Tschernig, Rolf; Yang, Lijian

Working Paper

Nonparametric lag selection for time series

SFB 373 Discussion Paper, No. 1997,59

Provided in Cooperation with:
Collaborative Research Center 373: Quantification and Simulation of Economic Processes, Humboldt University Berlin

Suggested Citation: Tschernig, Rolf; Yang, Lijian (1997) : Nonparametric lag selection for time series, SFB 373 Discussion Paper, No. 1997,59, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, Berlin, https://nbn-resolving.de/urn:nbn:de:kobv:11-10064444

This Version is available at:
https://hdl.handle.net/10419/66278

Nonparametric Lag Selection for Time Series

Rolf TSCHERNIG and Lijian YANG*

July 1997

Abstract

A nonparametric version of the Final Prediction Error (FPE) is proposed for lag selection in nonlinear autoregressive time series. We derive its consistency for both local constant and local linear estimators using a derived optimal bandwidth. Further asymptotic analysis suggests a greater probability of overfitting (too many lags) than underfitting (missing important lags). Thus a correction factor is proposed to increase correct fitting by reducing overfitting. Our Monte-Carlo study also corroborates that the correction factor generally improves the probability of correct lag selection for both linear and nonlinear processes. The proposed methods are successfully applied to the Canadian lynx data and daily returns of DM/US-Dollar exchange rates.

KEY WORDS: Consistency; Final Prediction Error; Foreign Exchange Rates; Lag Selection; Nonlinear Autoregression; Nonparametric Method.

1. INTRODUCTION

The past decade has witnessed the tremendous development of nonparametric modeling, 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, also often impractical. Following the successful adaption of nonparametric regression

*Rolf Tschernig is Research Associate, Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät, Humboldt-Universität zu Berlin, Spandauer Str.1, D-10178 Berlin, Germany. Lijian Yang is Assistant Professor, Department of Statistics and Probability, Michigan State University, East Lansing, Michigan 48824. 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. 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 visiting the Humboldt-Universität zu Berlin.

techniques to time series analysis (Györfi, Härdle, Sarda and Vieu 1989, Tjøstheim 1994), alternative lag selection criteria have been studied for nonlinear autoregressive processes by Cheng and Tong (1992), Vieu (1994), Yao and Tong (1994) using cross-validation, and by Auestad and Tjøstheim (1990), Tjøstheim and Auestad (1994) using FPE. Both the cross-validation and the FPE are substitutes of the naive mean squared error estimate which is known to be unsuitable for model selection. Other nonparametric lag selection methods were suggested by Chen and Tsay (1993) for additive nonlinear autoregressive models, a subclass of the nonlinear autoregressive models considered in this paper.

However, for the nonparametric FPE, neither the estimation properties are well investigated, nor a satisfactory bandwidth selection method has been derived. Both will be the topics of this paper. We derive consistency for the nonparametric FPE and give partial results of misspecification probabilities. As our calculation suggests that overfitting (too many lags) is more likely than underfitting (missing correct lags), a correction factor is used to reduce the probability of overfitting and hence increase correct fitting.

We also propose an optimal bandwidth for the FPE criterion by solving a type of bias-variance trade-off problem. Previously, the proposed bandwidths had an open range of orders and were selected by minimizing the specific criteria. Whatever bandwidth one decided to use did not necessarily approximate some optimal bandwidth. Our analysis takes Vieu (1994) as a starting point which gave some theoretical justification in the cross-validation case and pointed out problems of other methods.

Another innovation is the use of the local linear estimator in place of the Nadaraya-Watson estimator. The main reason for this is that the Nadaraya-Watson estimator has a poor bias rate when the density of the lagged variable is not sufficiently smooth, especially with nonlinear processes, while the local linear estimator needs only 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, Tsybakov and Yang (1997).

We also analyze the performance of the suggested methods in an extensive Monte-Carlo study and discuss implementation issues. Finally we apply these procedures to the 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.

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. Section 4 calculates the probabilities of over- and underfitting. The practical implementation of the nonparametric FPE estimators is discussed in Section 5. Section 6 consists of a comprehensive report of our Monte-Carlo study. The analysis of 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

Our idea of using a nonparametric FPE came from Auestad and Tjøstheim (1990) and Tjøstheim and Auestad (1994). Suppose one has a conditional heteroscedastic autoregressive time series $\{Y_t\}_{t\geq 0}$

$$Y_t = f(X_t) + \sigma^{1/2}(X_t)\xi_t \tag{2.1}$$

where $X_t = (Y_{t-i_1}, Y_{t-i_2}, ..., Y_{t-i_m})^T$ is the vector of all correct lagged values, $i_1 < \cdots < i_m$, and ξ_t are i.i.d. random variables with $E(\xi_t) = 0$, $E(\xi_t^2) = 1$, $t = i_m, i_m + 1, ...$ Here we assume that all lags $i_1, ..., i_m$ are needed for modelling $f(\bullet)$ but not necessarily for $\sigma(\bullet)$. The case in which $\sigma(\bullet)$ depends on lags not contained in $f(\bullet)$ is beyond this paper.

