Statistica Neerlandica (2003) Vol. 57, nr. 2, pp. 159–176

# On Conditional Density Estimation

Jan G. De Gooijer[1]

[1]*Department of Economic Statistics, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands*

Dawit Zerom[2]

[2]*Department of Quantitative Economics, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands*

With the aim of mitigating the possible problem of negativity in the estimation of the conditional density function, we introduce a so-called re-weighted Nadaraya-Watson (RNW) estimator. The proposed RNW estimator is constructed by a slight modification of the well-known Nadaraya-Watson smoother. With a detailed asymptotic analysis, we demonstrate that the RNW smoother preserves the superior large-sample bias property of the local linear smoother of the conditional density recently proposed in the literature. As a matter of independent statistical interest, the limit distribution of the RNW estimator is also derived.

*Key Words and Phrases:* $\alpha$-mixing, asymptotic properties, negativity, nonparametric, RNW.

## 1 Introduction

Let $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i);\ i \geqslant 1\}$ be a $\mathbb{R}^d \times \mathbb{R}$ valued strictly stationary process with a common probability density function $f(.,.)$ as $(\boldsymbol{X}, \boldsymbol{Y})$. Also assume that $\boldsymbol{X}$ admits a marginal density $g(.)$. Suppose we are given $n$ observations of $(X,Y)$ denoted by $(\boldsymbol{X}_1,Y_1),\ldots,(\boldsymbol{X}_{n,Yn})$. Of interest is estimating the conditional density of $Y$ given $\boldsymbol{X} = x$, i.e.

$$f(y|\boldsymbol{x}) = \frac{f(\boldsymbol{x}, y)}{g(\boldsymbol{x})}$$

where $g(.)$ is assumed positive at $x$. The conditional density can be a useful statistical tool in several ways. The most obvious need for estimating conditional densities arises when exploring relationships between a response and potential covariates.

**A motivating example:** Consider the bivariate data analysed by AZZALINI and BOWMAN (1990) on the waiting time between the starts of successive eruptions and the duration of the subsequent eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming. The data were collected from August 1st until August

[1]jandeg@fee.uva.nl

[2]zerom@fee.uva.nl

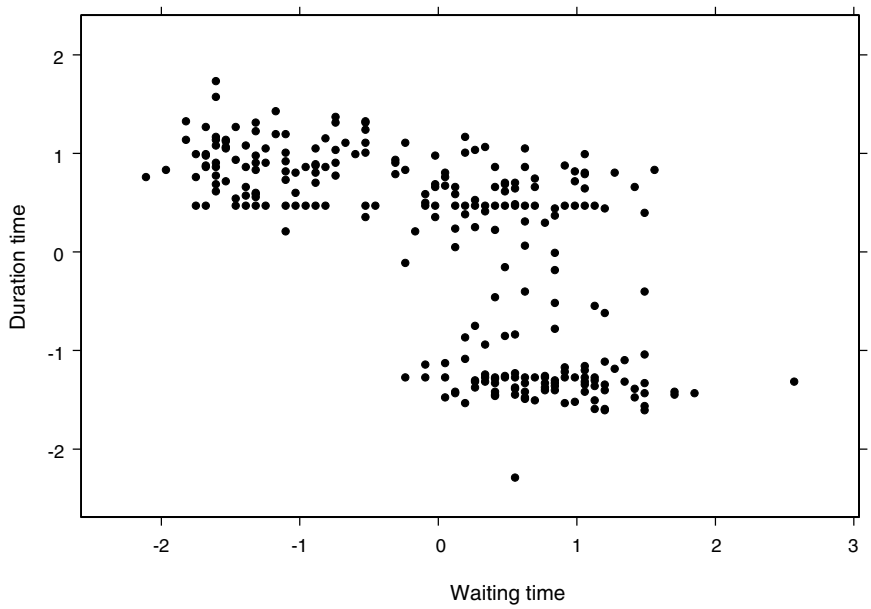


Fig. 1. Duration of eruption plotted against waiting time to eruption.

15th, 1985. There are a total of 299 observations. The times are measured in minutes. In Figure 1 we give a scatter plot of the data. Note that both variables are transformed to have mean zero and variance one. From the plot it is clear that when there has been a relatively short waiting time between eruptions, the duration of the next eruption is relatively long. But, when the waiting time between eruptions is longer than about −0.17 (or 70 minutes in the scale of the untransformed data), the duration of the next eruption is more or less a mixture of short and long durations. This interesting observation can be nicely summarized by the conditional density.

Figure 2 gives the estimated conditional density. Notice that when the waiting time to eruption is more than −0.17, the conditional density of eruption duration conditional on waiting time to eruption is bimodal. On the other hand, for waiting times below −0.17, the conditional density is unimodal. To appreciate visually how the shape of the conditional density evolves across the various values of the waiting time to eruption, the estimated conditional densities in Figure 2 are stacked side-by-side. This stacked conditional density plot is produced using HYNDMAN's (1996) S-Plus code which is freely available at the website: www-personal.buseco.monash.edu.au/hyndman.

Having recognized the role conditional densities could play in data analysis, the purpose of the present paper is to suggest a nonparametric estimator of the conditional density, $f(y|x)$. In particular, our suggestion adapts the conditional distribution smoother of HALL, WOLFF and YAO (1999). CAI (2001, 2002) also extended the smoother of Hall et al. (1999) to contexts other than the distribution function mainly to conditional quantiles.

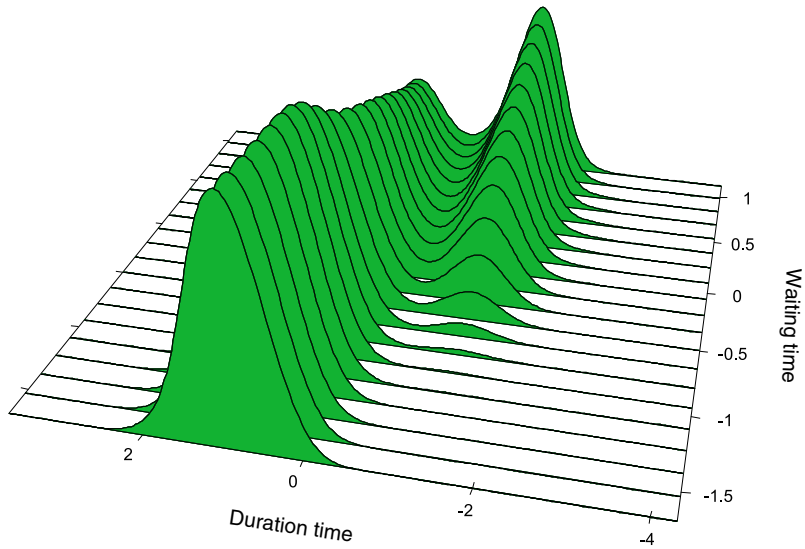


Fig. 2. Conditional density estimates of eruption duration conditional on the waiting times to eruption.

