

Nonparametric Multi-step Ahead Prediction in Time Series Analysis

Rong Chen Lijian Yang Christian Hafner *

February 6, 2002

Abstract

We consider the problem of multi-step ahead prediction in time series analysis using nonparametric smoothing techniques. Forecasting is always one of the main objectives in time series analysis. Recent research has shown that nonlinear time series models have certain advantages in multi-step ahead forecasting. Traditionally, nonparametric k -step ahead least squares prediction for nonlinear AR(d) models is done by estimating $E(X_{t+k} | X_t, \dots, X_{t-d+1})$ via nonparametric smoothing of X_{t+k} on (X_t, \dots, X_{t-d+1}) directly. In this paper we propose a multi-stage nonparametric predictor. We show that the new predictor has smaller asymptotic mean squared error than the direct smoother, though the convergence rate is the same. Hence, the proposed predictor is more efficient. Some simulation results, advice for practical bandwidth selection and a real data example are provided.

Key Words: Improvement ratio, local polynomial, multistage smoothing, optimal bandwidth, sunspot series.

*Rong Chen is professor, Department of Information and Decision Sciences, The University Illinois at Chicago, Chicago, IL 60607. Lijian Yang is associate professor, Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824. Christian Hafner is with Electrabel, Quantitative Analysis, Place de l'Université 16, B-1348 Louvain-la-Neuve and Sonderforschungsbereich 373, Humboldt-Universität zu Berlin, Spandauer Str.1, D-10178 Berlin. All three authors received support from Sonderforschungsbereich (SFB) 373 "Quantifikation und Simulation Ökonomischer Prozesse", Humboldt-Universität zu Berlin. Chen's research is also supported in part by NSF grant DMS 9626113 and Yang's research supported in part by NSF grant DMS 9971186. The authors would like to thank Michael H. Neumann for improving questions and comments. The authors also gratefully acknowledge the insightful comments from Editor David Firth, an Associate Editor, and two referees.

1 Introduction

Forecasting is always an important, if not the most important, objective in time series analysis. It has wide applications in the fields of economics, telecommunication, meteorology, etc. In this paper we consider multi-step ahead prediction, which is very different from and more difficult than one-step ahead prediction, as shown in Tiao and Tsay (1994). For linear models multi-step ahead prediction is relatively easy to perform. However, linear forecasts converge to the stationary mean quickly as the forecasting horizon increases (Box and Jenkins 1976). On the other hand, nonlinear models may possess long term nonlinear properties such as limit cycles. Recent research in nonlinear time series analysis (e.g. Tong, 1990 and Tjøstheim, 1994) has revealed the fact that nonlinear models usually perform better than linear models in multi-step ahead prediction.

With nonlinear parametric models, multi-step ahead predictions are usually done using iterative integration or multiple imputation methods. See Jones (1978), Pemberton (1987) and Tong (1990) for details. Guo, Bai & An (1999) also proposed an iterative integration procedure without noise distribution assumption. These procedures are based on parametric models. Their performance depends heavily on the correctness of the model and the accuracy of the estimated parameters.

Recently, nonparametric methods have drawn much attention in time series analysis. For a review, see Tjøstheim (1994), Györfi et al. (1989) and Härdle et al. (1997). This approach entertains the principle of ‘letting the data to speak for themselves’ and avoids the difficulty of identifying an appropriate parametric model, including the nonlinear functions and the error distributions. The existing nonparametric approaches for least squares multi-step ahead prediction (Robinson, 1983, Auestad and Tjøstheim, 1990 and Härdle and Vieu, 1992) estimate the conditional mean function using direct smoothing techniques. Consider a time series $\{X_t\}_{t=1}^{\infty}$ described by a general nonlinear AR(d) model

$$X_t = f(Y_{t-1}) + \sigma(Y_{t-1})\varepsilon_t \quad (1)$$

with $Y_{t-1} = (X_{t-1}, \dots, X_{t-d})^T$ denoting the predictor variables, f and σ the conditional mean and standard deviation functions, and $\{\varepsilon_t\}_{t=d+1}^{\infty}$ i.i.d. white noise with mean 0 and variance 1 independent of X_1, \dots, X_d . The conditional mean $E\{X_{t+k} | Y_t = (y_1, \dots, y_d)\}$ is then the least squares predictor for k -step ahead prediction. Auestad and Tjøstheim (1990) and Härdle and Vieu (1992) proposed to use the ordinary Nadaraya-Watson (N-W) estimator

$$\tilde{m}_{k,h}(y) = \frac{\sum_{t=d}^{n-k} K_h(y - Y_t) X_{t+k}}{\sum_{t=d}^{n-k} K_h(y - Y_t)} \quad (2)$$

where $y = (y_1, \dots, y_d)^T$ denote the conditioning values, K a kernel function, and the notation

$K_h(y) = h^{-d} \prod_{1 \leq i \leq d} K(y_i/h)$. For local linear estimation of vector AR models, see Härdle et al. (1998).

Note that the direct nonparametric estimation (2) ignores the substantial information about the conditional mean function $E(X_{t+k} | Y_t)$ contained in the intermediate variables $X_{t+1}, \dots, X_{t+k-1}$. In this paper, we propose a nonparametric multi-stage predictor which uses such information. The method is motivated by the following observations.

Consider two-step ahead forecasting under a first order nonlinear AR model $X_t = f(X_{t-1}) + \sigma(X_{t-1})\varepsilon_t$, i.e., setting $d = 1$ and $k = 2$. The least squares 2-step prediction of X_{t+2} given $X_t = x$ is the conditional mean

$$m_2(x) = E[X_{t+2} | X_t = x] = E[f(X_{t+1}) + \sigma(X_{t+1})\varepsilon_{t+2} | X_t = x] = E[f(X_{t+1}) | X_t = x].$$

Ideally, if we knew the function $f(\cdot)$, we would smooth on the pairs $(f(X_{t+1}), X_t)$, $t = 1, \dots, n-2$, to estimate $m_2(x)$. Note that the direct estimator (2) uses the pairs (X_{t+2}, X_t) . Since X_{t+2} is a noisier representative of $f(X_{t+1})$ with $O_p(1)$ error, we can improve the estimation by using a more accurate representative $\hat{f}(X_{t+1})$, where $\hat{f}(\cdot)$ is a nonparametric estimator of the function $f(\cdot)$. Under regularity conditions, we have $\hat{f}(X_{t+1}) - f(X_{t+1}) = o_p(1)$. This observation suggests that the ‘two-stage predictor’ which smoothes the pairs $(\hat{f}(X_{t+1}), X_t)$, performs as well as smoothing the pairs $(f(X_{t+1}), X_t)$.

To illustrate the effect of such two-stage smoothing, consider the following process

$$X_{t+1} = a \sin(bX_t) + \sigma\varepsilon_{t+1}, \quad (3)$$

where ε_t is Gaussian white noise with variance 1, and $a = 1$, $b = \pi/2$, $\sigma = 1$. We simulated a series of length 300 from this model. In Figure 1, the left panel shows the scatterplot of (X_t, X_{t+2}) . The dashed line is the estimated mean function $E(X_{t+2} | X_t)$ using the direct N-W estimator (2) and the solid line is the true mean function. The right panel of Figure 1 shows the scatterplot of $(X_t, \hat{f}(X_{t+1}))$, where \hat{f} is obtained by smoothing X_{t+2} on X_{t+1} (the first stage smoothing). We can see that the variation of $\hat{f}(X_{t+1})$ is much smaller than that of X_{t+2} . Of course the smoothing creates extra bias, but it can be controlled with a proper bandwidth. The dashed line is the estimated mean function by smoothing $(X_t, \hat{f}(X_{t+1}))$ (the second stage smoothing) and the solid line is the true mean function.

Throughout the paper we concentrate on the multi-step ahead prediction problem for the general nonlinear AR(d) model (1). This general model (1) can often be simplified by dropping out lags that are insignificant, and our procedure can be altered accordingly to take advantage of the less complicated model. For the exact identification of the lag structure, see Tschernig and Yang (2000).

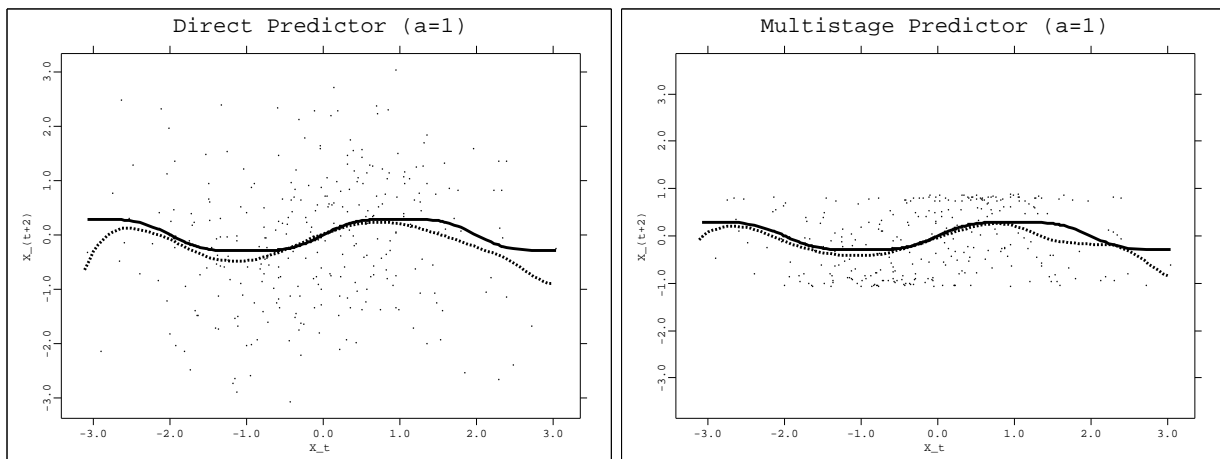


Figure 1: For a sample of $n = 300$ ($a = 1$), both panels show the true function $m_2(x) = \exp(-\pi^2/8) \sin(\pi/2 \sin(\frac{\pi}{2}x))$ (solid line). The left panel shows the direct predictor (dashed line) and the pairs $\{X_t, X_{t+2}\}$. The right panel shows the multistage predictor (dashed line), and the pairs $\{X_t, \hat{f}(X_{t+1})\}$, where \hat{f} is the first stage smoother for $f(x) = \sin(\frac{\pi}{2}x)$.

Chen (1996a,b) studied similar estimators for regression analysis using N-W estimators and showed that the multi-stage smoother does improve the estimation of the conditional mean function. In this paper, we extend the multi-stage smoothing idea to include multivariate predictors and local polynomial estimators, as well as time series instead of independent samples. We demonstrate the improvement in mean squared error of the multi-stage predictor over that of the direct predictor in these general settings.

Note that in the example we gave previously, undersmoothing $\hat{f}(\cdot)$ extremely yields $\hat{f}(X_{t+1}) \approx X_{t+2}$ and one gets back the direct smoother (2). Thus, heuristically, direct smoothing may be considered as a restricted case of two-stage smoothing. By optimizing the amount of smoothing, one can achieve a smaller error without the restriction. This also offers a great deal of flexibility in multi-stage smoothing when deciding whether to skip some stages, see our discussion in Section 4.

We also want to point out that the iterative integration procedures of Jones (1978), Pemberton (1987), Tong (1990) and Guo et al. (1999) can be extended to nonparametrically estimated models. It may be interesting to investigate the cumulating effect of error incurred with nonparametrically estimated functions and estimated empirical error distributions.

The paper is organized as follows. In Section 2, we formally introduce the two-stage predictor and show that it has smaller mean squared error than the direct predictor. Results are derived for both the N-W estimator and the local polynomial estimator. In Section 3, we investigate

implementation issues such as automatic selection of the bandwidth in finite samples and provide simulation evidence. Section 4 deals with multi-stage ($k > 2$) predictors. Results are provided for the performance of the multi-stage predictor with different numbers of iterations. To demonstrate the finite sample properties of the proposed predictor, results from simulation studies are presented within the sections. Finally, a real data example is provided in Section 5.

2 The two-stage predictor

In this section we consider the problem of predicting X_{t+2} based on X_t, X_{t-1}, \dots for the process (1). Since it is of d -th order Markovian, it is seen that the least square 2-step prediction is

$$\begin{aligned} E(X_{t+2}|X_t, X_{t-1}, \dots) &= E(X_{t+2}|X_t, X_{t-1}, \dots, X_{t-d+1}) \\ &= E\{f(X_{t+1}, X_t, \dots, X_{t-d+2})|X_t, X_{t-1}, \dots, X_{t-d+1}\} \\ &= E\{f(Y_{t+1})|Y_t\}. \end{aligned}$$

Thus the prediction problem is reduced to predicting X_{t+2} or equivalently, $f(Y_{t+1})$, from $Y_t = (X_t, \dots, X_{t-d+1})^T$. We will show that under regularity conditions, a 2-stage predictor using a pilot estimator $\hat{f}(Y_{t+1})$, has smaller mean squared error than the direct smoother in (2).