To define the Final Prediction Error (FPE), let $\left\{\widetilde{Y}_t\right\}$ be another series with exactly the same distribution as $\{Y_t\}$ but independent of $\{Y_t\}$. The FPE of an estimate $\widehat{f}$ of f is defined as the following functional

$$FPE(\widehat{f}) = \lim_{t\to\infty} E\left\{\widetilde{Y}_t - \widehat{f}(\widetilde{X}_t)\right\}^2 \tag{2.2}$$

where the expectation is taken over all the variables: $Y_0, Y_1, ..., Y_n, \widetilde{Y}_0, \widetilde{Y}_1, ..., \widetilde{Y}_t, ...$ This FPE measures the discrepancy between $\widehat{f}$ and the true functional relation of $\widetilde{Y}_t$ to $\widetilde{X}_t$, which is more easily understood conceptually than the cross-validation as it depends only on the estimator $\widehat{f}$ and the limiting distribution of the process. If the process $\{Y_t\}$ is a stationary linear AR process, $\widehat{f}$ 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 an ergodic nonlinear AR process and $\widehat{f}$ some nonparametric estimator, we have the nonparametric FPE.

To define the nonparametric FPE, we assume the following

(A1) There exists an integer $M \geq i_m$ such that the Markov chain $X_{M,t} = (Y_{M,t-1}, ..., Y_{M,t-M})^T$ defined by equation (2.1) is geometrically ergodic, i.e., it is ergodic, with stationary probability measure $\pi_M(\bullet)$ such that, for almost every $x \in I\!R^M$, as $k \to \infty$

$$||P^k(\bullet|x) - \pi(\bullet)||_{TV} = O(\rho^k),$$

for some $0 \leq \rho = \rho(x) < 1$. Here

$$P^k(B|x) = P\{X_{M,k} \in B | X_{M,M} = x\},$$

for a Borel subset $B \subset I\!R^M$, and $||\bullet||_{TV}$ is the total variation distance.

(A2) The stationary distribution of $X_{M,t}$ has a density function $\mu_M(x)$, which is compactly supported and bounded below from zero on its support. All the $X_{M,t}$'s take values within the support of $\mu_M(x)$.

Our assumptions here are similar to those of Yao and Tong (1994) and Vieu (1994). See Tweedie (1975), Nummelin and Tuominen (1982), Ango Nze (1992), Diebolt and Guégan (1993) for conditions that yield geometric ergodicity. For other assumptions that had been used, see Tjøstheim (1994), Tjøstheim and Auestad (1994). Note that we do not assume an identical distribution for the $X_{M,t}$'s, as Yao and Tong (1994) did, because geometric ergodicity suffices for our purpose here. Also, we have better mixing properties here as geometric ergodicity implies geometrical mixing under mild conditions:

Lemma 2.1 (Davydov(1973)).
A geometrically ergodic Markov chain whose initial variable is distributed with its stationary distribution is geometrically strongly mixing with the mixing coefficients satisfying $\alpha(n) \leq c_0\,\rho_0^n$ *for some* $0 < \rho_0 < 1$, $c_0 > 0$.

From now on, without loss of generality, we assume that the process $\{X_{M,t}\}$ has a stationary initial distribution and use $\mu(\bullet)$ to denote both $\mu_M(\bullet)$ and all of its marginal densities, and integration operations are carried out over the compact support of the appropriate $\mu(\bullet)$'s, although we will drop all such references. We assume further

(A3) The function $f(\bullet)$ is componentwise twice continuously differentiable at every point on the support of $\mu(\bullet)$ while $\sigma(\bullet)$ is continuous.

(A4) The density $\mu(\bullet)$ of the stationary distribution $\pi(\bullet)$ exists and is continuously differentiable on the support of $\mu(\bullet)$.

Assumption (A3) is a smoothness condition for the functions $f(\bullet)$ and $\sigma(\bullet)$. Assumption (A4) is necessary to compute the asymptotic bias and variance. However, as mentioned in the introduction, for the local linear estimator assumption (A4) can be relaxed to continuity of $\mu(\bullet)$.

Under assumptions (A1) to (A4), it is unnecessary to generate the process $\left\{\widetilde{Y}_t\right\}$ to compute the FPE. Denote $\mathbf{Y} = (Y_{i_m}, Y_{i_m+1}, ..., Y_n)^T$. For any $x \in I\!R^m$, write

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

in which

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

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

where

(A5) $K : I\!R^1 \longrightarrow I\!R^1$ is a symmetric positive kernel with $\int K(u)du = 1$ and

$$K_h(u) = 1/h^m \prod_{j=1}^m K(u_j/h)$$

for $u \in I\!R^m$; $h = h_n$ is a positive number (bandwidth), $h \to 0$, $nh^m \to \infty$ as $n \to \infty$.

The $\widehat{f}_1(x)$ and $\widehat{f}_2(x)$ 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, see Wand and Jones (1995). The kernel function K matters little here, so $\widehat{f}_1(x)$ and $\widehat{f}_2(x)$ depend primarily on h, and so do the FPEs. We therefore write for $a = 1, 2$

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

As in most kernel methods, these functions of h have simple approximations. Denoting $\|K\|_2^2 = \int K^2(u)du$ and $\sigma_K^2 = \int K(u)u^2du$ we obtain

Theorem 2.1 *Under assumptions (A1)-(A5), for $a = 1, 2$, as $n \to \infty$*

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

in which the Asymptotic FPE's are

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

where

$$A = \int \sigma(x)\mu(x)dx, \qquad B = \int \sigma(x)dx, \tag{2.4}$$

$$C_1 = \int \left[\mathrm{Tr}\left\{\nabla^2 f(x)\right\} + 2\nabla^T \mu(x)\nabla f(x)/\mu(x)\right]^2 \mu(x)dx, \tag{2.5}$$

$$C_2 = \int \left[\mathrm{Tr}\left\{\nabla^2 f(x)\right\}\right]^2 \mu(x)dx \tag{2.6}$$

