COVER SHEET

METHODS FOR ESTIMATING A CONDITIONAL DISTRIBUTION FUNCTION

Peter Hall[1] Rodney C.L. Wolff[1,2] Qiwei Yao[1,3]

Abstract. Motivated by the problem of setting prediction intervals in time series analysis, we suggest two new methods for conditional distribution estimation. The first 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 (1997), 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 Yu and Jones' work.

Keywords. 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.

Short title. Conditional distribution estimation

[1]Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia

[2]School of Mathematics, Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia

[3]Institute of Mathematics and Statistics, University of Kent at Canterbury, Canterbury, Kent CT2 7NF, UK

1 Introduction

In a variety of statistical problems, estimation of a conditional distribution function is a key aspect of inference. Consider, for example, estimation of the quantile function of Y given X, using a sample of independent data pairs $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. This problem has recently been tackled by Yu and Jones (1997) 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 employ the Nadaraya-Watson estimator, although it would suffer the excessive bias problems inherent in that approach; see for example 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 $x = Y_n$. 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 (1997).

While 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 generalisations of the Yu and Jones approach, such as methods based on local polynomial techniques (see *e.g.* Fan and Gijbels 1996), then the 'non-distribution properties' from which they suffer become even more of a problem.

In the present paper we suggest two new techniques that largely overcome these difficulties. The first, local logistic distribution estimation, produces estimators of arbitrarily high order which always lie strictly between 0 and 1. In spirit, this approach is related to recently-introduced local parametric methods for density estimation; see for example Copas (1995), Simonoff (1996, Section 3.4), Hjort and Jones (1997) and Loader (1997). 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) and Hall and Presnell (1997).

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

The paper is organised as follows. In Section 2 we introduce our methods for estimation of a conditional distribution. A bootstrap scheme for choosing bandwidths is also given. Numerical examples involving both simulated models and real-data applications are reported in Section 3. A case study with a multivariate predictor is included there. Section 4 describes convergence rates as well as asymptotic distribution properties of the proposed estimators. All technical arguments are given in the Appendix.

2 Methodology

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

To simplify discussion we introduce our methods and develop theory only in the case where X_i is a scalar (*i.e.* $d = 1$). The multivariate case will be illustrated 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 generalised local logistic model for $P(x)$ has the form $L(x, \theta) \equiv A(x, \theta)/\{1 + A(x, \theta)\}$,

where $A(\cdot, \theta)$ denotes a nonnegative function which depends on a vector of parameters $\theta = (\theta_1, \ldots, \theta_r)$ 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 $P(x), P^{(1)}(x), \ldots, P^{(r-1)}(x)$, respectively, there exist $\theta_1, \ldots, \theta_r$ such that

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

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

$$R(\theta; x, y) = \sum_{i=1}^{n} \{I(Y_i \le y) - L(X_i - x, \theta)\}^2 K_h(X_i - x)\,, \tag{2.1}$$

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, it also furnishes consistent estimators of the derivatives $\pi^{(i)}(y|x) \equiv (\partial/\partial x)^i \pi(y|x)$, in the form $\hat{\pi}^{(i)}(y|x) = L^{(i)}(0, \hat{\theta}_{xy})$ for $i = 1, \ldots, r-1$, where $L^{(i)}(x,\theta) \equiv (\partial/\partial x)^i L(x,\theta)$. In practice, $\hat{\theta}_{xy}$ may be computed using the 'downhill simplex' algorithm (see Section 10.4 in Press *et al.* 1992).

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 $n \to \infty$. A more detailed account of this property will be given in Section 4.

It is possible to fit the logistic model by matrix-weighted least-squares in place of the criterion at (2.1), reflecting the dependence structure of the process $\{X_i\}$. However, if the process is weakly dependent (*e.g.* absolutely regular, as assumed in Section 4), this has only a second-order effect on performance.

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

$$\sum_{i=1}^{n} p_i(x)\,(X_i - x)\,K_h(X_i - x) = 0\,. \tag{2.2}$$

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

$$\tilde{\pi}(y|x) = \frac{\sum_{i=1}^n I(Y_i \le y)\, p_i(x)\, K_h(X_i - x)}{\sum_{i=1}^n p_i(x)\, K_h(X_i - x)}\,. \tag{2.3}$$

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

Another way of viewing 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 (2.2), with the suggested change of weights, their estimator $\widehat{m}_0$ reduces to

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

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

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 (2.2). It is easily computed using a Newton-Raphson argument.

2.3 Bandwidth choice

Particularly in the time series case, deriving asymptotically optimal bandwidths for either the local logistic or 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 least because the target function $P(x) = \pi(y|x)$ is often approximately monotone and so has only limited opportunity

for complex behaviour. 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

$$Y_i = a_0 + a_1 X_i + \ldots + a_k X_i^k + \sigma \epsilon_i,$$

where ϵ_i is standard normal, $a_0, \ldots, a_k, \sigma$ are estimated from the data, and k is determined by AIC. We form a parametric estimator $\check{\pi}(y|x)$ 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 (2.1) 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

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

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

$$M(h; x) = \int M(h; x, y)\, \check{\pi}(y|x)\, dy. \tag{2.4}$$

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

In the event that we are working with time-series data (*e.g.* $X_i = Y_{i-m}$ for some $m \geq 1$), 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 formula above for $M(h; x, y)$.

