NONPARAMETRIC LAG SELECTION FOR TIME SERIES

By Rolf Tschernig* and Lijian Yang

Humboldt-Universität zu Berlin, Michigan State University

First version received January 1998

Abstract. A nonparametric version of the Final Prediction Error (FPE) is analysed for lag selection in nonlinear autoregressive time series under very general conditions including heteroskedasticity. We prove consistency and derive probabilities of incorrect selections that have been previously unavailable. Since it is more likely to overfit (have too many lags) than to underfit (miss some lags), a correction factor is proposed to reduce overfitting and hence increase correct fitting. For the FPE calculation, the local linear estimator is introduced in addition to the Nadaraya-Watson estimator in order to cover a very broad class of processes. To achieve faster computation, a plug-in bandwidth is suggested for the local linear estimator. Our Monte-Carlo study corroborates that the correction factor generally improves the probability of correct lag selection for both linear and nonlinear processes and that the plug-in bandwidth works at least as well as its commonly used competitor. The proposed methods are applied to the Canadian lynx data and daily returns of DM/US-Dollar exchange rates.

Keywords. Consistency; final prediction error; foreign exchange rates; heteroskedasticity; nonlinear autoregression; overfitting; plug-in bandwidth; underfitting.

1. INTRODUCTION

The past decade has witnessed an impressive development of nonparametric modelling in both theory and practice, with the flexibility of 'letting the data speak for themselves'. One area of recent interest is time series model identification, or more specifically, lag selection. Using linear lag selection methods based on classical criteria such as the Akaike Information Criterion (AIC), the Final Prediction Error (FPE) or the Schwarz Criterion for nonlinear stochastic processes is theoretically unjustifiable and as our simulation results indicate, often impractical.

Following the successful adaption of nonparametric techniques to time series analysis (Györfi *et al*., 1989; Tjøstheim, 1994; Härdle *et al*., 1997), alternative lag selection criteria have been studied for nonlinear autoregressive processes. Cheng and Tong (1992) suggested a method based on cross-validation. Assuming homoskedasticity Yao and Tong (1994) and Vieu (1994) were able to show consistency of the cross-validation approach. Alternatively, Auestad and Tjøstheim (1990) and Tjøstheim and Auestad (1994) suggested to use a

* Address for Correspondence: Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät, Humboldt-Universität zu Berlin, Spandauer Str. 1, D-10178 Berlin, Germany.

0143-9782/00/04 457–487 JOURNAL OF TIME SERIES ANALYSIS Vol. 21, No. 4

nonparametric version of the FPE. While they allowed for heteroskedasticity, which is a well known feature of financial and many other time series, they did not show consistency. In this paper we close this gap and prove consistency of the FPE based lag selection in the presence of heteroskedasticity.

More importantly, we derive the probabilities of incorrect lag selection for the nonparametric FPE criteria. Based on these calculated probabilities, which are new to this research area, we conclude that overfitting is more likely than underfitting. Here overfitting occurs if one chooses superfluous lags in addition to the correct ones, while missing correct lags is called underfitting. Consequently, we suggest a correction of the nonparametric FPE to reduce overfitting and hence increase correct fitting. Unlike the correction of Vieu (1994), ours incorporates asymptotic analysis. It is also found to substantially increase correct fitting in our simulation experiments.

Such calculations of over- and underfitting probabilities cannot be simply duplicated for cross-validation. One should also note that in some crude sense the general FPE as defined in (2.2) is 'equivalent' to the cross-validation, i.e. their difference is of higher order (Cheng and Tong, 1992). In the same way, they are both 'equivalent' to the data-driven asymptotic FPE defined in (3.4). These higher order terms are no longer negligible for the probability calculations. This is why we prefer the data-driven asymptotic FPE to the cross-validation. A second reason is that the plug-in method can be easily applied to the asymptotic FPE and gives a better rate of convergence than the cross-validation method. For such comparisons in density estimation see Jones *et al*. (1996). Therefore, we doubt cross-validation criteria can perform numerically as well as our asymptotic FPE criteria.

The other authors used exclusively the Nadaraya-Watson estimator for their lag selection procedures. However, the Nadaraya-Watson estimator has a poor bias rate if the density of the lagged variable is not sufficiently smooth, especially with nonlinear processes. In contrast, the local linear estimator only needs continuity of the density to have the optimal convergence rate (see, for example, Fan and Gijbels (1996), Ruppert and Wand (1994), Wand and Jones (1995), and Härdle *et al*. (1998)). This phenomenon is confirmed in our simulation study. Therefore our procedures include both types of estimators.

Another contribution of this paper, based on recent results of Härdle *et al*. (1998), is a closed formula of the optimal bandwidth used in the nonparametric FPE criteria. This allows one to use the plug-in bandwidth of Yang and Tschernig (1999) for the local linear FPE. Previously, the bandwidth was always obtained by minimizing the criteria over a pre-specified grid where only Vieu (1994) showed the optimality of the grid search procedure. In any case, the plug-in bandwidth requires much less computing than the grid search, and the performance is at least as good, as shown in our simulation study.

Our Monte-Carlo study is the first major investigation into the performance of nonparametric lag selection criteria. We compare our newly suggested methods and existing ones for a wide range of processes. Overall, we find our procedures to perform better than their competitors. Finally, we apply our

methods to the Canadian lynx data and the daily returns of DM/US-$ exchange rates. For the latter we also suggest a way to select lags of the conditional volatility function.

We want to mention that for additive nonlinear autoregressive models, a subclass of the nonlinear autoregressive models considered in this paper, other nonparametric lag selection methods were suggested by Chen and Tsay (1993).

The paper is organized as follows: Section 2 gives the asymptotic formula for the nonparametric FPE as a function of the bandwidth, and the formula of the optimal bandwidth which minimizes the FPE. Section 3 investigates the consistency of the criterion. In Section 4 we derive results on the probabilities of over- and underfitting and introduce the correction factor. The practical implementation of the nonparametric FPE estimators including a plug-in bandwidth is discussed in Section 5. Section 6 consists of a comprehensive report of our Monte-Carlo study. The analysis of the two real data sets is contained in Section 7. Section 8 concludes, while all technical proofs are in the Appendix. An examination of our proofs shows that the procedures developed here can be easily adapted to various regression settings, including those with exogenous variables.

2. THE NONPARAMETRIC FPE

Suppose one has a conditional heteroskedastic autoregressive time series $\{Y_t\}_{t\geqslant 0}$

$$Y_t = f(X_t) + \sigma(X_t)\xi_t \tag{2.1}$$

where $X_t = (Y_{t-i_1}, Y_{t-i_2}, \ldots, Y_{t-i_m})^T$ is the vector of all correct lagged values, $i_1 < \cdots < i_m$, and ξ_t are i.i.d random variables with $E(\xi_t) = 0$, $E(\xi_t^2) = 1$, $t = i_m, i_m + 1, \ldots$. Here we assume that all lags $i_1, \ldots, i_m$ are needed for modelling $f(\cdot)$ but not necessarily for $\sigma(\cdot)$. The case in which $\sigma(\cdot)$ depends on lags not contained in $f(\cdot)$ is beyond this paper. Previous works on nonparametric lag selection based on cross-validation assumed homoskedasticity, i.e. $\sigma(X_t) \equiv \sigma$ (Cheng and Tong, 1992; Yao and Tong, 1994; Vieu, 1994). We prefer the more general model (2.1) since financial and many other time series exhibit heteroskedasticity.

With regard to the process (2.1) we assume the following:

(A1) For some integer $M \geqslant i_m$, the vector process $X_{M,t} = (Y_{t-1}, \ldots, Y_{t-M})^T$ is strictly stationary and β-mixing with $\beta(n) \leqslant c_0 n^{-(2+\delta)/\delta}$ for some $\delta > 0$, $c_0 > 0$. Here

$$\beta(n) = E \sup\{|P(A|\mathcal{F}_M^k) - P(A)| : A \in \mathcal{F}_{n+k}^\infty\}$$

where $\mathcal{F}_t^{t'}$ is the σ-algebra generated by $X_{M,t}, X_{M,t+1}, \ldots, X_{M,t'}$.

(A2) The stationary distribution of the process $X_{M,t}$ has a density $\mu_M(x_M)$, $x_M \in \mathbb{R}^M$, which is continuous. Henceforth, we use $\mu(\cdot)$ to denote both $\mu_M(\cdot)$

and all of its marginal densities. If the Nadaraya-Watson estimator is used, $\mu_M(\cdot)$ has to be continuously differentiable.

(A3) The function $f(\cdot)$ is twice continuously differentiable while $\sigma(\cdot)$ is continuous and positive on the support of $\mu(\cdot)$.

(A4) The $\{\xi_t\}_{t \geqslant i_m}$ have a finite fourth moment m_4.

For conditions that guarantee (A1) and (A2) see Tweedie (1975), Nummelin and Tuominen (1982), Ango Nze (1992), Diebolt and Guégan (1993), and Doukhan (1994). Using e.g. Theorem 7 and Remarks 7 in Doukhan (1994, p. 102, 103), it is straightforward to verify that all processes except **NLAR4** presented in our Monte-Carlo study in Section 6 satisfy these assumptions.

The nonparametric FPE was introduced by Auestad and Tjøstheim (1990) and Tjøstheim and Auestad (1994). Let $\{\tilde{Y}_t\}$ be another series with exactly the same distribution as $\{Y_t\}$ but independent of $\{Y_t\}$. We define the FPE of an estimate $\hat{f}$ of f as the following functional

$$FPE(\hat{f}) = E[\{\tilde{Y}_t - \hat{f}(\tilde{X}_t)\}^2 w(\tilde{X}_{M,t})] \tag{2.2}$$

where the expectation is taken over all the variables: $Y_0,\ Y_1, \ldots, Y_n,\ \tilde{Y}_0, \tilde{Y}_1, \ldots, \tilde{Y}_t, \ldots$

For the weight function $w\colon \mathbb{R}^M \to \mathbb{R}$ we assume:

(A5) The support of w is compact with nonempty interior. The function w is continuous, nonnegative, and $\mu(x_M) > 0$ for $x_M \in \text{supp}(w)$.

Because we use a single weight function defined for the largest lag vector $X_{M,t}$, we can treat both bounded and unbounded time series. All the other authors were able to obtain consistency results only for bounded time series except Vieu (1994), whose weight function was a special case of ours.

The FPE measures the discrepancy between $\hat{f}$ and the true functional relation of $\tilde{Y}_t$ to $\tilde{X}_t$. If the process $\{Y_t\}$ is a stationary linear AR process, $\hat{f}$ a linear regressor, the FPE defined in (2.2) becomes the usual linear FPE introduced by Akaike (1969, 1971). If the process $\{Y_t\}$ is a stationary nonlinear AR process and $\hat{f}$ some nonparametric estimator, we have the nonparametric FPE.