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 now made possible by using instead the asymptotically equivalent AFPE. The term A represents the expected variance function of the data generating process with respect to its stationary distribution. The second and third term $b(h)B$ and $c(h)C_a$ come from estimation uncertainty and denote 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)-(A5) and the additional assumption that $C_a > 0$, $a = 1, 2$, the AFPE's are minimized by*

$$h_{a,opt} = \left\{m\|K\|_2^{2m} B(n - i_m + 1)^{-1} C_a^{-1}\sigma_K^{-4}\right\}^{1/(m+4)} \tag{2.7}$$

and the minimum AFPE is

$$AFPE_{a,opt} = A + \left(m^{-m/(m+4)} + \frac{1}{4}m^{4/(m+4)}\right)\left\{\|K\|_2^{8m} B^4 (n - i_m + 1)^{-4} C_a^m \sigma_K^{4m}\right\}^{1/(m+4)}. \tag{2.8}$$

From this point on, we refer to the bandwidths in (2.7) as the optimal bandwidths, although their optimality is only asymptotic.

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 mostly when one uses the local linear estimator for linear processes, in which case $\nabla^2 f(x) \equiv 0$ implies $C_2 = 0$ where the local linear estimator does not have a bias of order h^2. One may call this the "*curse of linearity*".

Note 2.2 If $C_a = +\infty$, the trade-off also fails. This occurs, for example, if one uses the Nadaraya-Watson estimator for a nonlinear process which violates the smoothness condition for $\mu(x)$ in assumption (A4) (i.e.$\nabla\mu(x)$ does not exist at some points), in which case $C_1 = +\infty$ (See the simulation example **NLAR4**).

Based on these discussions, we need a sixth assumption

(A6) For $a = 1, 2$, the C_a defined in (2.5) and (2.6) are positive.

Note that all the results of this section are based on the assumption that X_t is the vector of correct lagged values. Furthermore, (2.3) contains the unknown quantities A, B, 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. The main focus will be the consistency of the AFPE based lag selection rules.

3. THE CONSISTENCY

Formula (2.8) contains the unknown quantities A, B, and C_1 (C_2). We define the following estimates

$$\widehat{A}_a = (n - i_m + 1)^{-1} \sum_{i=i_m}^{n} \left\{ Y_i - \widehat{f}_a(X_i) \right\}^2, \tag{3.1}$$

$$\widehat{B}_a = (n - i_m + 1)^{-1} \sum_{i=i_m}^{n} \left\{ Y_i - \widehat{f}_a(X_i) \right\}^2 / \widehat{\mu}(X_i) \tag{3.2}$$

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

Theorem 3.1 *Under assumptions (A1)-(A6), for $a = 1, 2$, as $n \to \infty$*

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

Note here that the nonparametric estimate $\widehat{A}_a$ converges to A at the parametric $\sqrt{n}$ rate if $m \leq 4$, in which case the second and third term will be $O\left\{ (n - i_m + 1)^{-1/2} \right\}$.

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

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

in which $\widehat{A}_a$ is evaluated using the optimal bandwidth $h_{a,opt}$, while $\widehat{B}_a$ using any bandwidth of order $(n - i_m + 1)^{-1/(m+4)}$. Note that 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 MSE, while the second term serves as a penalty term to avoid noise fitting which would result by simply using $\widehat{A}_a$ alone.

Now one computes for every subset $\{i'_1, ..., i'_{m'}\}$ of $\{1, ..., M\}$ the $AFPE'_1$ and $AFPE'_2$ as discussed above. We propose the following

> **Lag Selection Rule I**: Select the subset $\left\{ \widehat{i}_1, ..., \widehat{i}_{\widehat{m}} \right\}$ with the smallest $AFPE'_1$ (or $AFPE'_2$).

Theorem 3.2 *Under assumptions (A1)-(A6), Lag Selection Rule I consistently selects the correct set of lags. I.e., if $\widehat{i}_1, ..., \widehat{i}_{\widehat{m}}$ are the selected lags, then as $n \to \infty$*

$$P\left[\widehat{m} = m, \widehat{i}_s = i_s, s = 1, 2, ..., m \right] \longrightarrow 1.$$

The theorem guarantees that the probability of Selection Rule I failing to completely identify the correct model diminishes with larger sample size. Our result bears similarity to Vieu (1994) and Yao and Tong (1994), except the use of AFPE instead of cross-validation. This theorem is obtained by investigating what happens to the AFPE if the model one uses in formula (3.4) is incorrect.

In the following, we denote by $AFPE_1'$, $AFPE_2'$ the statistics that one gets when using X', an arbitrary vector of lags, to calculate the AFPE's. We distinguish two cases where X' is different from X.

Definition 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 have the following result similar to Theorem 2.1.

