

where $\{\mathbf{X}_i\}$ and $\{\varepsilon_i\}$ are two independent sequences of independent random variables having a common distribution with density $1 - |x|$ on $[-1, 1]$. The true conditional distribution function is plotted in Figure 1(a). For each of the 400 samples of size $n = 600$ [one of which is depicted in Figure 1(b)], we calculated the MADEs with the optimal bandwidth $h_{\text{op}}(\cdot)$ (which is of size $n^{-1/9}$ for LG-3 and $n^{-1/5}$ for all of the other methods). We estimated $\pi(y|x)$ on a regular grid defined by steps .067 and .054 in x- and y-directions. The boxplots of MADEs are presented in Figure 1(c). The

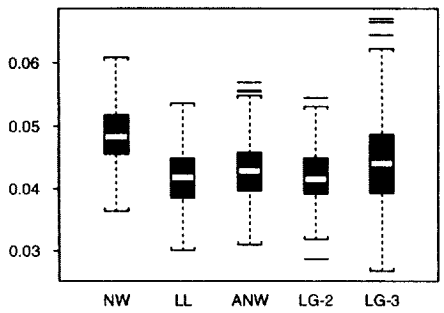
Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.



(a)



(b)



(c)

Figure 1. Simulation Results for Example 1: Estimating $\pi(y|x)$ Using Optimal Bandwidths. (a) The true conditional distribution function z = $\pi(y|x)$. (b) A typical sample of size n = 600 used in estimation, together with the curve y = 2 sin(3.1416x). (c) The boxplots of MADEs for the NW estimate, LL estimate, ANW estimate, LG-2 estimate, and LG-3 estimate.

variations of the MADEs for the NW, LL, ANW, and LG-2 methods are more or less the same, reflecting the fact that the (asymptotic) variances of those estimators are the same. Overall, both ANW and LG-2 provide competitive performance relative to the LL method. The larger MADE values for the NW method are due to its larger bias and poor boundary effect. The large variation of MADE for LG-3 is to be expected, as LG-3 has a larger asymptotic variance than the other estimators (see Fan and Gijbels 1996, remark 3, sec. 3.3.1). Note that its bias would be smaller than that of the others if a bandwidth of size $n^{-1/5}$ were used. For the sample size used in our simulation, and analogously to local polynomial regression, local logistic methods with $r > 2$ are not appealing for estimating the conditional distribution function itself.

For each of 200 random samples of size $n = 600$, we estimated $\pi(\cdot|x)$ using the bandwidth $\hat{h}(x)$ selected by the bootstrap scheme for $x = -.5, .1$, and $.7$. To this end, we fitted a parametric model for Y_i as a polynomial in X_i. In the 200 replications, the order determined by AIC was always 3. We replicated bootstrap resampling 40 times. We consider here only the ANW and LG-2 methods. For comparison, we calculated the estimates for the same data, using the optimal bandwidth $h_{\mathrm{op}}(x)$. For $x = -.5, .1$, and $.7$, $h_{\mathrm{op}}(x)$ is .062, .036, and .106 with the ANW method and .086, .055, and .085 with the LG-2 method respectively. Figure 2(a) presents boxplots of the differences of MADEs based on $\hat{h}(x)$ over the MADEs based on $h_{\mathrm{op}}(x)$. Figure 2(b) displays boxplots of $\hat{h}(x) - h_{\mathrm{op}}(x)$ in the simulation with 200 replications. The performance of estimates based on the bootstrap bandwidths is fairly consistent, although in most cases $\hat{h}(x)$ overestimates $h_{\mathrm{op}}(x)$. Figure 2(c)–(e) depict typical examples of the estimated conditional distribution functions $\hat{\pi}(\cdot|x)$ and $\tilde{\pi}(\cdot|x)$. The typical example was selected in such a way that the corresponding MADE was equal to its median in the simulation with 200 replications. Note that $\tilde{\pi}(\cdot|x)$ is monotonically increasing.

Example 2. Here we considered an AR(1) model,

$$Y_t = 3.76Y_{t-1} - 0.235Y_{t-1}^2 + 0.3\varepsilon_t, \tag{5}$$

where the errors ε_t were independent with common distribution $U[-\sqrt{3}, \sqrt{3}]$. We treated two- and three-step ahead prediction, by taking $X_t = Y_{t-m}$ for $m = 2$ and 3. For each of 400 samples of size $n = 600$, we calculated the MADEs with asymptotically optimal bandwidths on regular grid points with step .40 in the x-direction and steps .10 and .19 in the y-direction, for $m = 2$ and 3. Note that the conditional distributions no longer admit simple explicit forms. To calculate $h_{\mathrm{op}}(x)$, we evaluated the true values of $\pi(y|x)$ and its derivatives by simulation, as follows. We generated 50,000 random samples by iterating (5) two (or three) times, starting at a fixed value x. The relative frequency of the sample exceeding y was regarded as the true value of $\pi(y|x)$. The resulting conditional distribution functions are plotted in Figure 3, (a) and (b). We used kernel methods to estimate the marginal density function with a sample of size 100,000. Figure 3, (c) and (d), is the boxplots of MADEs for the 400 replications. Similar to Example 1, both the ANW and LG-2 methods provide competitive performance relative to the LL method, in terms of the absolute error of estimation, whereas the bias of the NW estimator is relatively larger.

For each of 200 random samples of size $n = 600$, we estimated the two-step-ahead predictive distribution $\pi(\cdot|x)$ using the bandwidth $\hat{h}(x)$ selected by the bootstrap scheme for $x = 4.99$ and 13.71. The bootstrap resampling was conducted as follows. We fitted a linear AR(1) model to the original data, and sampled time series (with length 600) from the fitted model. We replicated bootstrap sampling 40 times. As in the case of Example 1, we considered only the ANW estimator and the local logistic estimator with $r = 2$.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