To begin with, we define the two-stage predictor as

$$\hat{m}_2(y) = \hat{m}_{2;h,h'}(y) = \frac{\sum_{t=d}^{n-2} K_h(y - Y_t)w(y, Y_{t+1})\hat{f}_{h'}(Y_{t+1})}{\sum_{t=d}^{n-2} K_h(y - Y_t)}, \quad (4)$$

where $w(y, z)$ is a weight function introduced here for technical reasons and where

$$\hat{f}_{h'}(z) = \frac{\sum_{j=d}^{n-2} K_{h'}(z - Y_j)X_{j+1}}{\sum_{j=d}^{n-2} K_{h'}(z - Y_j)}. \quad (5)$$

In the following, the density of Y_t is denoted as $p(\cdot)$, the Laplacian operator is denoted as ∇^2 , while the gradient operator as ∇ . We also denote $Y_{t+1} = g(Y_t, \varepsilon_{t+1})$, hence $g(y, e) = \{f(y) + \sigma(y)e, \underline{y}_{d-1}\}$ in which \underline{y}_{d-1} denotes (y_1, \dots, y_{d-1}) . In particular,

$$m_2(y) = \int w\{y, g(y, e)\} f\{g(y, e)\} p_\varepsilon(e) de$$

where we denote the density of ε_t by $p_\varepsilon(e)$.

Theorem 1 *Under the conditions (A1)-(A5) listed in the appendix, and $h = \beta n^{-1/(d+4)}$ for some $\beta > 0$, $h' = o(h)$, $nh^{d'} \rightarrow \infty$, we have*

$$n^{2/(d+4)} \left\{ \hat{m}_2(y) - m_2(y) - \beta^2 B(y) \right\} \xrightarrow{\mathcal{D}} N \left\{ 0, \beta^{-d} \|K\|_2^{2d} s_w^2(y) / p(y) \right\}$$

where $s_w^2(y) = \text{Var}\{w(y, Y_{t+1})f(Y_{t+1}) \mid Y_t = y\} \leq \text{Var}\{f(Y_{t+1}) \mid Y_t = y\}$ and

$$B(y) = \mu_2(K)/2 \int \left[\nabla_z^2 w \{y, g(z, e)\} f \{g(y, e)\} \right] |_{z=y} p_\varepsilon(e) de + \mu_2(K) \int \nabla_z^T [w \{y, g(z, e)\} f \{g(y, e)\}] |_{z=y} \nabla p(y) p_\varepsilon(e) de / p(y) \quad (6)$$

and where $\mu_2(K) = \int u^2 K(u) du$ and $\|K\|_2^2 = \int K^2(u) du$.

A sketch of the proof is given in the appendix. All the same conclusions are true for local linear regressors as well. The only change is that the bias function $B(y)$ becomes

$$B(y) = \mu_2(K)/2 \int \left[\nabla_z^2 w \{y, g(z, e)\} f \{g(y, e)\} \right] |_{z=y} p_\varepsilon(e) de \quad (7)$$

Comparing with the direct smoother (2), we have the following corollary.

Corollary 1 *Under the conditions of Theorem 1, the ratio of the asymptotic optimal mean squared error of the 2-stage smoother (4) and the direct smoother (2) at a single point y is*

$$r(y) = \left(\frac{s^2(y)}{s^2(y) + v^2(y)} \right)^{4/(d+4)}$$

where $v^2(y) = E\{\sigma^2(Y_{t+1}) \mid Y_t = y\}$. The same ratio over a d -dimensional compact set \mathcal{K} that contains y is

$$r = \left(\frac{\int_{\mathcal{K}} s^2(y) dy}{\int_{\mathcal{K}} s^2(y) dy + \int_{\mathcal{K}} v^2(y) dy} \right)^{4/(d+4)}. \quad (8)$$

Remark 2.1: The theorem says that asymptotically the 2-stage predictor behaves the same as if we knew exactly the function f . The improvement of the MSE is due to the smaller asymptotic variance $\text{Var}\{w(y, Y_{t+1})f(Y_{t+1}) \mid Y_t = y\} = s_w^2(y) \leq s^2(y) = \text{Var}\{f(Y_{t+1}) \mid Y_t = y\}$, versus $\text{Var}(X_{t+2} \mid Y_t = y) = s^2(y) + v^2(y)$ for the direct predictor.

Remark 2.2: One can replace the N-W smoother by the local polynomial estimator of order $2p$ or $2p + 1$ and the asymptotic result will be similar. The ratio of improvement in terms of MSE becomes at most

$$r(y) = \left(\frac{s_w^2(y)}{s_w^2(y) + v^2(y)} \right)^{(4p+4)/(4p+4+d)},$$

where $p = 0$ represents the improvement for the N-W and local linear estimators.

Remark 2.3: The multistage predictor is sensitive to the correctness of the model specification, particularly the AR order. For example, examine the following equation:

$$E(X_{t+2} \mid X_t) = E[E(X_{t+2} \mid X_{t+1}, X_t) \mid X_t] = E[E(X_{t+2} \mid X_{t+1}) \mid X_t]$$

The first equality always holds, but the second holds only when X_t is of first order Markovian. Hence if the process is not a nonlinear AR(1) (NAR(1)), the multistage prediction based on the second equality would be incorrect, whilst the direct smoothing procedure does not have this problem. However, a multistage prediction based on NAR(2) (i.e. using the first equality) will still gain efficiency over the direct smoothing method. To identify the correct model structure, one can effectively use the nonparametric procedures of Auestad and Tjøstheim (1990) or Tschernig and Yang (2000).

Remark 2.4: Theorem 1 requires the first stage bandwidth to be of smaller order of the second stage bandwidth. This requirement basically ensures that the bias created in the first stage smoothing is of smaller order so that their effect on the second stage smoothing is negligible. Similar features have been found in other multistage nonparametric procedures (Fan and Zhang, 1999).

3 Bandwidth selection and simulation

Automatic bandwidth selection is always an important part of any nonparametric procedure. Cross validation and plug-in methods are commonly used. It is possible to perform cross-validation procedures to obtain the optimal combination of (h', h) for the two stage smoothing, though it requires a two dimensional search, which can be computationally intensive. A simpler and faster way is to obtain the optimal cross-validation bandwidth for each stage separately, which is justified by Theorem 2. Note that Theorem 1 requires that h' is of a smaller order than the optimal bandwidth. Hence, some adjustment is required, though our simulation shows that the final result is not very sensitive to the selection.

To find bandwidths that minimize the mean squared error, one can also use a plug-in method. Minimizing the highest and the second highest order terms, one obtains the optimal bandwidths.

Theorem 2 *Under the conditions of Theorem 1, as $nh'^d \rightarrow \infty$, $h \rightarrow 0$, and $h' = o(h)$, the optimal bandwidths h and h' for the two-step estimation at y are*

$$h_{opt}(y) = \left\{ \frac{d \|K\|_2^{2d} s_w^2(y)}{4nB^2(y)p(y)} \right\}^{1/(d+4)} \quad (9)$$

and

$$h'_{opt}(y) = \left[\frac{d \|K\|_2^{4d} \int w^2(y, x, \underline{y}_{d-1}) \sigma^2(x, \underline{y}_{d-1}) p(x, y) / p(x, \underline{y}_{d-1}) dx}{4n^2 h_{opt}^d(y) \left\{ \int B'(x, \underline{y}_{d-1}) w(y, x, \underline{y}_{d-1}) p(x, y) dx \right\}^2} \right]^{1/(d+4)} \quad (10)$$

where $B'(\cdot)$ is defined in (20), $p(x, \underline{y}_{d-1})$ = the density of Y_i at (x, y_1, \dots, y_{d-1}) and $p(x, y)$ = the density of (X_i, Y_{i-1}) at (x, y_1, \dots, y_d) . The optimal bandwidths h and h' for two-step estimation over

\mathcal{K} are

$$h_{opt}(\mathcal{K}) = \left\{ \frac{d \|K\|_2^{2d} \int_{\mathcal{K}} s_w^2(y) dy}{4n \int_{\mathcal{K}} B^2(y) p(y) dy} \right\}^{1/(d+4)}. \quad (11)$$

and

$$h'_{opt}(\mathcal{K}) = \left\{ \frac{d \|K\|_2^{4d} \int_{\mathcal{K}} \left\{ \int w^2(y, x, \underline{y}_{d-1}) \sigma^2(x, \underline{y}_{d-1}) p(x, y) / p(x, \underline{y}_{d-1}) dx \right\} / p(y) dy}{4n^2 h_{opt}^d(\mathcal{K}) \int_{\mathcal{K}} \left\{ \int B'(x, \underline{y}_{d-1}) w(y, x, \underline{y}_{d-1}) p(x, y) dx \right\}^2 / p(y) dy} \right\}^{1/(d+4)} \quad (12)$$

Note that, as $h_{opt} = O(n^{-1/(d+4)})$, the first stage smoothing uses bandwidth $h'_{opt} = O\left\{n^{-(d+8)/(d+4)^2}\right\}$. By replacing the unknown terms in the above expressions with their estimates from a preliminary procedure, we can obtain the plug-in optimal bandwidth for each stage. These plug-in bandwidths are used in the simulation and real data analysis. For more general results on plug-in bandwidth selection in multivariate setting, see Yang and Tschernig (1999).

In the following, we present simulation results for three processes, the first and third one with plug-in, the second one with cross-validation bandwidth selection.

Example 1: Let us first consider an extension of the process (3) given as

$$X_t = a \sin(bX_{t-1}) + \sigma(X_{t-1})\varepsilon_t \quad (13)$$

with

$$\sigma^2(X_{t-1}) = \omega + \alpha X_{t-1}^2,$$

where $\omega > 0, \alpha \geq 0$. We fix the parameters $b = \pi/2$, and the amplitude a can be 1 or 2. The process is geometrically ergodic by Cline and Pu (1995). In our study, we let α be alternatively 0, 0.2 and 0.5, and $\omega = 1 - \alpha$. Note that $\alpha = 0$ corresponds to the process (3).

The theoretically optimal bandwidths for $\alpha = 0, n = 300, 1000$ and $a = 1, 2$ are given in Table 1. We notice that h is smaller than h_{dir} . This is due to the fact that the second stage smoothing is performed on a sample with smaller variance. All optimal bandwidths decrease when n is increased. When, for a given sample size, a is changed from 1 to 2, we observe that h_{dir} and h decrease, which is caused by the higher curvature of $m(y)$ and thus a larger bias. On the other hand, h' increases slightly, which is explained by the inverse relationship to h , see equations (10) and (12). Figures 1 and 2 illustrate the direct and multistage smoothers versus the true function for one sample of size $n = 300$ with $\alpha = 0, a = 1$.

Table 2 provides summary statistics for the improvement rates for 200 replications of the process (13) with sample size n . Since the distribution of \hat{r} is highly skewed, we present their

n	a	h_{dir}	h'	h
300	1	1.133	0.434	0.896
1000	1	0.891	0.281	0.704
300	2	0.596	0.472	0.558
1000	2	0.468	0.307	0.430

Table 1: *Optimal bandwidths for the simulation study of the process given in (3). h_{dir} is the optimal bandwidth for direct smoothing, h' and h for respectively the first and second stage of the multistage smoother.*

quantiles. The theoretical (and optimal) improvement rate r in (8) in this case is

$$r = \left(\frac{\int_{-c}^c s_w^2(x) dx}{\int_{-c}^c [s_w^2(x) + v^2(x)] dx} \right)^{4/5} \quad (14)$$

where c determines the interval of interest. We have chosen $c = 4$ for the case $a = 1$ and $c = 5$ for the case $a = 2$. Those intervals $[-c, c]$ cover about all simulated data points. Functions $s_w^2(x)$ and $v^2(x)$, obtained through elementary calculation, are

$$s_w^2(x) = \text{Var}(E[X_{t+2} | X_{t+1}] | X_t = x) = \frac{1}{2} a^2 \left(1 - e^{-b^2(\omega + \alpha x^2)} \right) \left(1 + e^{-b^2(\omega + \alpha x^2)} \cos(2ab \sin(bx)) \right)$$

and

$$v^2(x) = E[\text{Var}(X_{t+2} | X_{t+1}) | X_t = x] = \omega(1 + \alpha) + \alpha a^2 \sin^2(bx) + \alpha^2 x^2$$

respectively. For each simulated series, we estimated m_2 using both the direct predictor (\tilde{m}) and the two-stage predictor (\hat{m}), and calculated

$$\hat{r} = \frac{\sum_{t=1}^{n-2} \{\hat{m}_2(Y_t) - m_2(Y_t)\}^2}{\sum_{t=1}^{n-2} \{\tilde{m}_2(Y_t) - m_2(Y_t)\}^2}. \quad (15)$$

For $\alpha = 0$, we used the theoretical optimal bandwidth for the first stage smoothing. For other cases, we used three bandwidths for the first stage smoothing: $h' = h^*$, $h' = h^*/5$, $h' = h^*/10$.