Theorem 3.3 *Let* $X_t' = (Y_{t-i_1}, Y_{t-i_2}, ..., Y_{t-i_m}, Y_{t-i_{m+1}}, ..., Y_{t-i_{m+l}})^T$ *where* $i_{m+1} < \cdots < i_{m+l}$ $(l > 0)$ *are different from but not necessarily larger than the correct lags, i.e.* $\{i_1, \ldots, i_m\} \cup \{i_{m+1}, \ldots, i_{m+l}\} = \emptyset$. *Define* $i_{m+l}^* = \max(i_m, i_{m+l})$. *Then under assumptions (A1)-(A6), for* $a = 1, 2$,

$$AFPE_a' = A + b(h_{a,opt}')B + c(h_{a,opt}')C_a' \tag{3.5}$$

where

$$C_1' = \int \left[\mathrm{Tr}\left\{\nabla^2 f(x)\right\} + 2\nabla^T \mu(x')\nabla f(x)/\mu(x')\right]^2 \mu(x')dx', \tag{3.6}$$

$$C_2' = \int \left[\mathrm{Tr}\left\{\nabla^2 f(x)\right\}\right]^2 \mu(x)dx = C_2 \tag{3.7}$$

in which

$$b(h_{a,opt}') = \|K\|_2^{2(m+l)} (n - i_{m+l}^* + 1)^{-1} \left(h_{a,opt}'\right)^{-(m+l)}, \quad c(h_{a,opt}') = \sigma_K^4 h_{a,opt}'^4/4,$$

x' *denotes the vector values at lags* $i_1, ..., i_{m+l}$, *and*

$$h_{a,opt}' = \left\{(m+l)\|K\|_2^{2(m+l)} B(n - i_{m+l}^* + 1)^{-1} C_a'^{-1} \sigma_K^{-4}\right\}^{1/(m+l+4)}$$

is the optimal bandwidth.

Corollary 3.1 *In the setting of Theorem 3.3,*

$$AFPE_{a,opt}' = A + \left[(m+l)^{-(m+l)/(m+l+4)} + \frac{1}{4}(m+l)^{4/(m+l+4)}\right]$$

$$\left\{\|K\|_2^{8(m+l)} B^4 (n - i_{m+l}^* + 1)^{-4} C_a'^{(m+l)} \sigma_K^{4(m+l)}\right\}^{1/(m+l+4)} \tag{3.8}$$

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 consider only the case of a proper subvector of the true lag vector for notational simplicity.

Theorem 3.4 *Let $X'_t = (Y_{t-i'_1}, ..., Y_{t-i'_{m'}})^T$ be any subvector of X_t ($0 < m' < m$). Under assumptions (A1)-(A6), there exists a constant $C' > 0$ (depending on $i'_1, ..., i'_{m'}$) 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' which is the squared error of projecting the process unto the submodel defined by X'.

The consistency result 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 takes 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 wrong lags. Our analysis here of the overfitting and underfitting probabilities gives insights into the quantitative aspects of the selction procedures. Such analysis should also be possible using cross-validation.

We first obtain a partial result on the asymptotic probability of overfitting

Theorem 4.1 *Let X'_t be defined as in Theorem 3.3. Under assumptions (A1)-(A6), there exist a constant $c'_a > 0$ and $\zeta'_a \xrightarrow{D} N(0,1)$ such that,*

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

The asymptotic probability of underfitting is given in

Theorem 4.2 *Let X'_t be as in Theorem 3.4. Under assumptions (A1)-(A6), there exists a $\zeta' \xrightarrow{D} N(0,1)$ such that, for $c' = C'/\Sigma'^{1/2} > 0$, where C' and Σ' are defined in (8.9) and (8.7), as $n \to \infty$*

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

Note 4.1 If heuristically, one assumes that the $\zeta'_a, a = 1, 2$ were *exactly* instead of asymptotically $N(0,1)$, then the overfitting probability in Theorem 4.1 would be $\Phi((n - i_m + 1)^{(m+l)/(2m+2l+8)} c'_a \left\{1 + o(1)\right\})$ where we denote by $\Phi(x) = 1/\sqrt{2\pi} \int_{-\infty}^{x} e^{-t^2/2} dt$ 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 $\Phi((n - i_{m'} + 1)^{1/2} c' \left\{1 + o(1)\right\})$. One may expect that 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 interested value to be of order no more than $n^{1/6}$, which is never fulfilled in our results except for

$P\left[\zeta_a' > (n - i_m + 1)^{(m+l)/(2m+2l+8)} c_a' \{1 + o(1)\}\right]$ with $m = l = 1$. This is why we had succeeded only in obtaining the partial results of Theorems 4.1 and 4.2, not the heuristics in Note 4.1.

Note 4.3 Since $1 - \Phi(x)$ goes to zero faster if x goes to $+\infty$ faster, Note 4.1 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 heuristic consideration is supported 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 needs to penalize overfitting more. We define a corrected AFPE as

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

which gets larger for larger models at a faster rate than $AFPE_a$. Correspondingly, one has a new lag selection rule

Lag Selection Rule II: Select the subset $\left\{\widehat{i}_1, ..., \widehat{i}_{\widehat{m}}\right\}$ 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)-(A6), let $\widehat{i}_1, ..., \widehat{i}_{\widehat{m}}$ be the lags selected according to the Lag Selection Rule II, then as $n \to \infty$*

$$P\left[\widehat{m} = m, \widehat{i}_s = i_s, s = 1, 2, ..., m\right] \longrightarrow 1.$$

Another interesting issue is what happens when one selects lags out of $\{1, 2, ..., M'\}$ 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 $(j = 1, ..., m)$ that are in $\{1, 2, ..., M'\}$ and no more. This is the case.