The plan of the paper is as follows. In Section 2 we introduce the RNW conditional density smoother. In the same section, two existing smoothers are also discussed so as to motivate the proposed smoother. In Section 3 we study the asymptotic behaviour of the suggested smoother. We close in Section 4 with some concluding remarks. Technical arguments and proofs are collected in the Appendix.

## 2 Methods

To help motivate the construction of the proposed conditional density estimator, we first discuss two existing kernel-based smoothers of the conditional density. To simplify the presentation, we shall consider the case of $d = 1$ throughout the paper.

### 2.1 *Nadaraya-Watson (NW) and local linear*

Let the kernel $K(.)$ be a symmetric density function on $\mathbb{R}$. Let $h_{1,n}$ and $h_{2,n}$ denote bandwidths (or smoothing parameters). As $h_{1,n} \to 0$, it is easy to see from a standard Taylor argument that

$$E\{K_{h_{1,n}}(y - Y)|X = x\} \simeq f(y|x)$$

where $K_{h_n}(.) = K(./h_n)/h_n$. This suggests that the estimation of $f(y|x)$ can be viewed as a nonparametric regression of $K_{h_n}(y - Y_i)$ on $\{X_i\}$. In fact, it is based on this particular idea that the well-known Nadaraya-Watson kernel smoother, here denoted by $\hat{f}_{NW}(y|x)$, was first proposed by ROSENBLATT (1969) and later extended by HYNDMAN, BASHTANNYK and GRUNWALD (1996). According to HYNDMAN et al. (1996), $\hat{f}_{NW}(y|x)$ is defined as

$$\hat{f}_{NW}(y|x) = \sum_{i=1}^{n} K_{h_{1,n}}(y - Y_i) w_i^{NW}(x) \tag{1}$$

where

$$w_i^{NW}(x) = \frac{K_{h_{2,n}}(x - X_i)}{\sum_{i=1}^{n} K_{h_{2,n}}(x - X_i)}.$$

Now suppose that the second derivative of $f(y|x)$ exists. Also introduce the shorthand notation, $f^{(i,j)}(y|x) = \partial^{i+j} f(y|x)/\partial x^i \partial y^j$. In a small neighbourhood of a point $x$, we can approximate $f(y|z)$ locally by a linear term

$$f(y|z) \simeq f(y|x) + f^{(1,0)}(y|x)(z - x) \equiv a + b(z - x).$$

In this sense, one can also regard the estimation of $f(y|x)$ as a nonparametric weighted regression of $K_{h_{1,n}}(y - Y_i)$ against $(1, (X_i - x))$ using weights $K_{h_{2,n}}(x - X_i)$. Accordingly, Fan, Yao and Tong (1996) proposed the so-called local linear smoother of $f(y|x)$. The local linear estimator, here denoted by $\hat{f}_{LL}(y|x)$, is defined as $\hat{a}(x)$, where $(\hat{a}, \hat{b})$ minimize

$$\sum_{i=1}^{n} \big(K_{h_{1,n}}(y - Y_i) - a - b(X_i - x)\big)^2 K_{h_{2,n}}(x - X_i).$$

Simple algebra (see Fan and Gijbels, 1996) shows that $\hat{f}_{LL}(y|x)$ can be expressed as

$$\hat{f}_{LL}(y|x) = \sum_{i=1}^{n} K_{h_{1,n}}(y - Y_i) w_i^{LL}(x),$$

where

$$w_i^{LL}(x) = \frac{K_{h_{2,n}}(x - X_i)\{T_{n,2} - (X_i - x)T_{n,1}\}}{(T_{n,0}T_{n,2} - T_{n,1}^2)}$$

with $T_{n,j} = \sum_{i=1}^{n} K_{h_{2,n}}(x - X_i)(X_i - x)^j$ $(j = 0, 1, 2)$.

From the definition of the two estimators, we can see that while $\hat{f}_{NW}(y|x)$ approximates $f(y|x)$ locally by a constant, say $a$, $\hat{f}_{LL}(y|x)$ approximates $f(y|x)$ locally by a linear model. To appreciate why the extension of the local constant fitting to the local linear alternative is interesting, we now compare the two estimators via their respective moments. To keep the presentation simple, we assume in the rest of the paper, without loss of generality, that $h_{1,n} = h_{2,n} = h_n$. When the process $\{(X_i, Y_i)\}$ is $\alpha$-mixing (see Section 3 for a definition of $\alpha$-mixing), Chen, Linton and Robinson (2001) showed the approximate asymptotic bias and variance of $\hat{f}_{NW}(y|x)$ to be

$$Bias(\hat{f}_{NW}(y|x)) = \frac{1}{2} k_1 h_n^2 \left[ f^{(2,0)}(y|x) + f^{(0,2)}(y|x) + 2\frac{g'(x)}{g(x)} f^{(1,0)}(y|x) \right], \tag{2}$$

and

$$Var(\hat{f}_{NW}(y|x)) = k_2^2 (n h_n^2)^{-1} \frac{f(y|x)}{g(x)} \tag{3}$$

where $k_1 = \int u^2 K(u) \mathrm{d}u$ and $k_2 = \int K^2(u) \mathrm{d}u$. Similarly, under $\rho$-mixing, Fan et al. (1996) gave approximate asymptotic bias and variance of $\hat{f}_{LL}(y|x)$, i.e.

$$Bias(\hat{f}_{LL}(y|x)) = \frac{1}{2}k_1 h_n^2 \left[ f^{(2,0)}(y|x) + f^{(0,2)}(y|x) \right], \tag{4}$$

$$Var(\hat{f}_{LL}(y|x)) = k_2^2 (nh_n^2)^{-1} \frac{f(y|x)}{g(x)}. \tag{5}$$

Some remarks about the above asymptotic bias and variance expressions are in order. We see that the two variances are identical. Therefore, the difference in the asymptotic mean squared errors (MSEs) between the two estimators depends only on their respective biases. Note that the bias of $\hat{f}_{NW}(y|x)$ has an extra term $(g'(x)/g(x))f^{(1,0)}(y|x)$. The bias of $\hat{f}_{NW}(y|x)$ is large if either $|g'(x)/g(x)|$ or $|f^{(1,0)}(y|x)|$ is large, but neither term appears in (4). For example, when the design density is highly clustered, the term $|g'(x)/g(x)|$ becomes large. Of course, when $g(x)$ is uniform the biases of the two estimators are the same. Thus, the fact that $\hat{f}_{LL}(y|x)$ does not depend on the density of $X$ makes it "design adaptive" (see FAN, 1992). Now, let's consider $|f^{(1,0)}(y|x)|$. For simplicity, suppose that the conditional density of $Y$ depends on $x$ only through a location parameter, say the conditional mean (denoted here by $m(x)$) and hence $f(y|x) = f(y-m(x))$. Then $f^{(1,0)}(y|x) = m^{(1)}(x)f^{(1,0)}(y-m(x)|x)$ where $m^{(1)}(.)$ denotes the first derivative of $m(.)$. In this set-up when, for example, $m(x)$ is linear $m(x) = a + bx$ with large coefficient $b$, the bias of $\hat{f}_{NW}(y|x)$ gets large. But, when $m(x)$ is flat or has maximum or minimum, or inflection point at $x$, the biases of the two estimators become the same.