Note here the interesting phenomenon that the improvement becomes more pronounced as heteroskedasticity is increased by changing α from 0 to 0.2 and then to 0.5. In all cases, the multistage predictor has substantial advantage over the direct predictor.

Example 2: Our second example uses cross-validation bandwidth selection. Here we use an exponential AR (EXPAR) model (Haggan and Ozaki, 1981)

$$X_t = (0.7 + 0.5e^{-cX_{t-1}^2})X_{t-1} + \sigma\varepsilon_t \quad (16)$$

(a, α)	h'	r	$n = 300$					$n = 1000$				
			Min	25%	50%	75%	Max	Min	25%	50%	75%	Max
(1,0)	$h' = h'_{opt}$	0.39	0.13	0.40	0.52	0.78	2.97	0.14	0.41	0.56	0.72	1.48
(2,0)	$h' = h'_{opt}$	0.71	0.42	0.66	0.79	0.92	1.61	0.36	0.67	0.80	0.96	1.76
(1,0.2)	$h' = h^*$	0.27	0.11	0.41	0.61	0.83	2.44	0.13	0.38	0.52	0.70	1.53
	$h' = h^*/5$	0.27	0.16	0.49	0.64	0.85	2.74	0.12	0.28	0.39	0.54	1.14
	$h' = h^*/10$	0.27	0.19	0.47	0.61	0.78	1.53	0.19	0.47	0.63	0.74	1.88
(1,0.5)	$h' = h^*$	0.17	0.08	0.24	0.35	0.48	1.02	0.12	0.34	0.46	0.60	1.53
	$h' = h^*/5$	0.17	0.10	0.47	0.64	0.82	1.51	0.10	0.37	0.50	0.67	1.45
	$h' = h^*/10$	0.17	0.08	0.23	0.33	0.44	0.84	0.14	0.29	0.39	0.49	1.18
(2,0.2)	$h' = h^*$	0.53	0.21	0.63	0.83	1.04	2.24	0.31	0.55	0.66	0.77	1.23
	$h' = h^*/5$	0.53	0.32	0.48	0.61	0.76	1.71	0.40	0.64	0.71	0.85	1.35
	$h' = h^*/10$	0.53	0.38	0.79	0.95	1.19	2.39	0.39	0.63	0.70	0.82	1.23
(2,0.5)	$h' = h^*$	0.33	0.38	0.73	0.94	1.40	13.77	0.39	0.51	0.57	0.67	1.21
	$h' = h^*/5$	0.33	0.22	0.66	0.77	0.90	1.53	0.34	0.51	0.58	0.66	1.12
	$h' = h^*/10$	0.33	0.05	0.99	1.03	1.08	2.31	0.26	0.49	0.61	0.71	1.19

Table 2: Summary statistics of the simulated ratios of improvement. The minima, maxima and quartiles of the simulated improvement rates \hat{r} are given. r is the theoretical ratio in (14).

with ε_t i.i.d. $\sim N(0, 1)$ and alternative (c, σ) combinations. Let h_1^* be the cross-validation bandwidth for smoothing X_t on X_{t-1} . Again, we tried three bandwidths for the first stage smoothing: $h_1 = h_1^*$, $h_1 = h_1^*/5$, $h_1 = h_1^*/10$. Two hundred series are generated, each of size 400. Table 3 shows the quartiles of the improvement rate r .

Example 3: To give an example for $d = 2$, consider the process

$$X_t = a_1 \sin(b_1 X_{t-1}) + a_2 \sin(b_2 X_{t-2}) + \sigma \varepsilon_t \quad (17)$$

with $\varepsilon_t \sim i.i.d.N(0, 1)$. For two step ahead prediction, the true conditional mean function is

$$\begin{aligned} m_2(x) &= E(X_{t+2} | X_t = x_1, X_{t-1} = x_2) \\ &= a_1 \sin\{a_1 b_1 \sin(b_1 x_1) + a_2 b_1 \sin(b_2 x_2)\} \exp(-b_1^2 \sigma^2 / 2) + a_2 \sin(b_2 x_1). \end{aligned}$$

This function is plotted in the lower left panel of Figure 3 with $a_1 = a_2 = 1/2$, $b_1 = \pi/4$, and $b_2 = \pi$. Using these parameters and $\sigma = 1/2$, we generated a series of (17) and estimated

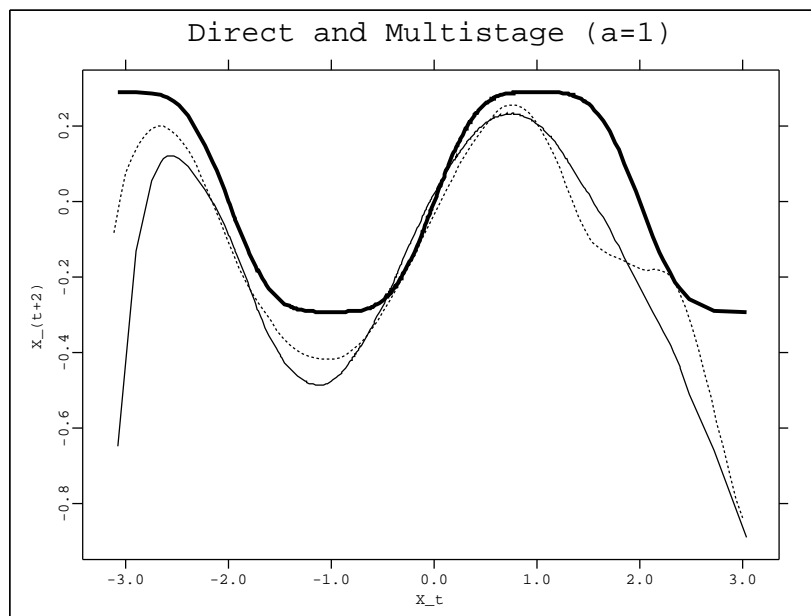


Figure 2: Notes as in Figure 1. Shown are the multistage predictor (dashed line), the direct predictor (solid line), and the true function (thick solid line). The improvement rate of this sample was 0.789.

m_2 using local linear estimates for the direct predictor and the multistage predictor, both shown also in Figure 3. To reduce the effect of outliers, we trimmed the generated series at the 2.5 and 97.5 percentiles. Plug-in bandwidths were used for the direct predictor and the second stage of the multistage predictor, whereas a grid search was performed for h' to minimize the mean squared error. All functions are shown for the range $(-1, 1) \times (-1, 1)$ which covers most data points. The mean squared errors were calculated also for this range. We generated 100 replications, each with $n = 1000$. The mean integrated squared errors of direct smoothing was 0.0093, while that of multistage smoothing 0.0064. The quartiles of the improvement ratios (15) were 0.6154, 0.6873, and 0.7616, respectively, with a minimum of 0.4319 and a maximum of 1.033.

4 Multi-stage predictor for multi-step ahead prediction

For nonlinear $AR(d)$ models in (1), multi-step prediction can be done recursively using the multi-stage smoother. Define $f_1(y) = E(X_{t+1} | Y_t = y)$ and for $j = 2, \dots, k$, recursively define $f_j(y) = E(f_{j-1}(Y_{t+1}) | Y_t = y)$. Then

$$m_k(y) = E(X_{t+k} | Y_t = y) = E\{f_1(Y_{t+k-1}) | Y_t = y\} = E\{f_2(Y_{t+k-2}) | Y_t = y\}$$

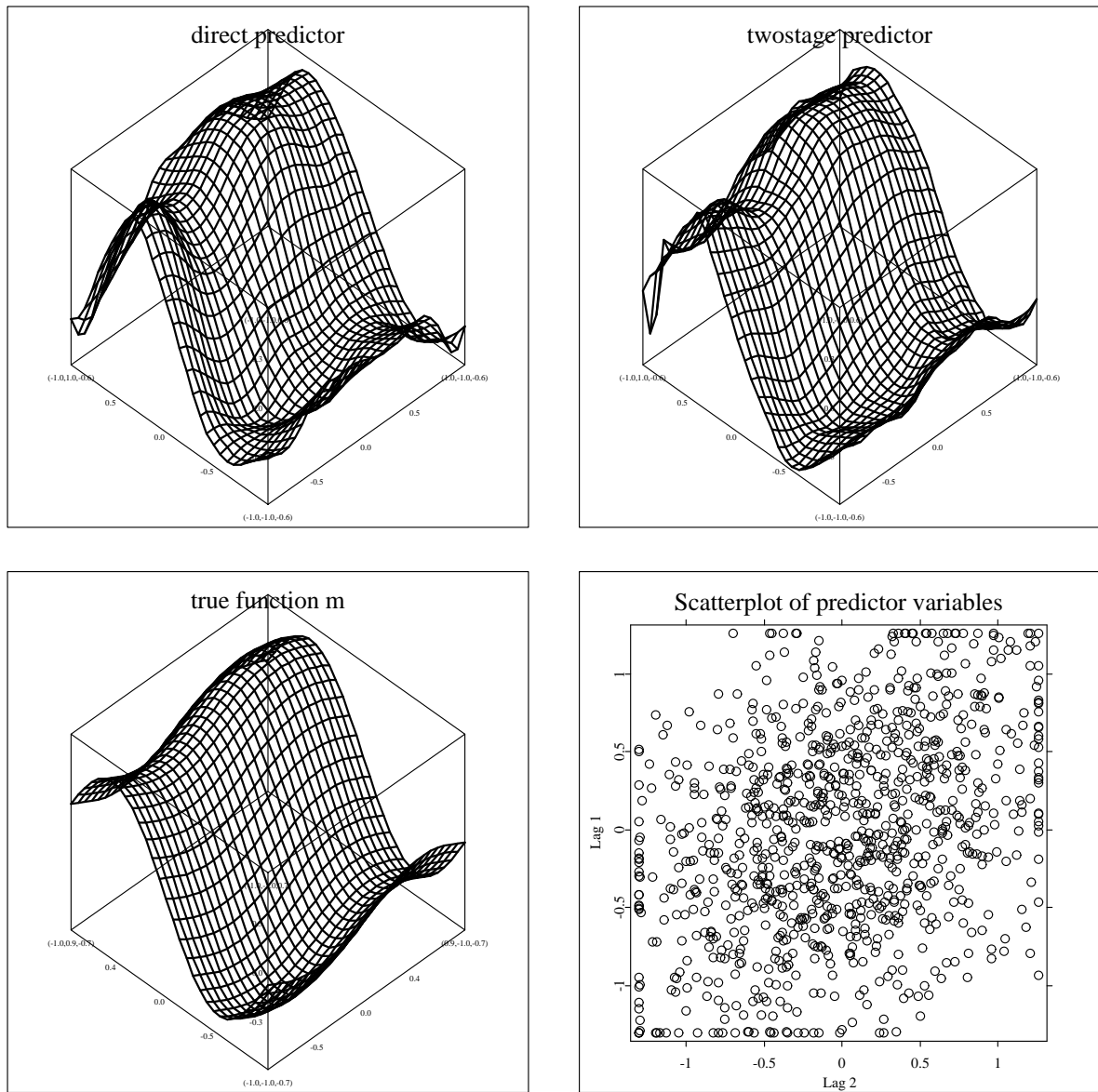


Figure 3: For a generated sample of 17 with $n = 1000$, the upper left plot shows the direct predictor, the upper right plot the multistage predictor, and the lower left plot the true function. The scatterplot of the predictors is given in the lower right plot. The ratio of the MSE for this sample was 0.50. Bandwidths were 0.4055 (direct), 0.2408 (first stage) and 0.2797 (second stage).

