J. R. Statist. Soc. B (2004)
66, *Part* 3, *pp.* 669–686

Nonparametric multistep-ahead prediction in time series analysis

Rong Chen,

University of Illinois, Chicago, USA, and Peking University, Beijing, China

Lijian Yang

Michigan State University, East Lansing, USA

and Christian Hafner

Erasmus University Rotterdam, the Netherlands

[Received November 1999. Final revision December 2003]

Summary. We consider the problem of multistep-ahead prediction in time series analysis by using nonparametric smoothing techniques. Forecasting is always one of the main objectives in time series analysis. Research has shown that non-linear time series models have certain advantages in multistep-ahead forecasting. Traditionally, nonparametric k-step-ahead least squares prediction for non-linear autoregressive AR(d) models is done by estimating $E(X_{t+k}|X_t, \ldots, X_{t-d+1})$ via nonparametric smoothing of X_{t+k} on $(X_t, \ldots, X_{t-d+1})$ directly. We propose a multistage 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 predictor proposed is more efficient. Some simulation results, advice for practical bandwidth selection and a real data example are provided.

Keywords: Improvement ratio; Local polynomial; Multistage smoothing; Optimal bandwidth; Sunspot series

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 multistep-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 multistep-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), but non-linear models may have long-term non-linear properties such as limit cycles. Recent research in non-linear time series analysis (e.g. Tong (1990) and Tjøstheim (1994)) has revealed the fact that non-linear models usually perform better than linear models in multistep-ahead prediction.

With non-linear parametric models, multistep-ahead predictions are usually done by using iterative integration or multiple-imputation methods. See Jones (1978), Pemberton (1987) and Tong (1990) for details. Guo *et al.* (1999) also proposed an iterative integration procedure without noise distribution assumptions. These procedures are based on parametric models. Their

Address for correspondence: Rong Chen, Department of Information and Decision Sciences, University of Illinois at Chicago, Chicago, IL 60607, USA.
E-mail: rongchen@uic.edu

 1369–7412/04/66669

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 speak for themselves' and avoids the difficulty of identifying an appropriate parametric model, including the non-linear functions and the error distributions. The existing nonparametric approaches for least squares multistep-ahead prediction (Robinson, 1983; Auestad and Tjøstheim, 1990; Härdle and Vieu, 1992) estimate the conditional mean function by using direct smoothing techniques. Consider a time series $\{X_t\}_{t=1}^n$ that is described by a general non-linear autoregressive AR(d) model

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

with $Y_{t-1} = (X_{t-1}, \ldots, X_{t-d})^{\mathrm{T}}$ denoting the predictor variables and f and σ the conditional mean and standard deviation functions. The noises $\{\varepsilon_t\}_{t=d+1}^n$ are independent and identically distributed (IID) with mean 0 and variance 1, independent of $X_1, \ldots, X_d$. The conditional mean $E\{X_{t+k}|Y_t = (y_1, \ldots, y_d)\}$ 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 (NW) estimator

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

where $y = (y_1, \ldots, y_d)^{\mathrm{T}}$ denote the conditioning values, K is a kernel function and the notation $K_h(y) = h^{-d} \prod_{1 \leqslant i \leqslant d} K(y_i/h)$. For local linear estimation of vector AR models, see Härdle *et al.* (1998).

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

Consider two-step ahead forecasting under a first-order non-linear 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 two-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 $(f(X_{t+1}), X_t)$, $t = 1, \ldots, n-2$, to estimate $m_2(x)$. 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 smooths 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 process

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

where ε_t is Gaussian white noise with variance 1 and $a = 1$, $b = \pi/2$ and $\sigma = 1$. We simulated a

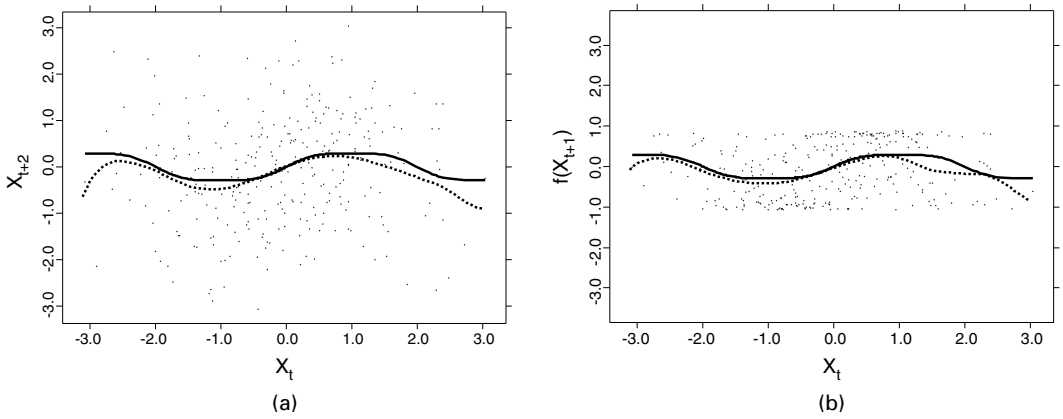

Fig. 1. True function $m(x) = \exp(-\pi^2/4)\sin\{(\pi/2)\sin(\pi x/2)\}$ (———) for a sample of $n = 300$ ($a = 1$): (a) direct predictor (·····) and the pairs $\{X_t, X_{t+2}\}$; (b) multistage predictor (·····) and the pairs $\{X_t, \hat{f}^*(X_{t+1})\}$, where $\hat{f}^*$ is the first-stage smoother for $f(x) = \sin(\pi x/2)$

series of length 300 from this model. Fig. 1(a) shows the scatterplot of (X_t, X_{t+2}). The dotted curve is the estimated mean function $E(X_{t+2}|X_t)$ using the direct NW estimator (2) and the full curve is the true mean function. Fig. 1(b) 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 dotted curve is the estimated mean function by smoothing $(X_t, \hat{f}^*(X_{t+1}))$ (the second-stage smoothing) and the full curve is the true mean function.

Throughout the paper we concentrate on the multistep-ahead prediction problem for the general non-linear 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).

Chen (1996a, b) studied similar estimators for regression analysis using NW estimators and showed that the multistage smoother does improve the estimation of the conditional mean function. In this paper, we extend the multistage 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 multistage predictor over that of the direct predictor in these general settings.