Under assumptions (A1) to (A5), it is unnecessary to generate the process $\{\tilde{Y}_t\}$ to compute the FPE. Denote $\mathbf{Y} = (Y_{i_m}, Y_{i_m+1}, \ldots, Y_n)^T$. For any $x \in \mathbb{R}^m$, write

$$\hat{f}_1(x) = (\mathbf{Z}_1^T W \mathbf{Z}_1)^{-1} \mathbf{Z}_1^T W \mathbf{Y}, \qquad \hat{f}_2(x) = e^T (\mathbf{Z}_2^T W \mathbf{Z}_2)^{-1} \mathbf{Z}_2^T W \mathbf{Y}$$

in which

$$\mathbf{Z}_1 = (1 \cdots 1)^T_{1 \times (n - i_m + 1)}, \quad \mathbf{Z}_2 = \begin{pmatrix} 1 & \cdots & 1 \\ X_{i_m} - x & \cdots & X_n - x \end{pmatrix}^T,$$

$$e = (1, 0_{1 \times m})^T, \quad W = \text{diag}\{K_h(X_i - x)/(n - i_m + 1)\}_{i=i_m}^n$$

where

(A6) $K\colon \mathbb{R}^1 \to \mathbb{R}^1$ is a symmetric probability density (kernel) and $h = h_n$ is a positive number (bandwidth) with $h \to 0$, $nh^m \to \infty$ as $n \to \infty$.

We denote $\|K\|_2^2 = \int K^2(u)\,du$, $\sigma_K^2 = \int K(u)u^2\,du$. For $x \in \mathbb{R}^m$, write

$$K_h(x) = 1/h^m \prod_{j=1}^{m} K(x_j/h).$$

The $\hat{f}_1(x)$ and $\hat{f}_2(x)$ are the Nadaraya-Watson and local linear estimates of $f(x)$, which are solutions to locally constant or locally linear least squares problems with kernel weights respectively. The estimation bias of $\hat{f}_a(x)$ is $r_a(x)\sigma_K^2 h^2/2$ where

$$r_1(x) = \mathrm{Tr}\{\nabla^2 f(x)\} + 2\nabla^T \mu(x)\nabla f(x)/\mu(x), \qquad r_2(x) = \mathrm{Tr}\{\nabla^2 f(x)\}.$$

The kernel function K matters little here, so $\hat{f}_1(x)$ and $\hat{f}_2(x)$ depend primarily on h, and so do the FPEs. We therefore write for $a = 1, 2$

$$FPE_a(h) = FPE(\hat{f}_a).$$

They have the following asymptotic expansions.

THEOREM 2.1. *Under assumptions (A1)–(A6), for* $a = 1, 2$, *as* $n \to \infty$

$$FPE_a(h) = AFPE_a(h) + o\{h^4 + (n - i_m + 1)^{-1} h^{-m}\},$$

in which the Asymptotic FPE's are

$$AFPE_a(h) = A + b(h)B + c(h)C_a \tag{2.3}$$

where

$$A = \int \sigma^2(x) w(x_M)\mu(x_M)dx_M,$$

$$B = \int \sigma^2(x) w(x_M)\mu(x_M)/\mu(x)dx_M, \tag{2.4}$$

$$C_a = \int r_a^2(x) w(x_M)\mu(x_M)dx_M,$$

and where

$$b(h) = \|K\|_2^{2m}(n - i_m + 1)^{-1} h^{-m}, \qquad c(h) = \sigma_K^4 h^4/4.$$

A closer analysis of the FPE is possible by using AFPE. The term A represents the final prediction error for the true function f. The terms $b(h)B$ and $c(h)C_a$ are the expected variance and squared bias of the estimator. As $n \to \infty$, both the FPE and AFPE tend to A as both $b(h)B$ and $c(h)C_a$ tend to zero. Solving a variance-bias trade-off between $b(h)B$ and $c(h)C_a$ one obtains

COROLLARY 2.1. *Under assumptions (A1)–(A6) and the additional assumption that* $0 < C_a < \infty$, $a = 1, 2$, *the AFPE's are minimized by the optimal bandwidth*

$$h_{a,opt} = \{m\|K\|_2^{2m} B(n - i_m + 1)^{-1} C_a^{-1} \sigma_K^{-4}\}^{1/(m+4)} \tag{2.5}$$

and the minimum AFPE is

$$AFPE_{a,opt} = A + (m^{-m/(m+4)} + \tfrac{1}{4} m^{4/(m+4)})\{\|K\|_2^{8m} B^4 (n - i_m + 1)^{-4} C_a^m \sigma_K^{4m}\}^{1/(m+4)}. \tag{2.6}$$

The closed form of the optimal bandwidth (2.5) is necessary to obtain a plug-in estimate for $h_{a,opt}$. For details, see Section 5.

Note 2.1 If $C_a = 0$, the trade-off fails. In that case, one would prefer a large bandwidth or heuristically, one has $h = +\infty$. This happens if one uses the local linear estimator for linear processes, in which case $\nabla^2 f(x) \equiv 0$ implies $C_2 = 0$. Then the local linear estimator does not have a bias of order h^2.

Note 2.2 If $C_a = +\infty$, the trade-off also fails. This occurs, for example, if one uses the Nadaraya-Watson estimator for processes which violates the smoothness condition for $\mu(x)$ in assumption (A2) (i.e. $\nabla\mu(x)$ does not exist at some points), in which case $C_1 = +\infty$ (See the simulation example **NLAR4** in Section 6).

Based on these discussions, we need a seventh assumption:

(A7) For $a = 1, 2$, the C_a's defined in (2.4) are positive and finite.

The expression for the Asymptotic FPE's (2.3) contains the unknown quantities A, B and C_a. In the next section we present a data-driven version of AFPE by introducing estimators of these quantities. We then study the behavior of the data-driven AFPE when one uses a set of lags different from those in X_t. After showing consistency of the AFPE based lag selection rule, we present in Section 4 results on the probabilities of selecting incorrect lag vectors. Based on these results we suggest a correction for the AFPE.

3. THE CONSISTENCY

For estimating the Asymptotic FPE's obtained in the previous section, the following estimates of A and B are needed

$$\hat{A}_a = (n - i_m + 1)^{-1} \sum_{i=i_m}^{n} \{Y_i - \hat{f}_a(X_i)\}^2 w(X_{M,i}) \tag{3.1}$$

$$\hat{B}_a = (n - i_m + 1)^{-1} \sum_{i=i_m}^{n} \{Y_i - \hat{f}_a(X_i)\}^2 w(X_{M,i})/\hat{\mu}(X_i) \tag{3.2}$$

in which the estimators $\hat{f}_a$ use bandwidths of the same order $(n - i_m + 1)^{-1/(m+4)}$ as the optimal $h_{a,opt}$, and $\hat{\mu}(X_i)$ is a kernel estimator of the density. As A is the dominant term in the AFPE expression, we look at the asymptotics of $\hat{A}_a$, which estimates the FPE for the true function f.

THEOREM 3.1. *Under assumptions (A1)–(A7), for* $a = 1, 2$ *as* $n \to \infty$

$$\hat{A}_a = A + \{\|K\|_2^{2m} - 2K(0)^m\}(n - i_m + 1)^{-1} h^{-m} B + C_a \sigma_K^4 h^4/4$$
$$+ o\{h^4 + (n - i_m + 1)^{-1} h^{-m}\} + O_p\{(n - i_m + 1)^{-1/2}\}. \tag{3.3}$$

Note here that with bandwidth h of the form $cn^{-1/(m+4)}$, the nonparametric estimate $\hat{A}_a$ converges to A at the parametric $\sqrt{n}$ rate if $m \leqslant 4$, in which case the second and third term will be $O\{(n - i_m + 1)^{-1/2}\}$.

Inserting (3.3) into (2.3), we obtain the following estimated FPE (for $a = 1, 2$)

$$AFPE_a = \hat{A}_a + 2K(0)^m (n - i_m + 1)^{-1} h_{a,opt}^{-m} \hat{B}_a \tag{3.4}$$

in which $\hat{A}_a$ is evaluated using the optimal bandwidth $h_{a,opt}$, while $\hat{B}_a$ using any bandwidth of order $(n - i_m + 1)^{-1/(m+4)}$. Note that $AFPE_a$ differs from $AFPE_a(h)$ since the former contains estimates for A and B using specified bandwidths. The FPE estimator (3.4) resembles in its structure traditional model selection criteria like the AIC or Schwarz criterion. The first term corresponds to the estimated mean squared error, while the second term serves as a penalty term to avoid noise fitting which would result by simply using $\hat{A}_a$ alone.

Now one computes an AFPE value according to (3.4) for every subset $\{i'_1, \ldots, i'_{m'}\}$ of $\{1, \ldots, M\}$ and denotes the result by $AFPE'_a$ to distinguish it from the unique $AFPE_a$ based on the correct lag set $\{i_1, \ldots, i_m\}$. We propose the following

Lag Selection Rule I: Select the subset $\{\hat{i}_1, \ldots, \hat{i}_{\hat{m}}\}$ with the smallest $AFPE'_1$ (or $AFPE'_2$).

THEOREM 3.2. *Under assumptions (A1)–(A7) and (A8) in the Appendix, Lag Selection Rule I consistently selects the correct set of lags, i.e. if* $\hat{i}_1, \ldots, \hat{i}_{\hat{m}}$ *are the selected lags, then as* $n \to \infty$

$$P[\hat{m} = m, \hat{i}_s = i_s, s = 1, 2, \ldots, m] \to 1.$$

Hence the probability of Selection Rule I failing to completely identify the correct model diminishes with larger sample size. Previous results on

consistency were only obtained for processes with homoskedastic errors using cross-validation (Vieu, 1994; Yao and Tong, 1994).

In what follows, we investigate what happens to the AFPE if the model one uses in formula (3.4) is incorrect, and derive Theorem 3.2 as a corollary. We distinguish two cases where X', an arbitrary vector of lags, is different from X.

DEFINITION 3.1. *A lag vector* underfits *if it does not include all correct lags. A lag vector* overfits *if it contains all correct lags plus some extra ones.*

Note that by this definition, a lag vector may underfit even when it contains more lags than the correct lag vector.

For an overfitting model, we denote the lag vector $X'_t = (Y_{t-i'_1}, Y_{t-i'_2}, \ldots, Y_{t-i'_{m+l}})^T$ with $i'_1 < \cdots < i'_{m+l}$ and $\{i_1, \ldots, i_m\} \subset \{i'_1, \ldots, i'_{m+l}\}$. We define

$$r'_1(x') = \mathrm{Tr}\{\nabla^2 f(x')\} + 2\nabla^T \mu(x')\nabla f(x')/\mu(x'), \qquad r'_2(x') = r_2(x)$$

where $f(x')$ denotes the function $f(x)$ regarded as a function of the larger vector variable x'. One has the following result similar to Theorem 2.1.

THEOREM 3.3. *Under assumptions (A1)–(A7)*

$$AFPE'_a = A + b(h'_{a,opt})B' + c(h'_{a,opt})C'_a \tag{3.5}$$

where

$$B' = \int \sigma^2(x) w(x_M)\mu(x_M)/\mu(x')\, dx_M, \tag{3.6}$$

$$C'_a = \int r'_a(x')^2 w(x_M)\mu(x_M)\, dx_M, \tag{3.7}$$

$$\begin{aligned} b(h'_{a,opt}) &= \|K\|_2^{2(m+l)}(n - i'_{m+l} + 1)^{-1}(h'_{a,opt})^{-(m+l)}, \\ c(h'_{a,opt}) &= \sigma_K^4 h'_{a,opt}/4, \end{aligned} \tag{3.8}$$

and the optimal bandwidth is

$$h'_{a,opt} = \{(m+l)\|K\|_2^{2(m+l)} B'(n - i'_{m+l} + 1)^{-1} C_a'^{-1} \sigma_K^{-4}\}^{1/(m+l+4)}.$$

COROLLARY 3.1. *In the setting of Theorem 3.3,*

$$AFPE'_{a,opt} = A + [(m+l)^{-(m+l)/(m+l+4)} + \tfrac{1}{4}(m+l)^{4/(m+l+4)}]$$
$$\{\|K\|_2^{8(m+l)} B'^4 (n - i'_{m+l} + 1)^{-4} C_a'^{(m+l)} \sigma_K^{4(m+l)}\}^{1/(m+l+4)}$$

and as $n \to \infty$

$$(AFPE'_a - A)/(AFPE_a - A) \xrightarrow{P} +\infty.$$

Thus, the overfitting $AFPE'_a$ is larger than the $AFPE_a$ because its infinitesimal part dies out more slowly than that of the $AFPE_a$: $n^{-1/(m+l+4)}$ versus $n^{-1/(m+4)}$.