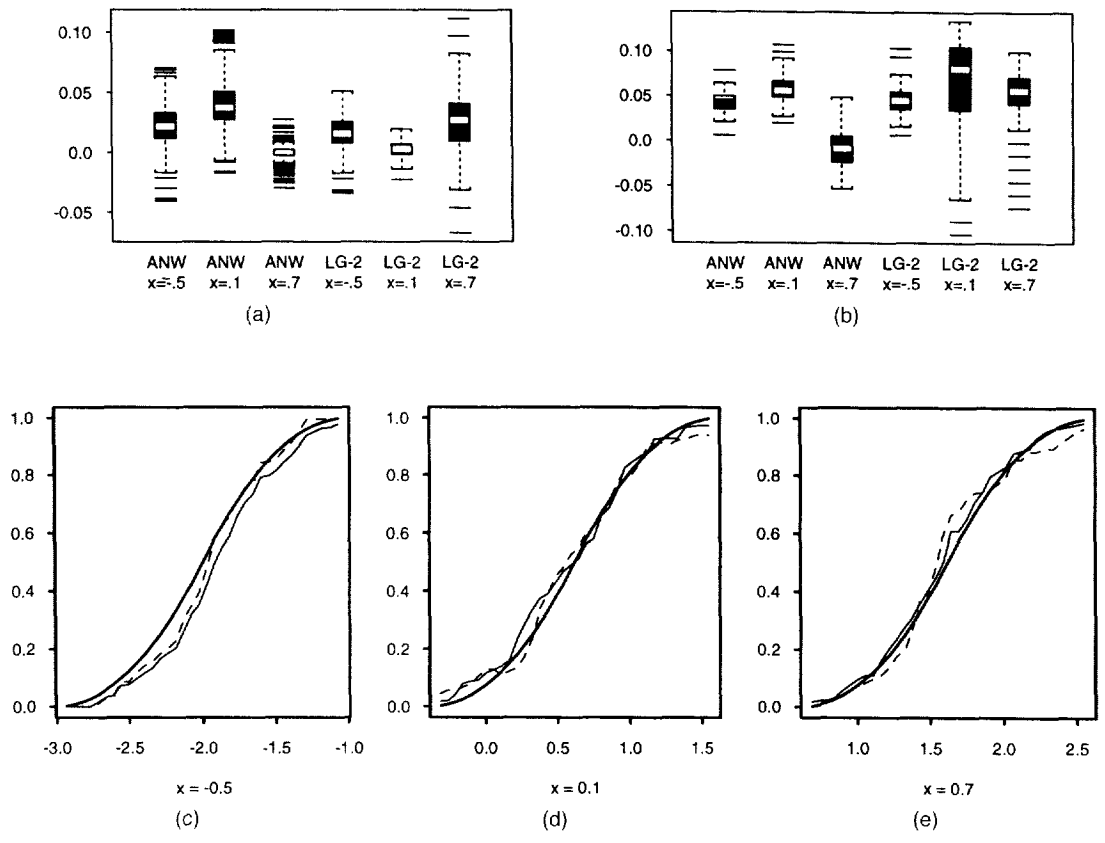


Figure 2. Simulation Results for Example 1: Estimating $\pi(y|x)$ With Bootstrap Bandwidths. (a) Boxplots of the MADEs based on $\hat{h}(x)$ minus the MADEs based on $h_{op}(x)$; (b) boxplots of $\hat{h}(x) - h_{op}(x)$; (c)–(e) curves representing conditional distribution functions $\pi(\cdot|x)$: thick line, $\pi(\cdot|x)$; thin line, ANW estimate, dashed line, LG-2 estimate.

We compared the estimates with those based on the optimal bandwidth $h_{\mathrm{op}}(x)$, which is equal to .182 for $x = 4.99$ and .216 for $x = 13.71$ in the case of the ANW estimate and equal to .241 for $x = 4.99$ and .168 for $x = 13.71$ in the case of the LG-2 estimate. Figure 4(a) presents boxplots of the differences of MADEs based on $\hat{h}(x)$ over the MADEs based on $h_{\mathrm{op}}(x)$. Figure 4(b) displays boxplots of $\hat{h}(x) - h_{\mathrm{op}}(x)$ in the simulation with 200 replications. Because we used a simple linear model to fit the nonlinear structure, it is not surprising that $\hat{h}(x)$ always overestimates $h_{\mathrm{op}}(x)$. But the estimates for $\pi(y|x)$ remain reasonably reliable. Figure 4, (c) and (d), depicts typical examples of the estimated conditional distribution functions $\hat{\pi}(\cdot|x)$ and $\tilde{\pi}(\cdot|x)$. The typical example was selected in such a way that the corresponding MADE was equal to its median in the simulation with 200 replications.

3.2 Case Study With Canadian Lynx Data

Finally, we illustrate our method with the Canadian lynx data (on a natural logarithmic scale) for the years 1821–1934. The time series data plot is presented in Figure 5(a). We estimated the conditional distribution of Y_t given Y_{t-1} by the ANW method. We selected the bandwidths using the bootstrap scheme based on resampling the entire time series from the best-fitted linear AR(1) model. We did 40 replications in the bootstrap resampling step. The estimated conditional distribution function is depicted in Figure 5(b).

As an alternative application, we constructed the predictive interval $[\pi^{-1}(\alpha/2|x), \pi^{-1}(1-\alpha/2|x)]$ for $\alpha \in (0,1)$, based on the estimated conditional distribution function. To check on performance, we used the data for 1821–1924 (i.e., $n = 104$) to estimate $\pi(y|x)$ and the last 10 data points to check the predicted values. This time we used the local logistic method with $r = 2$. The results with $\alpha = .1$ are reported in Table 1. All of the predictive intervals contain the corresponding true values. The average length of the intervals is 2.80, which is 53.9% of the dynamic range of the data.

Also included in Table 1 are the predictive intervals based on the estimated conditional distribution of Y_t given both Y_{t-1} and Y_{t-2}. To obtain these results, we used the local (linear) logistic method to estimate $\pi(y|x_1, x_2)$. To this end, let $L(x_1, x_2, \boldsymbol{\theta}) = A(x_1, x_2, \boldsymbol{\theta})/\{1 + A(x_1, x_2, \boldsymbol{\theta})\}$ with $A(x_1, x_2, \boldsymbol{\theta}) = \exp(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$. The estimator is defined as $\hat{\pi}(y|x_1, x_2) \equiv L(0, 0, \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ denotes the minimizer of

$$\sum_{t=3}^{n} \{I(Y_t \le y) - L(Y_{t-1} - x_1, Y_{t-2} - x_2, \boldsymbol{\theta})\}^2 \times K\left(\frac{Y_{t-1} - x_1}{h_1}, \frac{Y_{t-2} - x_2}{h_2}\right).$$