In the example that we gave previously, undersmoothing $\hat{f}^*(\cdot)$ extremely yields $\hat{f}^*(X_{t+1}) \approx X_{t+2}$ and we obtain again 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, we can achieve a smaller error without the restriction. This also offers a large amount of flexibility in multistage 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 cumulative 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 a smaller mean-squared error than the direct predictor. Results are derived

for both the NW 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 multistage ($k > 2$) predictors. Results are provided for the performance of the multistage predictor with various numbers of iterations. To demonstrate the finite sample properties of the predictor proposed, 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 $X_t, X_{t-1}, \ldots$ for the process (1). Since it is dth order Markovian, it is seen that the least square two-step prediction is

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

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

To begin with, we define the kernel estimator of the function f as

$$\hat{f}_{h'}(z) = \frac{\sum_{j=d}^{n-2} K_{h'}(z - Y_j) X_{j+1}}{\sum_{j=d}^{n-2} K_{h'}(z - Y_j)}. \tag{4}$$

On the basis of this estimator, let

$$\hat{f}^*_{h'}(Y_{t+1}) = w(Y_{t+1})\, \hat{f}_{h'}(Y_{t+1}) + \{1 - w(Y_{t+1})\} X_{t+2},$$

where $w(z) = I(z \in \mathcal{K})$ for an arbitrary large compact set $\mathcal{K}$. Alternatively, we can write

$$\hat{f}^*_{h'}(Y_{t+1}) = \begin{cases} \hat{f}_{h'}(Y_{t+1}) & \text{if } Y_{t+1} \in \mathcal{K}, \\ X_{t+2} & \text{if } Y_{t+1} \notin \mathcal{K}. \end{cases} \tag{5}$$

The use of weight function $w(z)$ is to avoid the estimation of $f(Y_{t+1})$ for large Y_{t+1}-values, since the smoothing estimator converges uniformly with optimal rate only on compact ranges. Such a technique of screening off extreme observations has been commonly used in the literature; see, for example, Tschernig and Yang (2000).

Then the two-stage predictor is defined as

$$\hat{m}(y) = \hat{m}_2(y) = \hat{m}_{2;h,h'}(y) = \frac{\sum_{t=d}^{n-2} K_h(y - Y_t)\, \hat{f}^*_{h'}(Y_{t+1})}{\sum_{t=d}^{n-2} K_h(y - Y_t)}, \tag{6}$$

In what follows, the density of Y_t is denoted as $p(\cdot)$, the gradient operator of a multivariate function as ∇ and the Laplacian operator is denoted as $\mathrm{tr}(\nabla^2)$, the trace of the Hessian matrix. Let $u^2(y) = \mathrm{var}\{f(Y_{t+1})|Y_t = y\}$ and $v^2(y, \mathcal{K}) = E[\{1 - w(Y_{t+1})\}\sigma^2(Y_{t+1})|Y_t = y]$.

First we list some conditions that are needed for the theorems.

(a) The noise ε_t is IID with mean 0 and variance 1. The function $\sigma(\cdot)$ is continuous and is positive on set $\mathcal{K}$.
(b) The process $\{X_t\}_{t\geqslant 0}$ is stationary and geometrically β mixing. Sets of sufficient conditions for geometric ergodicity can be found in Tjøstheim (1990) and Davydov (1973).
(c) The functions f and m are twice continuously differentiable.
(d) The stationary density $p(\cdot)$ of Y_t exists, is bounded, continuous and bounded from below on S with interior S^0 such that $S^0 \supset \mathcal{K} \supset \mathcal{K}_0$ and continuously differentiable on S^0.
(e) The kernel K is a compactly supported, symmetric probability density.

Theorem 1. Under conditions (a)–(e), and $h = \beta n^{-1/(d+4)}$ for some $\beta > 0$, $h' = o(h)$ and $h = o(h' n^{\delta'})$, $\forall \delta' > 0$, we have

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

where $s^2(y) = u^2(y) + v^2(y, \mathcal{K})$, and

$$B(y) = \frac{\mu_2(K)}{2}\mathrm{tr}\{\nabla_y^2 m_2(y)\} + \frac{\mu_2(K)}{p(y)}\nabla_y^T m_2(y)\, \nabla_y p(y) \tag{7}$$

and $\mu_2(K) = \int u^2\, K(u)\, \mathrm{d}u$ and $\|K\|_2^2 = \int K^2(u)\, \mathrm{d}u$.

To understand theorem 1, we introduce a 'genie' smoother

$$\tilde{m}(y) = \tilde{m}_2(y) = \frac{\sum_{t=d}^{n-2} K_h(y - Y_t)\, f^*(Y_{t+1})}{\sum_{t=d}^{n-2} K_h(y - Y_t)} \tag{8}$$

in which

$$f^*(Y_{t+1}) = f(Y_{t+1}) + \{1 - w(Y_{t+1})\}\, \sigma(Y_{t+1})\varepsilon_{t+2} = \begin{cases} f(Y_{t+1}) & \text{if } Y_{t+1} \in \mathcal{K}, \\ X_{t+2} & \text{if } Y_{t+1} \notin \mathcal{K}, \end{cases}$$

where $w(Y_{t+1}) = I(Y_{t+1} \in \mathcal{K})$, the indicator function. This estimator assumes perfect knowledge of the function f (except outside the set $\mathcal{K}$). Because of model (1), X_{t+2} is a noisier version of $f(Y_{t+1})$; hence the genie estimator should perform better than the director estimator in equation (2).

Of course in reality we do not have perfect knowledge of the function f, but the following lemma explains that the difference between the two-staged predictor $\hat{m}_2(y)$ and the genie estimator $\tilde{m}_2(y)$ is asymptotically negligible compared with that between $\tilde{m}_2(y)$ and the true function $m_2(y)$. Hence the two-stage predictor behaves the same, asymptotically, as the genie estimator.

Lemma 1 is based on the observation that, owing to the choice of bandwidths h and $h' = o(h)$, the asymptotic bias between $\hat{f}_{h'}^*(Y_{t+1})$ and $f^*(Y_{t+1})$ is of order h'^2, which is also the bias order between $\hat{m}_2(y)$ and $\tilde{m}_2(y)$. This bias is negligible compared with the bias between $\tilde{m}_2(y)$ and $m_2(y)$ (of order h^2). Meanwhile, the condition that $h/h' = a_n$ goes to ∞ slower than any power of n ensures that the asymptotic variance of approximating $\tilde{m}_2(y)$ with $\hat{m}_2(y)$ is of order $1/nh^{d-1}$, which is negligible compared with the variance order of approximating $m_2(y)$ with $\tilde{m}_2(y)$ (of order $1/nh^d$).