Theorem 4.4 *Let $i_1', ..., i_{m'}'$ be all the i_j's $(j = 1, ..., m)$ that are in $\{1, 2, ..., M'\}$. Under assumptions (A1)-(A6), let $\widehat{i}_1, ..., \widehat{i}_{\widehat{m}}$ be the lags selected according to the Lag Selection Rule I or II from among $1, 2, ..., M'$, then as $n \to \infty$*

$$P\left[\widehat{m} = m', \widehat{i}_s = i_s', s = 1, 2, ..., m'\right] \longrightarrow 1.$$

5. IMPLEMENTING THE FPE ESTIMATORS

Computing the FPE estimators (3.4) and (4.1) based on (3.1) and (3.2) requires suitable kernel and bandwidth choices. With respect to the former we decide to use the Gaussian kernel. To estimate the optimal bandwidth $h_{a,opt}$ given by (2.7) we estimate B by (3.2), while for C_2 (2.6) we use a consistent local quadratic estimator given in Yang and Tschernig (1997). For computing $\widehat{f}_a(\cdot)$ and $\widehat{\mu}(\cdot)$ in $\widehat{B}_a$ (3.2) 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 is applied. For estimating C_2 we use the bandwidth $2h_S(m + 10)$ plus the leave-one-out method.

Note that the above plug-in estimation of the "bias term" is harder for the local constant estimator C_1 (2.5) than for the local linear estimator C_2 since it also involves the first derivatives of the density. Therefore, we use a grid search procedure for the estimation of the optimal bandwidth $h_{1,opt}$ (2.7) which, of course, can also be applied to calculate $h_{2,opt}$. It is theoretically justified by Corollary 2.1 on the existence of an optimal bandwidth. The grid search is conducted by covering 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. This follows Tjøstheim and Auestad's (1994) specification of estimating $AFPE_1$.

We also implement two additional features of Tjøstheim and Auestad (1994) for robustification. First, all possible observations for estimating the density $\mu(x)$ are used by

$$\widetilde{\mu}(x) = (n - i_m + i_1 + 1)^{-1} \sum_{i=i_m}^{n+i_1} K_h(X_i - x) \tag{5.2}$$

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 $(Y_n, \ldots, Y_{n+i_1-i_m})^T$. This density estimate is used not only in the denominator of $\widehat{B}$ (3.2) but also in the denominator of the Nadaraya-Watson estimator. Second, for estimation 5% of the observations are screened off, i.e. those with the lowest density $\widetilde{\mu}(x)$.

We are now in the position to compute all $(C)AFPE_a$, $a = 1, 2$ 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 the finite sample properties of the $AFPE_a$ and $CAFPE_a$ 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 estimators $(C)AFPE_2$ do not obey Theorems 3.2 and 4.3 if the true DGP is linear in the conditional mean. As a consequence one may expect the local constant estimators $AFPE_1$ 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

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

are of order 2 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

$$\begin{aligned} Y_t &= -0.4(3 - Y_{t-1}^2)/(1 + Y_{t-1}^2) \\ &+ 0.6\left\{3 - (Y_{t-2} - 0.5)^3\right\} / \left\{1 + (Y_{t-2} - 0.5)^4\right\} + 0.1\xi_t, \quad \xi_t \sim i.i.d.N(0,1), \end{aligned}$$

NLAR2 Additive nonlinear AR process (exponential autoregression)

$$\begin{aligned} Y_t &= \left\{0.4 - 2\exp(-50Y_{t-6}^2)\right\} Y_{t-6} + \left\{0.5 - 0.5\exp(-50Y_{t-10}^2)\right\} Y_{t-10} + 0.1\xi_t, \\ &\quad \xi_t \sim i.i.d.N(0,1), \end{aligned}$$

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

$$\begin{aligned} Y_t &= (0.4 - 2\cos(40Y_{t-6})\exp(-30Y_{t-6}^2))Y_{t-6} \\ &+ (0.55 - 0.55\sin(40Y_{t-10})\exp(-10Y_{t-10}^2))Y_{t-10} + 0.1\xi_t, \quad \xi_t \sim i.i.d.N(0,1), \end{aligned}$$

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** have all additive nonlinear 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 (A4) in order to investigate the practical relevance of Note 2.2 for the local constant $(C)AFPE_1$ estimation. 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 model selection criteria and four versions of the nonparametric FPE lag selection criteria. The linear criteria are the FPE, AIC, Schwarz criterion and Hannan-Quinn criterion, abbreviated by $ARFPE$, $ARAIC$, $ARSC$ and $ARHQ$. See e.g. Lütkepohl (1991) for details. The nonparametric FPE criteria include: $AFPE_1$ (3.4), $CAFPE_1$, $CAFPE_2$ and $CAFPE_{2a}$ (4.1). They differ with respect to the use of the correction factor and the bandwidth selection method. We use the grid search procedure except for $CAFPE_{2a}$ where we use the plug-in bandwidth (2.7). Note that the $AFPE_1$ (3.4) was already suggested by Auestad and Tjøstheim (1990) and Tjøstheim and Auestad (1994). All nonparametric criteria were computed as described in section 5.

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

Linear AR(2) Processes

Figure 3 shows that 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 correct 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 is not 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 $CAFPE_{2(a)}$ criteria may fail for linear DGPs is practically relevant. The best nonparametric criterion $CAFPE_1$ is indeed based on the local constant estimator. It also has a much smaller overfitting probability than the $CAFPE_{2(a)}$ criteria which is a direct consequence of the non-existing finite optimal bandwidth for the latter criteria in the present case.

Note also the important finding 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). Furthermore, it reduces the probability of overfitting although underfitting becomes more likely.