K is a symmetric probability density on R^2, and h_1 and h_2 are bandwidths. In our calculation, we simply chose K to be the standard Gaussian kernel and $h_1 = h_2$. The bandwidths were selected by the bootstrap scheme based on resampling

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

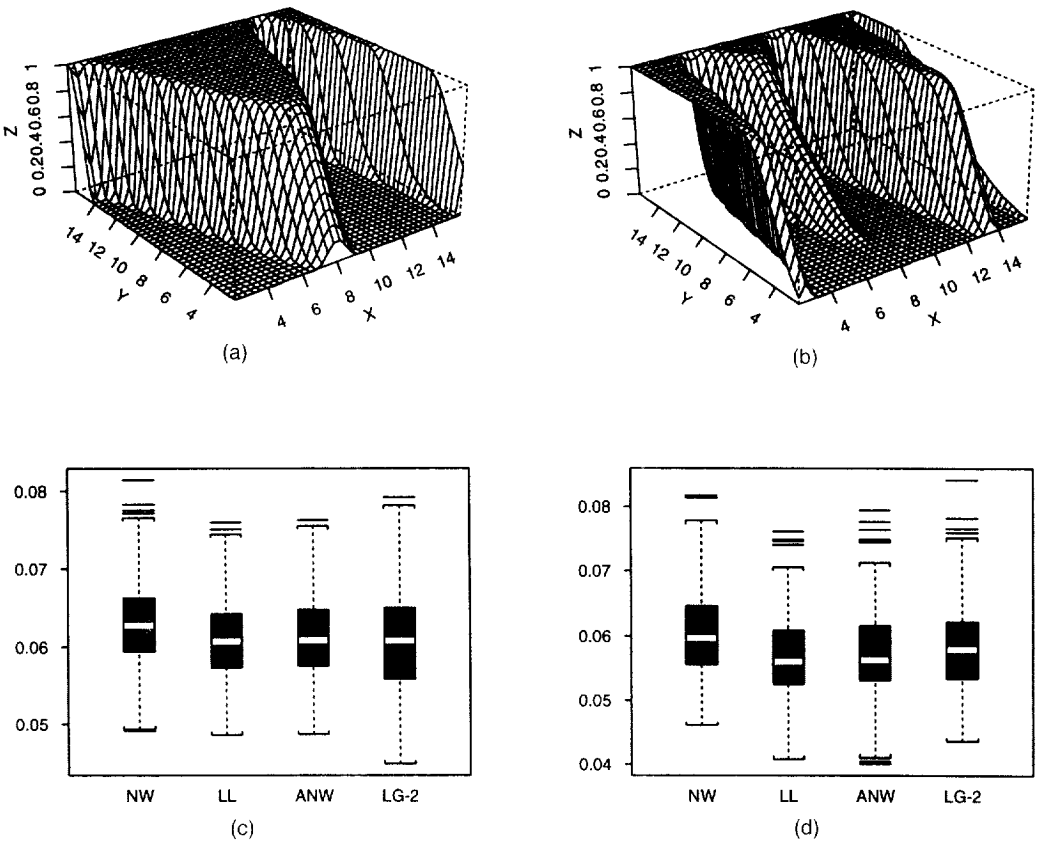


Figure 3. Simulation Results for Example 2: Estimating the Conditional Distribution of Y_t Given $X_t \equiv Y_{t-m}$ Using Optimal Bandwidths. (a) Conditional distribution function $z = \pi(y|x)$ for $m = 2$; (b) conditional distribution function $z = \pi(y|x)$ for $m = 3$; (c) boxplots of MADEs for the NW estimate, LL estimate, ANW estimate, LG-2 estimate when $m = 2$; (d) boxplots of MADEs for the NW estimate, LL estimate, ANW estimate, and LG-2 estimate when $m = 3$.

time series from the best-fitted linear AR(2) model. Out of 10 predictive intervals, only one (for the year 1929) missed the true value, and then only narrowly. The average length of the intervals is now reduced to 1.63, which is 32.8% of the dynamic range of the data.

4. THEORETICAL PROPERTIES

For the local logistic estimator $\hat{\pi}(y|x)$, we consider only functions A of the exponential-polynomial type, with $r \geq 2$: $A(x, \boldsymbol{\theta}) = \exp(\theta_1 x^0 + \cdots + \theta_r x^{r-1})$. Let f denote the marginal density of $\mathbf{X}_i$. We impose the following regularity conditions:

C1. For fixed y and x, $f(x) > 0, 0 < \pi(y|x) < 1$, f is continuous at x, and $\pi(y|\cdot)$ has $2[(r+1)/2]$ continuous derivatives in a neighbourhood of x, where $[t]$ denotes the integer part of t.

C2. The kernel K is a symmetric, compactly supported probability density satisfying $|K(x_1) - K(x_2)| \leq C|x_1 - x_2|$ for x_1, x_2.

C3. The process $\{(\mathbf{X}_i, Y_i)\}$ is absolutely regular; that is,

$$\beta(j) \equiv \sup_{i \geq 1} E \times \left\{ \sup_{A \in \mathcal{F}_{i+j}^{\infty}} |P(A|\mathcal{F}_1^i) - P(A)| \right\} \to 0 \quad \text{as} \quad j \to \infty,$$

where $\mathcal{F}_i^j$ denotes the σ field generated by $\{(X_k, Y_k): i \leq k \leq j\}$. Further, $\sum_{j \geq 1} j^2 \beta(j)^{\delta/(1+\delta)} < \infty$ for some $\delta \in [0, 1)$. (We define $a^b = 0$ when $a = b = 0$.)

C4. As $n \to \infty, h \to 0$ and $\liminf_{n \to \infty} nh^{2r} > 0$.

Remark 1: Discussion of Conditions. Assumption C3 holds with $\delta = 0$ if and only if the process $\{(X_i, Y_i)\}$ is m-dependent for some $m \geq 1$. The requirement in C2 that K be compactly supported is imposed for the sake of brevity of proofs, and can be removed at the cost of lengthier arguments. In particular, the Gaussian kernel is allowed. In C3, the assumption on the convergence rate of $\beta(j)$ is also not the weakest possible. The last condition in C4 may be relaxed if we are prepared to strengthen C3 somewhat. For example, if the process $\{(X_i, Y_i)\}$ is m dependent, then, for Theorem 1 (presented later), we need only $nh \to \infty$, not nh^{2r} bounded away from 0. However, because C4 is always satisfied by bandwidths of optimal size (i.e., $h \approx \text{const } n^{-1/(2r+1)}$), we do not concern ourselves with such refinements.