3 Numerical properties

3.1 Simulation studies

We compared various estimators of the conditional distribution function $\pi(.|.)$ through two simulated models, one with independent observations and one with nonlinear time

series. The estimators concerned 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):

$$\text{MADE} = \frac{\sum_i |\pi_e(y_i|x_i) - \pi(y_i|x_i)|\ I(0.001 \le \pi(y_i|x_i) \le 0.999)}{\sum_i I(0.001 \le \pi(y_i|x_i) \le 0.999)},$$

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

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

where $h_{op}(x, y)$ is the minimiser of the asymptotic mean squared error (up to first order) of the estimator. This guarantees a fair comparison among different methods. Secondly, 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 minimises (2.4). Throughout this section we used the Gaussian kernel.

Example 1. Let us consider the simple model

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

where $\{X_i\}$ and $\{\epsilon_i\}$ 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 them is depicted in Figure 1(b)), we calculated the MADEs with the optimal bandwidth $h_{op}(.)$ (which is of size $n^{-1/9}$ for LG-3, and $n^{-1/5}$ for all the other methods). We estimated $\pi(y|x)$ on a regular grid defined by steps 0.067 and 0.054 in x- and y-directions, respectively. The box-plots of MADEs are presented in Figure 1(c). The variations of the MADEs for the NW, LL, ANW and LG-2 methods are more or less the same, which reflects the fact that the (asymptotic) variances of those estimators are the same. Overall, both ANW and LG-2 provide competitive performance relative to

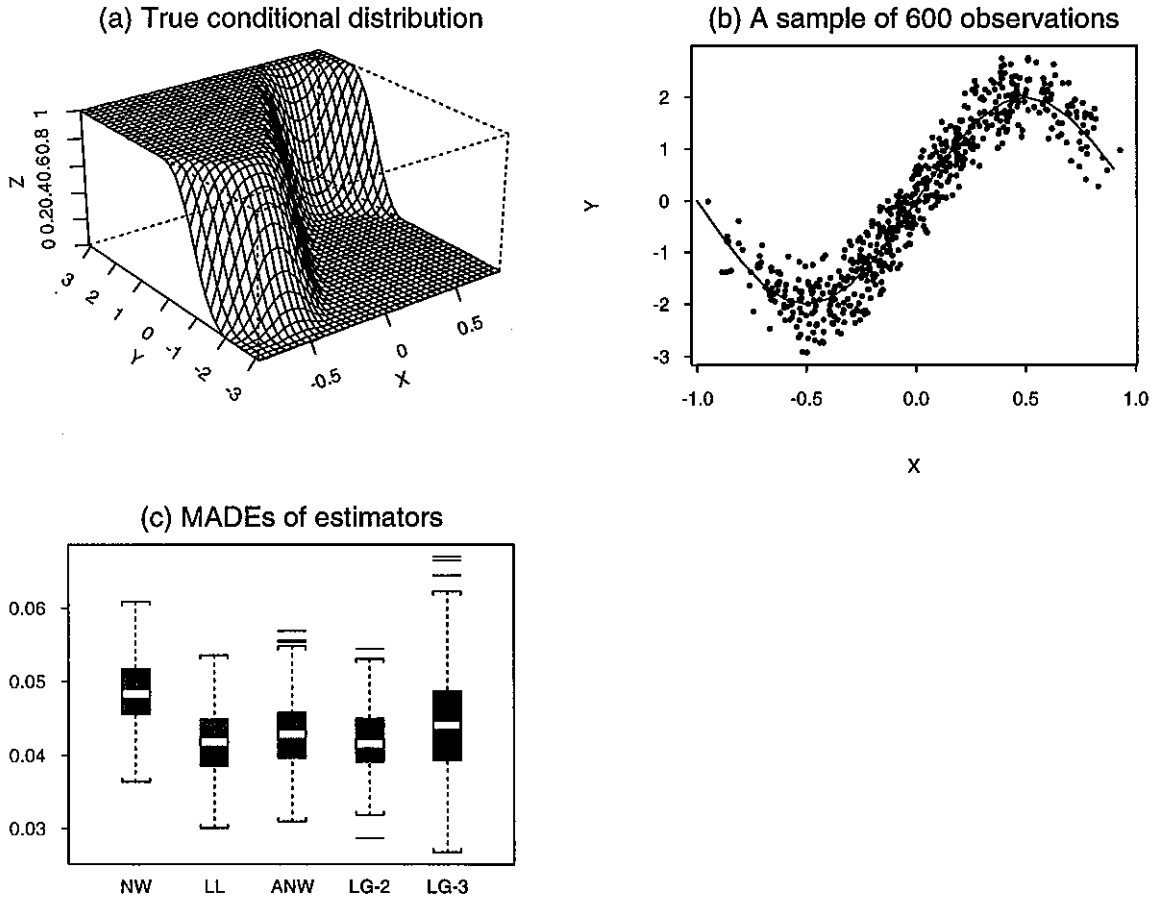

Figure 1: *Simulation results for Example 1: estimating* $\pi(y|x)$ *using optimal bandwidths. (a) The true conditional distribution function* $z = \pi(y|x)$. *(b) A typical sample of size* $n = 600$ *used in estimation, together with the curve* $y = 2\sin(3.1416\,x)$. *(c) The boxplots of MADEs for the Nadaraya-Watson estimate (NW), local linear regression estimate (LL), adjusted NW estimate (ANW), local logistic estimate with* $r = 2$ *(LG-2), and local logistic estimate with* $r = 3$ *(LG-3).*