	$h_1 = h_1^*$			$h_1 = h_1^*/5$			$h_1 = h_1^*/10$		
(c,s)	25%	50%	75%	25%	50%	75%	25%	50%	75%
(0.15, 1)	0.54	0.75	1.00	0.51	0.70	0.88	0.55	0.75	0.92
(0.15, 0.5)	1.07	1.26	1.44	0.78	0.91	1.04	0.79	0.91	1.03
(0.5, 1)	0.65	0.94	1.24	0.50	0.72	0.93	0.53	0.75	0.94
(0.5, 0.5)	0.77	1.09	1.41	0.65	0.83	1.01	0.68	0.85	1.02

Table 3: *Quartiles of the improvement rate for different combinations of (c, s) and bandwidth for the first stage smoothing using model (16)*

$$= \dots = E\{f_{k-1}(Y_{t+1}) \mid Y_t = y\} = f_k(y).$$

Note that $\text{Var}\{f_j(Y_{t+k-j})\} = \text{Var}\{f_{j+1}(Y_{t+k-j-1})\} + E[\text{Var}\{f_j(Y_{t+k-j}) \mid Y_{t+k-j-1}\}]$ and therefore $\text{Var}\{f_{j+1}(Y_{t+k-j-1})\} \leq \text{Var}\{f_j(Y_{t+k-j})\}$, which holds for all j . Applying this recursively leads to $\text{Var}\{f_k(Y_t)\} \leq \text{Var}\{f_1(Y_{t+k-1})\}$, which means that k -step smoothing has a smaller variance than smoothing $f_1(Y_{t+k-1})$ on Y_t . This is the motivation for doing more steps.

For clearer presentation, we will use the N-W smoother. The method can be immediately extended to using the local polynomial estimator. Starting with $X_t^{(0)} = X_t$, one repeats the following steps for $j = 1, \dots, k-1$.

Stage j: Estimate

$$\hat{f}_j(y) = \frac{\sum_{t=d}^{n-k} K_{h_j}(y - Y_t) w_j(y, Y_{t+1}) X_{t+j}^{(j-1)}}{\sum_{t=d}^{n-k} K_{h_j}(y - Y_t)},$$

and obtain the j -th smoothed version of X_{t+j} , $X_{t+j}^{(j)} = \hat{f}_j(Y_t)$.

Then, the conditional mean function $m_k(y)$ is estimated by

$$\hat{m}_k(y) = \frac{\sum_{t=d}^{n-k} K_{h_k}(y - Y_t) w_k(y, Y_{t+1}) X_{t+k}^{(k-1)}}{\sum_{t=d}^{n-k} K_{h_k}(y - Y_t)}. \quad (18)$$

Graphically, the above recursive method can be presented as

$$X_{t+k} \xrightarrow{(X_{t+k}, Y_{t+k-1})} X_{t+k}^{(1)} \xrightarrow{(X_{t+k}^{(1)}, Y_{t+k-2})} X_{t+k}^{(2)} \xrightarrow{(X_{t+k}^{(2)}, Y_{t+k-3})} \dots \xrightarrow{(X_{t+k}^{(k-2)}, Y_{t+1})} X_{t+k}^{(k-1)} \xrightarrow{(X_{t+k}^{(k-1)}, Y_t)} m_k(y)$$

In the above scheme, a series of weight functions $w_j(y, Y_{t+1})$, $j = 2, \dots, k$ are used, and $w_1(y, Y_{t+1}) \equiv 1$.

We have the following theorem.

Theorem 3 Under conditions (A1)-(A6) in the appendix, if $h_j = o(h_k)$, $nh_j^d \rightarrow \infty$ for $j = 1, \dots, k-1$, and $h_k = \beta n^{-1/(d+4)}$ for some $\beta > 0$, we have

$$n^{2/(d+4)}(\hat{m}_k(y) - m_k(y) - \beta^2 B_k(y)) \xrightarrow{\mathcal{D}} N \left\{ 0, \beta^{-d} \|K\|_2^{2d} s_{w,k}^2(y)/p(y) \right\}$$

where

$$B(y) = \mu_2(K)/2 \int \left[\nabla_z^2 w_k \{y, g(z, e)\} f_{k-1} \{g(y, e)\} \right] |_{z=y} p_\varepsilon(e) de + \\ \mu_2(K) \int \nabla_z^T [w_k \{y, g(z, e)\} f_{k-1} \{g(y, e)\}] |_{z=y} \nabla p(y) p_\varepsilon(e) de / p(y)$$

and

$$s_{w,k}^2(y) = \text{Var} \{w_k(y, Y_{t+1}) f_{k-1}(Y_{t+1}) \mid Y_t = y\}.$$

The proof of the theorem is very tedious and we will only show a sketch in the appendix.

Note that the asymptotic bias and variance are the same as if we knew exactly the function $f_{k-1}(\cdot)$ and smoothed $f_{k-1}(Y_t)$ on Y_{t-1} . Comparing with the direct smoother, the bias term is the same, while the estimator (18) has smaller variance, since $\text{Var} \{w_k(y, Y_{t+1}) f_{k-1}(Y_{t+1}) \mid Y_t = y\}$ is smaller than $\text{Var}(X_{t+k} \mid Y_t = y)$. The ratio of the asymptotic optimal mean squared error of the multi-stage smoother (18) and the direct smoother (2) at a single point is then

$$r(y) = \left(\frac{\text{Var} \{w_k(y, Y_{t+1}) f_{k-1}(Y_{t+1}) \mid Y_t = y\}}{\text{Var}(X_{t+k} \mid Y_t = y)} \right)^{4/(d+4)}$$

The improvement ratio over a compact set can be obtained similarly as (8).

Also note that the above asymptotic result is the same as a 2-step procedure: (i) estimate f_k by smoothing Y_{t+k} on Y_{t+1} and obtain $\hat{f}_{k-1}(Y_{t+1})$, then (ii) estimate m_k by smoothing $\hat{f}_{k-1}(Y_{t+1})$ on Y_t . The improvement using the extra intermediate steps are asymptotically of smaller order. However, the benefit of these intermediate steps can be seen with finite sample size, due to the fact that by inserting those intermediate steps, the estimation of f_{k-1} is more accurate. It is noted that the practical implementation of the estimator becomes more and more difficult as the number of steps k increases, due to the difficulties in selecting a bandwidth for each step. There is a large tendency to over-smooth, due to the large number of smoothing involved. Theoretically, we have shown that the bandwidths at the earlier stages h_1, \dots, h_{k-1} should be of smaller order than the optimal bandwidth (to keep the bias introduced in the early stages negligible) while the final stage uses the optimal rate. Simultaneous bandwidth selection of (h_1, \dots, h_{k-1}) using cross-validation is almost impossible computationally. It seems that the plug-in method may be computationally more feasible, as discussed in Section 2.2.

When k is large, it is reasonable to skip some steps in the recursion, i.e., setting some h_i to zero, since the intermediate steps are less important asymptotically. This enables us to control

the number of smoothing parameters used, while still benefiting from the multi-stage smoothing procedure. However, the second to the last step (obtaining \hat{f}_{k-1}) should not be skipped. This can be seen in the following simple example. For a nonlinear AR(1) model $X_{t+1} = f(X_t) + \sigma(X_t)\varepsilon_t$, we have

$$m_3(x) = E[X_{t+3} | X_t = x] = E[f(X_{t+2}) | X_t = x] = E[f_2(X_{t+1}) | X_t = x]$$

where $f_2(z) = E(f(X_{t+2}) | X_{t+1} = z)$. Since $Var(f_2(X_{t+1}) | X_t = x) \leq Var(f(X_{t+2}) | X_t = x)$, by Theorem 1 we should smooth X_{t+3} on X_{t+1} to obtain an estimate of f_2 , then smooth $\hat{f}_2(X_{t+1})$ on X_t to estimate m_3 . This is better than obtaining \hat{f} and then smoothing $\hat{f}(X_{t+2})$ on X_t to estimate m_3 .

Our limited simulation study shows that with sufficient sample size and carefully chosen bandwidth, performing smoothing in every step of recursion provides more accurate results. Experiments have also shown that the second to the last step should not be skipped.

Example 4: To check the performance of the recursive multistage prediction estimator, we generated 200 series from the exponential AR model (16) with different (c, σ) combinations. We tried three different recursive schemes for 4-step prediction: four stage smoothing (4): 4-3-2-1-0; three stage (3): 4-3-1-0 (i.e. $h_2 = 0$); and two stage (2): 4-2-0 (i.e. $h_1 = 0$ and $h_3 = 0$). Table 4 shows the improvement rate using different bandwidths, (the c-v optimal, the c-v optimal over 5 and the c-v optimal over 10), except for the last stage, which always uses the optimal cross-validation bandwidth.

We can see from the table that the multi-stage estimator has marked improvement over the direct smoothers. The use of $h_i^*/5$ as the early stage bandwidth seems to work the best. And the two-step estimator does not perform as good the other two since it skipped the most important step (setting $h_3 = 0$).

To see the effect of each stage of smoothing, we plotted one of the simulated series. In Figure 4, we show the scatterplot of X_{t+4} , $X_{t+4}^{(1)}$, $X_{t+4}^{(2)}$ and $X_{t+4}^{(3)}$ against X_t , where $X_{t+4}^{(i)}$ is the i -th smoothed version of X_{t+4} after the i -th stage of smoothing. We can see that the variation of $X_{t+4}^{(i)}$ becomes smaller and smaller after each stage of smoothing.

5 A real data example

The practical relevance of our results is seen by comparing the performance of the direct and the multistage smoothers on a real data set benchmark. We have chosen the famous sunspot data, i.e. the yearly average numbers of sunspots, as provided by the Royal Observatory of Belgium¹

¹in the internet at <http://www.oma.be>

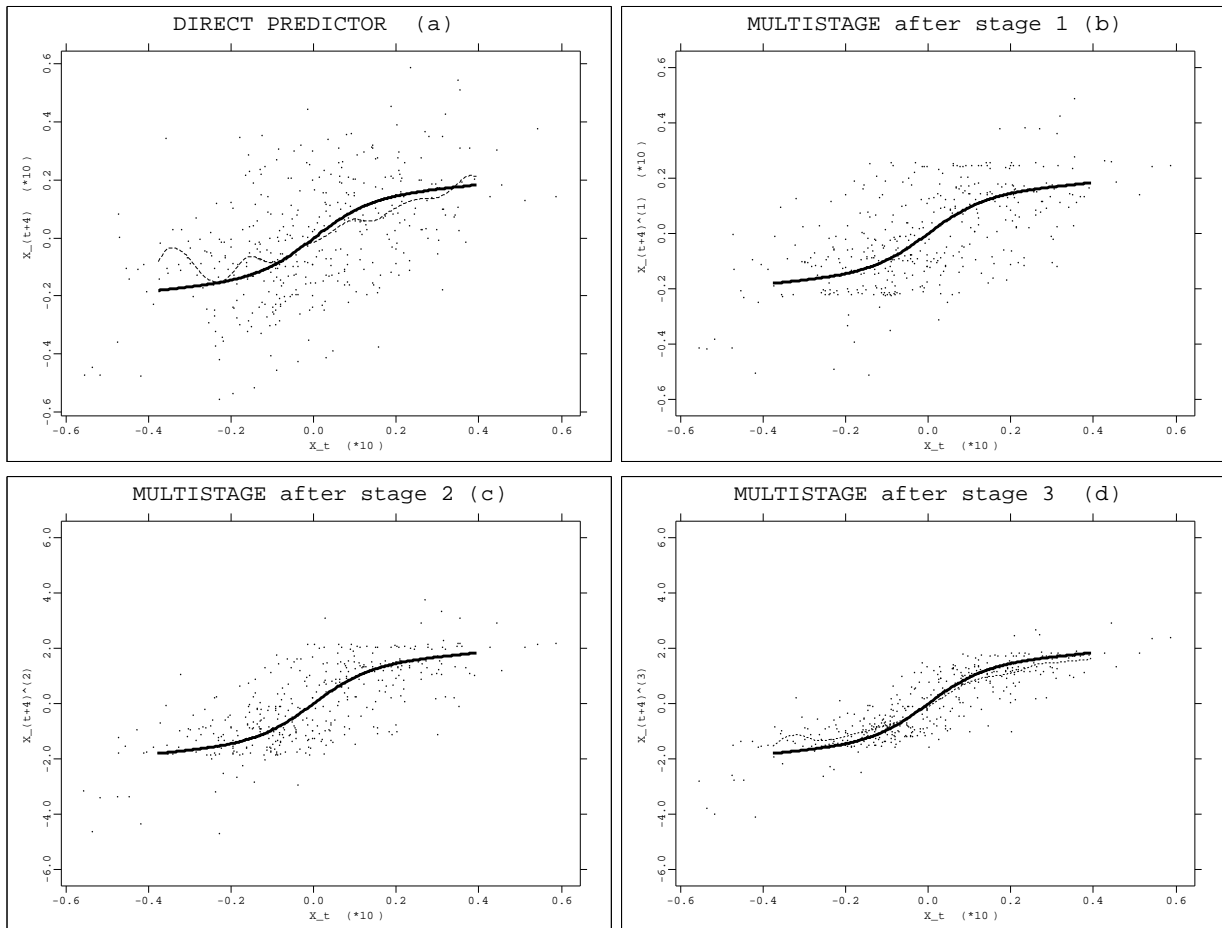


Figure 4: This figure demonstrates the process of four stage smoothing for four-step ahead prediction using model (16) with sample size 400. Panel (a) plots X_{t+4} against X_t . The solid line is the true conditional expectation $E(X_{t+4} | X_t)$ and the dashed line is the direct estimate. Panel (b) plots $X_{t+4}^{(1)}$ against X_t after the first stage smoothing. Panel (c) plots $X_{t+4}^{(2)}$ against X_t after the second stage smoothing. Panel (d) plots $X_{t+4}^{(3)}$ against X_t after the third stage smoothing. The dashed line is the final estimate.