Remark 2: Consistency. It may be proved that, under conditions C1–C4 and assuming that $r \geq 2, \hat{\boldsymbol{\theta}}_{xy} \to \boldsymbol{\theta}^0$ in probability, where $\boldsymbol{\theta}^0 = \boldsymbol{\theta}^0_{xy}$ is uniquely defined by

$$\pi^{(i)}(y|x) = L^{(i)}(0, \boldsymbol{\theta}^0), \qquad i = 0, 1, \ldots, r-1. \tag{6}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

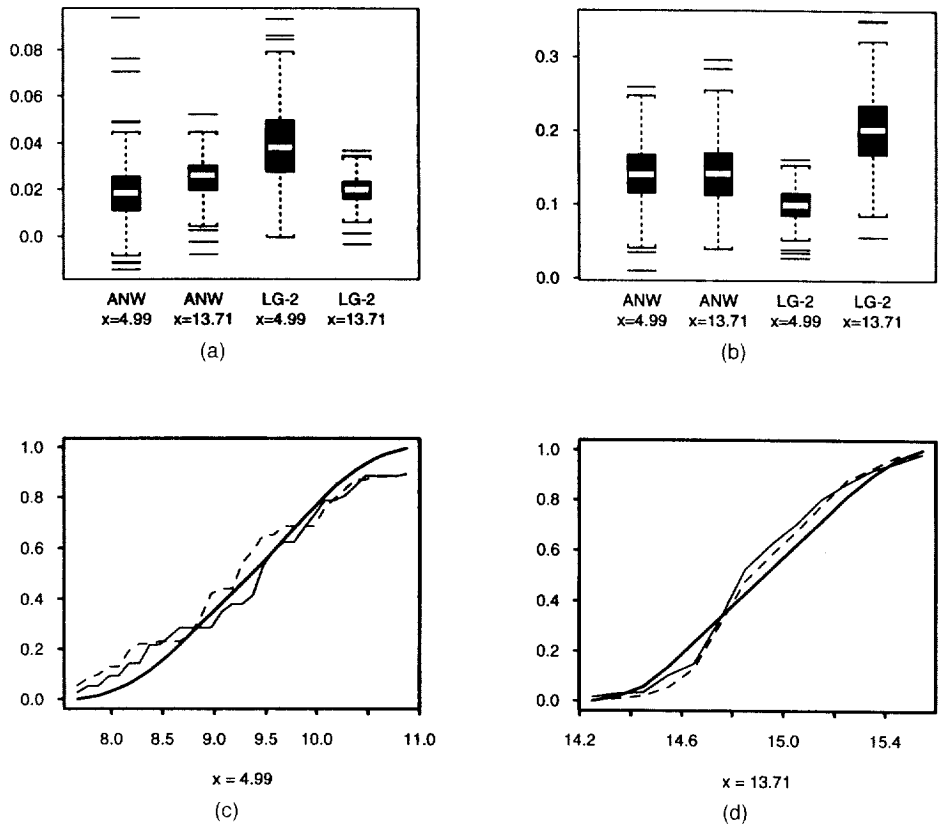


Figure 4. Simulation Results for Example 2: Estimating the Two-Step Ahead Predictive Distribution $\pi(y|x)$ With Bootstrap Bandwidths. (a) Boxplots of the MADEs based on $\hat{h}(x)$ minus the MADEs based on $h_{op}(x)$; (b) The boxplots of $\hat{h}(x) - h_{op}(x)$; (c)–(d) curves representing conditional distribution functions $\pi(\cdot|x)$: thick line, $\pi(\cdot|x)$: thin line, ANW estimate; dashed line, LG-2 estimate.

and $\pi^{(i)}, L^{(i)}$ are as in Section 2.1. It follows from this result that at x, $\hat{\pi}(y|\cdot)$ and its first $r-1$ derivatives are consistent for the corresponding derivatives of $\pi(y|\cdot)$. In the case of the "0th derivative," Theorem 1 extends this property to a detailed description of the stochastic and systematic errors of $\hat{\pi}(y|\cdot)$.

Define $\kappa_j = \int u^j K(u)\,du$ and $\nu_j = \int u^j K(u)^2\,du$. Let $\mathbf{S}$ denote the $r \times r$ matrix with (i,j)th element κ_{i+j-2}, and let $\kappa^{(i,j)}$ be the (i,j)th element of $\mathbf{S}^{-1}$. Let $r_1 = 2[(r+1)/2]$, and put $\tau(y|x)^2 = \pi(y|x)\{1-\pi(y|x)\}/f(x)$.

$$\mu_r(x) = (r!)^{-1}\{\pi^{(r_1)}(y|x) - L^{(r_1)}(0,\boldsymbol{\theta}^0)\}\sum_{i=1}^{r}\kappa^{(1,i)}\kappa_{r_1+i-1},$$

and

$$\tau_r^2 = \int\left(\sum_{i=1}^{r}\kappa^{(1,i)}u^{i-1}\right)^2 K(u)^2\,du.$$

Let N_{n1}, N_{n2}, and N_{n3} denote random variables with the standard normal distribution.