For underfitting, we only consider the case of a proper subvector of the true lag vector for notational simplicity. We need another assumption (A8) (see in the Appendix before Theorem A.2) which rules out the possibility that the restriction of f to the support of w reduces to a function of fewer variables. This can always be fulfilled by carefully choosing the support of the weight function. Let $X'_t = (Y_{t-i'_1}, \ldots, Y_{t-i'_{m'}})^T$ be any subvector of X_t $(0 < m' < m)$.

THEOREM 3.4. *Under assumptions (A1)–(A8) there exists a constant $C' > 0$ (depending on $i'_1, \ldots, i'_{m'}$) such that*

$$AFPE'_a - AFPE_a = C' + O_p(h'^2_{a,opt}).$$

Now in probability, $AFPE'_a$ is greater than $AFPE_a$ by a positive constant C' defined in (A.5) which is the weighted squared projection error of the submodel based on X'.

The consistency result of Theorem 3.2 is a corollary of Theorems 3.3 and 3.4 as any misspecified model is proved to have a larger $AFPE'_a$ than the true model, so asymptotically Lag Selection Rule I chooses the true model.

4. OVER- VERSUS UNDERFITTING

While the consistency result justifies the use of Lag Selection Rule I, it does not quantify the probabilities of selecting incorrect lags. Our analysis of the overfitting and underfitting probabilities provides new insights into the quantitative aspects of the selection procedures. Such analysis could be more difficult using cross-validation as mentioned in the introduction.

For the probability of overfitting we obtain

THEOREM 4.1. *In the setting of Theorem 3.3, there exists a constant $c'_a > 0$ and a random variable $\zeta'_a \xrightarrow{D} N(0, 1)$ such that,*

$$P[AFPE'_a < AFPE_a] = P[\zeta'_a > (n - i_m + 1)^{(m+l)/(2m+2l+8)} c'_a\{1 + o(1)\}].$$

In contrast, the probability of underfitting is given by

THEOREM 4.2. *In the setting of Theorem 3.4, there exists a random variable $\zeta' \xrightarrow{D} N(0, 1)$ such that, for $c' = C'/\Sigma'^{1/2} > 0$, where C' and Σ' are defined in (A.5) and (A.6), as $n \to \infty$*

$$P[AFPE'_a < AFPE_a] = P[\zeta' > (n - i'_{m'} + 1)^{1/2} c'\{1 + o(1)\}].$$

Note 4.1. If heuristically, one assumes that the ζ'_a, $a = 1, 2$ were *exactly* instead of asymptotically $N(0, 1)$, then the overfitting probability in Theorem 4.1 would

be $1 - \Phi((n - i_m + 1)^{(m+l)/2m+2l+8)} c'_a\{1 + o(1)\})$ where we denote by $\Phi(x) = 1/\sqrt{2\pi} \int_{-\infty}^{x} \exp(-t^2/2)\, dt$ the cumulative distribution function of the standard normal distribution. Similarly, if ζ' were *exactly* $N(0, 1)$, the underfitting probability in Theorem 4.2 would be $1 - \Phi((n - i'_{m'} + 1)^{1/2} c'\{1 + o(1)\})$. One may expect these to be asymptotically true when certain regularity conditions are met.

Note 4.2. All the probabilistic tools for handling large deviations that we are aware of, e.g., those contained in Saulis and Statulevičius (1991), require the value of interest to be of order no more than $n^{1/6}$, which is never fulfilled in our results except for $P[\zeta'_a > (n - i_m + 1)^{(m+l)/(2m+2l+8)} c'_a\{1 + o(1)\}]$ with $m = l = 1$. This is why we had been unable to prove the heuristics in Note 4.1.

Note 4.3. It is known that $1 - \Phi(x)$ goes to zero faster if x goes to $+\infty$ faster. Note 4.1 therefore suggests that the probabilities of overfitting go to zero slower than those of underfitting as

$$1/2 > (m + l)/(2m + 2l + 8).$$

Hence to increase correct fitting one can be more effective by reducing overfitting than underfitting. This consideration is suported by the fact that the $AFPE_a$ of an overfitting model is asymptotically smaller than that of an underfitting model, see Theorems 3.3 and 3.4. It is also validated by our simulation, see Section 6.

So to increase correct fitting, one should further penalize overfitting. We define a corrected AFPE as

$$CAFPE_a = \{\hat{A}_a + 2K(0)^m (n - i_m + 1)^{-1} h_{a,opt}^{-m} \hat{B}_a\}\{1 + m(n - i_m + 1)^{-4/(m+4)}\}, \tag{4.1}$$

which gets larger for models with more lags at a faster rate than $AFPE_a$. One then has

Lag Selection Rule II: Select the subset $\{\hat{i}_1, \ldots, \hat{i}_{\hat{m}}\}$ with the smallest $CAFPE'_1$ (or $CAFPE'_2$).

Notice that the extra term $m(n - i_m + 1)^{-4/(m+4)}$ in the correction has the same order as $(n - i_m + 1)^{-1} h_{a,opt}^{-m}$ and $h_{a,opt}^4$. Thus the asymptotics of $CAFPE_a$ and $AFPE_a$ have the same order, only different ratios. This entails

THEOREM 4.3. *Under assumptions (A1)–(A8), let* $\hat{i}_1, \ldots, \hat{i}_{\hat{m}}$ *be the lags selected according to the Lag Selection Rule II, then as* $n \to \infty$

$$P[\hat{m} = m, \hat{i}_s = i_s, s = 1, 2, \ldots, m] \to 1.$$

Another interesting issue is what happens when one selects lags out of $\{1, 2, \ldots, M'\}$ where $M' < i_m$. This becomes relevant when one deals, for

example, with finite moving average processes which invert into infinite autoregressive processes. In this case one always underfits, and ideally one should select the model that underfits the least, in other words, all the i_j's $(j = 1, \ldots, m)$ that are in $\{1, 2, \ldots, M'\}$ and no more. This is the case.

THEOREM 4.4. *Let* $i'_1, \ldots, i'_{m'}$ *be all the* i_j*'s* $(j = 1, \ldots, m)$ *that are in* $\{1, 2, \ldots, M'\}$. *Under assumptions (A1)–(A8), let* $\hat{i}_1, \ldots, \hat{i}_{\hat{m}}$ *be the lags selected according to the Lag Selection Rule I or II from among* $1, 2, \ldots, M'$, *then as* $n \to \infty$

$$P[\hat{m} = m', \hat{i}_s = i'_s, s = 1, 2, \ldots, m'] \to 1.$$

5. IMPLEMENTING THE FPE ESTIMATORS

Computing the (C)AFPE's in (3.4) and (4.1) requires suitable kernel and bandwidth choices in (3.1) and (3.2). With respect to the former we decide to use the Gaussian kernel. For computing $\hat{f}_a(\cdot)$ and $\hat{\mu}(\cdot)$ in $\hat{B}_a$ of (3.2) we apply the bandwidth

$$h_S(k) = \sqrt{\widehat{\mathrm{var}}(Y_t)}\{4/k\}^{1/(k+2)}\, n^{-1/(k+2)} \tag{5.1}$$

with $k = m + 2$ and additionally the leave-one-out method.

To estimate the optimal bandwidth $h_{a,opt}$ given by (2.5) which is used for computing $\hat{A}$ we apply either a grid search procedure or a plug-in method. We conduct the grid search over the interval $[0.2h_S, 2h_S]$ in 24 steps where h_S is given in (5.1). If the minimum occurs at the upper bound of the grid, the grid is extended by 16 additional steps of the previous step size. The (C)AFPEs calculated according to (3.4) and (4.1) with a grid search bandwidth are denoted by $AFPE_a$, $CAFPE_a$, $a = 1, 2$, respectively.

All existing studies have used a grid search procedure since it does not require the estimation of C_a in (2.5). Building on recent results by Yang and Tschernig (1999) we use a partial local quadratic estimator with bandwidth $h_C = 2h_S(m + 4)$ to estimate C_2 in (2.4) and thus to compute a plug-in bandwidth $\hat{h}_{a,opt}$ for the local linear estimator. Under additional smoothness assumptions, this plug-in bandwidth $\hat{h}_{a,opt}$ is optimal according to Yang and Tschernig (1999), i.e. $\hat{h}_{a,opt} = h_{a,opt}\{1 + O_p(n^{-2/(m+6)})\}$, thus using $\hat{h}_{a,opt}$ instead of $h_{a,opt}$ for the AFPE does not affect the asymptotics. We denote the CAFPE calculated according to (4.1) with a plug-in bandwidth by $CAFPE_{2a}$. The estimation of the plug-in bandwidth for the local constant estimator is more complicated since the 'bias term' C_1 in (2.4) also involves the first derivatives of the density. It is therefore omitted. The weight function $w(X_{M,i})$ in (3.1) and (3.2) is the indicator function on the range of the observed data.

We implement two additional features of Tjøstheim and Auestad (1994) for robustification. For $\hat{\mu}(x)$ in (3.2) we always use

$$\hat{\mu}(x) = (n - i_m + i_1 + 1)^{-1} \sum_{i=i_m}^{n+i_1} K_h(X_i - x)$$

where the vectors X_i, $i = n+1, \ldots, n+i_1$ are all available from the observations Y_t, $t = 0, 1, \ldots, n$. For example, X_{n+i_1} is given by $(Y_n, \ldots, Y_{n+i_1-i_m})^T$. Furthermore, 5% of those observations whose density values $\hat{\mu}(\cdot)$ are the lowest, are screened off. With these specifications, the $AFPE_1$ is exactly the same as in Tjøstheim and Auestad (1994).

We are now in the position to compute all CAFPE criteria. As a full search through all possible lag combinations will in general be computationally too costly, a directed search procedure is used instead as suggested by Tjøstheim and Auestad (1994): add lags as long as they reduce the selection criterion, and choose the lags with respect to their contribution to this reduction.

6. MONTE-CARLO STUDY

We investigate and compare the finite sample properties of the $AFPE_1$, $CAFPE_1$, $CAFPE_2$, and $CAFPE_{2a}$ criteria and four linear criteria by means of Monte-Carlo analysis.

6.1. *Setup*

We analyse three linear and four nonlinear data generating processes (DGP) with 100 observations each. The number of observations was chosen to be small so that the conditions are unfavorable to nonparametric analysis.