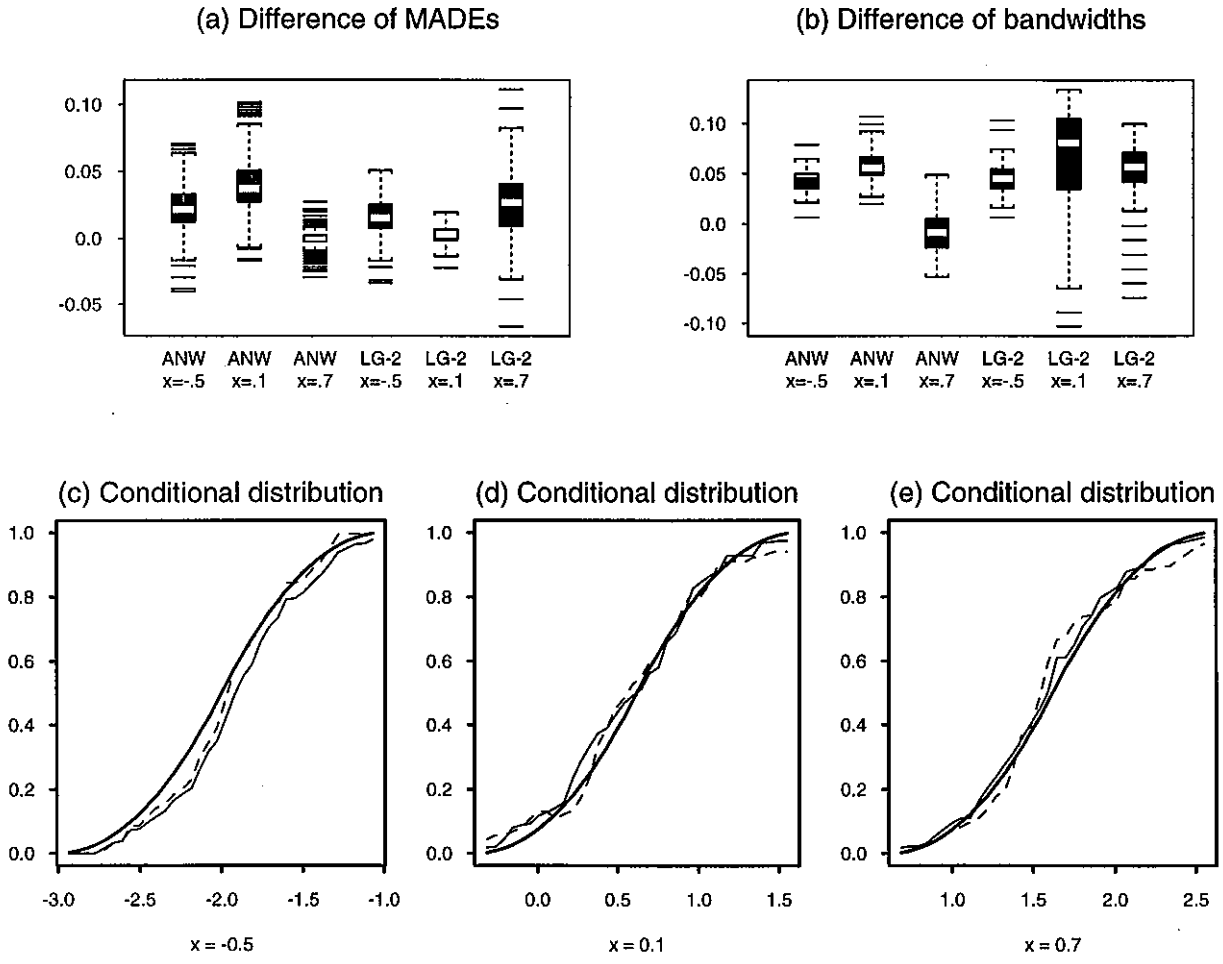

Figure 2: *Simulation results for Example 1: estimating* $\pi(y|x)$ *with bootstrap bandwidths. (a) The boxplots of the MADEs based on* $\hat{h}(x)$ *minus the MADEs based on* $h_{op}(x)$*. (b) The boxplots of* $\hat{h}(x) - h_{op}(x)$*. (c)–(e) The curves representing conditional distribution functions* $\pi(.|x)$*; thick line —* $\pi(.|x)$*, thin line — adjusted Nadaraya-Watson estimate, dashed line — local logistic estimate* $(r = 2)$*.*

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 the MADE for LG-3 is to be expected since it has larger asymptotic variance than the other estimators. (See Remark 3 and Section 3.3.1 in Fan and Gijbels 1996.) 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 $\pi(.|x)$ using the bandwidth $\hat{h}(x)$ selected by the bootstrap scheme for $x = -0.5$, 0.1 and 0.7. 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 only consider here the adjusted Nadaraya-Watson method and the local logistic method with $r = 2$. For the sake of comparison, we calculated the estimates, for the same data, using the optimal bandwidth $h_{op}(x)$. For $x = -0.5$, 0.1 and 0.7, $h_{op}(x)$ with the ANW method is 0.062, 0.036 and 0.106 respectively, and with the LG-2 method is 0.086, 0.055 and 0.085 respectively. Figure 2(a) presents boxplots of the differences of MADEs based on $\hat{h}(x)$ over the MADEs based on $h_{op}(x)$. Figure 2(b) displays boxplots of $\hat{h}(x) - h_{op}(x)$ 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 $h_{op}(x)$. Figures 2(c) – (e) depict typical examples of the estimated conditional distribution fuctions $\hat{\pi}(.|x)$ and $\tilde{\pi}(.|x)$. 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 $\tilde{\pi}(.|x)$ is monotonically increasing.