(c,s)		$h_i = h_i^*$			$h_i = h_i^*/5$			$h_i = h_i^*/10$		
		25%	50%	75%	25%	50%	75%	25%	50%	75%
(0.15,1)	4	0.30	0.46	0.70	0.29	0.45	0.64	0.33	0.48	0.66
	3	0.31	0.46	0.66	0.33	0.49	0.67	0.39	0.54	0.70
	2	0.41	0.58	0.74	0.47	0.62	0.79	0.53	0.67	0.80
(0.15,0.5)	4	0.50	0.71	0.94	0.46	0.62	0.88	0.49	0.65	0.90
	3	0.50	0.69	0.94	0.50	0.66	0.86	0.55	0.72	0.91
	2	0.57	0.75	0.90	0.62	0.76	0.91	0.67	0.80	0.96
(0.5,1)	4	0.17	0.31	0.47	0.19	0.33	0.53	0.24	0.39	0.55
	3	0.19	0.33	0.47	0.24	0.38	0.57	0.28	0.44	0.64
	2	0.28	0.42	0.57	0.36	0.49	0.70	0.43	0.57	0.79
(0.5,0.5)	4	0.31	0.46	0.65	0.30	0.45	0.61	0.34	0.48	0.64
	3	0.31	0.47	0.62	0.36	0.47	0.62	0.41	0.54	0.70
	2	0.41	0.58	0.74	0.47	0.61	0.78	0.52	0.66	0.83

Table 4: *Quartiles of the improvement rates of four step prediction for model (16) using different (c, s) combinations, different early stage bandwidth selection and different recursing schemes, based on two hundred simulated series, each of size 400.*

for the years 1700 to 1997 and analyzed e.g. by Tong (1990, pp.419) and Fan and Gijbels (1996, pp. 222). Following Fan and Gijbels, we regress linearly X_t on X_{t-10} with coefficient 0.903 to obtain the deseasonalized series Z_t . Then the object is to predict Z_t by X_{t-1} . We use local linear estimation with quartic kernel. The optimal bandwidths are obtained by leave-one-out cross-validation. Similar to Tong (1990, pp.425) we use the last twenty years (1978–1997) as the prediction period, which covers roughly two cycles. Hence, the function is estimated using data only until 1977. Then we performed two- and three-step ahead prediction within the prediction period, keeping the estimated function fixed. The optimal bandwidths for the direct smoother for 1 to 3-step ahead prediction are respectively 25.49, 22.02 and 25.49. For the i -th stage bandwidth of the multi-stage smoother we tried h_i^*/j , where h_i^* is the cross-validation optimal bandwidth of the i -th stage, $i < k$ and $j = 1, 2, 3, \dots, 10$. For the k th stage the cross-validation optimal bandwidth was used, $k = 2, 3$. Table 5 reports the results for the ratio of mean square prediction error, \tilde{r} .

Obviously, multi-stage smoothing substantially improves direct prediction for two-step ahead prediction. In Figure 5 the two predictors are visualized. It can be seen that the variance of the twostage smoother is much smaller. The drastic dip of the direct predictor at around $X_t = 1.6 \times 10^2$

k	h_{dir}	h_1	h_2	h_3	\tilde{r}
2	22.02	$h_1^*/2$	30.98		0.8033
2	22.02	$h_1^*/4$	30.98		0.7973
2	22.02	$h_1^*/5$	30.98		0.8393
3	25.49	$h_1^*/8$	$h_2^*/4$	22.13	0.9783
3	25.49	$h_1^*/7$	$h_2^*/6$	22.13	0.9754
3	25.49	$h_1^*/6$	$h_2^*/3$	21.08	0.9863

Table 5: h_{dir} is the bandwidth for direct smoothing, h_i is the bandwidth used at the i th stage of the multistage smoother, where h_i^* denotes the cross-validation optimal bandwidth of the i th stage. \tilde{r} is the ratio of mean square prediction errors.

apparently indicates that the bandwidth is small due to the large amount of noise in the data, whereas the reduced noise level in the pseudo data set $\{X_t, \hat{f}(X_{t+1})\}$ allows the use of a larger bandwidth. Bandwidth used in each plot is the plug-in optimal bandwidth.

For three-step ahead prediction the improvement is less. We also experimented with four-step ahead prediction where the improvement was even less. This may be due to the shape of the conditional mean function, which is nonlinear for one and two steps ahead, but quite linear for three- and four steps.

Appendix

First we list some conditions needed for the theorems.

- (A1) The noise ε_t is i.i.d. with mean zero and variance 1. The function $\sigma(\cdot)$ is continuous and is positive on set \mathcal{K} .
- (A2) The process $\{X_t\}_{t \geq 0}$ is stationary and geometrically strong mixing. Sets of sufficient conditions for geometric ergodicity can be found in Tjøstheim (1990) and Davydov (1973).
- (A3) The functions f and m are twice continuously differentiable.
- (A4) The stationary density $p(\cdot)$ of Y_t exists, is bounded, continuous and bounded from below on S with interior S° such that $S^\circ \supset \mathcal{K} \ni y$, and continuously differentiable on S° .
- (A5) The kernel K is a compactly supported, symmetric probability density.

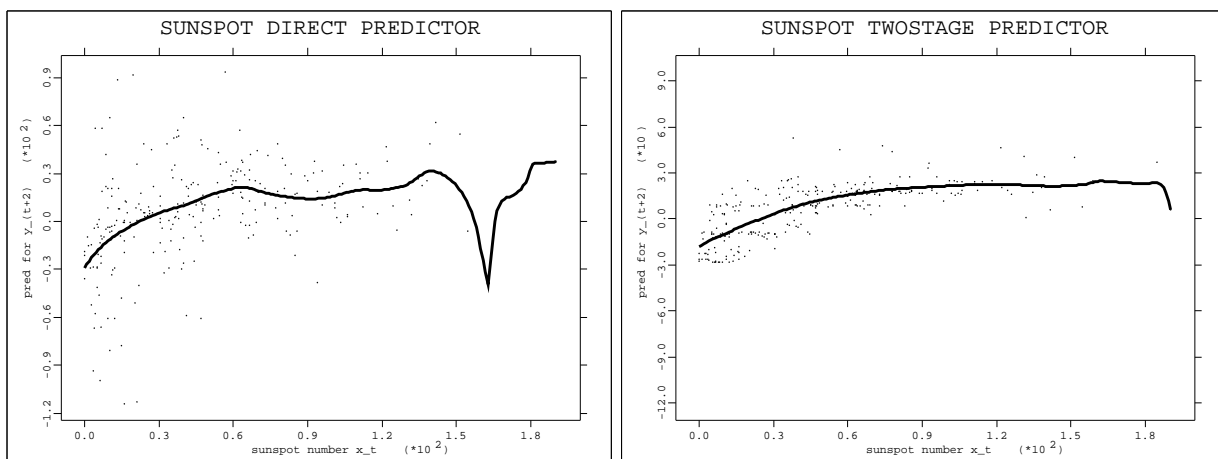


Figure 5: *Two-step ahead prediction of the sunspot numbers X_t . We predict the deseasonalized series $Z_t = X_t - 0.903X_{t-10}$ by X_{t-2} . The left panel shows the direct predictor and the pairs $\{X_t, Z_{t+2}\}$. The right panel shows the multistage predictor (dashed line), and the pairs $\{X_t, \hat{f}(X_{t+1})\}$, where \hat{f} is the first stage smoother for $E[Z_t | X_{t-1}]$.*

(A6) The weight function $w_k(y, z)$ satisfies

1. $E\{w_k(y, Y_{t+1})f_{k-1}(Y_{t+1}) | Y_t = y\} = E\{f_{k-1}(Y_{t+1}) | Y_t = y\}$ for all $y \in \mathcal{K}$,
2. $s_w^2(y) = \text{Var}\{w_k(y, Y_{t+1})f_{k-1}(Y_{t+1}) | Y_t = y\} \leq \text{Var}\{f_{k-1}(Y_{t+1}) | Y_t = y\}$ for all $y \in \mathcal{K}$,
3. there exists a compact set \mathcal{K}'_k such that $w_k(y, z) \equiv 0$ if $y \in \mathcal{K}, z \notin \mathcal{K}'_k$ (iv) $w_k(y, z) \geq 0, \int w_k(y, z)dz > 0$, for all $y \in \mathcal{K}$.

In particular, $w_2(y, z)$ is the weight function used in (4).

The proofs of Theorem 1 and Theorem 2 are closely related, and both make use of a few auxiliary lemmas that we present below. To illustrate the idea, denote

$$\tilde{m}(y) = \tilde{m}_2(y) = \frac{\sum_{t=d}^{n-2} K_h(y - Y_t)w(y, Y_{t+1})f(Y_{t+1})}{\sum_{t=d}^{n-2} K_h(y - Y_t)},$$

and recall from (4) and (5) that

$$\hat{m}(y) = \hat{m}_2(y) = \frac{\sum_{t=d}^{n-2} K_h(y - Y_t)w(y, Y_{t+1})\hat{f}'(Y_{t+1})}{\sum_{t=d}^{n-2} K_h(y - Y_t)},$$

where

$$\hat{f}'(z) = \hat{f}(z) = \frac{\sum_{j=d}^{n-2} K_{h'}(z - Y_j)X_{j+1}}{\sum_{j=d}^{n-2} K_{h'}(z - Y_j)}.$$

We want to decompose the difference $\hat{m}(y) - \tilde{m}(y)$ into bias and noise parts. These parts are then shown to be of higher order than that of $\tilde{m}(y) - m(y)$ and thus the difference between $\tilde{m}(y)$ and $\hat{m}(y)$ is negligible compared to that of $\tilde{m}(y)$ and $m(y)$, which proves Theorem 1 (see remark 2.1). To further obtain the optimal bandwidth h' and prove Theorem 2, one needs to solve a bias-variance trade-off between the bias and noise terms of $\hat{m}(y) - \tilde{m}(y)$. Lemma 1 concerns what would be the bias term while Lemma 2 what would be the noise term.

Lemma 1 *Define*

$$I(y) = n^{-1} \sum_{i=d}^{n-2} K_h(y - Y_i) w(y, Y_{i+1}) B'(Y_{i+1}) h'^2 \quad (19)$$

where

$$B'(z) = \left\{ \frac{1}{2} \nabla^2 f(z) + \nabla^T f(z) \frac{\nabla p(z)}{p(z)} \right\} \mu_2(K). \quad (20)$$

As $nh'^d \rightarrow \infty$, $h \rightarrow 0$, and $h' = o(h)$,

$$I(y) = \int B'(x, \underline{y}_{d-1}) w(y, x, \underline{y}_{d-1}) p(x, y) dx h'^2 + O_p \left(h'^2 h^2 + \frac{h'^2}{nh^d} \right). \quad (21)$$

Proof of this lemma is a straightforward application of geometric mixing properties.