Lemma 1. Under the conditions of theorem 1, the mean-squared error between $\hat{m}_2(y)$ and $\tilde{m}_2(y)$ is

$$E\{\hat{m}_2(y)-\tilde{m}_2(y)\}^2=\frac{\|K\|_2^{2(d-1)}}{nh^{d-1}\,p(y)^2}\int\frac{w^2(x,y_{-(d-1)})}{p(x,y_{-(d-1)})}\,\sigma^2(x,y_{-(d-1)})\,p^2(x,y)\,\mathrm{d}x$$
$$+\left\{\int B'(x,y_{-(d-1)})\,p(x,y)/p(y)\,\mathrm{d}x\right\}^2 h'^4+o(h'^4+1/nh^{d-1}), \quad (9)$$

in which $y_{-(d-1)}$ denotes $(y_1,\ldots,y_{d-1})$, the first $d-1$ elements of y, and $p(x,y_{-(d-1)})$ denotes the density of Y_i at $(x,y_1,\ldots,y_{d-1})$ and $p(x,y)$ the density of (X_i,Y_{i-1}) at $(x,y_1,\ldots,y_d)$ and function $B'(\cdot)$ is defined as

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

In particular

$$E\{\hat{m}_2(y)-\tilde{m}_2(y)\}^2=o[E\{m_2(y)-\tilde{m}_2(y)\}^2]. \quad (10)$$

Similarly, the mean integrated squared error over $\mathcal{K}_0$ is

$$\int_{\mathcal{K}_0}E\{\hat{m}(y)-\tilde{m}(y)\}^2\,p(y)\,\mathrm{d}y$$
$$=\frac{\|K\|_2^{2(d-1)}}{nh^{d-1}}\int_{\mathcal{K}_0}\left\{\int\frac{w^2(x,y_{-(d-1)})}{p(x,y_{-(d-1)})}\sigma^2(x,y_{-(d-1)})p^2(x,y)\,\mathrm{d}x\right\}\Big/p(y)\,\mathrm{d}y$$
$$+\int_{\mathcal{K}_0}\left\{\int B'(x,y_{-(d-1)})\,p(x,y)\,\mathrm{d}x\right\}^2\Big/p(y)\,\mathrm{d}y h'^4+o_p(h'^4+1/nh^{d-1}). \quad (11)$$

It is interesting to note from lemma 1 that, although two bandwidth parameters h and h' are used in $\hat{m}_2(y)$, asymptotically the effect of h' is only in the bias of approximating $\tilde{m}_2(y)$ with $\hat{m}_2(y)$; hence there is not an obvious way to balance the bias and variance due to the first-stage bandwidth h'. In contrast, according to theorem 1, the second-stage bandwidth h appears in both the bias and the variance of estimating $m_2(y)$ with $\hat{m}_2(y)$. Hence we have the following simple result.

Corollary 1. Under the conditions of theorem 1, the optimal bandwidth h for the two-step estimation at y is

$$h_{\mathrm{opt}}(y)=\left\{\frac{d}{4n}\|K\|_2^{2d}\,s^2(y)\,B^{-2}(y)\,p^{-1}(y)\right\}^{1/(d+4)}. \quad (12)$$

The optimal bandwidth h for two-step estimation over $\mathcal{K}_0$ is

$$h_{\mathrm{opt}}(\mathcal{K}_0)=\left[\frac{d}{4n}\|K\|_2^{2d}\int_{\mathcal{K}_0}s^2(y)\,dy\left\{\int_{\mathcal{K}_0}B^2(y)\,p(y)\,\mathrm{d}y\right\}^{-1}\right]^{1/(d+4)}. \quad (13)$$

In practice we could use $h'=h/\log(n)$ to satisfy $h'=o(h)$ and $h=o(h'n^{\delta})$, $\forall\delta>0$, which works quite well in simulations. 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 the second stage. For more general results on plug-in bandwidth selection in multivariate settings, see Yang and Tschernig (1999).

Detailed proofs of theorem 1 and lemma 1 are given in Chen *et al.* (2004). All the same conclusions are true for local linear regressors as well. The only change is that the bias function

$B(y)$ becomes

$$B(y) = \frac{\mu_2(K)}{2} \mathrm{tr}\{\nabla_y^2 m_2(y)\}. \tag{14}$$

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

Corollary 2. Under the conditions of theorem 1, at a single point y, the ratio of the asymptotic mean-squared error of the two-stage smoother (6) and the direct smoother (2), both using optimal bandwidths, is

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

where $v^2(y) = E\{\sigma^2(Y_{t+1})|Y_t = y\}$.

The same ratio over a d-dimensional compact set $\mathcal{K}_0$ that contains y is

$$r = \left\{ \frac{\int_{\mathcal{K}_0} u^2(y)\,\mathrm{d}y + \int_{\mathcal{K}_0} v^2(y, \mathcal{K})\,\mathrm{d}y}{\int_{\mathcal{K}_0} u^2(y)\,\mathrm{d}y + \int_{\mathcal{K}_0} v^2(y)\,\mathrm{d}y} \right\}^{4/(d+4)}. \tag{15}$$

Remark 1. Theorem 1 says that asymptotically the two-stage predictor behaves the same as if we knew exactly the function f in the compact set $\mathcal{K}$. The improvement of the mean-squared error over the direct predictor (2) is due to the smaller asymptotic variance $u^2(y) + v^2(y, \mathcal{K})$ *versus* $\mathrm{var}(X_{t+2}|Y_t = y) = u^2(y) + v^2(y)$. With large $\mathcal{K}$, we can achieve a significant improvement, since $v^2(y, \mathcal{K}) < v^2(y)$. If $\sigma(Y_{t+1}) = \sigma^2$, a constant, then $v^2(y, \mathcal{K}) = P(Y_{t+1} \notin \mathcal{K}|Y_t = y)\sigma^2$. The compact set $\mathcal{K}$ is used here for technical reasons in proving the asymptotics. In theory, as the sample size increases, we can also increase the size of $\mathcal{K}$, to increase the improvement. In practice we obtained satisfactory results with very large $\mathcal{K}$ that covers the entire data range.

Remark 2. We can replace the NW 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 the mean-squared error becomes at most

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

where $p = 0$ represents the improvement for the NW and local linear estimators.

Remark 3. The multistage predictor is sensitive to the correctness of the model specification, particularly the AR order. For example, examine the equation

$$E(X_{t+2}|X_t) = E\{E(X_{t+2}|X_{t+1}, X_t)|X_t\} = E\{E(X_{t+2}|X_{t+1})|X_t\}.$$

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

Remark 4. Theorem 1 requires the first-stage bandwidth to be of smaller order than the second-stage bandwidth. This requirement basically ensures that the bias that is created in the first-stage smoothing is of a 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, though our simulation shows that the final result is not very sensitive to the selection. Lastly, we can also use a plug-in bandwidth for the selection of second-stage bandwidth h as in corollary 1, together with $h' = h/\log(n)$.

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

3.1. Example 1

Let us first consider an extension of process (3) given as

$$X_t = a\sin(bX_{t-1}) + \sigma(X_{t-1})\varepsilon_t \tag{16}$$

with

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

where $\omega > 0$ and $\alpha \geqslant 0$. 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 (1999). In our study, we let α be alternatively 0, 0.2 and 0.5, and $\omega = 1 - \alpha$. Note that $\alpha = 0$ corresponds to process (3).