Linear AR processes are studied mainly for two reasons. First of all, one has to check the practical relevance of Note 2.1 which states that the local linear $(C)AFPE_2$ do not obey Theorems 3.2 and 4.3 if the true DGP is linear. As a consequence one may expect the local constant $AFPE_1$ and $CAFPE_1$ to be superior in this situation. Second, we want to evaluate the costs of extending the function class beyond linear functions if the true DGP is indeed linear.

All linear AR processes

$$Y_t = \phi_{i_1} Y_{t-i_1} + \phi_{i_2} Y_{t-i_2} + 0.1\xi_t, \qquad \xi_t \sim i.i.d.\ N(0, 1)$$

are of order 2 or 10 and parameterized as follows:

AR1 $\phi_1 = 0.5 \quad \phi_2 = 0.4,$

AR2 $\phi_1 = -0.5 \quad \phi_2 = 0.4$

AR3 $\phi_6 = -0.5 \quad \phi_{10} = 0.5.$

These linear processes differ with respect to their behavior in the frequency domain, their proximity to nonstationarity and their lag vector. With respect to

the latter properties, only the third AR process **AR3** is close to the border of nonstationarity and includes lag six and ten. We also chose the **AR3** process since Tjøstheim and Auestad (1994) used it to illustrate their $AFPE_1$ criterion.

The nonlinear processes were chosen as follows:

NLAR1. Additive nonlinear AR(2) model

$$Y_t = -0.4(3 - Y_{t-1}^2)/(1 + Y_{t-1}^2) + 0.6\{3 - (Y_{t-2} - 0.5)^3\}/\{1 + (Y_{t-2} - 0.5)^4\} + 0.1\xi_t,$$

$$\xi_t \sim i.i.d.N(0, 1),$$

NLAR2. Additive nonlinear AR process (exponential autoregression)

$$Y_t = \{0.4 - 2\exp(-50Y_{t-6}^2)\}Y_{t-6} + \{0.5 - 0.5\exp(-50Y_{t-10}^2)\}Y_{t-10} + 0.1\xi_t,$$

$$\xi_t \sim i.i.d.N(0, 1),$$

NLAR3. Additive nonlinear AR process (exponential autoregression with sine and cosine terms)

$$Y_t = (0.4 - 2\cos(40Y_{t-6})\exp(-30Y_{t-6}^2))Y_{t-6} + (0.55 - 0.55\sin(40Y_{t-10})\exp(-10Y_{t-10}^2))Y_{t-10} + 0.1\xi_t,$$

$$\xi_t \sim i.i.d.N(0, 1),$$

NLAR4 Fully nonlinear AR(2) model

$$Y_t = 0.9/(1 + Y_{t-1}^2 + Y_{t-2}^2) - 0.7 + 0.1\xi_t, \quad \xi_t \sim \text{i.i.d. triangular errors.}$$

These processes differ in the shape of the conditional mean function, the error distribution and the lag vector. The processes **NLAR1** to **NLAR3** are all generated from nonlinear additive mean functions which are shown in Figure 1. Each plot also exhibits the domain of one realization of the time series. Their inspection shows that the nonlinearities are in action. The functional shape of the fully nonlinear conditional mean of the **NLAR4** process is shown in Figure 2. This process is also driven by a triangular error density that violates the smoothness assumption (A2) in order to investigate the practical relevance of Note 2.2 for the local constant $(C)AFPE_1$. The triangular density is given by

$$p(x) = \left(\frac{1}{\sqrt{6}} - \frac{|x|}{6}\right)1_{\{|x| \leq \sqrt{6}\}}.$$

It has variance 1 and is not differentiable at 0.

We consider four linear criteria and four versions of the nonparametric FPE criteria. The linear criteria are the FPE, AIC, Schwarz criterion and Hannan-Quinn criterion, abbreviated by *ARFPE*, *ARAIC*, *ARSC* and *ARHQ*. See e.g.

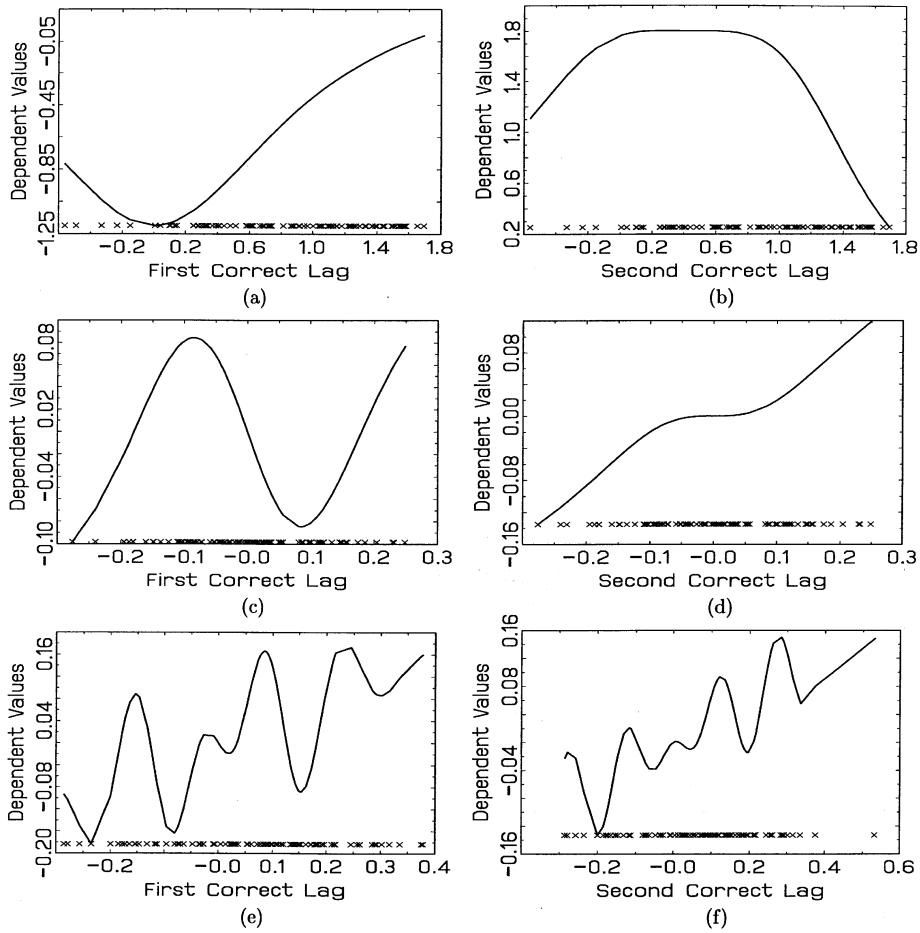

FIGURE 1. Additive nonlinear functions used in the Monte-Carlo experiments. The stars indicate one realization of the empirical distribution of 100 observations: (a) lag 1 in the **NLAR1** process; (b) lag 2 in the **NLAR1** process; (c) lag 6 in the **NLAR2** process; (d) lag 10 in the **NLAR2** process; (e) lag 6 in the **NLAR3** process; (f) lag 10 in the **NLAR3** process

Lütkephol (1991, Ch. 4.3) for details. The nonparametric FPE criteria include: $AFPE_1$, $CAFPE_1$, $CAFPE_2$ and $CAFPE_{2a}$.

In all cases the number of lags m is always smaller than 7 and the largest lag M to be considered is 15. For every experiment 100 replications are conducted with the same random numbers for each experiment. All procedures were programmed in UNIX GAUSS 3.2.7 and run on Sun workstations.

6.2. *Results*

The results of the Monte-Carlo experiments are shown in Figures 3 and 4 for the linear and nonlinear processes, respectively. Following Definition 3.1 they show

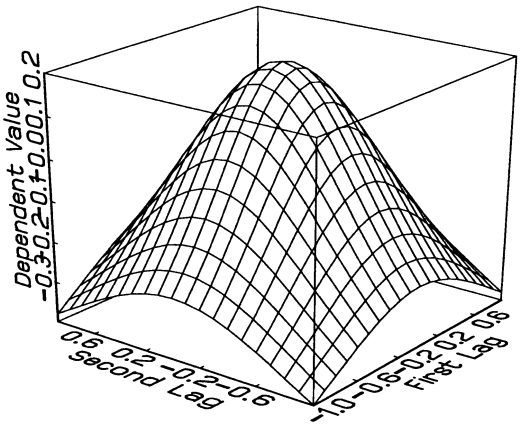

FIGURE 2. Regression function of the **NLAR4** process

for each investigated process the empirical frequencies of the eight criteria to underfit, correctly fit and overfit the true model.

Linear AR Processes

Figure 3 shows that the nonparametric criteria do not in general perform worse than linear ones for the linear DGPs. The best linear criterion *ARSC* and the best nonlinear criterion $CAFPE_1$ always cover rank one or two in terms of the correct selections. Except for the $AFPE_1$, all nonlinear criteria perform better than the linear FPE or AIC. As the results for **AR3** show, it can even happen that a nonlinear criterion performs best. The Nadaraya-Watson based $CAFPE_1$ has 30% more correction selections than the linear Schwarz criterion ranked second. On the other hand, for the processes **AR1** and **AR2** the nonlinear $CAFPE_1$ exhibits up to 20% fewer correct selections than the Schwarz criterion. Thus, extending the model class to nonlinear functions and using nonparametric lag selection criteria may not be too costly even for linear DGPs. They may, however, have a higher underfitting probability than the linear criteria while the latter have a strong tendency for overfitting.

The implication of Note 2.1 that the local linear CAFPE may fail for linear DGPs is practically relevant. The best nonparametric criterion is indeed the local constant $CAFPE_1$. It also has a much smaller overfitting probability than the $CAFPE_2$ and $CAFPE_{2a}$ criteria. This is a direct consequence of the non-existing finite optimal bandwidth for the latter criteria in the present case.

Note also that the correction factor suggested in Section 4 has substantially increased the probability of correct selection by comparing $CAFPE_1$ to the $AFPE_1$ of Tjøstheim and Auestad (1994). It reduces the probability of overfitting although underfitting becomes more likely.