Theorem 1.

a. Suppose that $r \geq 2$ and conditions C1–C4 hold. Then, as $n \to \infty$,

$$\hat{\pi}(y|x) - \pi(y|x) = (nh)^{-1/2}\tau(y|x)\tau_r \mathrm{N}_{n1} + h^{r_1}\mu_r(x) + o_p\{h^{r_1} + (nh)^{-1/2}\}. \quad (7)$$

Table 1. Predictive Intervals for Canadian Lynx in 1925–1934, Based on the Data in 1821–1924

Year	True value	Predictor from one lagged value	$\hat{h}(x)$	Predictor from two lagged values	$\hat{h}(x_1, x_2)$
1925	8.18	[5.89, 8.69]	.123	[6.86, 8.60]	.245
1926	7.98	[5.99, 8.81]	.340	[6.86, 8.81]	.570
1927	7.34	[5.94, 8.75]	.485	[6.40, 8.26]	.715
1928	6.27	[5.43, 8.35]	.195	[5.44, 6.86]	.715
1929	6.18	[4.69, 7.71]	.268	[4.60, 6.16]	1.095
1930	6.50	[4.65, 7.70]	.340	[5.43, 7.03]	.860
1931	6.91	[5.21, 7.72]	.268	[5.71, 7.50]	.860
1932	7.37	[5.37, 7.82]	.268	[6.38, 8.12]	.860
1933	7.88	[5.44, 8.38]	.123	[7.17, 8.25]	.715
1934	8.13	[5.89, 8.74]	.485	[7.26, 8.81]	1.205

NOTE: The nominal coverage probability is .9.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

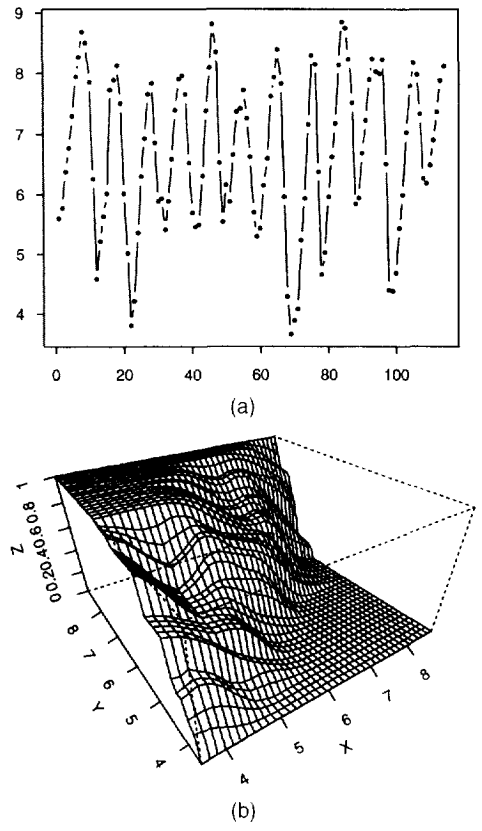


Figure 5. Canadian Lynx Data (a) and Estimated Conditional Distribution $z = \pi(y|x)$ of Y_t Given $Y_{t-1} = x$ (b).

b. Assume conditions C1–C4 with $r = 2$. Then, as $n \to \infty$,

$$\hat{\pi}(y|x) - \pi(y|x) = (nh)^{-1/2}\tau(y|x)\nu_0^{1/2}\mathrm{N}_{n2} + \frac{1}{2}h^2\kappa_2\pi^{(2)}(y|x) + o_p\{h^2 + (nh)^{-1/2}\}. \quad (8)$$

The first term on the right side in both (7) and (8) represents predominantly error about the mean, and the second term represents predominantly bias. It may be seen from the theorem that to first order, our methods enjoy the same convergence rates as those of Yu and Jones (1998), under similar regularity conditions. (Yu and Jones treated only the case of independent data, however.)

Remark 3: Comparison of $\hat{\pi}$ and Local Polynomial Estimator. To first order, and for general r, the asymptotic variance of $\hat{\pi}(y|x)$ is exactly as in the case of local polynomial regression estimators of order r. (For the latter, see, e.g., Ruppert and Wand 1994.) This similarity extends also to the bias term, to the extent that for both $\hat{\pi}$ and local polynomial estimators the bias is of order h^r for even r and h^{r+1} for odd r, and (to this order) does not depend on the design density f. However, the forms of bias as functionals of the "regression mean" π are quite different. This is due to the fact that, unlike a local polynomial estimator, $\hat{\pi}(y|x)$ is constrained to lie within (0, 1).

Remark 4: Comparison of $\tilde{\pi}$ and Local Linear Estimator. It can be shown that, assuming (C1)–(C4) for $r = 2$, the asymptotic formula (8) for $\tilde{\pi}(y|x)$ is shared exactly by the standard local linear estimator $\hat{\pi}_{\mathrm{LL}}(y|x)$, derived by minimizing

$$\sum_{i=1}^{n} \{I(Y_i \le y) - \alpha - \beta(X_i - x)\}^2 K_h(X_i - x)$$

with respect to (α, β) and taking $\hat{\pi}_{\mathrm{LL}}(y|x) = \hat{\alpha}$. Note, however, that unlike $\tilde{\pi}$, $\hat{\pi}_{\mathrm{LL}}$ is constrained neither to lie between 0 and 1 nor to be monotone in y. Additionally, $\tilde{\pi}$ is somewhat more resistant against data sparseness than $\hat{\pi}_{\mathrm{LL}}$. For example, it never assumes the form (nonzero number)/(zero).

Remark 5: Comparison of $\hat{\pi}$ and $\tilde{\pi}$. In the case where $r = 2$, (7) reduces to

$$\hat{\pi}(y|x) - \pi(y|x) = (nh)^{-1/2}\tau(y|x)\mathrm{N}_{n1} + \frac{1}{2}h^2\kappa_2\mu_2(y|x) + o_p\{h^2 + (nh)^{-1/2}\}, \quad (9)$$

where

$$\mu_2(y|x) = \pi^{(2)}(y|x) - \frac{\pi^{(1)}(y|x)^2\{1 - 2\pi(y|x)\}}{\pi(y|x)\{1 - \pi(y|x)\}}$$

and $\pi^{(i)}$ is defined as in Section 2.1. Comparing (8) and (9), we see that $\hat{\pi}(y|x)$ (with $r = 2$) and $\tilde{\pi}(y|x)$ have the same asymptotic variance, but that the first-order bias formula of the former contains an additional term. Consequently, if $\pi(y|x) < 1/2$, then $\hat{\pi}(y|x)$ is biased downward relative to $\tilde{\pi}(y|x)$, whereas if $\pi(y|x) > 1/2$, then it is biased upward.