To reduce the effect of outliers, each generated series was trimmed at its 0.5% and 99.5% quantile. We use cross-validation optimal bandwidths for the direct smoother and the second-stage of the multistage smoother. For the first-stage smoothing we should undersmooth. To see the sensitivity of the results with respect to the degree of undersmoothing, we use three bandwidths for the first-stage smoothing: $h' = h^*$, $h' = h^*/5$ and $h' = h^*/10$, in which h^* denotes the cross-validation optimal bandwidth for the NW smoother. Figs 1 and 2 illustrate the direct and multistage smoothers *versus* the true function for one sample of size $n = 300$ with $\alpha = 0$ and $a = 1$.

Table 1 provides summary statistics for the improvement rates for 200 replications of process (16) with sample size n. Since the distribution of $\hat{r}$ is highly skewed, we present their quantiles. The theoretical (and optimal) improvement rate r in equation (15) in this case is

$$r = \left(\frac{\int_{-c}^{c} \{u^2(x) + v^2(x, [-c, c])\}\, \mathrm{d}x}{\int_{-c}^{c} \{u^2(x) + v^2(x)\}\, \mathrm{d}x} \right)^{4/5} \tag{17}$$

where c determines the interval of interest. We have chosen $c = 4$ for the case $a = 1$ and $c = 5$ for the case $a = 2$. The intervals $[-c, c]$ cover about all the simulated data points, and hence

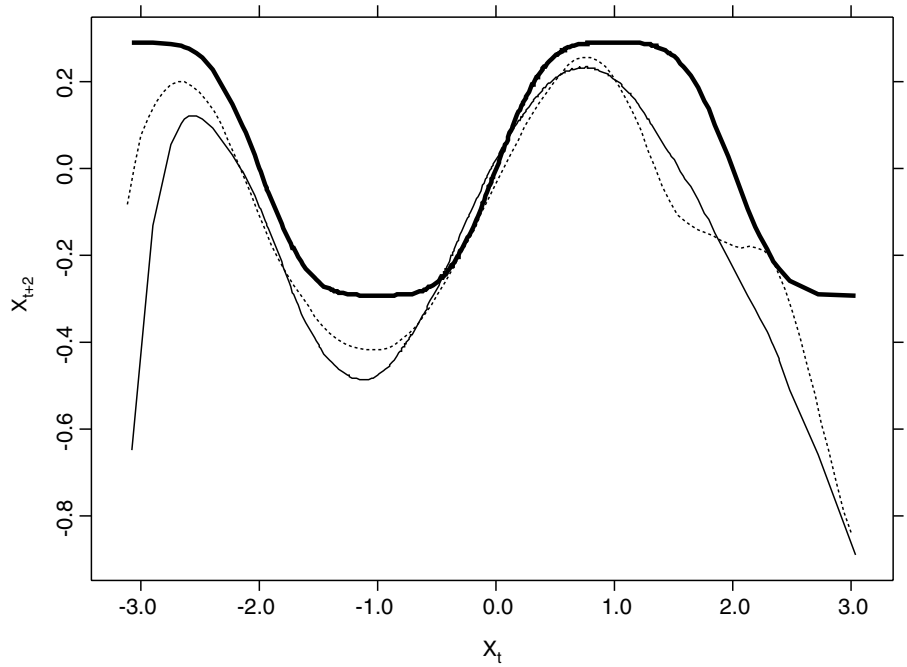

Fig. 2. True function $m(x) = \exp(-\pi^2/4)\sin\{(\pi/2)\sin(\pi x/2)\}$ (———) for a sample of $n = 300$ ($a = 1$; improvement rate 0.789), multistage predictor (· · · · · · ·) and direct predictor (———)

approximately we can take $v^2(x, [-c, c]) = 0$. Functions $u^2(x)$ and $v^2(x)$, which are obtained through elementary calculation, are

$$\begin{aligned} u^2(x) &= \operatorname{var}\{E(X_{t+2}|X_{t+1})|X_t = x\} \\ &= \tfrac{1}{2}a^2[1 - \exp\{-b^2(\omega + \alpha x^2)\}][1 + \exp\{-b^2(\omega + \alpha x^2)\}\cos\{2ab\sin(bx)\}] \end{aligned}$$

and

$$v^2(x) = E\{\operatorname{var}(X_{t+2}|X_{t+1})|X_t = x\} = \omega(1+\alpha) + \alpha a^2 \sin^2(bx) + \alpha^2 x^2$$

respectively. For each simulated series, we estimated m_2 by using both the direct predictor $\hat{m}^*$ in equation (2) and the two-stage predictor $\hat{m}$ in equation (6), and calculated

$$\hat{r} = \frac{\sum_{t=1}^{n-2}\{\hat{m}_2(Y_t) - m_2(Y_t)\}^2}{\sum_{t=1}^{n-2}\{\hat{m}_2^*(Y_t) - m_2(Y_t)\}^2}, \tag{18}$$

where the true prediction function is calculated to be

$$m_2(x) = a\sin\left\{\frac{\pi}{2}a\sin\left(\frac{\pi}{2}x\right)\right\}\exp\left\{-\frac{\pi^2}{8}(\omega + \alpha x^2)\right\}.$$

From the results in Table 1, the multistage predictor clearly outperforms the direct predictor. It appears that the improvement rate is not very sensitive to the degree of undersmoothing.

Table 1. Example 1: summary statistics of the simulated ratios of improvement†

(a, α)	h'	r	*Results for n = 300*					*Results for n = 1000*				
			Minimum	*25%*	*50%*	*75%*	*Maximum*	*Minimum*	*25%*	*50%*	*75%*	*Maximum*
(1, 0)	h^*	0.39	0.17	0.45	0.62	0.86	2.56	0.17	0.39	0.56	0.72	2.49
	$h^*/5$	0.39	0.29	0.51	0.58	0.65	0.92	0.13	0.39	0.51	0.70	1.41
	$h^*/10$	0.39	0.32	0.71	0.87	1.04	1.94	0.08	0.22	0.31	0.44	0.93
(1, 0.2)	h^*	0.27	0.11	0.41	0.61	0.83	2.44	0.13	0.38	0.52	0.70	1.53
	$h^*/5$	0.27	0.16	0.49	0.64	0.85	2.74	0.12	0.28	0.39	0.54	1.14
	$h^*/10$	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^*	0.17	0.08	0.24	0.35	0.48	1.02	0.12	0.34	0.46	0.60	1.53
	$h^*/5$	0.17	0.10	0.47	0.64	0.82	1.51	0.10	0.37	0.50	0.67	1.45
	$h^*/10$	0.17	0.08	0.23	0.33	0.44	0.84	0.14	0.29	0.39	0.49	1.18
(2, 0)	h^*	0.71	0.41	0.71	0.81	0.94	1.61	0.42	0.65	0.76	0.88	1.58
	$h^*/5$	0.71	0.86	0.98	0.99	1.01	1.06	0.22	0.44	0.52	0.61	0.98
	$h^*/10$	0.71	0.68	0.89	0.95	0.99	1.19	0.39	0.68	0.79	0.89	1.40
(2, 0.2)	h^*	0.53	0.21	0.63	0.83	1.04	2.24	0.31	0.55	0.66	0.77	1.23
	$h^*/5$	0.53	0.32	0.48	0.61	0.76	1.71	0.40	0.64	0.71	0.85	1.35
	$h^*/10$	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^*	0.33	0.38	0.73	0.94	1.40	13.77	0.39	0.51	0.57	0.67	1.21
	$h^*/5$	0.33	0.22	0.66	0.77	0.90	1.53	0.34	0.51	0.58	0.66	1.12
	$h^*/10$	0.33	0.05	0.99	1.03	1.08	2.31	0.26	0.49	0.61	0.71	1.19