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

$$Y_t = 3.76\, Y_{t-1} - 0.235\, Y_{t-1}^2 + 0.3\, \epsilon_t, \tag{3.1}$$

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 calculate the MADEs with asymptotically optimal bandwidths on regular grid points with step 0.40 in the x-direction and steps 0.10 and 0.19 in the y-direction, for $m = 2$ and 3 respectively. Note that the conditional distributions concerned no longer admit simple explicit forms. In order to calculate $h_{op}(x)$, we evaluated the true values of $\pi(y|x)$ and its derivatives by simulation, as follows.

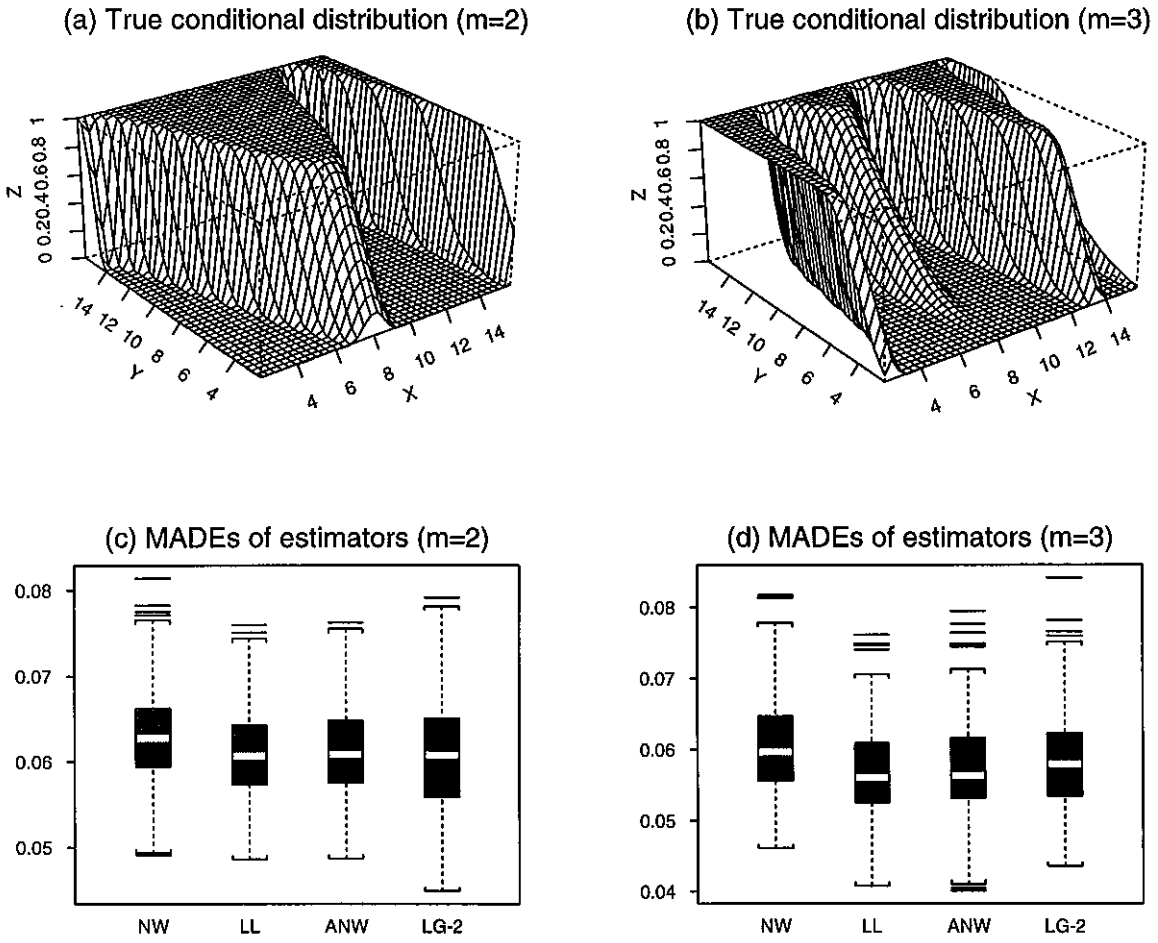

Figure 3: *Simulation results for Example 2: estimating the conditional distribution of* Y_t *given* $X_t \equiv Y_{t-m}$ *using optimal bandwidths. (a) The conditional distribution function* $z = \pi(y|x)$ *for* $m = 2$. *(b) The conditional distribution function* $z = \pi(y|x)$ *for* $m = 3$. *(c) The boxplots of MADEs for the Nadaraya-Watson estimate (NW), local linear regression estimate (LL), adjusted NW estimate (ANW), local logistic estimate with* $r = 2$ *(LG-2) when* $m = 2$. *(d) The boxplots of MADEs for the Nadaraya-Watson estimate (NW), local linear regression estimate (LL), adjusted NW estimate (ANW), and local logistic estimate with* $r = 2$ *(LG-2) when* $m = 3$.