The above theoretical comparisons suggest that the local linear estimator is more attractive than the local constant alternative because of its better bias performance and design adaptation. It is also possible to show that both in the interior and near the boundary of the support of $g(.)$, the asymptotic bias and the variance of $\hat{f}_{LL}(y|x)$ are of the same order of magnitude. On the other hand, $\hat{f}_{NW}(y|x)$ has a bias of order $h_n$ for $x$ in the boundary. So, at least in theory, the local linear smoother does not suffer from boundary effects and hence does not require modifications at the boundaries.

Although the local linear approach is more efficient in the sense already discussed, the smoother $\hat{f}_{LL}(y|x)$ may give conditional density function estimates that are not constrained to be nonnegative. On the other hand, $\hat{f}_{NW}(y|x)$ always gives nonnegative estimates. With these remarks in mind, we pass on to the suggestion of this paper.

### 2.2 *RNW estimator*

Now we introduce a simple kernel smoother called RNW which combines the better sides of the LL and NW smoothers. In other words, while sharing the nice sampling properties of the LL estimator, it is always nonnegative.

From least squares theory, it is easy to see that the local linear weights $w_i^{LL}(x)$ satisfy: $\sum_{i=1}^n (X_i - x)w_i^{LL}(x) = 0$. But for the Nadaraya-Watson weights $w_i^{NW}(x)$, this moment condition is not fulfilled. It is this observation that motivates the

introduction of the RNW smoother. Let $\tau_i(x)$ denote probability like weights with properties that $\tau_i(x) \geqslant 0$, $\sum_{i=1}^n \tau_i(x) = 1$, and

$$\sum_{i=1}^{n} \tau_i(x)(X_i - x)K_{h_n}(x - X_i) = 0. \tag{6}$$

Note how $\tau_i(x)$ is introduced to force the Nadaraya-Watson weights $w_i^{NW}(x)$ to resemble that of $w_i^{LL}(x)$ (see also HALL and PRESNELL, 1999). Following similar arguments as in OWEN (1988), we look for the unique solution of $\tau_i(x)$ by maximizing $\sum_{i=1}^n \log\{\tau_i(x)\}$ subject to the above constraints via Lagrange multipliers, i.e.

$$G = \sum_{i=1}^{n} \log\{\tau_i(x)\} + \kappa\left(1 - \sum_{i=1}^{n} \tau_i(x)\right) - n\lambda \sum_{i=1}^{n} \tau_i(x)(X_i - x)K_h(x - X_i).$$

Setting $\partial G/\partial \tau_i(x) = 0$, one obtains $\tau_i(x) = 1/\{\kappa + n\lambda(X_i - x)K_{h_n}(x - X_i)\}$. But, just summing $\partial G/\partial \tau_i(x)$ and employing (6), we can see that $\kappa = n$. Hence,

$$\tau_i(x) = n^{-1}\{1 + \lambda(X_i - x)K_{h_n}(x - X_i)\}^{-1}. \tag{7}$$

Now we show that $|\lambda| \leqslant O_p(h_n)$. This is a useful intermediate result in studying the asymptotic theory of the RNW smoother. Let $v_i = (X_i - x)K_{h_n}(x - X_i)$. Then from (6) and (7),

$$n^{-1} \sum_{i=1}^{n} v_i\{1 + \lambda v_i\}^{-1} = 0.$$

Rewriting this

$$\begin{aligned} 0 &= n^{-1} \sum_{i=1}^{n} \{\lambda v_i^2 (1 + \lambda v_i)^{-1} - v_i\} \\ &= n^{-1} \left| \sum_{i=1}^{n} \{\lambda v_i^2 (1 + \lambda v_i)^{-1} - v_i\} \right| \\ &\geqslant |\lambda n^{-1} \sum_{i=1}^{n} v_i^2 (1 + \lambda v_i)^{-1}| - |\bar{v}_1| \end{aligned}$$

where $\bar{v}_1 = n^{-1} \sum_{i=1}^n v_i$. But notice that

$$|1 + \lambda v_i|^{-1} \geqslant (1 + |\lambda| \max(|v_i|))^{-1}.$$

Thus continuing,

$$0 \geqslant |\lambda|(1 + |\lambda| C_2)^{-1} \bar{v}_2 - |\bar{v}_1|,$$

where $\bar{v}_2 = n^{-1} \sum_{i=1}^n v_i^2$ and $C_2$ denotes the upper bound to $v_i$. Hence,

$$|\lambda|(1 + C_2|\lambda|)^{-1} \bar{v}_2 \leqslant |\bar{v}_1|.$$

This implies,

$$|\lambda| \leqslant \frac{|\bar{v}_1|}{\bar{v}_2 - C_2|\bar{v}_1|}.$$

Now, by standard Taylor expansion, it follows easily that

$$\bar{v}_1 = O_p(h_n^2), \quad \text{and} \quad \bar{v}_2 = O_p(h_n).$$

Therefore, $|\lambda| \leqslant O_p(h_n)$.

**Definition and computation.** The RNW smoother looks very much like that of NW smoother. The only difference is that it involves re-weighting the NW weights by $\tau_i(x)$. The role of $\tau_i(x)$ is to adjust the NW weights such that resulting conditional density estimates resemble that from the LL smoother. We define the RNW conditional density estimator as follows

$$\hat{f}_{RNW}(y|x) = \sum_{i=1}^{n} K_{h_n}(y - Y_i) w_i^{RNW}(x) \tag{8}$$

where

$$w_i^{RNW}(x) = \frac{\tau_i(x) K_{h_n}(x - X_i)}{\sum_{i=1}^{n} \tau_i(x) K_{h_n}(x - X_i)}.$$

From computational perspective the RNW smoother is easy to implement. To see that, let's substitute (7) into (6). Upon doing this, we obtain

$$0 = \sum_{i=1}^{n} \frac{(X_i - x) K_{h_n}(x - X_i)}{1 + \lambda(X_i - x) K_{h_n}(x - X_i)} \equiv g(\lambda).$$

Now notice that $-g(.)$ is just the gradient with respect to $\lambda$ of

$$L(\lambda) = -\sum_{i=1}^{n} \log\{1 + \lambda(X_i - x) K_{h_n}(x - X_i)\}.$$

So a zero of $g(.)$ is a stationary point of $L(.)$. The implication is that, in practice, one can compute $\lambda$ as the unique minimizer of $L(.)$. Our experience suggests that a line search algorithm is a suitable choice to compute $\lambda$. The conditional densities displayed in Section 1 are computed via the RNW smoother.