†The minima, maxima and quartiles of the simulated improvement rates $\hat{r}$ are given. r is the theoretical ratio in equation (17).

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

3.2. Example 2

Our second example uses an exponential AR model (Haggan and Ozaki, 1981), given by

$$X_t = \{0.7 + 0.5 \exp(-cX_{t-1}^2)\} X_{t-1} + \sigma\varepsilon_t \tag{19}$$

with ε_t IID as $N(0, 1)$ and alternative (c, σ) combinations. As in example 1, we use cross-validation optimal bandwidths. 200 series are generated, each of size 400. Table 2 shows the quartiles of the improvement rate r. Again, the multistage predictor outperforms the direct predictor. Note that, as the noise variance σ^2 is increased, the improvement increases. The same happens when the parameter c is increased.

3.3. Example 3

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

$$X_t = a_1 \sin(b_1 X_{t-1}) + a_2 \sin(b_2 X_{t-2}) + \sigma\varepsilon_t \tag{20}$$

with ε_t IID as $N(0, 1)$. For two-step-ahead prediction, the true conditional mean function is

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

Table 2. Example 2: theoretical improvement rate r and quartiles of the improvement rate $\hat{r}$ of simulated series for various combinations of (c, σ) and bandwidth for the first-stage smoothing using model (19)

(c, σ)	r	*Results for* $h_1 = h_1^*$					*Results for* $h_1 = h_1^*/5$					*Results for* $h_1 = h_1^*/10$				
		Minimum	*25%*	*50%*	*75%*	*Maximum*	*Minimum*	*25%*	*50%*	*75%*	*Maximum*	*Minimum*	*25%*	*50%*	*75%*	*Maximum*
(0.15, 1)	0.52	0.16	0.54	0.75	1.00	3.80	0.22	0.51	0.70	0.88	2.30	0.24	0.55	0.75	0.92	2.00
(0.15, 0.5)	0.56	0.58	1.10	1.30	1.40	2.60	0.39	0.78	0.91	1.00	1.50	0.39	0.79	0.91	1.00	1.50
(0.5, 1)	0.43	0.10	0.65	0.94	1.20	4.40	0.17	0.50	0.72	0.93	2.10	0.19	0.53	0.75	0.94	2.60
(0.5, 0.5)	0.52	0.21	0.77	1.10	1.40	3.20	0.16	0.65	0.83	1.00	2.40	0.21	0.68	0.85	1.00	2.30

This function is plotted in Fig. 3(c) with $a_1 = a_2 = \frac{1}{2}$, $b_1 = \pi/4$ and $b_2 = \pi$. Using these parameters and $\sigma = \frac{1}{2}$, we generated a series from model (20) and estimated m_2 by using local linear estimates for the direct predictor and the multistage predictor, both shown also in Fig. 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 error of direct smoothing was 0.0093, whereas that of multistage smoothing was 0.0064. The quartiles of the improvement ratios defined in equation (18) were 0.6154, 0.6873 and 0.7616, with a minimum of 0.4319 and a maximum of 1.033.

4. Multistage predictor for multistep-ahead prediction

For non-linear AR(d) models in equation (1), multistep prediction can be done recursively by using the multistage smoother. Define $f_1(y) = E(X_{t+1}|Y_t = y)$ and for $j = 2, \ldots, k$ recursively define $f_j(y) = E\{f_{j-1}(Y_{t+1})|Y_t = y\}$. Then

$$\begin{aligned} m_k(y) &= E(X_{t+k}|Y_t = y) = E\{f_1(Y_{t+k-1})|Y_t = y\} = E\{f_2(Y_{t+k-2})|Y_t = y\} \\ &= \ldots = E\{f_{k-1}(Y_{t+1})|Y_t = y\} = f_k(y). \end{aligned}$$

Note that $\mathrm{var}\{f_j(Y_{t+k-j})\} = \mathrm{var}\{f_{j+1}(Y_{t+k-j-1})\} + E[\mathrm{var}\{f_j(Y_{t+k-j})|Y_{t+k-j-1}\}]$ and therefore $\mathrm{var}\{f_{j+1}(Y_{t+k-j-1})\} \leqslant \mathrm{var}\{f_j(Y_{t+k-j})\}$, which holds for all j. Applying this recursively leads to $\mathrm{var}\{f_k(Y_t)\} \leqslant \mathrm{var}\{f_1(Y_{t+k-1})\}$, 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.

In what follows, we write

$$X_{t+k} = f_{k-1}(Y_{t+1}) + \sigma_{k-1}(Y_{t+1})\varepsilon_{t+k,k-1} \tag{21}$$

where $f_{k-1}(z) \equiv E(X_{t+k}|Y_{t+1} = z)$ and $\sigma^2_{k-1}(z) \equiv \mathrm{var}(X_{t+k}|Y_{t+1} = z)$, and $E(\varepsilon_{t+k,k-1}|Y_{t+1} = z) = 0$ and $\mathrm{var}(\varepsilon_{t+k,k-1}|Y_{t+1} = z) = 1$. For clearer presentation, we shall use the NW smoother. The method can be immediately extended to using the local polynomial estimator. Starting with $X_t^{(0)} = X_t$, we repeat the following steps for $j = 1, \ldots, k-1$.