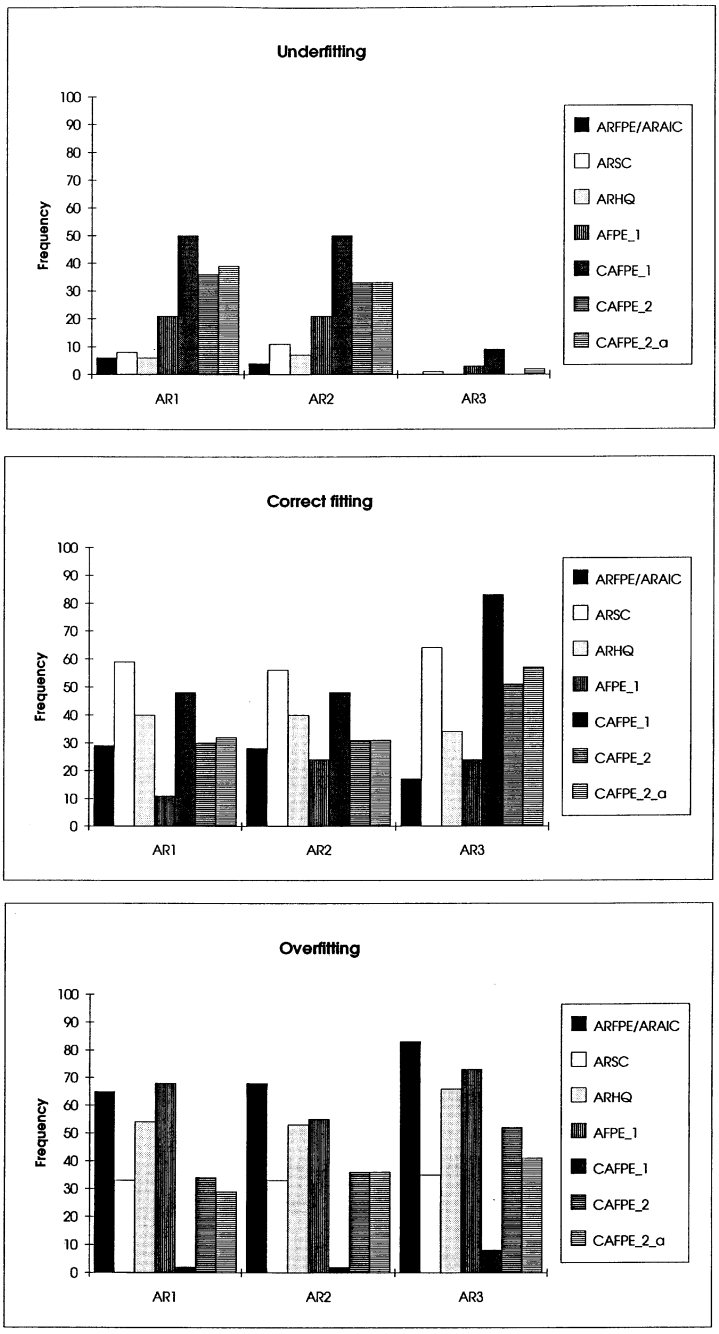

FIGURE 3. Empirical frequencies of underfitting, correct fitting and overfitting for linear AR models

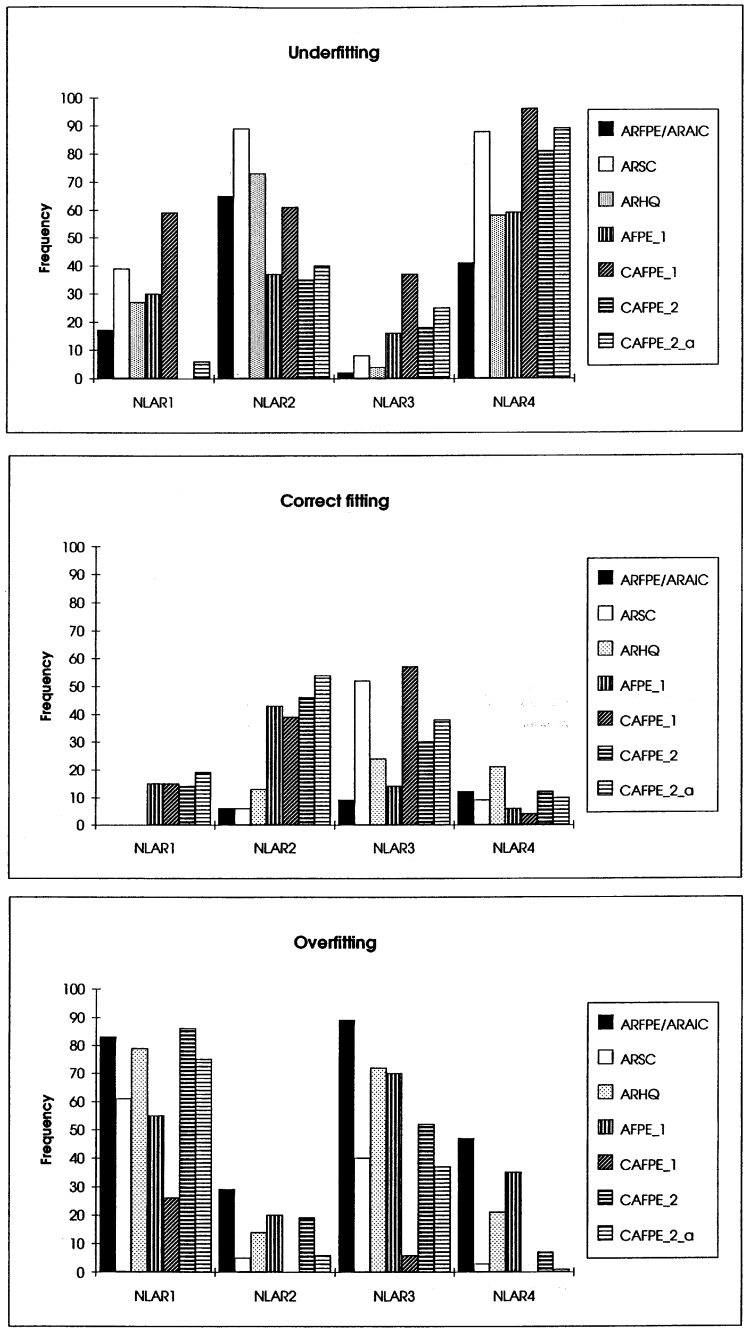

FIGURE 4. Empirical frequencies of underfitting, correct fitting and overfitting for nonlinear AR models

Nonlinear AR Processes

In the presence of nonlinear DGPs some of these results may change drastically. Figure 4 shows that it may happen that all linear criteria fail as the results for the processes **NLAR1** and **NLAR2** indicate. On the other hand, it also may happen that the linear criteria perform as well as the nonliner ones like for the **NLAR4** process. In any case, comparing again the best linear and best nonlinear criterion in terms of correct fitting, they do no longer always rank one or two.

In constrast to the case of linear DGPs the $CAFPE_2$, $CAFPE_{2a}$ criteria generally perform at least as good as or better than those based on the local constant estimator. The only exception is the **NLAR3** process. A possible explanation for this is that the strong nonlinearity of its functional shape (Figure 1e and 1f) cannot be distinguished from noise due to the small number of 100 observations. Therefore, the procedure tries to fit linear models for which Note 2.1 applies.

Recall from Note 2.2 that in the situation of a nonsmooth density, $C_1 = +\infty$, the local constant criteria $(C)AFPE_1$ do not obey Theorem 3.2 and Theorem 4.3. In such a case one might prefer to use $CAFPE_2$, $CAFPE_{2a}$ as corroborated by the results for the **NLAR4** process. There, $CAFPE_2$ and $CAFPE_{2a}$ do better than $CAFPE_1$.

For nonlinear DGPs the correction factor either changes little or improves the probability of correct selection. This can be seen by comparing the $AFPE_1$ and the $CAFPE_1$ in Figure 4. Note also that correct selection is higher for additive models **NLAR1** through **NLAR3** than for the non-additive **NLAR4**. It seems that detecting the right lag set becomes easier with simpler model structures, as one would expect. Finally, one observes that for the complex nonlinear processes we selected, overall the correct selection frequencies are quite high based on only 100 observations.

All Processes

Using the plug-in bandwidth in (2.5) leads to at least as many correct selections as using the grid search bandwidth. This can be seen by comparing $CAFPE_2$ and $CAFPE_{2a}$ in Figures 3 and 4. This allows to save a very large amount of computing time. Furthermore, the correction factor should always be used.

Evaluating the results for all processes, it seems that the Nadaraya-Watson based $CAFPE_1$ criterion has slight advantages over the local linear $CAFPE_{2a}$ criterion in terms of correct fitting since the former is less sensitive to linearity in the DGP. However, the $CAFPE_1$ has the drawback of having a higher underfitting probability, while the risk of using the $CAFPE_{2a}$ consists mainly in overfitting.

From these results we suggest the following procedure for empirical work. Using the $CAFPE_{2a}$ criterion seems best for reducing the initial set of potential

lags to a smaller set which is likely to include the correct lags. Eliminating possible irrelevant lags has then to be done by investigating the properties of the submodels of the proposed model and their residuals. One should also employ the Nadaraya-Watson based $CAFPE_1$, which, due to its tendency to underfit, might give a different set of lags. Two examples of this procedure are presented in the next section.

7. EMPIRICAL EXAMPLES

We now apply our proposed methods to the Canadian lynx data and daily returns of the DM/US-$ exchange rate from January 2, 1980 to October 30, 1992. These data sets differ in their number of observations and structure.

The lynx data set consists of 114 observations which roughly corresponds to the number of observations in the Monte-Carlo study. We use the estimation setup of Section 5 and logs with base 10 were taken of the original data. We follow the suggested procedure of the last section and use only the $CAFPE_1$ and the $CAFPE_{2a}$ criteria and for reasons of comparison, the linear Schwarz criterion *ARSC*.

Table 1 summarizes the results for the lynx data. Except for the $CAFPE_1$ criterion all criteria include lag 1 and 2 in their selection. However, there is no agreement on additional lags. Only the $CAFPE_{2a}$ additionally suggests lags 3 and 4. Recalling the results of the previous section, these lags for the $CAFPE_{2a}$ may be due to overfitting. To decide whether the more parsimonious model is sufficient, we investigated the residuals of all suggested models using the bandwidths of Table 1 and conclude that lags 1 and 2 are sufficient. A plot of the estimated regression function on a relevant grid is shown in Figure 5. We dismiss the model with lag 1 and 3 since its residuals exhibit more remaining autocorrelation than the competing model. Tjøstheim and Auestad (1994) found

TABLE I

NONPARAMETRIC LAG SELECTION FOR LYNX DATA

Estimation method	Max. # lags	Selected lags	Crit. value	$\hat{h}_{a,opt}$
ARSC	6	1,2	−2.828	
$CAFPE_1$	6	1,3	0.0780	0.241
$CAFPE_{2a}$	6	1,2,3,4	0.0433	0.353
	3	1,2,3	0.0448	0.347
	2	1,2	0.0471	0.331

Notes: The highest lag considered is 15. The second column displays the maximal number of lags to be allowed in the specific model. The last three rows contain the vector of selected lags, the corresponding selection criterion value and the underlying bandwidth.

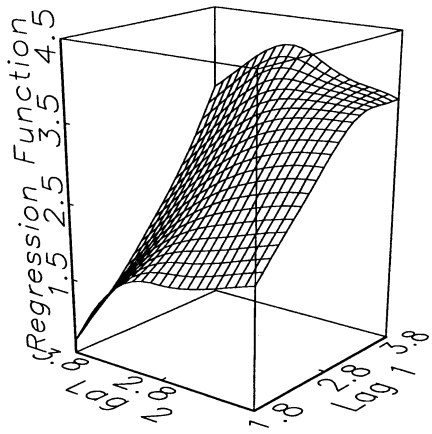

FIGURE 5. Regression function for the log lynx data obtained with the local linear estimator

lags 1 and 3 using $AFPE_1$ while Yao and Tong (1994) found lags 1, 3 and 6 using cross-validation.

Applying our methods to daily exchange rate data poses a different challenge. While there are plenty of data (3212 observations), this benefit is compromised as the data is known to be highly dependent (although only weakly correlated) and therefore asymptotics kick in very slowly.

By applying the $CAFPE_{2a}$ criterion and conducting a full search up to lag 6 we find lags 1, 3 and 4 with $h_{2,opt} = 0.0056$. Using lags 1 and 3 the autocorrelation function of the estimated residuals in Figure 6a does not indicate any remaining autocorrelation. This figure also contains the corresponding autocorrelations of the original data and a 95% confidence interval for white noise. Figure 6b contains a plot of the estimated conditional mean function on an appropriate grid of the data. It is consistent with the general finding that for this data set $f(x)$ is very close to zero. Note that the steep increase in one corner is likely to be caused by boundary effects. We therefore assume in the following that $f(x)$ is zero. This is also the result of the lag selection using the Schwarz criterion.