## 3 Asymptotic behaviour

In this section, our aim is to study the asymptotic properties of the RNW conditional density estimator $\hat{f}_{RNW}(y|x)$ under a reasonably weak mixing condition. In particular, we consider the so-called strong mixing ($\alpha$-mixing). This mixing condition ensures an asymptotically vanishing memory of the strictly stationary process. The $\alpha$-mixing condition (Rosenblatt, 1956) is satisfied if there exists a sequence of nonnegative numbers called mixing coefficients ($\alpha(k)$) such

that $\lim_{k\to\infty}\alpha(k)=0$ and for any $A$ in $\mathcal{F}_1^n=\sigma\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ and any set $B$ in $\mathcal{F}_{n+k}^{\infty}=\sigma\{(X_{n+k},Y_{n+k}),\ldots\}$, we have $|P(A\cap B)-P(A)P(B)|\leqslant\alpha(k)$.

The $\alpha$-mixing condition is weaker than many other mixing modes and dependence conditions, for example, $m$-dependent, $\phi$-mixing, absolute regular, and $\rho$-mixing. Further, it is known that $\alpha$-mixing is fulfilled for many stochastic processes, including many time series models. For example, under mild assumptions, linear AR and bilinear time series models are strongly mixing, with mixing coefficients decaying exponentially. For more details on mixing conditions, we refer the interested reader to, for example, Roussas and Ioannides (1987).

Before we state the results of this paper, we first provide a list of regularity conditions that are useful in asymptotic theory of the RNW smoother. For brevity, theoretical results will be given for $x$ in the interior of the support of $X$. All through this paper $C$ will denote a generic constant.

A.1 The kernel $K(.)$ is a symmetric and bounded probability density function such that $|u|K(u)\to 0$ as $|u|\to\infty$ and $\int u^2K(u)du<\infty$.

A.2

(i) The marginal density $g(x)$ is continuous and is bounded from below by a positive constant.

(ii) The function $f(y|x)$ has bounded continuous second order derivative with respect to $x$ at $(x,y)$.

A.3 The joint conditional density $f_{(Y_1,Y_j)|(X_1,X_j)}$ of $(Y_1,Y_j)$ given $(X_1,X_j)$ satisfies, for all $j>1$ and all values of arguments involved,

$$f_{(Y_1,Y_j)|(X_1,X_j)}(y_1,y_j|u,v)\leqslant C<\infty.$$

A.4

As $n\to\infty$, $h_n\to 0$ and $nh_n^2\to\infty$.

A.5

(i) There exists a sequence of positive integers $\{d_n\}$ such that $d_n\to\infty$ and $d_nh_n^2\to 0$.

(ii) For some constant $\delta$, $0<\delta<1$, and $a>\delta$, $\sum_{j=1}^{\infty}j^a\alpha^{\delta}(j)<\infty$.

A.6 Assume that there exists a sequence of positive integers, $q_n$ such that $q_n\to\infty$, $q_n=o((nh_n^2)^{1/2})$, and $(n/h_n^2)^{1/2}\alpha(q_n)\to 0$ as $n\to\infty$.

Remark 1. We provide a sufficient condition for the mixing coefficient $\alpha(n)$ to satisfy Conditions A.5(ii) and A.6. Suppose that $h_n=An^{-0.5\rho}$ $(0<\rho<1,\ A>0)$, $q_n=(nh_n^2/\log n)^{1/2}$ and $\alpha(n)=O(n^{-\theta})$ for some $\theta>0$. Note that such choice of $\alpha(n)$ encompasses a large class of strongly mixing random variables with mixing coefficients decaying moderately fast. Then Condition A.5(ii) is satisfied for $\theta>(\delta+1)/\delta$ and Condition A.6 is satisfied if $\theta>(1+\rho)/(1-\rho)$. Hence, both conditions are satisfied if

$$\alpha(n) = O(n^{-\theta}), \quad \text{and} \quad \theta > \max\left\{\frac{1+\rho}{1-\rho}, \frac{\delta+1}{\delta}\right\}.$$

THEOREM 1. *Assume that Conditions* A.1–A.6 *are satisfied and suppose that* $nh_n^6 \to c$ *for some* $c \neq 0$. *Then, as* $n \to \infty$, *we have*

*(i)*

$$\hat{f}_{RNW}(y|x) - f(y|x) = \text{Bias}(\hat{f}_{RNW}(y|x)) + O_p((nh_n^2)^{-1/2}),$$

*(ii)*

$$(nh_n^2)^{1/2}\left[\hat{f}_{RNW}(y|x) - f(y|x) - \text{Bias}(\hat{f}_{RNW}(y|x))\right] \xrightarrow{D} \mathcal{N}\left(0, \frac{k_2^2 f(y|x)}{g(x)}\right)$$

$$\text{Bias}(\hat{f}_{RNW}(y|x)) = \frac{1}{2}k_1 h_n^2[f^{(2,0)}(y|x) + f^{(0,2)}(y|x)]. \tag{9}$$

REMARK 2. From Theorem 1 (i), it may be seen that to the first order, the RNW smoother enjoys the same convergence rates as the LL smoother of FAN et al. (1996). However, they employed the $\rho$-mixing condition which is stronger than the $\alpha$-mixing.

REMARK 3. From Theorem 1 (ii), the asymptotic variance is given as

$$Var(\hat{f}_{RNW}(y|x)) = k_2^2(nh_n^2)^{-1}\frac{f(y|x)}{g(x)}.$$

Note that to the first order, $\hat{f}_{RNW}(y|x)$ matches both the bias and the variance of the local linear smoother $\hat{f}_{LL}(y|x)$ (see (4)). Thus, the RNW smoother shares the better bias behaviour of the LL smoother.