Stage j: estimate

$$X_{t+k}^{(j)} = \hat{f}_j^*(Y_{t+k-j}) = \begin{cases} \hat{f}_j(Y_{t+k-j}) & \text{if } Y_{t+k-j} \in \mathcal{K}, \\ X_{t+j} & \text{if } Y_{t+k-j} \notin \mathcal{K} \end{cases}$$

where

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

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

$$\hat{m}_k(y) = \frac{\sum_{t=d}^{n-k} K_{h_k}(y - Y_t) X_{t+k}^{(k-1)}}{\sum_{t=d}^{n-k} K_{h_k}(y - Y_t)}. \tag{22}$$

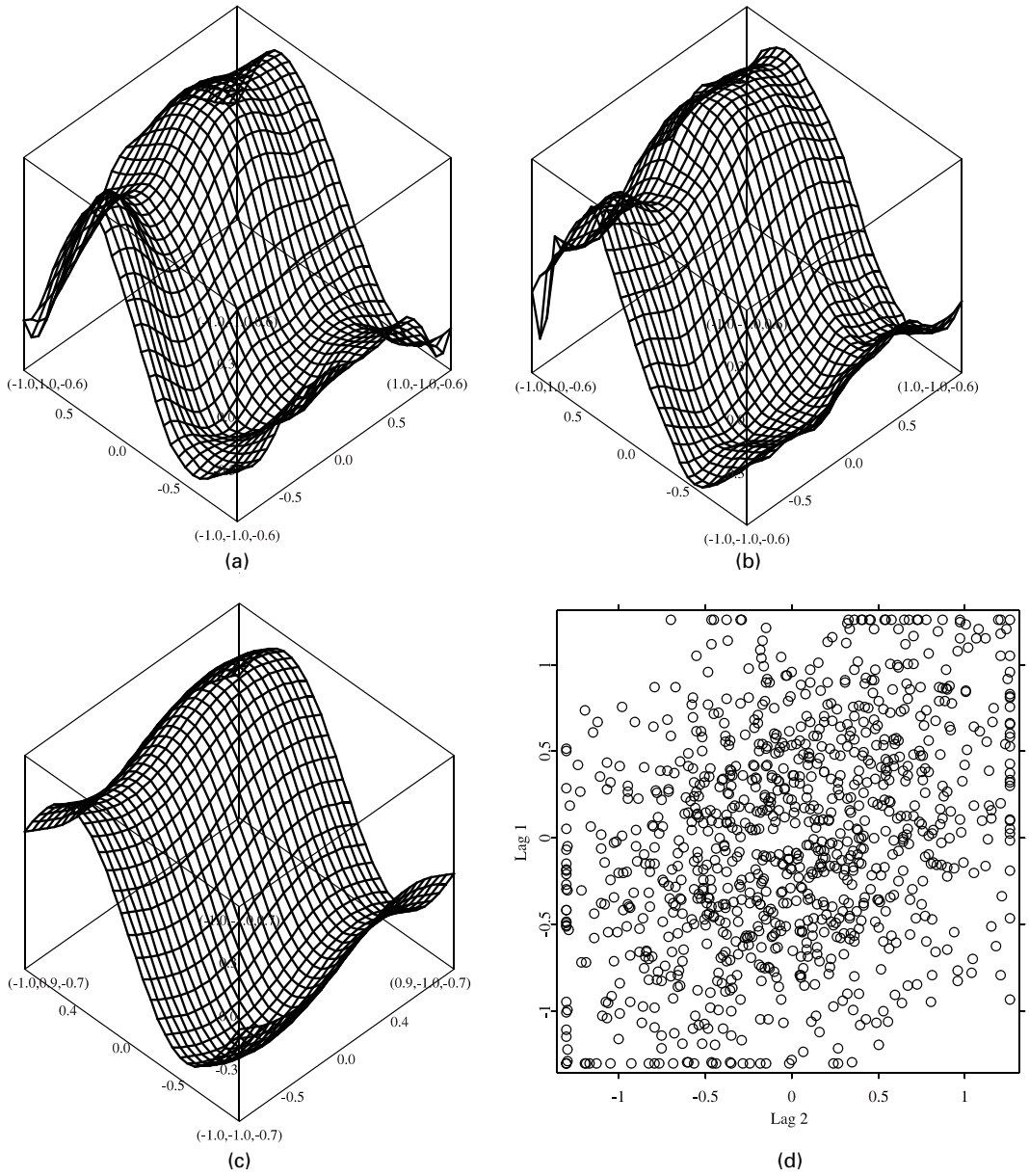

Fig. 3. Results for a generated sample of model (20) with $n = 1000$ (the ratio of the mean-squared error for this sample was 0.50; the bandwidths were 0.4055 for the direct predictor, 0.2408 for the first stage and 0.2797 for the second stage): (a) direct predictor; (b) multistage predictor; (c) true function; (d) scatterplot of the predictors (in (a)–(c) the left-hand axis is the first lag (x_{t-1}) and the right-hand axis is the second lag (x_{t-2}))

Graphically, this recursive method can be presented as

$$X_{t+k} \overset{(X_{t+k},Y_{t+k-1})}{\Longrightarrow} X_{t+k}^{(1)} \overset{(X_{t+k}^{(1)},Y_{t+k-2})}{\Longrightarrow} X_{t+k}^{(2)} \overset{(X_{t+k}^{(2)},Y_{t+k-3})}{\Longrightarrow} \cdots \overset{(X_{t+k}^{(k-2)},Y_{t+1})}{\Longrightarrow} X_{t+k}^{(k-1)} \overset{(X_{t+k}^{(k-1)},Y_t)}{\Longrightarrow} m_k(y).$$

We have the following theorem.

Theorem 2. Under the conditions of theorem 1, if $h_k = \beta n^{-1/(d+4)}$ for some $\beta > 0$, $h'_j = o(h_k)$ and $h_k = o(h'_j n^{\delta'})$, $\forall \delta' > 0$ for $j = 1, \ldots, k-1$, we have

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

where

$$B(y) = \frac{\mu_2(K)}{2}\mathrm{tr}\{\nabla_y^2 m_k(y)\} + \mu_2(K)\nabla_y^T m_k(y)\nabla p(y)/p(y),$$

and

$$\begin{aligned} s_k^2(y) &= \mathrm{var}\{f_{k-1}(Y_{t+1})|Y_t = y\} + \mathrm{var}[\{1 - w(Y_{t+1})\}\{X_{t+k} - f_{k-1}(Y_{t+1})\}|Y_t = y] \\ &= \mathrm{var}\{f_{k-1}(Y_{t+1})|Y_t = y\} + \mathrm{var}[\{1 - w(Y_{t+1})\}\,\sigma_{k-1}^2(Y_{t+1})|Y_t = y]. \end{aligned}$$

The proof of the theorem is very tedious and we show only a sketch in Chen *et al.* (2004).

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, whereas the estimator (22) has smaller variance. The ratio of the asymptotic optimal mean squared error of the multistage smoother (22) and the direct smoother (2) at a single point is then

$$\begin{aligned} r(y) &= \left(\frac{\mathrm{var}\{f_{k-1}(Y_{t+1})|Y_t = y\} + \mathrm{var}[\{1 - w(Y_{t+1})\}\{X_{t+k} - f_{k-1}(Y_{t+1})\}|Y_t = y]}{\mathrm{var}(X_{t+k}|Y_t = y)}\right)^{4/(d+4)} \\ &= \left(\frac{\mathrm{var}\{f_{k-1}(Y_{t+1})|Y_t = y\} + \mathrm{var}[\{1 - w(Y_{t+1})\}\,\sigma_{k-1}^2(Y_{t+1})|Y_t = y]}{\mathrm{var}(X_{t+k}|Y_t = y)}\right)^{4/(d+4)}. \end{aligned}$$

The improvement ratio over a compact set can be obtained similarly as equation (15).