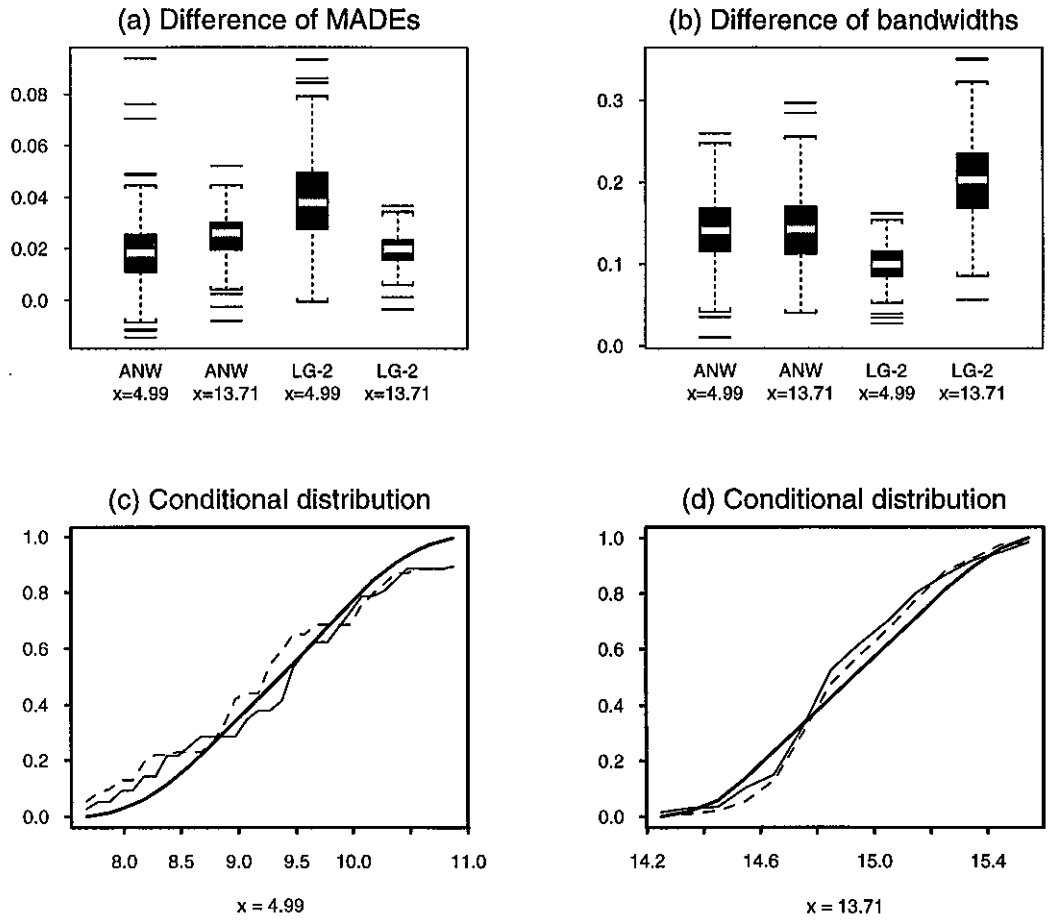

Figure 4: *Simulation results for Example 2: estimating the two-step ahead predictive distribution* $\pi(y|x)$ *with bootstrap bandwidths. (a) The boxplots of the MADEs based on* $\hat{h}(x)$ *minus the MADEs based on* $h_{op}(x)$*. (b) The boxplots of* $\hat{h}(x) - h_{op}(x)$*. (c) — (d) The curves representing conditional distribution functions* $\pi(.|x)$*: thick line —* $\pi(.|x)$*, thin line — adjusted Nadaraya-Watson estimate, dashed line — local logistic estimate* $(r = 2)$*.*

We generated 50,000 random samples by iterating (3.1) 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 Figures 3(a) and (b). We used kernel methods to estimate the marginal density function with a sample of size 100,000. Figures 3(c) and (d) are the boxplots of MADEs for the 400 replications. Similar to Example 1, both ANW and LG-2 methods provide competitive performance relative to the LL method, in terms of the absolute error of estimation; while 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 $\pi(.|x)$ 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 adjusted Nadaraya-Watson estimator and the local logistic estimator with $r = 2$. We compared the estimates with those based on the optimal bandwidth $h_{op}(x)$, which is equal to 0.182 for $x = 4.99$ and 0.216 for $x = 13.71$ in the case of the ANW estimate, and equal to 0.241 for $x = 4.99$ and 0.168 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 $h_{op}(x)$. Figure 4(b) displays boxplots of $\hat{h}(x) - h_{op}(x)$ in the simulation with 200 replications. Since we used a simple linear model to fit the nonlinear structure, it is not surprising that $\hat{h}(x)$ always overestimates $h_{op}(x)$. But the estimates for $\pi(y|x)$ remain reasonably reliable. Figure 4(c) – (d) depicts typical examples of the estimated conditional distribution functions $\hat{\pi}(.|x)$ and $\tilde{\pi}(.|x)$. 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 adjusted Nadaraya-Watson

method. The bandwidths were selected by the bootstrap scheme based on resampling the whole 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).