REMARK 4. If one chooses the optimal bandwidth, say $h_n^*$, such that it minimizes the asymptotic MSE of $\hat{f}_{RNW}(y|x)$, it is easy to see that

$$h_n^* = Bn^{-1/6}$$

where $B$ is a function of some unknowns such as $f(y|x)$. In practice, $B$ may be replaced by consistent estimates in order to construct a feasible, approximately optimal bandwidth. Unlike the $n^{-1/5}$ rate familiar from the univariate density estimation, notice that $h_n^* \sim n^{-1/6}$ as one needs to smooth in both $x$ and $y$ directions.

REMARK 5. As seen in Remark 4, the optimal bandwidth choice under the MSE criterion gives $h_n$ satisfying $nh_n^6 \to c$. If $nh_n^6 \to 0$, the bias will be negligible, the

asymptotic MSE will be dominated by the variance and hence we are not in the optimal case. So the imposed condition $nh_n^6 \to c$ allows us to give an asymptotic normality theorem under optimal conditions of convergence.

## 4 Concluding remarks

In defining the RNW smoother, we have used the same bandwidth $h_n$ in both $x$ and $y$ directions, i.e. $h_{1,n} = h_{2,n} = h_n$. As was mentioned, our use of a single bandwidth is aimed at simplifying the theoretical results of the paper. In practice there may indeed arise a need to have different levels of smoothing for each direction. For example, in the eruption-waiting time illustration, it is not advisable to have the same bandwidth for both variables because they have different levels of variability. In fact that was the reason for standardizing the variables before using a single bandwidth for both. If the approach of pre-standardizing the data is found inadequate, the RNW smoother can easily be re-defined to involve two bandwidths.

To help appreciate the value-added of the RNW smoother in practical applications, studying the finite sample size performance of the RNW smoother using Monte Carlo methods is crucial. This would include the investigation of the accuracy (in some sense) of the RNW smoother vis-à-vis existing conditional density smoothers such as the traditional Nadaraya-Watson, the local linear smoother, and so on. This paper does not go into such empirical comparison, it remains a likely topic for future investigations.

In conclusion we like to note that HYNDMAN and YAO (2002) also introduced two alternative kernel smoothers of the conditional density, both aimed at removing negativity. Apart from the computation of $\lambda$, which can be done independently, the RNW smoother is explicitly defined in terms of data observations. In this sense, our approach is computationally more feasible than that of HYNDMAN and YAO (2002).

## Appendix: Proofs

Throughout the proof we re-denote $\hat{f}_{RNW}(y|x)$ by $\hat{f}(y|x)$. In the course of the proof of the theorem, we also derive some auxiliary results which are necessary to establish the theorem.

*Proof of the theorem*

The first step in the proof is to get an arbitrary good approximation to the value of $\lambda$. Recall that $|\lambda| \leqslant O_p(h_n)$. After replacing $\tau_i(x)$ by (7), we Taylor expand (6) about $\lambda = 0$. This gives

$$\lambda = \frac{h_n k_1 g'(x)}{k_3 g(x)} + o_p(h_n) \tag{10}$$

where $k_3 = \int u^2 K^2(u)\mathrm{d}u$. Now substituting (10) into (7),

$$\tau_i(x) = n^{-1} b_i(x)(1 + o_p(1)) \tag{11}$$

where

$$b_i(x) = \left(1 + \frac{h_n k_1 g'(x)}{k_3 g(x)}(X_i - x)K_{h_n}(x - X_i)\right)^{-1}.$$

Let $m(x,y) = E\{K_{h_n}(y-Y)|X = x\}$. Also define $\varepsilon_i = K_{h_n}(x-X_i) - m(X_i,y)$. Using (8) and (11),

$$\begin{aligned}\hat{f}(y|x) - f(y|x) &= \frac{n^{-1}\sum_{i=1}^n [\varepsilon_i + m(X_i,y) - f(y|x)] b_i(x) K_{h_n}(x-X_i)}{n^{-1}\sum_{i=1}^n b_i(x) K_{h_n}(x-X_i)} \left\{1 + o_p(1)\right\} \\ &\equiv \{(nh_n^2)^{-1/2} J_1 + J_2\} J_3^{-1}\{1 + o_p(1)\}\end{aligned} \tag{12}$$

where

$$\begin{aligned} J_1 &= h_n n^{-1/2} \sum_{i=1}^n b_i(x)\varepsilon_i K_{h_n}(x - X_i), \\ J_2 &= n^{-1} \sum_{i=1}^n [m(X_i, y) - f(y|x)] b_i(x) K_{h_n}(x - X_i), \quad \text{and} \\ J_3 &= n^{-1} \sum_{i=1}^n b_i(x) K_{h_n}(x - X_i). \end{aligned}$$

By Condition A.2(ii) and (A.6) as well as the Taylor expansion, $J_2$ becomes

$$J_2 = \frac{1}{2} k_1 h_n^2 g(x) \left\{ f^{(2,0)}(y|x) + f^{(0,2)}(y|x) \right\} + o_p(h_n^2).$$

Similar manipulation applied to $J_3$ gives $J_3 = g(x) + o_p(1)$. Substituting the evaluated $J_2$ and $J_3$, (12) becomes

$$\begin{aligned}&(nh_n^2)^{1/2}\left[\hat{f}(y|x) - f(y|x) - Bias(\hat{f}(y|x)) + o_p(h_n^2)\right] \\ &\quad = g^{-1}(x) J_1 + o_p(1)\end{aligned} \tag{13}$$

where $Bias(\hat{f}(y|x))$ is as defined in (9). Note that since the condition $nh_n^6 \to c$ implies $(nh_n^2)^{1/2} o_p(h_n^2) = o_p(1)$, (13) reduces to

$$(nh_n^2)^{1/2}\left[\hat{f}(y|x) - f(y|x) - Bias(\hat{f}(y|x))\right] = g^{-1}(x) J_1 + o_p(1). \tag{14}$$

To deal with $J_1$, we evaluate $E(J_1)$ and $Var(J_1)$. Set $\Delta_i = h_n \varepsilon_i b_i(x) K_{h_n}(x-X_i)$, then

$$J_1 = n^{-1/2} \sum_{i=1}^{n} \Delta_i.$$

Note that $E(\Delta_i) = 0$. Thus $E(J_1) = 0$. Exploiting stationarity

$$Var(J_1) = E(\Delta_1^2) + 2 \sum_{j=2}^{n} \left(1 - \frac{j-1}{n}\right) E(\Delta_1 \Delta_j). \tag{15}$$