Also note that the above asymptotic result is the same as a two-step procedure:

(a) estimate f_k by smoothing Y_{t+k} on Y_{t+1} and obtain $\hat{f}_{k-1}^*(Y_{t+1})$; then
(b) estimate m_k by smoothing $\hat{f}_{k-1}^*(Y_{t+1})$ on Y_t.

The improvements in using the extra intermediate steps are asymptotically of smaller order. However, the benefit of these intermediate steps can be seen with a finite sample size, because 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 increasingly difficult as the number of steps k increases, owing to the difficulties in selecting a bandwidth for each step. There is a strong tendency to oversmooth, owing to the large amount of smoothing that is involved. Theoretically, we have shown that the bandwidths at the earlier stages $h_1, \ldots, h_{k-1}$ should be of smaller order than the optimal bandwidth (to keep the bias that is introduced in the early stages negligible) whereas the final stage uses the optimal rate. Simultaneous bandwidth selection of $(h_1, \ldots, h_{k-1})$ by 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 0, since the intermediate steps are less important asymptotically. This enables us to control the number of smoothing parameters that are used, while still benefiting from the multistage smoothing procedure. However, the second-to-last step (obtaining $\hat{f}_{k-1}^*$) should not be skipped. This can be seen in the following simple example. For a non-linear 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 $\mathrm{var}\{f_2(X_{t+1})|X_t = x\} \leqslant \mathrm{var}\{f(X_{t+2})|X_t = x\}$, by theorem 1 we should smooth X_{t+3} on X_{t+1} to obtain an estimate of f_2, and 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 a 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-last step should not be skipped.

4.1. Example 4

To check the performance of the recursive multistage prediction estimator, we generated 200 series from the exponential AR model (19) with different (c, σ) combinations. We tried three different recursive schemes for four-step prediction: four-stage smoothing (4), 4–3–2–1–0; three-stage (3), 4–3–1–0 (i.e. $h_2 = 0$); two-stage (2), 4–2–0 (i.e. $h_1 = 0$ and $h_3 = 0$). Table 3 shows the improvement rate by using different bandwidths (the cross-validation optimal, the cross-validation optimal over 5 and the cross-validation optimal over 10), except for the last stage, which always uses the optimal cross-validation bandwidth.

We can see from Table 3 that the multistage estimator has a 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 well as 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 Fig. 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 ith smoothed version of X_{t+4} after the ith stage of smoothing. We can see that the variation of $X_{t+4}^{(i)}$ becomes increasingly 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 bench-mark. We have chosen the famous sunspot

Table 3. Quartiles of the improvement rates of four-step prediction for model (19) using various (c, s) combinations, various early stage bandwidth selection and various recursion schemes, based on 200 simulated series, each of size 400

(c, s)		*Results for* $h_i = h_i^*$			*Results for* $h_i = h_i^*/5$			*Results for* $h_i = h_i^*/10$		
		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

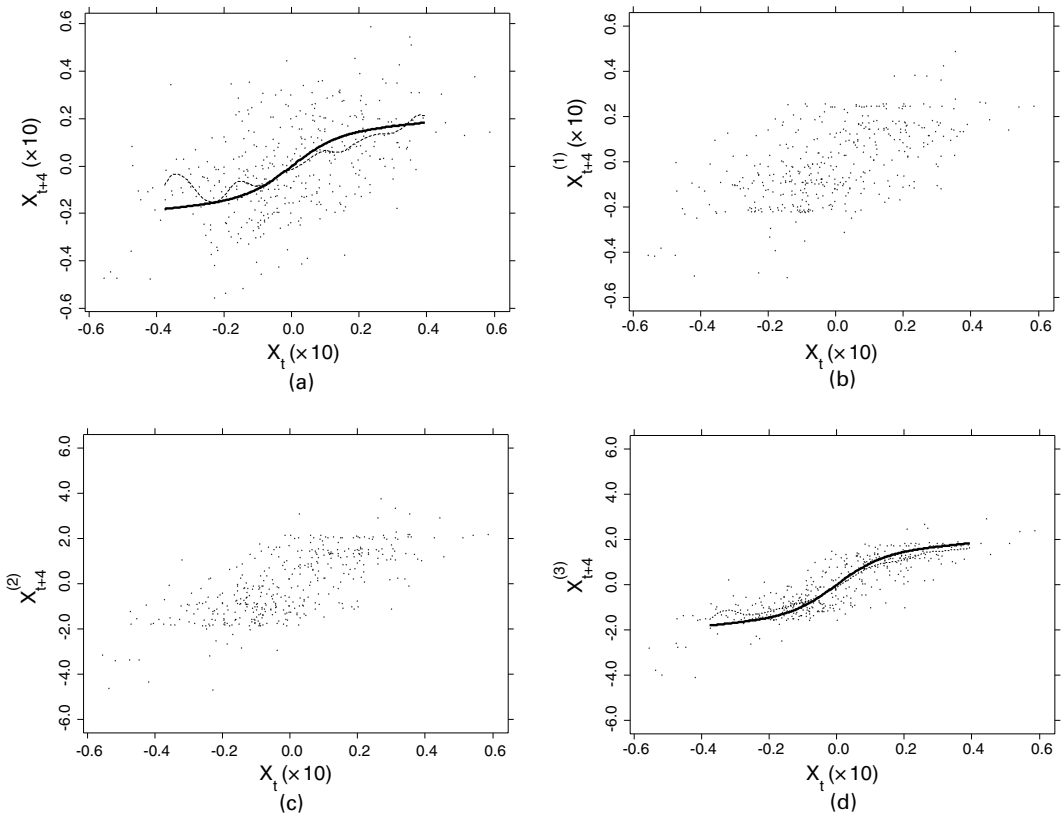

Fig. 4. Process of four-stage smoothing for four-step-ahead prediction using model (19) with a sample size of 400: (a) X_{t+4} *versus* X_t (———, true conditional expectation $E(X_{t+4}|X_t)$; ·····, direct estimate); (b) $X^{(1)}_{t+4}$ *versus* X_t after the first-stage smoothing; (c) $X^{(2)}_{t+4}$ *versus* X_t after the second-stage smoothing; (d) $X^{(3)}_{t+4}$ *versus* X_t after the third-stage smoothing (·····, final estimate)

data, i.e. the yearly average numbers of sunspots, as provided by the Royal Observatory of Belgium (`http://www.oma.be`) for the years 1700–1997. A short version of this series has been analysed for example by Tong (1990), page 419, and Fan and Gijbels (1996), page 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 a quartic kernel. The optimal bandwidths are obtained by leave-one-out cross-validation. Similarly to Tong (1990), page 425, we use the last 20 years (1978–1997) as the prediction period, which covers roughly two cycles. Hence, the function is estimated by 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 one- to three-step-ahead prediction are 25.49, 22.02 and 25.49. For the ith stage bandwidth of the multistage smoother we tried h_i^*/j, where h_i^* is the cross-validation optimal bandwidth of the ith stage, $i<k$ and $j=1,2,3,\ldots,10$. For the kth stage the cross-validation optimal bandwidth was used, $k=2,3$. Table 4 reports the results for the ratio of the mean-square prediction error $\tilde{r}$.