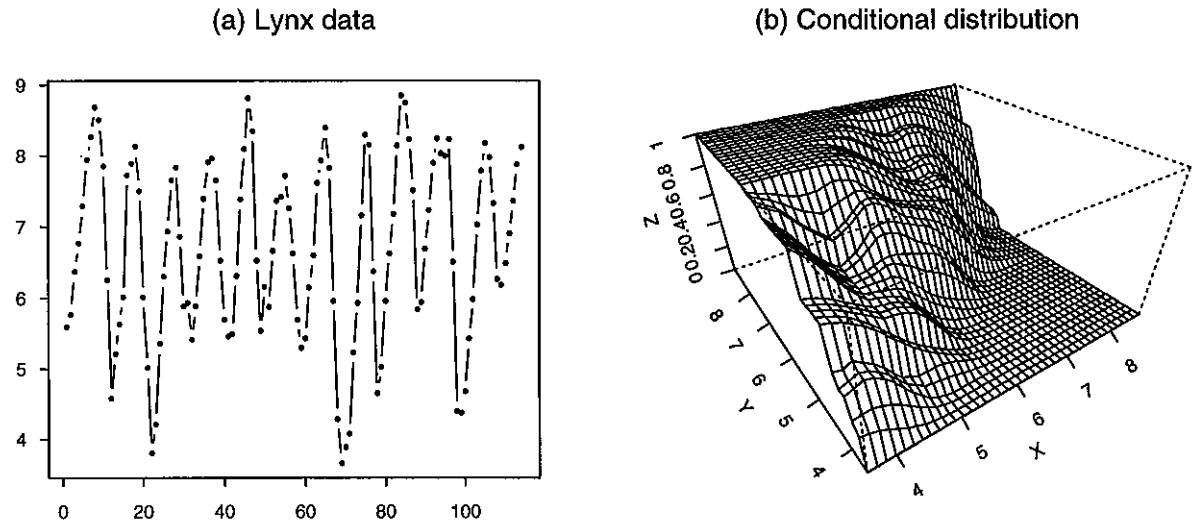

Figure 5: *(a) Canadian lynx data. (b) Estimated conditional distribution* $z = \pi(y|x)$ *of* Y_t *given* $Y_{t-1} = x$.

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 $\alpha = 0.1$ are reported in Table 1. All 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.

We also include in Table 1 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 $L(x_1, x_2, \theta) = A(x_1, x_2, \theta)/\{1 + A(x_1, x_2, \theta)\}$ with $A(x_1, x_2, \theta) = \exp(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$. The estimator is defined as $\hat{\pi}(y|x_1, x_2) \equiv L(0, 0, \hat{\theta})$, where $\hat{\theta}$ denotes the minimiser of

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

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 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.

Table 1: Predictive intervals for Canadian lynx in 1925-1934, based on the data in 1821-1924. The nominal coverage probability is 0.9.

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]	0.123	[6.86, 8.60]	0.245
1926	7.98	[5.99, 8.81]	0.340	[6.86, 8.81]	0.570
1927	7.34	[5.94, 8.75]	0.485	[6.40, 8.26]	0.715
1928	6.27	[5.43, 8.35]	0.195	[5.44, 6.86]	0.715
1929	6.18	[4.69, 7.71]	0.268	[4.60, 6.16]	1.095
1930	6.50	[4.65, 7.70]	0.340	[5.43, 7.03]	0.860
1931	6.91	[5.21, 7.72]	0.268	[5.71, 7.50]	0.860
1932	7.37	[5.37, 7.82]	0.268	[6.38, 8.12]	0.860
1933	7.88	[5.44, 8.38]	0.123	[7.17, 8.25]	0.715
1934	8.13	[5.89, 8.74]	0.485	[7.26, 8.81]	1.205

4 Theoretical properties

For the local logistic estimator $\hat{\pi}(y|x)$ we only consider functions A of exponential-polynomial type, with $r \geq 2$: $A(x, \theta) = \exp(\theta_1 x^0 + \ldots + \theta_r x^{r-1})$. Let f denote the marginal density of 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 $\{(X_i, Y_i)\}$ is absolutely regular, *i.e.*

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

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 below, we need only $nh \to \infty$, not nh^{2r} bounded away from 0. However, since (C4) is always satisfied by bandwidths of optimal size (*i.e.* $h \approx \text{const.} n^{-1/(2r+1)}$), we shall not concern ourselves with such refinements.

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

$$\pi^{(i)}(y|x) = L^{(i)}(0, \theta^0)\,, \quad i = 0, 1, \ldots, r-1\,, \tag{4.1}$$

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 'zeroth derivative,' Theorem 1 will extend 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 S denote the $r \times r$ matrix with (i, j)'th element κ_{i+j-2}, and $\kappa^{(i,j)}$ be the (i, j)'th element of S^{-1}. Let $r_1 = 2[(r+1)/2]$ and put $\tau(y|x)^2 = \pi(y|x)\,\{1 - \pi(y|x)\}/f(x)$,

$$\begin{aligned} \mu_r(x) &= (r!)^{-1} \left\{ \pi^{(r_1)}(y|x) - L^{(r_1)}(0, \theta^0) \right\} \sum_{i=1}^{r} \kappa^{(1,i)}\, \kappa_{r_1+i-1}\,, \\ \tau_r^2 &= \int \Big(\sum_{i=1}^{r} \kappa^{(1,i)}\, u^{i-1} \Big)^2 K(u)^2\, du\,. \end{aligned}$$