Lemma 2 *Define*

$$II(y) = n^{-1} \sum_{i=d}^{n-2} K_h(y - Y_{i-1}) \frac{w(y, Y_i)}{np(Y_i)} \sum_{j=d}^{n-2} K_{h'}(Y_j - Y_i) \sigma(Y_j) \varepsilon_{j+1}. \quad (22)$$

As $nh'^d \rightarrow \infty$, $h \rightarrow 0$ and $h' = o(h)$,

$$E \{II(y)\}^2 = \frac{\|K\|_2^{4d}}{n^2 h^d h'^d} \int \frac{w^2(y, x, \underline{y}_{d-1}) \sigma^2(x, \underline{y}_{d-1}) p(x, y)}{p(x, \underline{y}_{d-1})} dx + o \left(\frac{1}{n^2 h^d h'^d} \right). \quad (23)$$

Sketch of a proof of Lemma 2:

The noise term $II(y)$ is more complicated than bias term $I(y)$. We first rewrite it as

$$II(y) = n^{-2} \sum_{i=d}^{n-2} \sum_{j=d}^{n-2} T_{ij},$$

in which

$$T_{ij} = \frac{K_h(y - Y_{i-1}) w(y, Y_i) K_{h'}(Y_j - Y_i) \sigma(Y_j) \varepsilon_{j+1}}{p(Y_i)}.$$

First, it is easily seen that

$$n^{-2} \sum_{i=1}^{n-2} T_{i,i} = O_p \left(\frac{1}{\sqrt{n^3 h'^{2d} h^d}} \right).$$

So now it is clear that

$$II(y) = II_1(y) + II_2(y) + O_p\left(\frac{1}{\sqrt{n^3 h^{2d} h^d}}\right),$$

in which

$$II_1(y) = n^{-2} \sum_{d \leq i < j \leq n-2} T_{ij}, \quad II_2(y) = n^{-2} \sum_{d \leq j < i \leq n-2} T_{ij}.$$

For the term $II_1(y)$, the variance consists mainly of a sum of individual variances as

$$E\{II_1(y)\}^2 = n^{-4} \sum_{d \leq i < j \leq n-2} \sum_{d \leq i' < j' \leq n-2} E(T_{ij} T_{i'j'}),$$

and noting that $E(T_{ij} T_{i'j'}) = 0$ for $j \neq j'$, it becomes

$$\begin{aligned} E\{II_1(y)\}^2 &= n^{-4} \sum_{d \leq i < j \leq n-2} \sum_{d \leq i' < j \leq n-2} E(T_{ij} T_{i'j}) \\ &= n^{-4} \sum_{d \leq i < j \leq n-2} E(T_{ij}^2) + n^{-4} \sum_{d \leq i < j \leq n-2, d \leq i' < j \leq n-2, i \neq i'} E(T_{ij} T_{i'j}). \end{aligned}$$

One then notes that the second term contains $O(n^3)$ terms and, by geometric mixing, it is easy to show that $E(T_{ij} T_{i'j}) = O(1)$ uniformly for all $d \leq i < j \leq n-2, d \leq i' < j \leq n-2, i \neq i'$. Thus

$$n^{-4} \sum_{d \leq i < j \leq n-2, d \leq i' < j \leq n-2, i \neq i'} E(T_{ij} T_{i'j}) = O(n^{-1}).$$

Similarly,

$$\begin{aligned} n^{-4} \sum_{d \leq i < j \leq n-2} E(T_{ij}^2) &= n^{-4} \sum_{d \leq i < j \leq n-2} E\left\{\frac{w^2(y, Y_i) K_h^2(y - Y_{i-1}) K_{h'}^2(Y_j - Y_i) \sigma^2(Y_j) \varepsilon_{j+1}^2}{p^2(Y_i)}\right\} \\ &= n^{-4} \sum_{d \leq i < j \leq n-2} E\left\{\frac{w^2(y, Y_i) K_h^2(y - Y_{i-1}) K_{h'}^2(Y_j - Y_i) \sigma^2(Y_j)}{p^2(Y_i)}\right\}, \end{aligned}$$

which, after doing a change of variable $Y_{i-1} = y + hu, Y_j = Y_i + h'v$ and using geometric mixing properties, becomes

$$\begin{aligned} &\frac{\{1 + o(1)\}}{2n^2} \int \frac{p(x, y + hu) p\left\{\left(x, \underline{y}_{d-1} + h\underline{u}_{d-1}\right) + h'v\right\}}{p^2\left(x, \underline{y}_{d-1} + h\underline{u}_{d-1}\right) h^d h'^d} \times \\ &w^2\left(y, x, \underline{y}_{d-1} + h\underline{u}_{d-1}\right) K^2(u) K^2(v) \sigma^2\left\{\left(x, \underline{y}_{d-1} + h\underline{u}_{d-1}\right) + h'v\right\} dudvdx \\ &= \frac{\|K\|_2^{4d}}{2n^2 h^d h'^d} \int \frac{w^2\left(y, x, \underline{y}_{d-1}\right) \sigma^2\left(x, \underline{y}_{d-1}\right) p(x, y)}{p\left(x, \underline{y}_{d-1}\right)} dx + o\left(\frac{1}{n^2 h^d h'^d}\right). \end{aligned}$$

This has proved that

$$E\{II_1(y)\}^2 = \frac{\|K\|_2^{4d}}{2n^2 h^d h'^d} \int \frac{w^2\left(y, x, \underline{y}_{d-1}\right) \sigma^2\left(x, \underline{y}_{d-1}\right) p(x, y)}{p\left(x, \underline{y}_{d-1}\right)} dx + o\left(\frac{1}{n^2 h^d h'^d}\right).$$

The term $II_2(y) = n^{-2} \sum_{d \leq j < i \leq n-2} T_{ij}$ is more complicated as one no longer has $E(T_{ij}T_{i'j'}) = 0$ for $j \neq j'$. We want to show here that the variance still consists mainly of a sum of individual variances as for $II_1(x)$. By definition

$$E\{II_2(y)\}^2 = n^{-4} \sum_{d \leq j < i \leq n-2} \sum_{d \leq j' < i' \leq n-2} E(T_{ij}T_{i'j'}).$$

The rest of the proof is completed by the following two lemmas.

Lemma 3 *As $nh^{d'} \rightarrow \infty$, $h \rightarrow 0$ and $h' = o(h)$,*

$$n^{-4} \sum_{d \leq j < i \leq n-2} E(T_{ij}^2) = \frac{\|K\|_2^{4d}}{2n^2 h^d h'^d} \int \frac{w^2(y, x, \underline{y}_{d-1}) \sigma^2(x, \underline{y}_{d-1}) p(x, y)}{p(x, \underline{y}_{d-1})} dx + o\left(\frac{1}{n^2 h^d h'^d}\right).$$

Proof:

$$\begin{aligned} n^{-4} \sum_{d \leq j < i \leq n-2} E(T_{ij}^2) &= n^{-4} \sum_{d \leq j < i \leq n-2} E\left\{ \frac{w^2(y, Y_i) K_h^2(y - Y_{i-1}) K_{h'}^2(Y_j - Y_i) \sigma^2(Y_j) \varepsilon_{j+1}^2}{p^2(Y_i)} \right\} \\ &= P_1 + P_2, \end{aligned}$$

where

$$\begin{aligned} P_1 &= n^{-4} \sum_{i-j > c \ln n} E\left\{ \frac{w^2(y, Y_i) K_h^2(y - Y_{i-1}) K_{h'}^2(Y_j - Y_i) \sigma^2(Y_j) \varepsilon_{j+1}^2}{p^2(Y_i)} \right\}, \\ P_2 &= n^{-4} \sum_{0 < i-j \leq c \ln n} E\left\{ \frac{w^2(y, Y_i) K_h^2(y - Y_{i-1}) K_{h'}^2(Y_j - Y_i) \sigma^2(Y_j) \varepsilon_{j+1}^2}{p^2(Y_i)} \right\}, \end{aligned}$$

where one assumes that when $i - j > c \ln n$, the mixing coefficient between the σ -fields $\sigma\{X_t\}_{t=0}^{j+1}$ and $\sigma\{X_t\}_{t=i}^{\infty}$ is smaller than $C\rho^{c \ln n} \leq n^{-4}$. After a change of variable, we can easily show that

$$P_1 = \frac{\|K\|_2^{4d}}{2n^2 h^d h'^d} \int \frac{w^2(y, x, \underline{y}_{d-1}) \sigma^2(x, \underline{y}_{d-1}) p(x, y)}{p(x, \underline{y}_{d-1})} dx + o\left(\frac{1}{n^2 h^d h'^d}\right).$$

For P_2 , the estimate is

$$\begin{aligned} P_2 &= n^{-4} \sum_{0 < i-j \leq c \ln n} E\left[\frac{w^2(y, Y_i) K_h^2(y - Y_{i-1}) K_{h'}^2(Y_j - Y_i) \{X_{j+1} - f(Y_j)\}^2}{p^2(Y_i)} \right] \\ &\leq n^{-4} n c \ln n h^{-2d} h'^{-2d} h'^d h = O(n^{-3} \ln n h^{1-2d} h'^{-d}) = o(n^{-2} h^{-d} h'^{-d}). \end{aligned}$$

Lemma 4 *As $nh^{d'} \rightarrow \infty$, $h \rightarrow 0$ and $h' = o(h)$,*

$$S = n^{-4} \sum_{d \leq j < i \leq n-2, d \leq j' < i' \leq n-2, j < j'} E(T_{ij}T_{i'j'}) = o(n^{-2} h^{-d} h'^{-d}).$$

Proof: The lemma can be proved by parting the summation into the following eight cases, to group the observations with large correlation and small correlations. We will skip the detailed calculation here.

Case 1 $j' < i - c \ln n$.

Case 2 $j < \min(i, j') - c \ln n$.

Case 3 $i - c \ln n \leq j' < i$, $j' - c \ln n \leq j < j'$, and $i' - c \ln n \leq i < i'$.

Case 4 $i - c \ln n \leq j' < i$, $j' - c \ln n \leq j < j'$, and $i < i' - c \ln n$.

Case 5 $i - c \ln n \leq j < i < j' - c \ln n$, and $i' - c \ln n \leq j' < i'$.

Case 6 $i - c \ln n \leq j < i < j' - c \ln n$, and $j' < i' - c \ln n$.

Case 7 $i - c \ln n \leq j < i$, $j' - c \ln n \leq i < j'$, and $j' < i' - c \ln n$.

Case 8 $i - c \ln n \leq j < i$, $j' - c \ln n \leq i < j'$, and $i' - c \ln n \leq j' < i'$.

Sketch of proofs of Theorem 1 and Theorem 2:

We note that

$$\hat{m}(y) - \tilde{m}(y) = \frac{\sum_{i=d}^{n-2} K_h(y - Y_i) w(y, Y_{i+1}) \{ \hat{f}(Y_{i+1}) - f(Y_{i+1}) \}}{\sum_{t=d}^{n-2} K_h(y - Y_t)}$$

where

$$\hat{f}(Y_{i+1}) - f(Y_{i+1}) = \frac{1}{n \hat{p}(Y_{i+1})} \sum_{j=d}^{n-2} K_{h'}(Y_j - Y_{i+1}) \{ X_{j+1} - f(Y_{i+1}) \}$$

with

$$\hat{p}(y') = \frac{1}{n} \sum_{j=d}^{n-2} K_{h'}(Y_j - y').$$

We want to prove first that by replacing the denominator $\hat{p}(Y_{i+1})$ by $p(Y_{i+1})$ in $\hat{f}(Y_{i+1}) - f(Y_{i+1})$ and hence also in $\hat{m}(y) - \tilde{m}(y)$, the deviation is an negligible $o_p(h'^2)$.

Lemma 5

$$\begin{aligned} & \frac{1}{n^2} \sum_{i,j=d}^{n-2} K_h(y - Y_i) w(y, Y_{i+1}) \left\{ \frac{1}{\hat{p}(Y_{i+1})} - \frac{1}{p(Y_{i+1})} \right\} K_{h'}(Y_j - Y_{i+1}) \{ X_{j+1} - f(Y_{i+1}) \} \\ &= o_p(h'^2). \end{aligned} \tag{24}$$

Proof. One first notes that for any integer $T > 0$

$$\begin{aligned} \left\{ \frac{1}{\widehat{p}(z)} - \frac{1}{p(z)} \right\} &= \frac{1}{p(z)} \left[\left\{ 1 - \frac{\widehat{p}(z)}{p(z)} \right\} + \cdots + \left\{ 1 - \frac{\widehat{p}(z)}{p(z)} \right\}^T \right] \\ &+ \frac{1}{p(z)} \left\{ 1 - \frac{\widehat{p}(z)}{p(z)} \right\}^{T+1} \left[1 - \left\{ 1 - \frac{\widehat{p}(z)}{p(z)} \right\} \right]^{-1} \end{aligned} \quad (25)$$

and hence

$$\begin{aligned} &\frac{1}{n^2} \sum_{i,j=d}^{n-2} K_h(y - Y_i) w(y, Y_{i+1}) \left\{ \frac{1}{\widehat{p}(Y_{i+1})} - \frac{1}{p(Y_{i+1})} \right\} K_{h'}(Y_j - Y_{i+1}) \{X_{j+1} - f(Y_{i+1})\} \\ &= \sum_{t=1}^T \frac{1}{n^2} \sum_{i,j=d}^{n-2} K_h(y - Y_i) \frac{w(y, Y_{i+1})}{p(Y_{i+1})} \left\{ 1 - \frac{\widehat{p}(Y_{i+1})}{p(Y_{i+1})} \right\}^t K_{h'}(Y_j - Y_{i+1}) \{X_{j+1} - f(Y_{i+1})\} \\ &\quad + \frac{1}{n^2} \sum_{i,j=d}^{n-2} K_h(y - Y_i) \frac{w(y, Y_{i+1})}{p(Y_{i+1})} \left\{ 1 - \frac{\widehat{p}(Y_{i+1})}{p(Y_{i+1})} \right\}^{T+1} \left[1 - \left\{ 1 - \frac{\widehat{p}(Y_{i+1})}{p(Y_{i+1})} \right\} \right]^{-1} \times \\ &\quad K_{h'}(Y_j - Y_{i+1}) \{X_{j+1} - f(Y_{i+1})\}. \end{aligned} \quad (26)$$