Nonlinear AR(2) 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 comparatively or even superior to the nonlinear 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 contrast to the case of linear DGPs the $CAFPE_{2(a)}$ criteria now perform in general at least as good 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 f)) 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 a situation of a nonsmooth density $C_1 = +\infty$ and therefore $(C)AFPE_1$ do not obey Theorem 3.2 and Theorem 4.3. In such a case one might prefer to use $CAFPE_{2(a)}$ as corroborated by the results for the **NLAR4** process. There $CAFPE_{2(a)}$ 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. Finally, one observes that overall the correct selection frequencies are higher than what one might have expected for nonlinear processes based on only 100 observations.

All Processes

Using the plug-in bandwidth (2.7) leads to at least as many correct selections than using the grid search procedure. This can be seen by always comparing the performance of the $CAFPE_2$ and $CAFPE_{2a}$ criteria in Figures 3 and 4. This result allows to save an enormous amout of computer time.

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. On the other hand, the risk of using the $CAFPE_{2a}$ criterion consists mainly in overfitting the true model. Furthermore, the correction factor should always be used and the optimal bandwidth estimated if possible.

From these results we suggest the following procedure for empirical work. Using the $CAFPE_{2a}$ criterion is ideal for reducing the initial set of potential lags to a smaller set which

Table 1: Nonparametric lag selection for lynx data

Est. method	max. # lags	Selected lags	crit. value	$h_n, h_{a,opt}$
$ARSC$	6	1,2	-2.828	
$CAFPE_1$	6	1,3	0.0780	0.241
$CAFPE_{2a}$	6	1,2,5,8	0.0420	0.429
	3	1,2,5	0.0434	0.363
	2	1,2	0.0457	0.335

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.

is likely to include the correct lags. Eliminating possible irrelevant lags has then to be done by investigating the properties of the proposed model and included submodels as well as of the corresponding 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 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 same estimation setup as in the Monte-Carlo study and logs 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 5 and 8. 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 dismissed the model with lag 1 and 3 since its residuals exhibit more remaining autocorrelation than the competing model. Tjøstheim and Auestad (1994) found lags 1 and 3 using $AFPE_1$ and 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 of the large sample size is compromised as the data is known to be highly dependent and therefore asymptotics are expected to kick in very slowly.

By applying the $CAFPE_2$ criterion we find lags 1 and 3 with $h_{2,opt} = 0.0064$. 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 belief that $f(x) = 0$. 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 $\sigma(x)$ we square the model (2.1) with $f(x) = 0$. This gives

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

which can be estimated with the tools developed in this paper by simply replacing the dependent variable Y_t by its squares. Using the $CAFPE_2$ criterion we obtain again lag 1 and 3 with a bandwidth estimate of 0.0040. Investigating 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 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 and Härdle (1997) where lags recommended by our analysis were used.

8. CONCLUSIONS

In this paper we looked closely at the nonparametric FPE using either the Tjøstheim and Auestad (1994) local constant estimates or local linear estimates. We derived consistency and asymptotic probabilities for underfitting and overfitting. Based on these results we proposed a correction factor to increase correct fitting. The new criteria were compared to some 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 obtain a good fit with a parsimonious model. For the daily DM/US-$ exchange rate returns we find a highly nonlinear and asymmetric volatility function of lag 1 and 3, which presents interesting new challenges for the parametric modelling of this highly investigated series.

If nonlinearity is considered in empirical research, our corrected nonparametric FPE criteria provide some helpful tools for both detecting the correct lags and modelling.

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 adopted the original notation to ours)

$$\lim_{t\to\infty} E\left\{\widehat{f}(\widetilde{X}_t) - f(\widetilde{X}_t)\right\}^2$$

$$= \lim_{t\to\infty} E\left\{\widehat{f}(\widetilde{X}_t) - E\widehat{f}(\widetilde{X}_t) + E\widehat{f}(\widetilde{X}_t) - f(\widetilde{X}_t)\right\}^2 = \lim_{t\to\infty} E(I' + II')^2.$$

As one sees from that paper, II' is the bias term of $\widehat{f}(\widetilde{X}_t)$. Härdle, Tsybakov and Yang (1997) gave an explicit formula of the bias for the local linear estimator $\widehat{f}_2(x)$, which is

$$\sigma_K^2 h^2/2 \,\mathrm{Tr}\left\{\nabla^2 f(x)\right\}.$$

Thus

$$\lim_{t\to\infty} E(II')^2 = \sigma_K^4 h^4/4 \int \left[\mathrm{Tr}\left\{\nabla^2 f(x)\right\}\right]^2 \mu(x)dx + O\left\{h^4(n - i_m + 1)^{-1/2}\right\}$$

$$= c(h)C_2 + O\left\{h^4(n - i_m + 1)^{-1/2}\right\}$$

by applying the mixing property and an array type central limit theorem. Similarly, one derives that if the NW estimator $\widehat{f}_1(x)$ is used instead, then

$$\lim_{t\to\infty} E(II')^2 = c(h)C_1 + O\left\{h^4(n - i_m + 1)^{-1/2}\right\}.$$

For the NW estimator, the term $\lim_{t\to\infty} E(I'II')$ was shown by Tjøstheim *et.al.* (1994) to be negligible by a standard U-statistic argument, which remains equally true for a local linear estimator.