Remark 6: Comparison With the Nadaraya–Watson Estimator. The analog of (8) and (9) in the case of the NW estimator,

$$\hat{\pi}_{\mathrm{NW}}(y|x) = \left\{\sum_{i=1}^{n} I(Y_i \le y)K_h(X_i - x)\right\} \div \left\{\sum_{i=1}^{n} K_h(X_i - x)\right\},$$

is

$$\hat{\pi}_{\mathrm{NW}}(y|x) - \pi(y|x) = (nh)^{-1/2}\tau(y|x)\nu_0^{1/2}\mathrm{N}_{n3} + \frac{1}{2}h^2\kappa_2\mu(y|x) + o_p\{h^2 + (nh)^{-1/2}\},$$

where $\mu(y|x) = \pi^{(2)}(y|x) + 2f(x)^{-1}f'(x)\pi^{(1)}(y|x)$. Note in particular that, unlike any of $\hat{\pi}$, $\tilde{\pi}$, and $\hat{\pi}_{\mathrm{LL}}$, $\hat{\pi}_{\mathrm{NW}}$ has a bias that depends to first order on the density f of X_i. However, the variances of all four estimators ($\tilde{\pi}$ with $r = 2$) are identical to first order.

Remark 7: Continuity of $\pi(y|x)$ With Respect to y. Conditions C1–C4 require continuity of $\pi(y|x)$ with respect only to x, not to y. In principle, we could exploit the smoothness of $\pi(y|x)$ in y by, for example, taking the integral average of $\hat{\pi}(\cdot|x)$ or $\tilde{\pi}(\cdot|x)$ in the neighborhood of y, thereby obtaining an estimator that had potentially lower variance. However, any improvement in performance is available only to second order. To appreciate why, note

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

that if $y_1 \le y_2$, then, to first order, the covariance of $\hat\pi(y_1|x)$ and $\hat\pi(y_2|x)$ equals $(nh)^{-1}\pi(y_1|x)\{1-\pi(y_2|x)\}\tau_r^2$, which, as $y_1, y_2 \to y$, converges to the first-order term in the variance of $\hat\pi(y|x)$. It follows that no first-order reductions in variance are obtainable by averaging over values of $\hat\pi(z|x)$ for z in a decreasingly small neighborhood of y. The same argument applies to $\tilde\pi$.

Improvements in the convergence rate generally cannot be obtained even with parametric knowledge of $\pi(y|x)$ as a smooth function of y. This is perhaps most clearly shown by considering the following example. Suppose that the distribution of (X, Y) is known to be a mixture of the distributions of $(X^{(1)}, Y^{(1)})$ and $(X^{(2)}, Y^{(2)})$, where in each case $X^{(i)}$ is independent of $Y^{(i)}$, and also suppose that the mixing proportion and the distributions of $X^{(1)}, Y^{(1)}$, and $Y^{(2)}$ are all known. Then, provided that the mixing proportion is not 0 or 1, X and Y are not independent, and $\pi(y|x)$ is a known, smooth functional of the density ψ of $X^{(2)}$. It follows that the rate at which $\pi(y|x)$ can be estimated (for any fixed y) is identical to that at which $\psi(x)$ can be estimated. Under conditions C1–C4, this is the same as the rate implicit in Theorem 1, regardless of any smoothness of $\pi(y|x)$ in y that might be available through assumptions made about smoothness of the distributions of $Y^{(1)}$ and $Y^{(2)}$.

APPENDIX: PROOFS

We derive only (7), noting that a proof of (8) is similar but simpler. For any $\varepsilon \in (0,1)$, it follows from Remark 2 that there exists $\eta \in (0,\infty)$ such that $P\{|\hat{\boldsymbol\theta}_{xy} - \boldsymbol\theta^{(0)}| \le \eta\} \ge 1-\varepsilon$ for all sufficiently large n. Let $G = G(\eta)$ denote the closed ball centered at $\boldsymbol\theta^{(0)}$ and with radius η. Let $\hat{\boldsymbol\theta}_{xy,G}$ be the minimizer of (1), with $\boldsymbol\theta$ restricted to G. Define $\hat\pi_G(y|x) = L(0|\hat{\boldsymbol\theta}_{xy,G})$. Then $P\{\hat\pi_G(y|x) \ne \hat\pi(y|x)\} < \varepsilon$ when n is sufficiently large. This argument indicates that we need only establish (7) for $\hat\pi_G(y|x)$. Thus we may develop the proof by assuming that $\hat{\boldsymbol\theta}_{xy}$ is always contained within a compact set G.

We consider only the case of even r. By simple Taylor expansion of L in (1), we may show that

$$R(\boldsymbol\theta; x, y) = \sum_{i=1}^n \left[I(Y_i \le y) - \sum_{j=0}^{r-1} (j!)^{-1} L^{(j)}(0,\boldsymbol\theta)(X_i - x)^j - (r!)^{-1} L^{(r)}\{c_i(X_i - x), \boldsymbol\theta\}(X_i - x)^r \right]^2 K_h(X_i - x),$$

where $c_i \in [0,1]$. Define $R^*(\boldsymbol\theta; x, y)$ as $R(\boldsymbol\theta; x, y)$, with $\boldsymbol\theta$ in $L^{(r)}\{c_i(X_i - x), \theta\}$ replaced by $\hat{\boldsymbol\theta}_{xy}$. Let $\hat{\boldsymbol\theta}^*_{xy}$ denote the minimizer of $R^*(\boldsymbol\theta; x, y)$, and put $\hat\pi^*(y|x) = L(0, \hat{\boldsymbol\theta}^*_{xy})$. To derive (7), it suffices to show that the result holds for $\hat\pi^*(y|x)$, and also that

$$\hat\pi(y|x) = \hat\pi^*(y|x) - o_p(h^r). \tag{A.1}$$

Define

$$s_j(x) = (nh^j)^{-1} \sum_{i=1}^n K_h(X_i - x)(X_i - x)^j,$$

let $\mathbf{S}_n(x)$ denote the $r \times r$ matrix with $s_{i+j-2}(x)$ as its (i,j)th element, and put

$$W_n(u, x) = (1, 0, \ldots, 0)\mathbf{S}_n(x)^{-1}(1, u, \ldots, u^{r-1})^T K(u)$$

and $W_{nh}(u, x) = W_n(u/h, x)$. In this notation we have, by the definition of $\hat\pi^*(y|x)$,