By Theorem of Bosq (1998), one has

$$\sup_{z \in \mathcal{K}'} \left| 1 - \frac{\widehat{p}(z)}{p(z)} \right| = O_p \left(h'^2 + \frac{\log n}{\sqrt{nh'^d}} \right). \quad (27)$$

so for large enough T , one has

$$\begin{aligned} &\frac{1}{n^2} \sum_{i,j=d}^{n-2} K_h(y - Y_i) \frac{w(y, Y_{i+1})}{p(Y_{i+1})} \left\{ 1 - \frac{\widehat{p}(Y_{i+1})}{p(Y_{i+1})} \right\}^T \left[1 - \left\{ 1 - \frac{\widehat{p}(Y_{i+1})}{p(Y_{i+1})} \right\} \right]^{-1} \times \\ &K_{h'}(Y_j - Y_{i+1}) \{X_{j+1} - f(Y_{i+1})\} = O_p \left(h'^2 h^2 \right). \end{aligned} \quad (28)$$

The Lemma will be completed if we can show that

Lemma 6 For any integer $t > 0$

$$\frac{1}{n^2} \sum_{i,j=d}^{n-2} K_h(y - Y_i) \frac{w(y, Y_{i+1})}{p(Y_{i+1})} \left\{ 1 - \frac{\widehat{p}(Y_{i+1})}{p(Y_{i+1})} \right\}^t K_{h'}(Y_j - Y_{i+1}) \{X_{j+1} - f(Y_{i+1})\} = o_p \left(h'^2 \right).$$

Proof. To avoid complicated notations, we prove the lemma only for the case of $t = 1$. One has

$$\begin{aligned} &\frac{1}{n^2} \sum_{i,j=d}^{n-2} K_h(y - Y_i) \frac{w(y, Y_{i+1})}{p(Y_{i+1})} \left\{ 1 - \frac{\widehat{p}(Y_{i+1})}{p(Y_{i+1})} \right\} K_{h'}(Y_j - Y_{i+1}) \{X_{j+1} - f(Y_{i+1})\} \\ &= B + V \end{aligned}$$

where

$$B = \frac{1}{n^2} \sum_{i,j=d}^{n-2} K_h(y - Y_i) \frac{w(y, Y_{i+1})}{p(Y_{i+1})} \left\{ 1 - \frac{\widehat{p}(Y_{i+1})}{p(Y_{i+1})} \right\} K_{h'}(Y_j - Y_{i+1}) \{f(Y_j) - f(Y_{i+1})\},$$

$$V = \frac{1}{n^2} \sum_{i,j=d}^{n-2} K_h(y - Y_i) \frac{w(y, Y_{i+1})}{p(Y_{i+1})} \left\{ 1 - \frac{\hat{p}(Y_{i+1})}{p(Y_{i+1})} \right\} K_{h'}(Y_j - Y_{i+1}) \{ \sigma(Y_j) \varepsilon_j \}.$$

One can show that both B and V are of order $o_p(h'^2)$. We rewrite B and V as

$$B = \frac{1}{n^3} \sum_{i,j,k=d}^{n-2} B_{ijk}, V = \frac{1}{n^3} \sum_{i,j,k=d}^{n-2} V_{ijk}$$

where

$$B_{ijk} = K_h(y - Y_i) \frac{w(y, Y_{i+1})}{p^2(Y_{i+1})} \{ p(Y_{i+1}) - K_{h'}(Y_k - Y_{i+1}) \} K_{h'}(Y_j - Y_{i+1}) \{ f(Y_j) - f(Y_{i+1}) \}$$

$$V_{ijk} = K_h(y - Y_i) \frac{w(y, Y_{i+1})}{p^2(Y_{i+1})} \{ p(Y_{i+1}) - K_{h'}(Y_k - Y_{i+1}) \} K_{h'}(Y_j - Y_{i+1}) \sigma(Y_j) \varepsilon_j.$$

In order to apply Lemma 2 of Yoshihara (1976), one first calculates the expectation of the term B_{ijk} under the condition of independence (which was denoted as $\theta(F)$ in the paper). That expectation is a $(3d+1)$ -dimensional integral

$$\int K_h \left\{ y - (y_{32}, \dots, y_{3d}, x)^T \right\} \frac{w(y, y_3)}{p^2(y_3)} \{ p(y_3) - K_{h'}(y_1 - y_3) \} \\ K_{h'}(y_2 - y_3) \{ f(y_2) - f(y_3) \} p(y_1) p(y_2) p(y_3, x) dy_1 dy_2 dy_3 dx$$

in which $y_3 = (y_{31}, \dots, y_{3d})^T$, and y_1, y_2 are likewise. Using changes of variables $y_1 = y_3 + h'u_1, y_2 = y_3 + h'u_2$, it becomes

$$\int K_h \left\{ y - (y_{32}, \dots, y_{3d}, x)^T \right\} \frac{w(y, y_3)}{p^2(y_3)} \left\{ p(y_3) - \frac{1}{h'^d} K(u_1) \right\} \\ K(u_2) \{ f(y_3 + h'u_2) - f(y_3) \} p(y_3 + h'u_1) p(y_3 + h'u_2) p(y_3, x) h'^d du_1 du_2 dy_3 dx = \\ \int K_h \left\{ y - (y_{32}, \dots, y_{3d}, x)^T \right\} \frac{w(y, y_3)}{p^2(y_3)} p(y_3, x) \left[\int \left\{ p(y_3) - \frac{1}{h'^d} K(u_1) \right\} p(y_3 + h'u_1) h'^d du_1 \right] \\ \left[\int K(u_2) \{ f(y_3 + h'u_2) - f(y_3) \} p(y_3 + h'u_2) du_2 \right] dy_3 dx = \\ \int K_h \left\{ y - (y_{32}, \dots, y_{3d}, x)^T \right\} \frac{w(y, y_3)}{p^2(y_3)} p(y_3, x) \left[p(y_3) - \int K(u_1) p(y_3 + h'u_1) du_1 \right] \\ \left[\int K(u_2) \{ f(y_3 + h'u_2) - f(y_3) \} p(y_3 + h'u_2) du_2 \right] dy_3 dx = \\ \int K_h \left\{ y - (y_{32}, \dots, y_{3d}, x)^T \right\} \frac{w(y, y_3)}{p^2(y_3)} p(y_3, x) \left[p(y_3) - p(y_3) + O(h'^2) \right] \\ \left[\int K(u_2) \{ f(y_3 + h'u_2) - f(y_3) \} p(y_3 + h'u_2) du_2 \right] dy_3 dx \\ = O(h'^2) \times O(h'^2) = O(h'^4) = o(h'^2).$$

We then need to calculate the r -th absolute moment of the term B_{ijk} ($r = 2 + \delta, \delta > 0$)

$$\int \left| K_h \left\{ y - (y_{32}, \dots, y_{3d}, x)^T \right\} \right|^r \frac{w^r(y, y_3)}{p^{2r}(y_3)} |p(y_3) - K_{h'}(y_1 - y_3)|^r \\ \{K_{h'}(y_2 - y_3)\}^r |f(y_2) - f(y_3)|^r p(y_1)p(y_2)p(y_3, x) dy_1 dy_2 dy_3 dx.$$

Using changes of variables $y_1 = y_3 + h'u_1, y_2 = y_3 + h'u_2$, it becomes

$$\int \left| K_h \left\{ y - (y_{32}, \dots, y_{3d}, x)^T \right\} \right|^r \frac{w^r(y, y_3)}{p^{2r}(y_3)} \left| p(y_3) - \frac{1}{h'^d} K(u_1) \right|^r \\ h'^{-rd} K^r(u_2) |f(y_3 + h'u_2) - f(y_3)|^r p(y_3 + h'u_1)p(y_3 + h'u_2)p(y_3, x) h'^{2d} du_1 du_2 dy_3 dx = \\ h'^{(2-r)d} \int \left| K_h \left\{ y - (y_{32}, \dots, y_{3d}, x)^T \right\} \right|^r \frac{w^r(y, y_3)}{p^{2r}(y_3)} \left[\int \left| p(y_3) - \frac{1}{h'^d} K(u_1) \right|^r p(y_3 + h'u_1) du_1 \right] \\ \left[\int K^r(u_2) |f(y_3 + h'u_2) - f(y_3)|^r p(y_3 + h'u_2) du_2 \right] p(y_3, x) dy_3 dx = \\ O \left(h'^{(2-r)d} h'^{-rd} h'^r h^{(1-r)d} \right) = O \left\{ h^{(1-r)d} (h')^{(r+2d-2rd)} \right\}.$$

Hence the third degree symmetrization of B has second moment bounded by

$$O \left[\left\{ h^{(1-r)d} (h')^{(r+2d-2rd)} \right\}^{\frac{2}{r}} n^{-3} \right] = O \left(h'^{-d} h^{-d} n^{-2} \right) \times O \left\{ h^{(2-r)d/r} (h')^{(2r+4d-3rd)/r} n^{-1} \right\} \\ = o \left(h'^{-d} h^{-d} n^{-2} \right) = o \left(h'^4 \right)$$

by taking r sufficiently close to 2. This is seen because if $r = 2$, the term

$$h^{(2-r)d/r} (h')^{(2r+4d-3rd)/r} n^{-1} = n^{-(d+8)/(d+4)^2(4+4d-6d)/2} n^{-1} \\ = n^{(d-2)(d+8)/(d+4)^2-1} = n^{-2(d+16)/(d+4)^2} \rightarrow 0.$$

Similar calculation can be done for term V_{ijk} as well. The conditional expectation of V_{ijk} under the condition of independence is 0., while the r -th absolute moment of the term V_{ijk} ($r = 2 + \delta, \delta > 0$) has order $O \left\{ h^{(1-r)d} (h')^{(2d-2rd)} \right\}$ and so the third degree symmetrization of V has second moment bounded by

$$O \left[\left\{ h^{(1-r)d} (h')^{(2d-2rd)} \right\}^{\frac{2}{r}} n^{-3} \right] = O \left(h'^{-d} h^{-d} n^{-2} \right) \times O \left\{ h^{(2-r)d/r} (h')^{(4d-3rd)/r} n^{-1} \right\} \\ = o \left(h'^{-d} h^{-d} n^{-2} \right) = o \left(h'^4 \right)$$

by taking r sufficiently close to 2. This is seen because if $r = 2$, then

$$h^{(2-r)d/r} (h')^{(4d-3rd)/r} n^{-1} = n^{-(d+8)/(d+4)^2(4d-6d)/2} n^{-1} \\ = n^{d(d+8)/(d+4)^2-1} = n^{-16/(d+4)^2} \rightarrow 0.$$

Hence, one can conclude by Lemma 2 of Yoshihara (1976) that $B = o(h'^2)$, $V = o(h'^2)$, which finishes the proof of this lemma.

Based on Lemma 5, one can replace $\widehat{f}(Y_{i+1}) - f(Y_{i+1})$ by

$$\begin{aligned} & \frac{1}{np(Y_{i+1})} \sum_{j=d}^{n-2} K'_{h'}(Y_j - Y_{i+1}) \{X_{j+1} - f(Y_{i+1})\} = \\ & B'(Y_{i+1})h'^2 + \frac{1}{np(Y_{i+1})} \sum_{j=d}^{n-2} K_{h'}(Y_j - Y_{i+1})\sigma(Y_j)\varepsilon_i + o(h'^2) \end{aligned}$$

with function B' as defined in (20). One then has

$$\widehat{m}(y) - \widetilde{m}(y) = \frac{I(y) + II(y) + \text{higher order terms}}{n^{-1} \sum_{i=d}^{n-2} K_h(y - Y_i)} = \frac{I(y) + II(y)}{p(y)} + \text{higher order terms}$$

where $I(y), II(y)$ are defined in (19) and (22).