From routine calculations, it follows that $E(\Delta_1^2) = g(x)k_2^2 f(y|x) + o_p(1)$. It remains to evaluate the second term of (15). For notational convenience, we shall denote this term by $B$. We follow the technique by Masry (MASRY, 1986). Namely, define the sets $S_1$, and $S_2$ by

$$S_1 = \{(1, j) : j \in \{1, \ldots, n\}, \quad 1 \leqslant j - 1 \leqslant d_n\}$$
$$S_2 = \{(1, j) : j \in \{1, \ldots, n\}, \quad d_n + 1 \leqslant j - 1 \leqslant n - 1\}$$

where $\{d_n\}$ is as defined in Condition A.5(i). From the above splitting, notice that

$$B = 2 \sum_{j \in S_1} \left(1 - \frac{j-1}{n}\right) E(\Delta_1 \Delta_j) + 2 \sum_{j \in S_2} \left(1 - \frac{j-1}{n}\right) E(\Delta_1 \Delta_j). \tag{16}$$

Consider the first term on the right-hand side of $B$ (or (16)), i.e.

$$\begin{aligned} 2 \sum_{j \in S_1} \left(1 - \frac{j-1}{n}\right) E(\Delta_1 \Delta_j) &\leqslant \sum_{j=1}^{d_n} |E(\Delta_1 \Delta_j)| \\ &\leqslant \sum_{j=1}^{d_n} C h_n^2, \text{ by Lemma 1 } (ii) \text{ (in the Auxiliary results)} \\ &= C d_n h_n^2 = o(1). \end{aligned} \tag{17}$$

The last step follows from Condition A.5(i). For the second term on the right-hand side of (16), note that

$$2 \sum_{j \in S_2} \left(1 - \frac{j-1}{n}\right) E(\Delta_1 \Delta_j) \leqslant \sum_{j \in S_2} |E(\Delta_1 \Delta_j)|.$$

Applying DAVYDOV's (1970) inequality, we have that

$$\sum_{j \in S_2} |E(\Delta_1 \Delta_j)| \leqslant \sum_{j \in S_2} 8 \alpha^{\delta}(j-1) E^{1/s}(|\Delta_1|^s) E^{1/t}(|\Delta_1|^t)$$

where $s > 1$, $t > 1$ and $1/s + 1/t = 1 - \delta$ with $\delta$ is as defined in Condition A.5(ii). Now setting $s = t = \ell$,

$$\sum_{j\in S_2}|E(\Delta_1\Delta_j)| \leqslant \sum_{j\in S_2} 8\alpha^{\delta}(j-1)E^{(1-\delta)}(|\Delta_1|^{\ell})$$
$$\leqslant 8\sum_{j\in S_2}\alpha^{\delta}(j-1)(Ch_n)^{1-\delta}, \text{ by Lemma 1 } (i) \text{ (in the Auxiliary results)}$$
$$\leqslant Ch_n^{1-\delta}d_n^{-a}\sum_{j=d_n}^{\infty} j^a\alpha^{\delta}(j).$$

Let's choose $d_n$ such that $h_n^{\delta}d_n^a = O(h_n)$. Then, using Condition A.5(ii), we have $\sum_{j\in S_2}|E(\Delta_1\Delta_j)| = o(1)$. Observe that under the above choice of $d_n$, the condition $d_n h_n^2 \to 0$ is satisfied. Substituting the above evaluated terms into (15), we can see that

$$Var(J_1) \to g(x)k_2^2 f(y|x). \tag{18}$$

Finally, recalling (14),

$$\hat{f}(y|x) - f(y|x) = Bias(\hat{f}(y|x)) + O_p((nh_n^2)^{-1/2}).$$

This completes the proof of the first part of the theorem.

Denoting $Var(J_1)$ by $\sigma^2(x,y)$, we now move to the second part of the theorem, i.e. to show that the left-hand side of (14) is asymptotically normally distributed. To achieve this, it is sufficient to establish that $J_1$ is $\mathcal{N}(0,\sigma^2(x,y))$ distributed. For the proof we make use of Doob's technique (see Doob, 1953, pp. 228–232) according to which the sum $\sum_{i=1}^{n}\Delta_i$ is split into large and small blocks. Specifically, we partition $\{1,\ldots,n\}$ into $2r_n+1$ subsets with large block of size $p_n$ and small block of size $q_n$. Set

$$r_n = \left[\frac{n}{p_n+q_n}\right]$$

where [.] denotes the integer part. Thus, we can write $J_1$ as,

$$J_1 = n^{-1/2}\sum_{i=1}^{n}\Delta_i = n^{-1/2}\{S_{1,n} + S_{2,n} + S_{3,n}\}$$

where

$$S_{1,n} = \sum_{j=1}^{r_n}\eta_j, \quad S_{2,n} = \sum_{j=1}^{r_n}\epsilon_j, \quad S_{3,n} = \omega_{r_n}$$

with

$$\eta_j = \sum_{i=k_j}^{k_j+p_n-1}\Delta_i, \quad \text{where} \quad k_j = (j-1)(p_n+q_n)+1,$$

$$\epsilon_j = \sum_{i=l_j}^{l_j+q_n-1}\Delta_i, \quad \text{where} \quad l_j = (j-1)(p_n+q_n)+p_n+1,$$

$$\omega_{r_n} = \sum_{i=r_n(p_n+q_n)+1}^{n}\Delta_i.$$

Before continuing with the proof, we first show some consequences of Condition A.6. This condition implies that there is a sequence of positive constants $\beta_n \to 0$ such that

$$\beta_n q_n = o((nh^2)^{1/2}) \quad \text{and} \quad \beta_n (n/h^2)^{1/2} \alpha(q_n) \to 0. \tag{19}$$

Now define $p_n$ by $p_n = [(nh^2)^{1/2}/\beta_n]$. Then it follows easily from (19) that as $n \to \infty$ ,

$$q_n/p_n \to 0, \quad p_n/n \to 0, \quad p_n(nh^2)^{-1/2} \to 0, \quad \text{and} \qquad (n/p_n)\alpha(q_n) \to 0. \tag{20}$$

Now we exploit Lemma 2. This Lemma tells us that $S_{2,n}$ and $S_{3,n}$ are asymptotically negligible. Then showing the asymptotic normality of $J_1$ reduces to proving that $n^{-1/2}S_{1,n}$ converges to $\mathcal{N}(0,\sigma^2(x,y))$. The main idea of the proof is to approximate $n^{-1/2}S_{1,n}$ by a sum of independent random variables (r.v.'s). For each $n$, let $z_{n,1},\ldots,z_{n,r}$ denote independent r.v.'s with the distribution that of

$$n^{-1/2}\eta_1 = n^{-1/2}\sum_{j=1}^{p_n} \Delta_j.$$

Then, the characteristic function (cf.) of $\sum_{m=1}^{r} z_{n,m}$ is $\Phi_{p_n}^{r_n}(tn^{-1/2})$, where $\Phi_{p_n}(tn^{-1/2})$ is the cf. of $n^{-1/2}\eta_1$. Notice that $\eta_a$ is $\mathcal{F}_{i_a}^{j_a}$-measurable with $i_a = (a-1)(p_n+q_n)+1$ and $j_a = i_a + p_n - 1$. Let $V_j = \exp(it\eta_j/\sqrt{n})$, then using Lemma 4,

$$\begin{aligned}\left|E[\exp(itn^{-1/2}S_{1,n})] - \Phi_p^r(tn^{-1/2})\right| &= \left|E[\exp(itn^{-1/2}S_{1,n})] - \prod_{j=1}^{r} E[\exp(itn^{-1/2}\eta_j)]\right| \\ &\leqslant 16r\alpha(q_n+1) \to 0.\end{aligned}$$