$$\begin{aligned}
&\hat\pi^*(y|x) - \pi(y|x) \\
&\quad = (nh)^{-1} \sum_{i=1}^n W_{nh}(X_i - x, x) \\
&\qquad \times \left\{ I(Y_i \le y) - \sum_{j=0}^{r-1} (j!)^{-1} \pi^{(j)}(y|x) \right. \\
&\qquad\qquad \left. \times (X_i - x)^j - (r!)^{-1} L^{(r)}\{c_i(X_i - x), \hat{\boldsymbol\theta}_{xy}\}(X_i - x)^r \right\} \\
&\quad = (nh)^{-1} \sum_{i=1}^n W_{nh}(X_i - x, x) \\
&\qquad \times (\varepsilon_i + r!)^{-1}[\pi^{(r)}\{y|x + c_i'(X_i - x)\} \\
&\qquad - L^{(r)}\{c_i(X_i - x), \hat{\boldsymbol\theta}_{xy}\}](X_i - x)^r).
\end{aligned} \tag{A.2}$$

where $\varepsilon_i = I(Y_i \le y) - \pi(y|X_i)$ and $c_i' \in [0,1]$. (See, e.g., formula 3.11 of Fan and Gijbels 1996.) By the ergodic theorem, $\mathbf{S}_n(x) \to f(x)\mathbf{S}$ in probability, where $\mathbf{S}$ denotes the $r \times r$ matrix with κ_{i+j-2} in position (i,j).

Define $\xi_i = \sum_{1 \le j \le r} \kappa^{(1,j)}\{(X_i - x)/h\}^{j-1}$ and

$$R_i = (r!)^{-1}\{\pi^{(r)}\{y|x + c_i'(X_i - x)\} - L^{(r)}\{c_i(X_i - x), \hat{\boldsymbol\theta}_{xy}\}].$$

Noting the representation (A.2) for $\hat\pi^*(y|x) - \pi(y|x)$, and also lemmas 1 and 2 of Yao and Tong (1997), and recalling that $\hat{\boldsymbol\theta}_{xy} \in G$, we may prove that the ratio of $\hat\pi^*(y|x) - \pi(y|x)$ and

$$(nh)^{-1} f(x)^{-1} \sum_{i=1}^n \xi_i K_h(X_i - x)\{\varepsilon_i + R_i(X_i - x)^r\}$$

converges in probability to 1. By the ergodic theorem,

$$(nh^{r+1})^{-1} f(x)^{-1} \sum_{i=1}^n \xi_i K_h(X_i - x) R_i (X_i - x)^r = \mu_r(x) + o_p(1).$$

If the δ in C3 is strictly positive, then it follows from theorem 1.7 of Peligrad (1986) that $(nh)^{-1/2}\sum_i \xi_i K_h(X_i - x)\varepsilon_i$ is asymptotically normal with mean 0 and variance

$$h^{-1} E\{K_h(X_1 - x)\varepsilon_1\}^2 + h^{-1} \times \sum_{i=2}^n E\{\xi_1 K_h(X_1 - x)\varepsilon_1 \xi_i K_h(X_i - x)\varepsilon_i\}.$$

The first term in this expression converges to $f(x)\pi(y|x)\{1 - \pi(y|x)\}\tau_r^2$. By lemma 1 of Yoshihara (1976), the second term is bounded above by a constant multiple of

$$h^{(1-\delta)/(1+\delta)} \sum_{i=1}^n \beta(i)^{\delta/(1+\delta)},$$

which, in view of C3, converges to 0. Thus, the second term is asymptotically negligible. Combining the results in this paragraph, we obtain, in the case where $\delta > 0$, the version of (7) that arises if $\hat\pi^*(y|x)$ is replaced by $\hat\pi(y|x)$. This result continues to hold in the case where $\delta = 0$, and indeed is relatively easy to prove there; see Remark 1 in Section 4.

The final step is to prove (A.1). Formula (A.2) provides an explicit expression for $L(0, \hat{\boldsymbol\theta}^*_{xy}) - L(0, \boldsymbol\theta^{(0)})$, and we may derive an expression for $L^{(j)}(0, \hat{\boldsymbol\theta}^*_{xy}) - L^{(j)}(0, \boldsymbol\theta^{(0)})$ similarly. Arguing thus,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

we can prove that $L^{(i)}(0, \hat{\boldsymbol{\theta}}^*_{xy}) \to L^{(i)}(0, \boldsymbol{\theta}^0)$ in probability. Therefore, because $\boldsymbol{\theta}^0$ is uniquely determined by (6), $\hat{\boldsymbol{\theta}}^*_{xy} \to \boldsymbol{\theta}^0$ in probability. Hence $|\hat{\boldsymbol{\theta}}^*_{xy} - \hat{\boldsymbol{\theta}}_{xy}| \to 0$. Because all of the first derivatives of $R^*(\boldsymbol{\theta}; x, y)$ (with respect to components of $\boldsymbol{\theta}$) vanish at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^*_{xy}$, this implies that $R(\hat{\boldsymbol{\theta}}^*_{xy}; x, y) = R^*(\hat{\boldsymbol{\theta}}^*_{xy}; x, y) + o_p(nh^{2r})$. Now, $R(\hat{\boldsymbol{\theta}}_{xy}; x, y) = R^*(\hat{\boldsymbol{\theta}}_{xy}; x, y)$. Hence, because $\hat{\boldsymbol{\theta}}_{xy}$ and $\hat{\boldsymbol{\theta}}^*_{xy}$ minimize R and R^*,

$$0 \le R(\hat{\boldsymbol{\theta}}^*_{xy}; x, y) - R(\hat{\boldsymbol{\theta}}_{xy}; x, y)$$
$$= R^*(\hat{\boldsymbol{\theta}}^*_{xy}; x, y) - R^*(\hat{\boldsymbol{\theta}}_{xy}; x, y) + o_p(nh^{2r}) \le o_p(nh^{2r}).$$