To conduct an explicit lag selection for the conditional volatility function $\sigma^2(x)$ we square the model (2.1) with $f(x) = 0$. This gives

$$Y_t^2 = \sigma^2(X_t) + \sigma^2(X_t)(\xi_t^2 - 1) \tag{7.1}$$

which can be estimated with the tools developed in this paper by simply replacing the dependent variable Y_t by its squares. Using the $CAFPE_{2a}$ criterion we obtain lag 1, 3 and 6 with a bandwidth estimate of 0.0051. Choosing lags 1 and 3 and investigating the autocorrelations of the residuals of (7.1) and of the squared observations in Figure 6c indicates that most of the conditional heteroskedasticity has been removed.

Figure 6d shows the standard deviation function on the relevant grid using

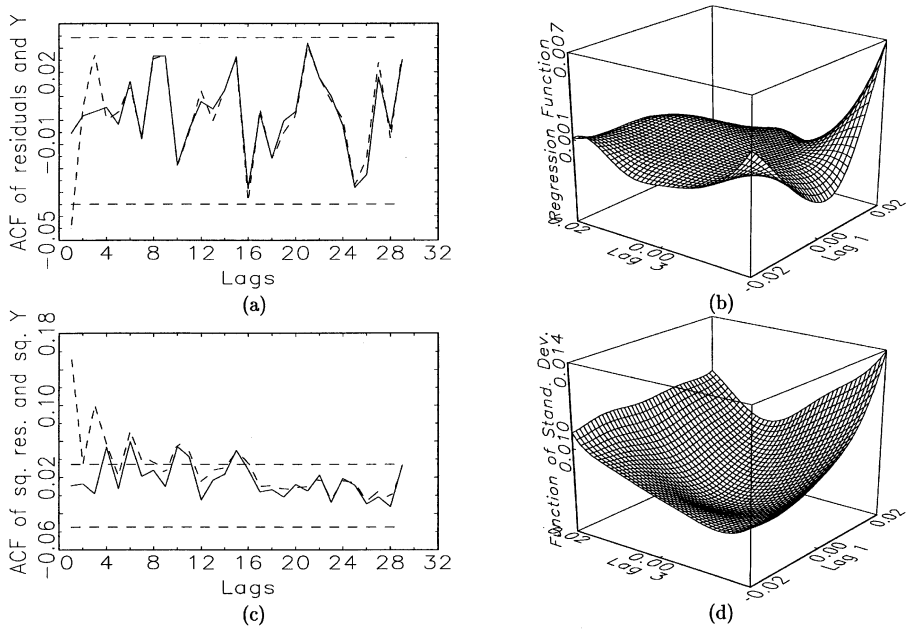

FIGURE 6. Local linear estimates for daily DM/US-Dollar series: (a) ACF of estimated residuals (solid line) and of observations (dashed line); (b) regression function; (c) ACF of squared estimated residuals (solid line) and of squared observations (dashed line); (d) conditional standard deviation

the bandwidth 0.0080. Its plot appears to be asymmetric and highly nonlinear. It also suggests that the conditional volatility increases sharply if the previous observations are large in absolute value and of opposite sign. Further investigation of this feature can be modelled within the context of parametric ARCH models as in Engle (1982), or the nonparametric additive/multiplicative CHARN models as in Yang, Härdle and Nielsen (1999) where lags recommended by our analysis were used.

8. CONCLUSIONS

In this paper we looked closely at the nonparametric FPE using either the local constant estimates of Tjøstheim and Auestad (1994) or local linear estimates. Under very general conditions we derived consistency and probabilities for underfitting as well as overfitting. Based on these results we proposed a correction factor to increase correct fitting. The new criteria were compared to existing ones in a large Monte-Carlo study including linear and nonlinear DGPs. It was found that including the correction factor leads to considerable improvement in the number of correct selections, especially for linear DGPs.

The nonparametric FPE criteria can select the correct lags for nonlinear

processes while linear criteria may fail completely. Also for linear processes, the corrected nonparametric FPE based on the Nadaraya-Watson estimator always ranked at least second. The criteria based on the local linear estimator perform somewhat worse for linear processes due to the lack of an estimation bias of a proper order. For nonlinear processes, however, the local linear criteria seem to be the best. Our plug-in estimation of the optimal bandwidth performs as well as the grid search method and saves substantial computation time.

We applied our procedure to two real data sets of different size and properties. For the lynx data we obtained a good fit with a parsimonious model. For the daily DM/US-$ exchange rate returns we found a highly nonlinear and asymmetric volatility function of lags 1, 3 and 6 which presents interesting challanges for the parametric modelling of this much investigated series.

We agree with a referee's comment that more effective lag selection criteria may be designed for special multidimensional models, such as additive models. Based on our Monte-Carlo study of the nonlinear processes, which shows that even our generic method performs better when additive structure is present, we can expect our general idea of using a local linear (instead of the Nadaraya-Watson) estimator together with a plug-in (instead of a cross-validation) bandwidth and the correction factor to improve the existing method of Chen and Tsay (1993).

We also concur with a cautionary note of the Associate Editor that our nonparametric methods could suffer from the 'curse of dimensionality' when relatively many lags are involved, whereas this problem is not present for parametric methods based on linear models. Further numerical work is needed to properly address this issue.

APPENDIX

PROOF OF THEOREM 2.1. We note that the second term of the FPE in formula (7) of Tjøstheim and Auestad (1994) was decomposed as the following (here we have changed the original notation to ours)

$$
\begin{aligned}
&E[\{\hat{f}(\tilde{X}_t) - f(\tilde{X}_t)\}^2 w(\tilde{X}_{M,t})] \\
&\qquad = E[\{\hat{f}(\tilde{X}_t) - E\hat{f}(\tilde{X}_t) + E\hat{f}(\tilde{X}_t) - f(\tilde{X}_t)\}^2 w(\tilde{X}_{M,t})] \\
&\qquad = E[(I' + II')^2 w(\tilde{X}_{M,t})].
\end{aligned}
$$

As one sees from that paper, II' is the bias tem of $\hat{f}(\tilde{X}_t)$. Härdle *et al*. (1998) gave an explicit formula of the bias for the local linear estimator $\hat{f}_2(x)$, which is

$$\sigma_K^2 h^2/2 \operatorname{Tr}\{\nabla^2 f(x)\}.$$

Thus

$$E[(II')^2 w(\tilde{X}_{M,t})] = \sigma_K^2 h^4/4 \int [\mathrm{Tr}\{\nabla^2 f(x)\}]^2 w(x_M)\mu(x_M)\,dx_M$$
$$+ O\{h^4(n - i_m + 1)^{-1/2}\}$$
$$= c(h)C_2 + O\{h^4(n - i_m + 1)^{-1/2}\}$$

by applying the beta-mixing property. Similarly, one derives that if the local constant estimator $\hat{f}_1(x)$ is used instead, then

$$E[(II')^2 w(\tilde{X}_{M,t})] = c(h)C_1 + O\{h^4(n - i_m + 1)^{-1/2}\}.$$

The term $E(I'II'w(\tilde{X}_{M,t}))$ is negligible by a standard U-statistic argument, using our beta-mixing assumption (A1) and Lemma 1 of Yoshihara (1976), similar to Tjøstheim and Auestad (1994).

Now we derive the term $E[(I')^2 w(\tilde{X}_{M,t})]$. Using the result of the same paper by Härdle *et al.* (1998)

$$E[(I')^2 w(\tilde{X}_{M,t})]$$
$$= E\int \left[\mu(x)^{-1}(n - i_m + 1)^{-1}\{1 + o_p(1)\}\sum_{i=i_m}^{n} K_h(X_i - x)\sigma(X_i)\xi_i\right]^2 w(x_M)\mu(x_M)\,dx_M$$

which becomes

$$E\int \mu(x)^{-2}(n - i_m + 1)^{-2}\{1 + o_p(1)\}\sum_{i=i_m}^{n}\{K_h(X_i - x)\sigma(X_i)\}^2 w(x_M)\mu(x_M)\,dx_M,$$

where the cross terms are left out only by a U-statistic argument as in Tjøstheim and Auestad (1994). The above expression can be written as

$$\int \mu(x)^{-2}(n - i_m + 1)^{-1}\{1 + o_p(1)\}\{K_h(y - x)\sigma(y)\}^2 w(x_M)\mu(x_M)\mu(y)\,dx_M\,dy$$
$$= \int \mu(x)^{-2}(n - i_m + 1)^{-1}h^{-m}\{1 + o_p(1)\}\{K(u)\sigma(x + hu)\}^2 w(x_M)\mu(x_M)\mu(x + hu)\,dx_M\,du$$
$$= \|K\|_2^{2m}(n - i_m + 1)^{-1}h^{-m}\int \sigma^2(x)w(x_M)\mu(x_M)/\mu(x)\,dx_M\{1 + o_p(1)\}$$
$$= b(h)B\{1 + o_p(1)\},$$

which has completed the proof of the formula (2.3).

The following theorem extends Theorem 3.1.

THEOREM A.1 *Let* $Z = (n - i_m + 1)^{-1}\sum_{i=i_m}^{n}\sigma^2(X_i)\xi_i^2 w(X_{M,i}) - A$, *then under assumptions (A1)–(A7), for* $a = 1, 2$, *as* $n \to \infty$

$$\hat{A}_a = A + \{\|K\|_2^{2m} - 2K(0)^m\}(n - i_m + 1)^{-1}h^{-m}B + C_a\sigma_K^4 h^4/4$$
$$+ Z + o\{h^4 + (n - i_m + 1)^{-1}h^{-m}\} + o\{(n - i_m + 1)^{-1/2}\} \qquad \text{(A.1)}$$

with

$$\sqrt{n - i_m + 1}\,Z \xrightarrow{D} N(0, \Sigma) \qquad \text{(A.2)}$$

$$\Sigma = m_4 \int \sigma^4(x) w^2(x_M) \mu(x_M)\, dx_M - A^2$$

$$+ 2 \sum_{i=1}^{\infty} [E\{\sigma^2(X_{M,M}) \xi_M^2 w(X_{M,M}) \sigma^2(X_{M,M+i}) w(X_{M,M+i})\} - A^2]. \qquad \text{(A.3)}$$

A similar result exists for the overfitting case.

PROOF OF THEOREM 3.1 AND THEOREM A.1. To prove (A.1), note that one obtains (A.2) and (A.3) by the central limit theorem for mixing processes, see Doukhan (1994), Theorem 1, p. 46. We then note that by (3.1), $\hat{A}_a$ is

$$(n - i_m + 1)^{-1} \sum_{i=i_m}^{n} \{f(X_i) - \hat{f}_a(X_i) + \sigma(X_i)\xi_i\}^2 w(X_{M,i})$$

$$= (n - i_m + 1)^{-1} \sum_{i=i_m}^{n} \sigma^2(X_i)\xi_i^2 w(X_{M,i}) + (n - i_m + 1)^{-1} \sum_{i=i_m}^{n} \{f(X_i) - \hat{f}_a(X_i)\}^2 w(X_{M,i})$$

$$+ (n - i_m + 1)^{-1} \sum_{i=i_m}^{n} 2\{f(X_i) - \hat{f}_a(X_i)\}\sigma(X_i)\xi_i w(X_{M,i}) \qquad \text{(A.4)}$$

in which the second term contributes to the $\|K\|_2^{2m}(n - i_m + 1)^{-1} h^{-m} B + C_a \sigma_K^4 h^4/4$ just as in the proof of Theorem 2.1, while the last term contributes the $-2K(0)^m (n - i_m + 1)^{-1} h^{-m} B$, see Tjøstheim and Auestad (1994) for proof.