Now we derive the term $\lim_{t\to\infty} E(I'^2)$. Using the result of the same paper by Härdle *et.al.* (1997)

$$\lim_{t\to\infty} E(I'^2) = E\int \left[\mu(x)^{-1}(n - i_m + 1)^{-1}\left\{1 + o_p(1)\right\} \sum_{i=i_m}^{n} K_h(X_i - x)\sigma^{1/2}(X_i)\xi_i\right]^2 \mu(x)dx$$

which becomes

$$E\int \mu(x)^{-2}(n - i_m + 1)^{-2}\left\{1 + o_p(1)\right\} \sum_{i=i_m}^{n} \left[K_h(X_i - x)\sigma^{1/2}(X_i)\right]^2 \mu(x)dx,$$

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

$$\int \mu(x)^{-2}(n - i_m + 1)^{-1}\left\{1 + o_p(1)\right\}\left[K_h(y - x)\sigma^{1/2}(y)\right]^2 \mu(x)\mu(y)dxdy$$

$$= \int \mu(x)^{-2}(n-i_m+1)^{-1}h^{-m}\left\{1+o_p(1)\right\}\left[K(u)\sigma^{1/2}(x+hu)\right]^2\mu(x)\mu(x+hu)dxdu$$

$$= \|K\|_2^{2m}(n-i_m+1)^{-1}h^{-m}\int\sigma(x)dx\left\{1+o_p(1)\right\} = b(h)B\left\{1+o_p(1)\right\},$$

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

We denote the fourth moment of the errors $\{\xi_t\}_{t=1}^{\infty}$ by m_4, which is finite as the ξ_t's have compact support by (A2). The following theorem extends Theorem 3.1.

Theorem 8.1 *Let* $Z = (n-i_m+1)^{-1}\sum_{i=i_m}^{n}\sigma(X_i)\xi_i^2 - A$, *then under assumptions (A1)-(A6), for* $a = 1, 2$, *as* $n \to \infty$

$$\widehat{A}_a = A + \left\{\|K\|_2^{2m} - 2K(0)^m\right\}(n-i_m+1)^{-1}h^{-m}B + C_a\sigma_K^4h^4/4$$

$$+Z + o\left\{h^4 + (n-i_m+1)^{-1}h^{-m}\right\} + o\left\{(n-i_m+1)^{-1/2}\right\} \tag{8.1}$$

with

$$\sqrt{n-i_m+1}Z \xrightarrow{D} N(0,\Sigma), \quad \Sigma = m_4\int\sigma^2(x)\mu(x)dx - A^2. \tag{8.2}$$

A similar result exists for the overfitting case

Theorem 8.2 *Under assumptions (A1)-(A6), for* $a = 1, 2$, *as* $n \to \infty$

$$\widehat{A}'_a = A + \left\{\|K\|_2^{2(m+l)} - 2K(0)^{(m+l)}\right\}(n-i^*_{m+l}+1)^{-1}h'^{-(m+l)}B + C'_a\sigma_K^4h'^4/4 +$$

$$Z' + o\left\{h'^4 + (n-i^*_{m+l}+1)^{-1}h'^{-(m+l)} + (n-i^*_{m+l}+1)^{-1/2}\right\} \tag{8.3}$$

where

$$Z' = (n-i^*_{m+l}+1)^{-1}\sum_{i=i^*_{m+l}}^{n}\sigma(X_i)\xi_i^2 - A, \quad \sqrt{n-i^*_{m+l}+1}Z' \xrightarrow{D} N(0,\Sigma). \tag{8.4}$$

Proof of Theorem 3.1 and Theorem 8.1. To prove (8.1), note that by the Central Limit Theorem

$$\sqrt{n-i_m+1}Z \xrightarrow{D} N(0,\Sigma), \quad \Sigma = m_4\int\sigma^2(x)\mu(x)dx - A^2.$$

We then note that by (3.1), $\widehat{A}_a$ is

$$(n-i_m+1)^{-1}\sum_{i=i_m}^{n}\left\{f(X_i) - \widehat{f}_a(X_i) + \sigma^{1/2}(X_i)\xi_i\right\}^2$$

$$= (n-i_m+1)^{-1}\sum_{i=i_m}^{n}\sigma(X_i)\xi_i^2 + (n-i_m+1)^{-1}\sum_{i=i_m}^{n}\left\{f(X_i) - \widehat{f}_a(X_i)\right\}^2$$

$$+(n-i_m+1)^{-1}\sum_{i=i_m}^{n}2\left\{f(X_i) - \widehat{f}_a(X_i)\right\}\sigma^{1/2}(X_i)\xi_i \tag{8.5}$$

in which the second term contributes to the $\|K\|_2^{2m}(n-i_m+1)^{-1}h^{-m}B + C_a\sigma_K^4h^4/4$ 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 *et.al.* (1994) for proof.

Proof of Theorem 3.3 and Theorem 8.2. To illustrate the kind of argument we use, note that if one writes $x' = (x, x'')$, where x represents the m-dimensional vector of correct lagged values and x'' the extra l lags, then

$$\int \sigma(x)\mu(x')dx' = \int \sigma(x)\mu(x,x'')dxdx''$$

$$= \int \sigma(x)dx \left\{ \int \mu(x,x'')dx'' \right\} = \int \sigma(x)\mu(x)dx = A.$$

Similar arguments give the expression for B and C'_a and therefore (3.5), (3.6) and (3.7).

The following is a refined version of Theorem 3.4.