The last step follows from (20), i.e. $r_n\alpha(q_n) \leqslant (n/p_n)\alpha(q_n) \to 0$. Therefore, it suffices to establish that $\Phi_{p_n}^{r_n}(tn^{-1/2})$ converges to the cf. of the $\mathcal{N}(0,\sigma^2(x,y))$. Equivalently, it would suffice to show that $\sum_{m=1}^{r_n} Z_{n,m}$ is asymptotically $\mathcal{N}(0,1)$, where

$$Z_{n,m} = z_{n,m}/s_n, \quad s_n^2 = \sum_{m=1}^{r_n} E(z_{n,m}^2) = \frac{r_n}{n}E(\eta_1^2).$$

Now $\sum_{m=1}^{r_n} Z_{n,m}$ will converge to $\mathcal{N}(0,1)$ provided that, for every $\varepsilon > 0$,

$$g_n(\varepsilon) = \sum_{m=1}^{r_n} \int_{|x|>\varepsilon} x^2 dF_{n,m}(x) \to 0, \quad \text{as } n \to \infty,$$

where $F_{n,m}(.)$ is the distribution function of $Z_{n,m}$. This is the well-known Lindeberg condition. But, since $F_{n,m}(.)$ is the same for $m = 1,\ldots,r_n$,

$$\begin{aligned} g_n(\varepsilon) &= r_n E\left(Z_{n,1}^2 \mathbf{1}_{\{|\mathbf{Z_{n,1}}| \geqslant \varepsilon\}}\right), \\ &\leqslant \frac{r_n p_n^2 C^2}{n s_n^2} P(|\eta_1| \geqslant \varepsilon\sqrt{n} s_n) \quad \text{since} \quad |\eta_1| \leqslant p_n C, \\ &\leqslant \frac{C^2}{\varepsilon^2}\frac{p_n^2}{n}\frac{1}{s_n^2}. \end{aligned}$$

By Lemma 3, below, $E(\eta_1^2) \to p_n\sigma^2(x,y)$. Further, from (20), it is easy to see that

$p_{nrn}/n \to 1$ (see the proof of Lemma 2, below). Thus, $s_n^2 \to \sigma^2(x,y) \neq 0$. Again from (20), $p_n^2/n \to 0$. Therefore, $g_n(\varepsilon) \to 0$. Hence $\sum_{m=1}^{r_n} Z_{n,m}$ will converge to $\mathcal{N}(0,1)$, or equivalently $\sum_{m=1}^{r_n} z_{n,m}$ will converge to $\mathcal{N}(0,\sigma^2(x,y))$. This completes the proof of the theorem.

**Auxiliary results**

LEMMA 1. *Under the conditions of Theorem* 1,

*(i)* $E(|\Delta_1|^\ell) \leqslant Ch_n$;
*(ii)* $|E(\Delta_1\Delta_j)| \leqslant Ch_n^2$.

PROOF: (i) Recall that $\Delta_1 = h_n\varepsilon_1 b_1(x)K_{h_n}(x-X_1)$. Note that

$$E(|\Delta_1|^\ell) = h_n^\ell \int_{\mathbb{R}}\int_{\mathbb{R}} |\varepsilon_1 b_1(x)K_{h_n}(x-X_1)|^\ell f(X_1,Y_1)\mathrm{d}X_1\mathrm{d}Y_1.$$

Now conditioning on $X_1 = u$, and using Conditions A.1 and A.2(i), we can see that

$$E(|\Delta_1|^\ell) \leqslant Ch_n^\ell \int_{\mathbb{R}} |b_1(x)K_{h_n}(x-u)|^\ell du \leqslant Ch_n.$$

(ii) Clearly

$$E(\Delta_1\Delta_j) = h_n^2 \int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}} \epsilon_1 b_1(x)K_{h_n}(x-X_1)\varepsilon_j \times b_j(x)K_h(x-X_j)f_{(X_1,Y_1,X_j,Y_j)}(x_1,y_1,x_j,y_j)dx_1dy_1dx_jdy_j.$$

Conditioning on $(X_1,X_j) = (u,v)$ and using Condition A.3,

$$\begin{aligned}|E(\Delta_1\Delta_j)| &\leqslant Ch_n^2 \int_{\mathbb{R}}\int_{\mathbb{R}} |\varepsilon_1 b_1(x)K_{h_n}(x-u)\varepsilon_j b_j(x)K_{h_n}(x-v)|f_{(X_1,X_j)}(u,v)\mathrm{d}u\mathrm{d}v \\ &\leqslant Ch_n^2\left(\int_{\mathbb{R}} |b_1(x)K_{h_n}(x-u)|\mathrm{d}u\right)^2 \\ &\leqslant Ch_n^2.\end{aligned}$$

LEMMA 2. *Under conditions of Theorem* 1, $n^{-1}E(S_{2,n}^2) = o(1)$ *and* $n^{-1}E(S_{3,n}^2) = o(1)$.

PROOF: Here we only prove $n^{-1}E(S_{n,2}^2) = o(1)$. The proof of $n^{-1}E(S_{n,3}^2) = o(1)$ can be done along the same lines. Observe that

$$E(S_{n,2}^2) = \sum_{j=1}^{r_n} E(\epsilon_j^2) + 2\sum_{1\leqslant u<j\leqslant r_n} E(\epsilon_u\epsilon_j). \tag{21}$$

Consider the first term on the right-hand side of (21). In the rest of the proof, the stationarity property will be repeatedly used. We can see that

$$\begin{aligned}\sum_{j=1}^{r_n} E(\epsilon_j^2) &= \sum_{j=1}^{r_n}\left[\sum_{i=\ell_j}^{\ell_j+q_n-1} E(\Delta_i^2) + 2\sum_{\ell_j\leqslant m<i\leqslant \ell_j+q_n-1} E(\Delta_m\Delta_i)\right]\\ &= \sum_{j=1}^{r_n}\left[\sum_{i=\ell_j}^{\ell_j+q_n-1} E(\Delta_i^2) + 2q_n\sum_{i=2}^{q_n}\left(1-\frac{i}{q_n}\right)E(\Delta_1\Delta_i)\right]\\ &\leqslant \sum_{j=1}^{r_n}\left[q_nE(\Delta_1^2) + 2q_n\sum_{i=2}^{q_n}\left(1-\frac{i}{q_n}\right)|E(\Delta_1\Delta_i)|\right]\\ &\leqslant \sum_{j=1}^{r_n}\left[q_nE(\Delta_1^2) + 2q_n\sum_{i=2}^{q_n}|E(\Delta_1\Delta_i)|\right]\\ &\leqslant r_nq_n\left(E(\Delta_1^2)+o(1)\right). \end{aligned} \tag{22}$$