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

Theorem 1. *(i) Suppose $r \geq 2$ and conditions (C1)–(C4) hold. Then as $n \to \infty$,*

$$\hat{\pi}(y|x) - \pi(y|x) = (nh)^{-1/2}\,\tau(y|x)\,\tau_r\,N_{n1} + h^{r_1}\mu_r(x) + o_p\left\{h^{r_1} + (nh)^{-1/2}\right\}, \qquad (4.2)$$

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

$$\tilde{\pi}(y|x) - \pi(y|x) = (nh)^{-1/2}\,\tau(y|x)\,\nu_0^{1/2}\,N_{n2} + \tfrac{1}{2}\,h^2\,\kappa_2\,\pi^{(2)}(y|x) + o_p\left\{h^2 + (nh)^{-1/2}\right\}. \qquad (4.3)$$

The first term on the right-hand side in both (4.2) and (4.3) represents predominantly error about the mean, and the second term represents predominantly bias.

Remark 3: *Comparison of $\hat{\pi}$ and local polynomial estimator.* To first order, and for general x, 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 for example 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 a consequence of 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 (4.3) for $\tilde{\pi}(y|x)$ is shared exactly by the standard local linear estimator $\hat{\pi}_{LL}(y|x)$, derived by minimising

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

with respect to (α, β) and taking $\hat{\pi}_{LL}(y|x) = \hat{\alpha}$. Note, however, that unlike $\tilde{\pi}$, $\hat{\pi}_{LL}$ 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 $\hat{\pi}_{LL}$. For example, it never assumes the form (nonzero number)/(zero).

Remark 5: *Comparison of $\hat{\pi}$ and $\tilde{\pi}$.* In the case $r = 2$, (4.2) reduces to

$$\hat{\pi}(y|x) - \pi(y|x) = (nh)^{-1/2}\,\tau(y|x)\,N_{n1} + \tfrac{1}{2}\,h^2\,\kappa_2\,\mu_2(y|x) + o_p\left\{h^2 + (nh)^{-1/2}\right\}, \qquad (4.4)$$

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 (4.3) and (4.4) 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. In consequence, if $\pi(y|x) < \frac{1}{2}$ then $\hat{\pi}(y|x)$ is biased downwards relative to $\tilde{\pi}(y|x)$, while if $\pi(y|x) > \frac{1}{2}$ then it is biased upwards.

Remark 6: *Comparison with Nadaraya-Watson estimator.* The analogue of (4.3) and (4.4) in the case of the Nadaraya-Watson estimator,

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

is

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

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

5 Appendix: Proofs

We derive only (4.2), noting that a proof of (4.3) is similar but simpler. For any $\varepsilon \in (0, 1)$, it follows from Remark 2 that there exists $\eta \in (0, \infty)$ such that $P\{\|\hat{\theta}_{xy} - \theta^0\| \le \eta\} \ge 1 - \varepsilon$ for all sufficiently large n. Let $G = G(\eta)$ denote the closed ball centred at θ^0 and with radius η. Let $\hat{\theta}_{xy,G}$ be the minimiser of (2.1) with θ restricted to G. Define $\hat{\pi}_G(y|x) = L(0|\hat{\theta}_{xy,G})$. 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 (4.2) for $\hat{\pi}_G(y|x)$. Therefore, we may develop the proof below by assuming $\hat{\theta}_{xy}$ is always contained within a compact set G.

We consider only the case of even r. By simple Taylor expansion of L in (2.1) we may show that

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

where $c_i \in [0,1]$. Define $R^*(\theta; x, y)$ as $R(\theta; x, y)$ with θ in $L^{(r)}\{c_i(X_i - x), \theta\}$ replaced by $\hat{\theta}_{xy}$. Let $\hat{\theta}^*_{xy}$ denote the minimiser of $R^*(\theta; x, y)$, and put $\hat{\pi}^*(y|x) = L(0, \hat{\theta}^*_{xy})$. To derive (4.2) it suffices to show that the result holds for $\hat{\pi}^*(y|x)$, and additionally that

$$\hat{\pi}(y|x) = \hat{\pi}^*(y|x) + o_p(h^r)\,. \tag{5.1}$$

Define

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

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

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

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

$$\begin{aligned}
\hat{\pi}^*(y|x) - \pi(y|x) \;&=\; (nh)^{-1} \sum_{i=1}^{n} W_{nh}(X_i - x, x) \Big\{ I(Y_i \le y) - \sum_{j=0}^{r-1} (j!)^{-1}\, \pi^{(j)}(y|x) \\
&\qquad \times (X_i - x)^j - (r!)^{-1}\, L^{(r)}\{c_i(X_i - x), \hat{\theta}_{xy}\}\,(X_i - x)^r \Big\} \\
&=\; (nh)^{-1} \sum_{i=1}^{n} W_{nh}(X_i - x, x) \Big(\epsilon_i + (r!)^{-1} \Big[\pi^{(r)}\{y|x + c'_i(X_x - x)\} \\
&\qquad - L^{(r)}\{c_i(X_i - x), \hat{\theta}_{xy}\} \Big]\,(X_i - x)^r \Big)\,,
\end{aligned} \tag{5.2}$$

where $\epsilon_i = I(Y_i \le y) - \pi(y|X_i)$ and $c'_i \in [0,1]$. (See, for example, formula (3.11) of Fan and Gijbels (1996).) By the ergodic theorem, $S_n(x) \to f(x)\, S$ in probability, where 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