PROOF OF THEOREM 3.3. Similar arguments as in the proofs of Theorem 2.1 and Theorem A.1 give the expression for B' and C_a' in (3.6) and (3.7) and therefore (3.8) and (3.5).

To study the undefitting case, one denotes the discrepency between $f(x)$ and its conditional expectation on x' as

$$f^{\perp}(x) = f(x) - \mu(x')^{-1} \int f(x', u'') \mu(x', u'') du'' = f(x) - E\{f(x)|x'\}$$

and the weighted squared projection error

$$C' = \int f^{\perp}(x)^2 w(x_M) \mu(x_M)\, dx_M$$

$$= \int f(x)^2 w(x_M) \mu(x_M)\, dx_M - \int E^2\{f(x)|x'\} w(x_M) \mu(x_M)\, dx_M. \qquad \text{(A.5)}$$

We assume that

(A8) Every function $f^{\perp}(x)$ has at least one nonzero point in the interior of the support of w, and hence the projection error C' defined in (A.5) is positive.

This is satisfied if one simply enlarges the support of w so that it includes in its interior at least one nonzero point from each $f^{\perp}(x)$, which is easy as all the $f^{\perp}(x)$'s are nonzero functions on the support of μ.

The following is a refined version of Theorem 3.4.

THEOREM A.2. *Let* $X_t' = (Y_{t-i_1'}, \ldots, Y_{t-i_{m'}'})^T$ *be as in Theorem 3.4. Then under assumptions (A1)–(A8), for*

$$Z'_a = (n - i_{m'} + 1)^{-1} \sum_{i=i_{m'}}^{n} \{f(X_i) - \hat{f}_a(X'_i)\}^2 w(X_{M,i}) - C'$$

$$+ (n - i_{m'} + 1)^{-1} \sum_{i=i_{m'}}^{n} 2\{f(X_i) - \hat{f}_a(X'_i)\}\sigma(X_i)\xi_i w(X_{M,i})$$

one has

$$\sqrt{n - i_{m'} + 1}\, Z'_a \xrightarrow{D} N(0, \Sigma')$$

where

$$\Sigma' = \int f^{\perp}(x)^4 w^2(x_M)\mu(x_M)\, dx_M - \left\{\int f^{\perp}(x)^2 w(x_M)\mu(x_M)\, dx_M\right\}^2$$

$$+ 2\sum_{i=1}^{\infty}\left[E\{f^{\perp 2}(X_{M,M})w(X_{M,M})f^{\perp 2}(X_{M,M+i})w(X_{M,M+i})\}\right.$$

$$\left. - \left\{\int f^{\perp}(x)^2 w(x_M)\mu(x_M)\, dx_M\right\}^2\right]$$

$$+ 4\int f^{\perp}(x)^2\sigma^2(x)w^2(x_M)\mu(x_M)\, dx_M \tag{A.6}$$

and also

$$AFPE'_a - AFPE_a = Z'_a + C' + O(h'^2_{a,opt}).$$

PROOF OF THEOREM A.2 AND THEOREM 4.2. Like in the proof of Theorem 3.3, write $x = (x', x'')$, where x represents the vector of m correct lags and x' the subvector of m' lags, and x'' the other correct lags. As in the proof of Theorem 3.1, one writes $\hat{A}'_a$ as

$$(n - i_{m'} + 1)^{-1} \sum_{i=i_{m'}}^{n} \{f(X_i) - \hat{f}_a(X'_i) + \sigma(X_i)\xi_i\}^2 w(X_{M,i})$$

$$= (n - i_{m'} + 1)^{-1} \sum_{i=i_{m'}}^{n} \sigma^2(X_i)\xi_i^2 w(X_{M,i}) + (n - i_{m'} + 1)^{-1} \sum_{i=i_{m'}}^{n} \{f(X_i) - \hat{f}_a(X'_i)\}^2 w(X_{M,i})$$

$$+ (n - i_{m'} + 1)^{-1} \sum_{i=i_{m'}}^{n} 2\{f(X_i) - \hat{f}_a(X'_i)\}\sigma(X_i)\xi_i w(X_{M,i}).$$

It is obvious that

$$(n - i_{m'} + 1)^{-1} \sum_{i=i_{m'}}^{n} \sigma^2(X_i)\xi_i^2 w(X_{M,i}) = (n - i_m + 1)^{-1} \sum_{i=i_m}^{n} \sigma^2(X_i)\xi_i^2 w(X_{M,i}) + O_p\left(\frac{1}{n}\right).$$

Next

$$\hat{f}_1(x') - f(x) =$$

$$\mu(x')^{-1}(n - i_{m'} + 1)^{-1}\{1 + o_p(1)\} \sum_{i=i_{m'}}^{n} K_h(X'_i - x')\{f(X_i) - f(x) + \sigma(X_i)\xi_i\} = T_1 + T_2$$

where

$$T_1 = \mu(x')^{-1}(n - i_{m'} + 1)^{-1}\{1 + o_p(1)\} \sum_{i=i_{m'}}^{n} K_h(X_i' - x')\{f(X_i) - f(x)\},$$

$$T_2 = \mu(x')^{-1}(n - i_{m'} + 1)^{-1}\{1 + o_p(1)\} \sum_{i=i_{m'}}^{n} K_h(X_i' - x')\sigma(X_i)\xi_i.$$

The variance of T_2 is calculated as

$$\mu(x')^{-2}(n - i_{m'} + 1)^{-1}\{1 + o(1)\}\int K_h(u' - x')^2\sigma^2(u)\mu(u)\,du$$

which is (using $u' = x' + hv'$)

$$\mu(x')^{-2}(n - i_{m'} + 1)^{-1}h^{-m'}\{1 + o(1)\}\int K(v')^2\sigma^2(x' + hv',\, u'')\mu(x' + hv',\, u'')\,dv'\,du''$$

$$= \mu(x')^{-2}(n - i_{m'} + 1)^{-1}h^{-m'}\|K\|_2^{2m'}\int \sigma^2(x',\, u'')\mu(x',\, u'')\,du''\{1 + o(1)\}.$$

Similarly, the bias from T_1 is

$$\mu(x')^{-1}\{1 + o(1)\}\int K_h(u' - x')f(u)\mu(u)\,du - f(x)$$

$$= \mu(x')^{-1}\{1 + o(1)\}\int K(v')f(x' + hv',\, u'')\mu(x' + hv',\, u'')\,dv'\,du'') - f(x)$$

$$= \mu(x')^{-1}\{1 + o(1)\}\int K(v')\left\{f(x',\, u'') + hv'^T\nabla_{x'}f(x',\, u'') + h^2\frac{1}{2}v^T\nabla_{x'}^2 f(x',\, u'')v\right\}$$

$$\times\left\{\mu(x',\, u'') + hv'^T\nabla_{x'}\mu(x',\, u'') + h^2\frac{1}{2}v^T\nabla_{x'}^2\mu(x',\, u'')v\right\}dv'\,du'' - f(x)$$

$$= \mu(x')^{-1}\int\{f(x',\, u'') - f(x)\}\mu(x',\, u'')\,du'' + O_p(h^2) = -f^{\perp}(x) + O_p(h'^2).$$

One can derive a similar formula for $\hat{f}_2(x') - f(x)$, thus we have

$$\hat{f}_a(x') - f(x) = -f^{\perp}(x) + O_p(h'^2). \tag{A.7}$$

Because x' is a proper subvector of x, the true model, we know that $f^{\perp}(x) \not\equiv 0$. Now

$$(n - i_{m'} + 1)^{-1}\sum_{i=i_{m'}}^{n}\{f(X_i) - \hat{f}_a(X_i')\}^2 w(X_{M,i})$$

has asymptotic mean

$$E[\{f(X_i) - \hat{f}_a(X_i')\}^2 w(X_{M,i})] = \int f^{\perp}(x)^2 w(x_M)\mu(x_M)dx_M + O(h'^2)$$

by using (A.7), and its asymptotic variance is $\Sigma' - 4\int f^{\perp}(x)^2\sigma^2(x)w^2(x_M)\mu(x_M)\,dx_M + O(h'^2)$. Similarly

$$(n - i_{m'} + 1)^{-1}\sum_{i=i_{m'}}^{n} 2\{f(X_i) - \hat{f}_a(X_i')\}\sigma(X_i)\xi_i w(X_{M,i})$$

has mean 0 and asymptotic variance

$$(n - i_{m'} + 1)^{-1} 4E[\{f(X_i) - \hat{f}_a(X_i')\}^2 \sigma^2(X_i) w^2(X_{M,i})]$$

which, by using (A.7), is

$$(n - i_{m'} + 1)^{-1} 4 \int f^{\perp}(x)^2 \sigma^2(x) w^2(x_M) \mu(x_M)\, dx_M.$$

Thus

$$AFPE_a' - AFPE_a = Z_a' + C' + O(h_{a,opt}'^2)$$

with

$$\sqrt{n - i_{m'} + 1}\, Z_a' \xrightarrow{D} N(0, \Sigma')$$

where Σ' and C' are as in (A.6) and (A.5). Then we have

$$P[AFPE_a' < AFPE_a] = P[Z_a' + C' + O(h_{a,opt}'^2) < 0]$$

$$= P[\zeta' > (n - i_{m'} + 1)^{1/2} c' \{1 + o(1)\}]$$

where

$$\zeta' = -\sqrt{n - i_{m'} + 1}\, Z_a' / \Sigma'^{1/2}.$$

To prove Theorem 4.1, one needs to have an auxiliary result. Note that if $h_{a,opt} = \beta(n - i_m + 1)^{-1/(m+4)}$, the variance of the third term in (A.4) is asymptotically

$$(n - i_m + 1)^{-1} E \int 4\{f(x) - \hat{f}_a(x)\}^2 \sigma^2(x) w^2(x_M) \mu(x_M)\, dx_M$$

which, by writing $\{f(x) - \hat{f}_a(x)\}^2$ as bias and stochastic parts, equals

$$(n - i_m + 1)^{-(m+8)/(m+4)} \Sigma_a \{1 + o(1)\}$$

where

$$\Sigma_a = \sigma_K^4 \beta^4 \int \sigma^2(x) r_a^2(x) w^2(x_M) \mu(x_M)\, dx_M$$

$$+ 4\|K\|_2^{2m} \beta^{-m} \int \sigma^4(x) w^2(x_M) \mu(x_M)/\mu(x)\, dx_M.$$

Meanwhile, the variance of the second term is asymptotically smaller than

$$(n - i_m + 1)^{-1} E \int \{f(x) - \hat{f}_a(x)\}^4 w^2(x_M) \mu(x_M)\, dx_M$$

$$= \sigma_K^8 h^8 \int r_a^4(x) w^2(x_M)\mu(x_M)\,dx_M/\{16(n-i_m+1)\}$$

$$+ 6\sigma_K^4 h^4\{1+o_p(1)\}E\int r_a^2(x)\sum_{i=i_m}^{n}\frac{K_h(X_i-x)^2\sigma^2(X_i)\xi_i^2}{4(n-i_m+1)^3\mu(x)^2}w^2(x_M)\mu(x_M)\,dx_M$$

$$+ 4\sigma_K^2 h^2\{1+o_p(1)\}E\int r_a(x)\sum_{i=i_m}^{n}\frac{K_h(X_i-x)^3\sigma^3(X_i)\xi_i^3}{2(n-i_m+1)^4\mu(x)^3}w^2(x_M)\mu(x_M)\,dx_M$$

$$+ \{1+o_p(1)\}E\int \sum_{i=i_m}^{n}\frac{K_h(X_i-x)^4\sigma^4(X_i)\xi_i^4}{\mu(x)^4(n-i_m+1)^5}w^2(x_M)\mu(x_M)\,dx_M$$

$$+ \{1+o_p(1)\}E\int \sum_{i,j=i_m,j\neq i}^{n}\frac{K_h(X_i-x)^2K_h(X_j-x)^2\sigma^2(X_i)\sigma^2(X_j)\xi_i^2\xi_j^2}{\mu(x)^4(n-i_m+1)^5}w^2(x_M)\mu(x_M)\,dx_M$$

$$= O_p\{h^8/(n-i_m+1)+h^4(n-i_m+1)^{-2}h^{-m}$$

$$+ h^2(n-i_m+1)^{-3}h^{-2m}+(n-i_m+1)^{-4}h^{-3m}\}$$

$$= O_p(h^8n^{-1}+h^8n^{-1}+h^{10}n^{-1}+\mathrm{h}^{12}n^{-1}) = o_p(h^4n^{-1}).$$