The last step follows from applying similar arguments used in deriving the variance of $J_1$. Now we deal with the second term of (21). When $u \neq k$,

$$\sum_{1\leqslant u<j\leqslant r_n} E(\epsilon_u\epsilon_j) = \sum_{u=1}^{r_n}\sum_{k=1}^{r_n}\sum_{i=1}^{q_n}\sum_{j=1}^{q_n} E(\Delta_{u(p_n+q_n)+p_n+i}\Delta_{k(p_n+q_n)+p_n+j}).$$

But since, $|u(p_n+q_n)+p_n+i-(k(p_n+q_n)+p_n+j)| \geqslant p_n$,

$$\begin{aligned} 2\sum_{1\leqslant u<j\leqslant r_n} E(\epsilon_u\epsilon_j) &\leqslant 2\sum_{i=1}^{n-p_n}\sum_{j=i+p_n}^{n}|E(\Delta_i\Delta_j)|\\ &\leqslant 2n\sum_{j=1+p_n}^{n}|E(\Delta_1\Delta_j)|\\ &= o(n). \end{aligned} \tag{23}$$

Note that $\sum_{j=1+p_n}^{n}|E(\Delta_1\Delta_j)| = o(1)$. Now combining (22) and (23),

$$n^{-1}E(S_{n,2}^2) \leqslant n^{-1}\left(r_nq_nE(\Delta_1^2)+o(1)+o(n)\right).$$

But from (19) $r_{nqn}/n \leqslant q_n/(p_n+q_n)$. Further, from (20), $q_n/p_n \to 0$. Therefore, $r_nq_n/n \to 0$. Thus the proof of the lemma is complete.

LEMMA 3. *Under conditions of Theorem* 1,

$$E(\eta_1^2) \to p_n\sigma^2(x,y).$$

PROOF: Recall that $\eta_1 = \sum_{j=1}^{p_n}\Delta_j$. Then

$$E(\eta_1^2) = \sum_{j=1}^{p_n} E(\Delta_j^2) + 2 \sum_{1 \leqslant j < m \leqslant p_n} E(\Delta_m \Delta_j).$$

Proceeding in a similar fashion as in deriving the variance of $J_1$, the lemma follows.

LEMMA 4. (VOLKONSKII and ROZANOV, 1959) *Let $V_1,\ldots,V_L$ be strongly mixing random variables with respect to the $\sigma$-algebras $\mathcal{F}_{i_1}^{j_1}, \ldots, \mathcal{F}_{i_L}^{j_L}$ respectively with $1 \leqslant i_1 < j_1 < i_2 < \ldots < j_L \leqslant n$, $i_{l+1} - j_l \geqslant w \geqslant 1$ and $|V_j| \leqslant 1$ for $j = 1,\ldots,L$.*

*Then*

$$\left| E\left( \prod_{j=1}^{L} V_j \right) - \prod_{j=1}^{L} E(V_j) \right| \leqslant 16(L-1)\alpha(w)$$

*where $\alpha(w)$ is the strongly mixing coefficient.*

**Acknowledgement**

We acknowledge comments from two referees, which were useful in improving the presentation.

**References**

AZZALINI, A. and A.W. BOWMAN, (1990), A look at some data on the Old Faithful geyser, *Applied Statistics* **39**, 357–365.

CAI, Z. (2001), Weighted Nadaraya-Watson regression estimation, *Statistics & Probability Letters* **51**, 307–318.

CAI, Z. (2002), Regression quantiles for time series, *Econometric Theory* **18**, 169–192.

CHEN, X., O.B. LINTON and P. ROBINSON (2001), The estimation of conditional densities, *STICERD Econometrics Discussion Paper*, No. EM/01/415.

DAVYDOV, A. (1970), The invariance principle for stationary processes, *Theory of Probability and its Applications* **15**, 487–498.

DOOB, J. (1953), *Stochastic processes*, John Wiley and Sons, New York.

FAN, J. (1992), Design-adaptive nonparametric regression, *Journal of the American Statistical Association* **87**, 998–1004.

FAN, J. and I. GIJBELS (1996), *Local polynomial modelling and its applications. Monographs on statistics and applied probability 66* Chapman & Hall, London.

FAN, J., Q. YAO and H. TONG (1996), Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems, *Biometrika* **83**, 189–206.

HALL, P. and B. PRESNELL (1999), Intentionally biased bootstrap methods. *Journal of the Royal Statistical Society* B **61**, 143–158.

HALL, P., R.C.L. WOLFF and Q. YAO (1999), Methods for estimating a conditional distribution function. *Journal of the American Statistical Association* **94**, 154–163.

HYNDMAN, R.J. (1996), Computing and graphing highest density regions, *The American Statistician* **50**, 120–126.

HYNDMAN, R.J., D.M. BASHTANNYK and G.K. GRUNWALD (1996), Estimating and visualizing conditional densities, *Journal of Computational and Graphical Statistics* **5**, 315–336.

Hyndman, R.J. and Q. Yao (2002), Nonparametric estimation and symmetry tests for conditional density functions, *Journal of Nonparametric Statistics* **14**, 259–278.

Masry, E. (1986), Recursive probability density estimation for weakly dependent processes, *IEEE Transactions on Information Theory* **32**, 254–267.

Owen, A.B. (1988), Empirical likelihood ratio confidence intervals for a single functional, *Biometrika* **75**, 237–249.

Rosenblatt, M. (1956), A central limit theorem and a strong mixing condition, *Proceedings of The National Academy of Science of The United States of America* **42**, 43–47.

Rosenblatt, M. (1969), Conditional probability density and regression estimators, in: *Multivariate Analysis II*, Academic Press, New York, 25–31.

Roussas, G.G. and D.A. Ioannides (1987), Moment inequalities for mixing sequences of random variables, *Stochastic Analysis and Applications* **5**, 61–120.

Volkonskii, V.A. and Y.A. Rozanov (1959), Some limit theorems for random functions, *Theory of Probability and its Applications* **4**, 178–197.

Received: February 2002. Revised: July 2002.