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

Noting the representation (5.2) for $\hat{\pi}^*(y|x) - \pi(y|x)$, and also Lemmas 1 and 2 of Yao and Tong (1997); and recalling that $\hat{\theta}_{xy} \in G$; we my prove that the ratio of $\hat{\pi}^*(y|x) - \pi(y|x)$ and

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

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 $(nh)^{-1/2} \sum_i \xi_i K_h(X_i - x) \epsilon_i$ is asymptotically Normal with mean 0 and variance

$$h^{-1} E\{K_h(X_1 - x) \epsilon_1\}^2 + h^{-1} \sum_{i=2}^{n} E\{\xi_1 K_h(X_1 - x) \epsilon_1 \xi_i K_h(X_i - x) \epsilon_i\} .$$

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. Therefore, the second term is asymptotically negligible. Combining the results in this paragraph we obtain, in the case $\delta > 0$, the version of (4.2) that arises if $\hat{\pi}^*(y|x)$ is replaced by $\hat{\pi}(y|x)$. This result continues to hold in the case $\delta = 0$, and indeed is relatively easy to prove there; see Remark 1 in Section 4.

The final step is to prove (5.1). Formula (5.2) provides an explicit expression for $L(0, \hat{\theta}^*_{xy}) - L(0, \theta^0)$, and similarly we may derive an expression for $L^{(i)}(0, \hat{\theta}^*_{xy}) - L^{(i)}(0, \theta^0)$. Arguing thus we may prove that $L^{(i)}(0, \hat{\theta}^*_{xy}) \to L^{(i)}(0, \theta^0)$ in probability. Therefore, since θ^0 is uniquely determined by (4.1), $\hat{\theta}^*_{xy} \to \theta^0$ in probability. Hence, $|\hat{\theta}^*_{xy} - \hat{\theta}_{xy}| \to 0$. Since all the first derivatives of $R^*(\theta; x, y)$ (with respect to components of θ) vanish at $\theta = \hat{\theta}^*_{xy}$, this implies that $R(\hat{\theta}^*_{xy}; x, y) = R^*(\hat{\theta}^*_{xy}; x, y) + o_p(nh^{2r})$. Now, $R(\hat{\theta}_{xy}; x, y) = R^*(\hat{\theta}_{xy}; x, y)$. Hence, since $\hat{\theta}_{xy}$ and $\hat{\theta}^*_{xy}$ minimise R and R^*, respectively,

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

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

$$h^{-2r} (\hat{\theta}_{xy} - \hat{\theta}^*_{xy})^T \tilde{R}(\hat{\theta}_{xy}) (\hat{\theta}_{xy}) (\hat{\theta}_{xy} - \hat{\theta}^*_{xy}) \to 0 \tag{5.3}$$

in probability, where $\tilde{R}(\theta)$ equals the $r \times r$ matrix of second derivatives, with respect to components of θ, of $R(\theta; x, y)$. The left-hand side of (5.3) may be written as $V^T \bar{R}(\hat{\theta}_{xy}) V$, where V denotes the r-vector whose i'th element is $(\hat{\theta}_{xy} - \hat{\theta}^*_{xy})^{(i)} / h^{r-i+1}$, and

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

It may be proved that $\bar{R} \to f(x) \pi(y|x) \{1 - \pi(y|x)\} S$ in probability, where S is the positive-definite matrix defined earlier in the proof. Hence, the i'th element of $\hat{\theta}_{xy} - \hat{\theta}^*_{xy}$

equals $o_p(h^{r-i+1})$ for $1 \leq i \leq r$. The desired result (5.1) follows from this formula and the fact that $\hat{\pi}(y|x) = \exp(\hat{\theta}_{xy}^{(1)})/\{1 + \exp(\hat{\theta}_{xy}^{(1)})\}$, where $\hat{\theta}_{xy}^{(1)}$ denotes the first element of $\hat{\theta}_{xy}$.

References

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

Chu, C.-K. and Marron, J.S. (1991). Choosing a kernel regression estimator. *Statist. Science* **6**, 404–436.

Copas, J.B. (1995). Local likelihood based on kernel censoring. *J. Roy. Statist. Soc. Ser. B* **57**, 221–235.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196–216.

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

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. Manuscript.

Hjort, N.L. and Jones, M.C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.*, **24**, 1619–1647.

Loader, C.R. (1996). Local likelihood density estimation. *Ann. Statist.*, **24**, 1602–1618.

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*, Ed. E. Eberlein and M.S. Taqqu. Birkhäuser, Boston, 193-223.

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

Ruppert, D. and Wand, M.P. (1994). Multivariate weighted least squares regression. *Ann. Statist.*, **22**, 1346–1370.

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

Yao, Q. and Tong, H. (1997). Nonparametric estimation of ratios of noise to signal in stochastic regressions. Manuscript.

Yoshihara, K. (1976). Limiting behaviour of U-statistics for stationary absolutely regular processes. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, **35**, 237-252.

Yu, K. and Jones, M.C. (1997). Local linear quantile regression. *J. Amer. Statist. Assoc.*