Hence

$$(n-i_m+1)^{(m+8)/(2m+8)}$$

$$\left[\hat{A}_a - c(K,B,C_a)(n-i_m+1)^{-4/(m+4)} - (n-i_m+1)^{-1}\sum_{i=i_m}^{n}\sigma^2(X_i)\xi_i^2 w(X_{M,i})\right]$$

is asymptotically normal with mean zero variance Σ_a, where $c(K, B, C_a)$ is a positive constant. Similarly

$$(n-i_m+1)^{(m+8)/(2m+8)}$$

$$\left[\hat{B}_a - \tilde{c}(K,B,C_a)(n-i_m+1)^{-4/(m+4)} - (n-i_m+1)^{-1}\sum_{i=i_m}^{n}\sigma^2(X_i)\xi_i^2 w(X_{M,i})/\hat{\mu}(X_i)\right]$$

is also asymptotically normal with mean zero and some positive constant variance. By equation (3.4)

$$AFPE_a = \hat{A}_a + 2K(0)^m(n-i_m+1)^{-1}h_{a,opt}^{-m}\hat{B}_a$$

which now gives the following proposition.

PROPOSITION A.1. *Under assumptions (A1)–(A7), for* $a = 1, 2$ *as* $n \to \infty$, *define*

$$Z_a = \hat{A}_a - c(K,B,C_a)(n-i_m+1)^{-4/(m+4)} - (n-i_m+1)^{-1}\sum_{i=i_m}^{n}\sigma^2(X_i)\xi_i^2 w(X_{M,i})$$

$$+ 2K(0)^m(n-i_m+1)^{-1}h_{a,opt}^{-m}\hat{B}_a$$

then

$$(n-i_m+1)^{(m+8)/(2m+8)}Z_a \xrightarrow{D} N(0,\Sigma_a) \tag{A.8}$$

and

$$AFPE_a = Z_a + c(K, B, C_a)(n - i_m + 1)^{-4/(m+4)} + (n - i_m + 1)^{-1} \sum_{i=i_m}^{n} \sigma^2(X_i)\xi_i^2 w(X_{M,i}). \tag{A.9}$$

A similar result exists for the overfitting case, i.e.

$$AFPE'_a = Z'_a + c'(K, B', C'_a)(n - i'_{m+l} + 1)^{-4/(m+l+4)} + (n - i'_{m+l} + 1)^{-1} \sum_{i=i_{m+l}}^{n} \sigma^2(X_i)\xi_i^2 w(X_{M,i}), \tag{A.10}$$

$$(n - i'_{m+l} + 1)^{(m+l+8)/(2m+2l+8)} Z'_a \xrightarrow{D} N(0, \Sigma'_a). \tag{A.11}$$

PROOF OF THEOREM 4.1. We note that

$$\delta = (n - i'_{m+l} + 1)^{-1} \sum_{i=i'_{m+l}}^{4n} \sigma^2(X_i)\xi_i^2 w(X_{M,i}) - (n - i_m + 1)^{-1} \sum_{i=i_m}^{n} \sigma^2(X_i)\xi_i^2 w(X_{M,i}) = O_p(n^{-1}).$$

Thus, using equations (A.9), (A.8), (A.10) and (A.11) one obtains

$$P[AFPE'_a < AFPE_a] = P[Z'_a - Z_a + \delta < -c'(K, B', C'_a)(n - i'_{m+l} + 1)^{-4/(m+l+4)} + c(K, B, C_a)(n - i_m + 1)^{-4/(m+4)}].$$

Note that

$$(n - i'_{m+l} + 1)^{(m+l+8)/(2m+2l+8)} Z_a = \{(n - i_m + 1)^{(m+8)/(2m+8)} Z_a\} \times O\{n^{-2l/\{(m+l+4)(m+4)\}}\} \xrightarrow{P} 0,$$

$$(n - i'_{m+l} + 1)^{(m+l+8)/(2m+2l+8)}(n - i_m + 1)^{-4/(m+4)} = o\{(n - i'_{m+l} + 1)^{(m+l)/(2m+2l+8)}\}$$

which give

$$P[AFPE'_a < AFPE_a] = P[\zeta'_a > (n - i_m + 1)^{(m+l)/(2m+2l+8)} c'_a\{1 + o(1)\}]$$

where

$$\zeta'_a = (n - i'_{m+l} + 1)^{(m+l+8)/(2m+2l+8)}(Z'_a - Z_a + \delta).$$

PROOF OF THEOREM 4.4. Using arguments as before, one needs only to show that if x'' is a proper subvector of $x' = (x_{i'_1}, \ldots, x_{i'_{m'}})$, then

$$C'' > C'$$

where C' is as in (A.5) and

$$C'' = \int f(x)^2 w(x_M)\mu(x_M)\,dx_M - \int E^2\{f(x)|x''\}w(x_M)\mu(x_M)\,dx_M$$

which yields

$$C'' - C' = \int [E\{f(x)|x''\} - E\{f(x)|x'\}]^2 w(x_M)\mu(x_M)\,dx_M > 0$$

as we assume that the true model includes all the lags $i'_1, \ldots, i'_{m'}$.

ACKNOWLEDGEMENTS

The authors thank Björn Auestad, Olaf Bunke, Christian Hafner, Wolfgang Härdle, Joel Horowitz, Helmut Lütkepohl, Michael Neumann, Franz Palm, Dag Tjøstheim, Howell Tong and Alexander Tsybakov for many helpful discussions and comments. The comments from the Associate Editor and a referee also helped us to improve our paper significantly. Versions of this work have been presented in seminars at the Georgia Institute of Technology in Atlanta, the Chinese Academy of Sciences and Peking University in Beijing, LIFE of the University of Maastricht, CREST in Paris, Charles University in Prague, Tinbergen Institute in Rotterdam, University of California at Santa Barbara, the Stockholm School of Economics and CentER at Tilburg University. We gladly acknowledge the constructive comments of the seminar participants. This research was financially supported by the Sonderforschungsbereich 373 'Quantifikation und Simulation Ökonomischer Prozesse' which was funded by the Deutsche Forschungsgemeinschaft and was mostly done while Lijian Yang was visting the Humboldt-Universität zu Berlin.

REFERENCES

Akaike, H. (1969) Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* 21, 243–247.

Akaike, H. (1971) Autoregressive model fitting for control. *Annals of the Institute of Statistical Mathematics* 23, 163–180.

Ango Nze, P. (1992) Critères d'ergodicité de quelques modèles à représentation Markovienne. *C. R. Acad. Sci. Paris*, sér I, 315, 1301–1304.

Auestad, B. and Tjøstheim, D. (1990) Identification of nonlinear time series: first order characterization and order determination. *Biometrika* 77, 669–687.

Chen, R. and Tsay, R. S. (1993) Nonlinear additive ARX models. *Journal of the American Statistical Association* 88, 955–967.

Cheng, B. and Tong, H. (1992) On consistent non-parametric order determination and chaos (with discussion). *Journal of the Royal Statistical Society, Series B* 54, 427–474.

Diebolt, J. and Guégan, D. (1993) Tail behaviour of the stationary density of general nonlinear autoregressive processes of order one. *Journal of Applied Probability* 30, 315–329.

Doukhan, P. (1994) *Mixing, Properties and Examples*. New York: Springer-Verlag.

Engle, R. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987–1008.

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

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

Härdle, W., Lütkepohl, H. and Chen, R. (1997) A review of nonparametric time series analysis. *International Statistical Review* 65, 49–72.

HÄRDLE, W., TSYBAKOV, A. B. and YANG, L. (1998) Nonparametric vector autoregression. *Journal of Statistical Planning and Inference* 68, 221–245.
JONES, M. C., MARRON, J. S. and SHEATHER, S. J. (1996) A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* 91, 401–407.
LÜTKEPOHL, H. (1991) *Introduction to Multiple Time Series Analysis*. Heidelberg: Springer-Verlag.
NUMMELIN, E. and TUOMINEN, P. (1982) Geometric erodicity of Harris–recurrent Markov chains with applications to renewal theory. *Stochastic Processes and their Applications* 12, 187–202.
RUPPERT, D. and WAND, M. P. (1994) Multivariate locally weighted least squares regression. *Annals of Statistics* 21, 1346–1370.
SAULIS, L. and STATULEVIČIUS, V. A. (1991) *Limit Theorems for Large Deviations*. Kluwer.
TJØSTHEIM, D. (1994) Non-linear time series analysis: a selective review. *Scandinavian Journal of Statistics* 21, 97–130
TJØSTHEIM, D. and AUESTAD, B. (1994) Nonparametric identification of nonlinear time series: selecting significant lags. *Journal of the American Statistical Association* 89, 1410–1419.
TWEEDIE, R. L. (1975) Sufficient conditions for ergodicity and recurrence of Markov chains on a general state space. *Stochastic Processes and their Applications* 3, 385–403.
VIEU, P. (1994) Order choice in nonlinear autoregressive models. *Statistics* 24, 1–22.
WAND, M. P. and JONES, M. C. (1995) *Kernel Smoothing*. London: Chapman and Hall.
YANG, L., HÄRDLE, W. and NIELSEN, J. P. (1999) Nonparametric autoregression with multiplicative volatility and additive mean. *Journal of Time Series Analysis*, 20, 579–604.
YANG, L. and TSCHERNIG, R. (1999) Multivariate bandwidth selection for local linear regression. *Journal of the Royal Statistical Society, Series B*, 61, 793–815.
YAO, Q. and TONG, H. (1994) On subset selection in non-parametric stochastic regression. *Statistica Sinica* 4, 51–70.
YOSHIHARA, K. (1976) Limiting behaviour of *U*-statistics for stationary, absolutely regular processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 35, 237–252.