This establishes that $(nh^{2r})^{-1}\{R(\hat{\boldsymbol{\theta}}^*_{xy}; x, y) - R(\hat{\boldsymbol{\theta}}_{xy}; x, y)\} \to 0$ in probability. Because all of the first derivatives of $R(\boldsymbol{\theta}; x, y)$ (with respect to components of $\boldsymbol{\theta}$) vanish at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{xy}$, this implies that

$$h^{-2r}(\hat{\boldsymbol{\theta}}_{xy} - \hat{\boldsymbol{\theta}}^*_{xy})^T \tilde{\mathbf{R}}(\hat{\boldsymbol{\theta}}_{xy})(\hat{\boldsymbol{\theta}}_{xy} - \hat{\boldsymbol{\theta}}^*_{xy}) \to 0 \qquad \text{(A.3)}$$

in probability, where $\tilde{\mathbf{R}}(\boldsymbol{\theta})$ equals the $r \times r$ matrix of second derivatives with respect to components of $\boldsymbol{\theta}$, or $\mathbf{R}(\boldsymbol{\theta}; x, y)$. The left-side of (A.3) may be written as $\mathbf{V}^T \bar{R}(\hat{\boldsymbol{\theta}}_{xy})\mathbf{V}$, where $\mathbf{V}$ denotes the r-vector whose ith element is $(\hat{\boldsymbol{\theta}}_{xy} - \hat{\boldsymbol{\theta}}^*_{xy})^{(i)}/h^{r-i+1}$, and

$$\bar{\mathbf{R}} = \operatorname{diag}(1, h^{-1}, \ldots, h^{-(r-1)})\tilde{\mathbf{R}}(\hat{\boldsymbol{\theta}}_{xy})\operatorname{diag}(1, h^{-1}, \ldots, h^{-(r-1)}).$$

It may be proved that $\bar{\mathbf{R}} \to f(x)\pi(y|x)\{1 - \pi(y|x)\}\mathbf{S}$ in probability, where $\mathbf{S}$ is the positive-definite matrix defined earlier in the proof. Hence the ith element of $\hat{\boldsymbol{\theta}}_{xy} - \hat{\boldsymbol{\theta}}^*_{xy}$ equals $o_p(h^{r-i+1})$ for $1 \le i \le r$. The desired result (A.1) follows from this formula and the fact that $\hat{\pi}(y|x) = \exp(\hat{\boldsymbol{\theta}}^{(1)}_{xy})/\{1 + \exp(\hat{\boldsymbol{\theta}}^{(1)}_{xy})\}$, where $\hat{\boldsymbol{\theta}}^{(1)}_{xy}$ denotes the first element of $\hat{\boldsymbol{\theta}}_{xy}$.

[Received October 1997. Revised July 1998.]

REFERENCES

Ahn, H. (1995), "Nonparametric Two-Stage Estimation of Conditional Choice Probabilities in a Binary Choice Model Under Uncertainty," *Journal of Econometrics*, 67, 337–378.

Ahn, H., and Manski, C. F. (1993), "Distribution Theory for the Analysis of Binary Choice Under Uncertainty With Nonparametric Estimation of Expectations," *Journal of Econometrics*, 56, 291–321.

Barbe, P., and Bertail, P. (1995), *The Weighted Bootstrap*, Berlin: Springer-Verlag.

Bhattacharya, P. K., and Gangopadhyay, A. K. (1990), "Kernel and Nearest-Neighbor Estimation of a Conditional Quantile," *The Annals of Statistics*, 18, 1400–1415.

Chu, C-K., and Cheng, K. F. (1995), "Nonparametric Regression Estimates Using Misclassified Binary Responses," *Biometrika*, 82, 315–325.

Chu, C-K., and Marron, J. S. (1991), "Choosing a Kernel Regression Estimator," *Statistical Science*, 6, 404–436.

Copas, J. B. (1995), "Local Likelihood Based on Kernel Censoring," *Journal of the Royal Statistical Society*, Ser. B, 57, 221–235.

Fan, J. (1993), "Local Linear Regression Smoothers and Their Minimax Efficiencies," *The Annals of Statistics*, 21, 196–216.

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

Fan, J., Yao, Q., and Tong, H. (1996), "Estimation of Conditional Densities and Sensitivity Measures in Nonlinear Dynamical Systems," *Biometrika*, 83, 189–206.

Hall, P., and Presnell, B. (1997), "Intentionally Biased Bootstrap Methods," unpublished manuscript.

Hall, P., and Turlach, B. W. (1997), "Interpolation Methods for Adapting to Sparse Design in Nonparametric Regression" (with discussion), *Journal of the American Statistical Association*, 92, 466–476.

Hjort, N. L., and Jones, M. C. (1996), "Locally Parametric Nonparametric Density Estimation," *Annals of Statistics*, 24, 1619–1647.

Jones, M. C., and Hall, P. (1990), "Mean Squared Error Properties of Kernel Estimates of Regression," *Statistics and Probability Letters*, 10, 283–289.

Klein, R. W. (1993), "Specification Tests for Binary Choice Models Based on Index Quantiles," *Journal of Econometrics*, 59, 343–375.

Loader, C. R. (1996), "Local Likelihood Density Estimation," *Annals of Statistics*, 24, 1602–1618.

Matzkin, R. L. (1992), "Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models," *Econometrica*, 60, 239–270.

Peligrad, M. (1986), "Recent Advances in the Central Limit Theorem and Its Weak Invariance Principle for Mixing Sequences of Random Variables," in *Dependence in Probability and Statistics*, eds. E. Eberlein and M. S. Taqqu, Boston, MA: Birkhäuser, pp. 193–223.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992), *Numerical Recipes in C*, Cambridge, MA: Cambridge University Press.

Ruppert, D., and Wand, M. P. (1994), "Multivariate Weighted Least Squares Regression," *Annals of Statistics*, 22, 1346–1370.

Seifert, B., and Gasser, T. (1996a), "Finite-Sample Analysis of Local Polynomials: Analysis and Solutions," *Journal of the American Statistical Association*, 91, 267–275.

——— (1996b), "Variance Properties of Local Polynomials and Ensuing Modifications (with discussion), in *Statistical Theory and Computational Aspects of Smoothing*, eds. W. Härdle and M. G. Schimek, Heidelberg, Germany: Physica-Verlag, pp. 50–79.

Sheather, S. J., and Marron, J. S. (1990), "Kernel Quantile Estimators," *Journal of the American Statistical Association*, 85, 410–416.

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

Yao, Q. (1995), "Conditional Predictive Regions for Stochastic Processes," Technical Report UKC/IMS/96/17, University of Kent.

Yao, Q., and Tong, H. (1997), "Nonparametric Estimation of Ratios of Noise to Signal in Stochastic Regressions," unpublished manuscript.

Yoshihara, K. (1976), "Limiting Behaviour of U-Statistics for Stationary Absolutely Regular Processes," *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 35, 237–252.

Yu, K., and Jones, M. C. (1998), "Local Linear Quantile Regression," *Journal of the American Statistical Association*, 93, 228–237.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.