Now equations (21) and (23) entails

$$\begin{aligned} \frac{I(y)}{p(y)} &= \left\{ \int B'(x, \underline{y}_{d-1})p(x, y)/p(y)dx \right\} h'^2 + O_p \left(h'^2 h^2 + \frac{h'^2}{nh^d} \right), \\ \frac{E \{II(y)\}^2}{\{p(y)\}^2} &= \frac{\|K\|_2^{4d}}{n^2 h^d h'^d p(y)^2} \int \frac{\sigma^2(x, \underline{y}_{d-1})p(x, y)}{p(x, \underline{y}_{d-1})} dx + o \left(\frac{1}{n^2 h^d h'^d} \right) \end{aligned}$$

leading one to conclude that the asymptotic bias between $\widehat{m}(y)$ and $\widetilde{m}(y)$ is

$$E \{\widehat{m}(y) - \widetilde{m}(y)\} = \left\{ \int B'(x, \underline{y}_{d-1})p(x, y)/p(y)dx \right\} h'^2 + O_p \left(h'^2 h^2 + \frac{h'^2}{nh^d} \right)$$

while the asymptotic variance of $\widehat{m}(y) - \widetilde{m}(y)$ is

$$\text{var} \{\widehat{m}(y) - \widetilde{m}(y)\} = \frac{\|K\|_2^{4d}}{n^2 h^d h'^d p(y)^2} \int \frac{\sigma^2(x, \underline{y}_{d-1})p(x, y)}{p(x, \underline{y}_{d-1})} dx + o \left(\frac{1}{n^2 h^d h'^d} \right).$$

Note here that $nh'^d \rightarrow \infty$, $h \rightarrow 0$ and $h' = o(h)$ implies that

$$\begin{aligned} E \{\widehat{m}(y) - \widetilde{m}(y)\} &= o(h^2) \\ \text{var} \{\widehat{m}(y) - \widetilde{m}(y)\} &= o \left(\frac{1}{nh^d} \right) \end{aligned}$$

and therefore

$$E \{\widehat{m}(y) - \widetilde{m}(y)\}^2 = o \left[E \{m(y) - \widetilde{m}(y)\}^2 \right]$$

which entails that

$$E \{\widehat{m}(y) - m(y)\}^2 = E \{\widetilde{m}(y) - m(y)\}^2.$$

In other words, \widehat{m} estimates m as efficiently as \widetilde{m} , the would-be estimator if one knew the true conditional mean function f . This is Theorem 1, and it is the same type of results as in Chen

(1996a,b). Here, however, we have also shown Theorem 2 which specifies the optimal choice of the bandwidth h' based on the asymptotic bias and variance formulae of $\hat{m}(y) - \tilde{m}(y)$, which is a new contribution. This is achieved by balancing the squared bias and the variance:

$$E \{ \hat{m}(y) - \tilde{m}(y) \}^2 = \left\{ \int B'(x, \underline{y}_{d-1}) p(x, y) / p(y) dx \right\}^2 h'^4 + \frac{\|K\|_2^{4d}}{n^2 h^d h'^d p(y)^2} \int \frac{\sigma^2(x, \underline{y}_{d-1}) p(x, y)}{p(x, \underline{y}_{d-1})} dx + o \left(h'^4 + \frac{1}{n^2 h^d h'^d} \right).$$

Given h being the optimal bandwidth in (11), the h' that minimizes the above expression is $h'_{opt}(y)$ given by (10), and hence the minimal mean squared error $E \{ \hat{m}(y) - \tilde{m}(y) \}^2$ is of the exact order $h'^4_{opt}(y) = C n^{-4(d+8)/(d+4)^2} = o \left(n^{-4/(d+4)} \right)$. Similarly, the mean integrated squared error over \mathcal{K} is

$$\int_{\mathcal{K}} E \{ \hat{m}(y) - \tilde{m}(y) \}^2 p(y) dy$$

which comprises of the mean integrated squared bias

$$\int_{\mathcal{K}} \frac{I^2(y)}{\{p(y)\}^2} p(y) dy = \int_{\mathcal{K}} \left\{ \int B'(x, \underline{y}_{d-1}) p(x, y) dx \right\}^2 / p(y) dy h'^4 + O_p \left(h'^4 h^2 + \frac{h'^4}{n h^d} \right)$$

and the mean integrated squared variance

$$\int_{\mathcal{K}} \frac{E \{ II(y) \}^2}{\{p(y)\}^2} p(y) dy = \frac{\|K\|_2^{4d}}{n^2 h^d h'^d} \int_{\mathcal{K}} \left\{ \int \frac{\sigma^2(x, \underline{y}_{d-1}) p(x, y)}{p(x, \underline{y}_{d-1})} dx \right\} / p(y) dy + o \left(\frac{1}{n^2 h^d h'^d} \right).$$

and the h' that minimizes the sum of these two terms is the $h'_{opt}(\mathcal{K})$ given in (12). The explicit formula (12) of the optimal bandwidth allows us to select h' automatically via a plug-in formula, whereas Chen (1996a,b) had used cross-validation method to select h' . The advantage of plug-in over cross-validation bandwidth selection is well-known, see for example, the discussion of Yang and Tschernig (1999).

Sketch of a proof of Theorem 3: The proof for the case of $k = 2$ has been shown in Theorem 1 in detail. Here we only present a sketch of a proof for $k = 3$. To prove the theorem, we only need to show that

$$A = \frac{\sum_i K_{h_3}(Y_{i-1} - y) (\hat{f}_2(Y_i) - f_2(Y_i))}{\sum_i K_{h_3}(Y_{i-1} - y)} = o(h_3^2).$$

Using arguments similar to those used in the proof of Theorem 1, one has

$$\begin{aligned} A &= \frac{1}{n p(y)} \sum_i K_{h_3}(Y_{i-1} - y) \left[\frac{\sum_j K_{h_2}(Y_j - Y_i) \{ \hat{f}_1(Y_{j+1}) - f_2(Y_i) \}}{\sum_j K_{h_2}(Y_j - Y_i)} \right] \{1 + o_p(1)\} \\ &= I + II \end{aligned}$$

where

$$\begin{aligned}
I &= \frac{1}{n^2 p(y)} \sum_i \frac{K_{h_3}(Y_{i-1} - y)}{p(Y_i)} \left[\sum_j K_{h_2}(Y_j - Y_i) \{ \hat{f}_1(Y_{j+1}) - f_1(Y_{j+1}) \} \right] \{1 + o_p(1)\} \\
II &= \frac{1}{n^2 p(y)} \sum_i \frac{K_{h_3}(Y_{i-1} - y)}{p(Y_i)} \left[\sum_j K_{h_2}(Y_j - Y_i) \{ f_1(Y_{j+1}) - f_2(Y_i) \} \right] \{1 + o_p(1)\}
\end{aligned}$$

As in the proof of Theorem 1, we can show $II = O(h_2^2) + O\left\{(n^2 h_2^d h_3^d)^{-1/2}\right\}$. Under the condition that $h_2 = o(h_3)$ and $nh_2 \rightarrow \infty$, we have $II = o(h_3^2) + o\left\{(nh_3^d)^{-1/2}\right\}$. We now concentrate on I .

$$\begin{aligned}
I &= \frac{1}{n^2 p(y)} \sum_i \frac{K_{h_3}(Y_{i-1} - y)}{p(Y_i)} \sum_j K_{h_2}(Y_j - Y_i) \left[\frac{\sum_s K_{h_1}(Y_{s+1} - Y_{j+1}) \{X_{s+2} - f_1(Y_{j+1})\}}{\sum_s K_{h_1}(Y_{s+1} - Y_{j+1})} \right] \\
&= I_1 + I_2
\end{aligned}$$

in which

$$\begin{aligned}
I_1 &= \frac{\{1 + o_p(1)\}}{n^3 p(y)} \sum_{i,j,s} \frac{K_{h_3}(Y_{i-1} - y)}{p(Y_i)} \frac{K_{h_2}(Y_j - Y_i)}{p(Y_{j+1})} [K_{h_1}(Y_{s+1} - Y_{j+1}) \{X_{s+2} - f_1(Y_{s+1})\}] \\
I_2 &= \frac{\{1 + o_p(1)\}}{n^3 p(y)} \sum_{i,j,s} \frac{K_{h_3}(Y_{i-1} - y)}{p(Y_i)} \frac{K_{h_2}(Y_j - Y_i)}{p(Y_{j+1})} [K_{h_1}(Y_{s+1} - Y_{j+1}) \{f_1(Y_{s+1}) - f_1(Y_{j+1})\}]
\end{aligned}$$

It is easy to show that $I_2 = O(h_1^2)$. Note also that $\sigma^2(Y_{s+1}) = \text{Var}(X_{s+2} | Y_{s+1})$, then considering all different cases such as $\{i_1 = i_2, j_1 = j_2\}$ as we did in the proof of Lemma 5.2, and using the strong mixing condition, we have

$$\begin{aligned}
EI_1^2 &= \frac{1 + o(1)}{n^6 p^2(y)} \sum_{i,j,s} E \left\{ \sigma^2(Y_{j+1}) \frac{K_{h_3}^2(Y_{i-1} - y)}{p^2(Y_i)} \frac{K_{h_2}^2(Y_j - Y_i)}{p^2(Y_{j+1})} K_{h_1}^2(Y_{s+1} - Y_{j+1}) \right\} \\
&= O\left\{n^{-1} + (n^2 h_3^d)^{-1} + (n^3 h_1^d h_2^d h_3^d)^{-1} + (n^2 h_1^d)^{-1} + (n^3 h_2^{2d})^{-1}\right\}.
\end{aligned}$$

Hence, under the condition of the theorem, $I_1 = o(h_3^2)$. The theorem follows.

References

- Auestad, B. and Tjøstheim, D. (1990), Identification of nonlinear time series: First order characterization and order determination, *Biometrika*, **77**, 669–687.
- Box, G.E.P. and Jenkins, G.M. (1976), *Time Series Analysis: Forecasting and Control*, Holden Day, San Francisco.
- Chen, R. (1996a), Incorporating extra information in nonparametric smoothing,” *J. Multivariate Analysis*, **58**, 133-150

- Chen, R. (1996b), A nonparametric multi-step prediction estimator in Markovian structures, *Statistica Sinica*, **6**, 603–615.
- Cleveland, W.S. and Devlin, S.J. (1988), Locally weighted regression: An approach to regression analysis by local fitting, *Journal of American Statistical Association*, **83**, 596–610
- Cline, D.B.H. and Pu, H.H. (1995), Geometric ergodicity of nonparametric nonlinear time series, *Technical Report, Dept. of Statistics, Texas A&M Univ.*
- Davydov, Yu.A. (1973), Mixing conditions for Markov chains, *Theory of Probability and its Application*, **18**, 312-328
- Fan, J. (1992), Design-adaptive nonparametric regression, *Journal of American Statistical Association* **87**, 998-1004
- Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and its Applications*, Monographs on Statistics and Applied Probability 66, Chapman & Hall.
- Fan, J. and Zhang, W.Y. (1999), Statistical estimation in varying-coefficient models, *Ann. Statist.*, **27**, 1491-1518
- Guo, M., Bai, Z. and An, H.Z. (1999) Multi-step prediction for nonlinear autoregressive models based on empirical distributions. *Statistica Sinica*, **9**, 559-570
- Györfi, L., Härdle, W., Sarda, P. and Vieu, P. (1989), *Nonparametric Curve Estimation from Time Series*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg.
- Haggan, V. and Ozaki, T. (1981), Modeling nonlinear vibrations using an amplitude-dependent autoregressive time series model, *Biometrika*, **68**, 189-196.
- Härdle, W., Lütkepohl, H. and Chen, R. (1997), A review of nonparametric time series analysis, *International Statistics Review*, **65**, 49-72
- Härdle, W., Tsybakov, A. and Yang, L. (1998) Nonparametric vector autoregression, *Journal of Statistical Planning and Inference*, **68**, 221-245.
- Härdle, W. and Vieu, P. (1992), Kernel regression smoothing for time series, *J. Time Series Analysis*, **13**, 209-232
- Jones, D.A. (1978), Nonlinear autoregressive processes, *Proceedings of the Royal Statistical Society, London*, **A, 360**, 71–95.

- Pemberton, J. (1987). Exact least squares multi-step prediction from nonlinear autoregressive models, *J. Time Series Analysis* **8**, 443-448.
- Robinson, P.M. (1983), Non-parametric estimation for time series models, *J. Time Series Analysis*, **4**, 185-208
- Tiao, G.C. and Tsay, R.S. (1994), Some advances in non-linear and adaptive modelling in time series, *Journal of Forecasting*, **13**, 109-131.
- Tjøstheim, D. (1990), Non-linear time series and Markov chains, *Adv. App. Prob.*, **22**, 587-611
- Tjøstheim, D. (1994), Nonlinear time series, a selective review, *Scand. J. Statist*, **21**, 97-130.
- Tong, H. (1983), *Threshold Models in Nonlinear Time Series Analysis*, Lecture Notes in Statistics. New York: Springer-Verlag, **21**.
- Tong, H. (1990), *Nonlinear Time Series Analysis: A Dynamical System Approach*, London: Oxford University Press.
- Tschernig, R. and Yang, L. (2000), Nonparametric lag selection for time series, *J. Time Series Analysis*, **21**, 457-487.
- Yang, L. and Tschernig, R. (1999), Multivariate bandwidth selection for local linear regression, *Journal of the Royal Statistical Society, Series B*, **61**, 793-815.
- Yoshihara, K. (1976), Limiting behavior of U-statistics for stationary absolutely regular processes, *Z. Wahrsch. Verw. Gebiete* **35**, 237-252.