Obviously, multistage smoothing substantially improves direct prediction for two-step-ahead prediction. In Fig. 5 the two predictors are visualized. It can be seen that the variance of the two-stage smoother is much smaller. The drastic dip in the direct predictor at around $X_t=1.6\times10^2$ apparently indicates that the bandwidth is small owing to the large amount of noise in the

Table 4. Results from the sunspot analysis†

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

†h_{dir} is the bandwidth for direct smoothing, h_i is the bandwidth used at the ith stage of the multistage smoother, where h_i^* denotes the cross-validation optimal bandwidth of the ith stage, and $\tilde{r}$ is the ratio of mean-square prediction errors.

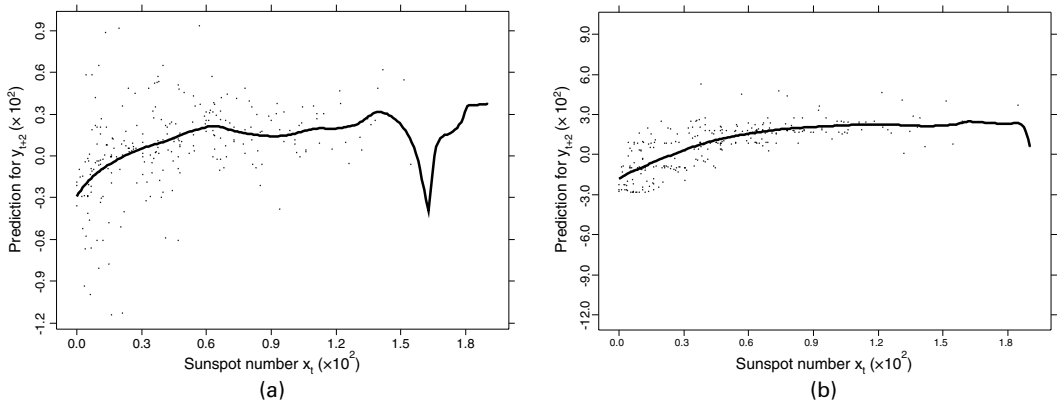

Fig. 5. Two-step-ahead prediction of the sunspot numbers X_t, predicting the deseasonalized series $Z_t = X_t - 0.903X_{t-10}$ by X_{t-2}: (a) direct predictor and the pairs $\{X_t, Z_{t+2}\}$; (b) multistage predictor and the pairs $\{X_t, \hat{f}^*(X_{t+1})\}$, where $\hat{f}^*$ is the first-stage smoother for $E(Z_t|X_{t-1})$

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. The bandwidth that was 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 non-linear for one and two steps ahead but quite linear for three and four steps.

Acknowledgements

All three authors received support from Sonderforschungsbereich 373 'Quantifikation und Simulation Ökonomischer Prozesse', Humboldt-Universität zu Berlin. Chen's research is also supported in part by National Science Foundation grants DMS 9626113, 9982846, 0073601 and CCR 9980599 and Yang's research was supported in part by National Science Foundation grant DMS 9971186. The authors thank Michael H. Neumann for questions and comments which improved the paper. The authors also gratefully acknowledge the insightful comments from the Joint Editors, an Associate Editor and two referees.

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*. San Francisco: Holden-Day.

Chen, R. (1996a) Incorporating extra information in nonparametric smoothing. *J. Multiv. Anal.*, **58**, 133–150.

Chen, R. (1996b) A nonparametric multi-step prediction estimator in Markovian structures. *Statist. Sin.*, **6**, 603–615.

Chen, R., Yang, L. and Hafner, C. (2004) Technical Supplements to 'Nonparametric multistep-ahead prediction in time series analysis' (2004, JRSSB). *Preprint*. University of Illinos, Chicago. (Available from `http://www.uic.edu/~rongchen`.)

Cline, D. B. H. and Pu, H. H. (1999) Geometric ergodicity of nonlinear time series. *Statist. Sin.*, **9**, 1103–1118.

Davydov, Yu. A. (1973) Mixing conditions for Markov chains. *Theory Probab. Applic.*, **18**, 312–328.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*. London: Chapman and 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. *Statist. Sin.*, **9**, 559–570.

Györfi, L., Härdle, W., Sarda, P. and Vieu, P. (1989) *Nonparametric Curve Estimation from Time Series*. Heidelberg: Springer.

Haggan, V. and Ozaki, T. (1981) Modelling 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. *Int. Statist. Rev.*, **65**, 49–72.

Härdle, W., Tsybakov, A. and Yang, L. (1998) Nonparametric vector autoregression. *J. Statist. Planng Inf.*, **68**, 221–245.

Härdle, W. and Vieu, P. (1992) Kernel regression smoothing for time series. *J. Time Ser. Anal.*, **13**, 209–232.

Jones, D. A. (1978) Nonlinear autoregressive processes. *Proc. R. Soc. Lond.* A, **360**, 71–95.

Pemberton, J. (1987) Exact least squares multi-step prediction from nonlinear autoregressive models. *J. Time Ser. Anal.*, **8**, 443–448.

Robinson, P. M. (1983) Non-parametric estimation for time series models. *J. Time Ser. Anal.*, **4**, 185–208.

Tiao, G. C. and Tsay, R. S. (1994) Some advances in non-linear and adaptive modelling in time series. *J. Forecast.*, **13**, 109–131.

Tjøstheim, D. (1990) Non-linear time series and Markov chains. *Adv. Appl. Probab.*, **22**, 587–611.

Tjøstheim, D. (1994) Nonlinear time series, a selective review. *Scand. J. Statist.*, **21**, 97–130.

Tong, H. (1990) *Nonlinear Time Series Analysis: a Dynamical System Approach*. Oxford: Oxford University Press.

Tschernig, R. and Yang, L. (2000) Nonparametric lag selection for time series. *J. Time Ser. Anal.*, **21**, 457–487.

Yang, L. and Tschernig, R. (1999) Multivariate bandwidth selection for local linear regression. *J. R. Statist. Soc.* B, **61**, 793–815.