Theorem 8.3 *Let $X'_t = (Y_{t-i'_1}, ..., Y_{t-i'_{m'}})^T$ be as in Theorem 3.4. Define the discrepancy between $f(x)$ and its conditional expectation on x' as*

$$f^{\perp}(x) = f(x) - \mu(x')^{-1} \int f(x', u'')\mu(x', u'')du'' = f(x) - E\left\{ f(x) \mid x' \right\} \tag{8.6}$$

and the squared projection error

$$C' = \int f^{\perp}(x)^2 \mu(x)dx = \int f(x)^2 \mu(x)dx - \int E^2\left\{ f(x) \mid x' \right\} \mu(x')dx'. \tag{8.7}$$

Then under assumptions (A1)-(A6), for

$$Z'_a = (n - i_{m'} + 1)^{-1} \sum_{i=i_{m'}}^{n} \left\{ f(X_i) - \widehat{f}_a(X'_i) \right\}^2 - C' \tag{8.8}$$

one has

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

where

$$\Sigma' = \int f^{\perp}(x)^4 \mu(x)dx - \left\{ \int f^{\perp}(x)^2 \mu(x)dx \right\}^2 + 4 \int f^{\perp}(x)^2 \sigma(x)\mu(x)dx \tag{8.9}$$

and also

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

Proof of Theorem 8.3 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 $\widehat{A}'_a$ as

$$(n - i_{m'} + 1)^{-1} \sum_{i=i_{m'}}^{n} \left\{ f(X_i) - \widehat{f}_a(X'_i) + \sigma^{1/2}(X_i)\xi_i \right\}^2$$

$$= (n - i_{m'} + 1)^{-1} \sum_{i=i_{m'}}^{n} \sigma(X_i)\xi_i^2 + (n - i_{m'} + 1)^{-1} \sum_{i=i_{m'}}^{n} \left\{ f(X_i) - \widehat{f}_a(X'_i) \right\}^2$$

$$+ (n - i_{m'} + 1)^{-1} \sum_{i=i_{m'}}^{n} 2 \left\{ f(X_i) - \widehat{f}_a(X'_i) \right\} \sigma^{1/2}(X_i)\xi_i.$$

It is straightforward to check that

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

Next

$$\widehat{f_1}(x') - f(x) =$$

$$\mu(x')^{-1}(n - i_{m'} + 1)^{-1} \{1 + o_p(1)\} \sum_{i=i_{m'}}^{n} K_h(X_i' - x') \left\{f(X_i) - f(x) + \sigma^{1/2}(X_i)\xi_i\right\} = T_1 + T_2$$

where

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

$$T_2 = \mu(x')^{-1}(n - i_{m'} + 1)^{-1} \{1 + o_p(1)\} \sum_{i=i_{m'}}^{n} K_h(X_i' - x')\sigma^{1/2}(X_i)\xi_i.$$

The variance of T_2 is calculated as

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

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

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

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

Similarly, the bias from T_1 is

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

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

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

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

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

One can derive a similar formula for $\widehat{f_2}(x') - f(x)$, thus we have

$$\widehat{f_a}(x') - f(x) = -f^{\perp}(x) + O_p(h'^2). \tag{8.10}$$

Because x' is a proper subvector of x, the true model, we know that $f^{\perp}(x) \neq 0$. Now

$$(n - i_{m'} + 1)^{-1} \sum_{i=i_{m'}}^{n} \left\{ f(X_i) - \widehat{f}_a(X_i') \right\}^2$$

has asymptotic mean

$$E \left\{ f(X_i) - \widehat{f}_a(X_i') \right\}^2$$

and asymptotic variance

$$(n - i_{m'} + 1)^{-1} E \left\{ f(X_i) - \widehat{f}_a(X_i') \right\}^4 - (n - i_{m'} + 1)^{-1} \left[E \left\{ f(X_i) - \widehat{f}_a(X_i') \right\}^2 \right]^2$$

which, by using (8.10), are

$$\int f^{\perp}(x)^2 \mu(x) dx + O(h'^2) = C' + O(h'^2)$$

and

$$(n - i_{m'} + 1)^{-1} \left[\int f^{\perp}(x)^4 \mu(x) dx - \left\{ \int f^{\perp}(x)^2 \mu(x) dx \right\}^2 \right]$$

respectively. Similarly

$$(n - i_{m'} + 1)^{-1} \sum_{i=i_{m'}}^{n} 2 \left\{ f(X_i) - \widehat{f}_a(X_i') \right\} \sigma^{1/2}(X_i) \xi_i$$

has mean 0 and asymptotic variance

$$(n - i_{m'} + 1)^{-1} 4 E \left\{ f(X_i) - \widehat{f}_a(X_i') \right\}^2 \sigma(X_i)$$

which, by using (8.10), is

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

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 (8.9) and (8.7) . Then we have

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

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

where

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

To prove Theorem 4.1, one needs to have an auxilliary result

Proposition 8.1 *Under assumptions (A1)-(A6), for $a = 1, 2$, as $n \to \infty$, if $h = \beta(n - i_m + 1)^{-1/(m+4)}$, and one defines*

$$Z_a = \widehat{A}_a - \left\{ \|K\|_2^{2m} - 2K(0)^m \right\} (n - i_m + 1)^{-1} h^{-m} B - C_a \sigma_K^4 h^4/4 - (n - i_m + 1)^{-1} \sum_{i=i_m}^{n} \sigma(X_i)\xi_i^2$$

then

$$(n - i_m + 1)^{(m+8)/(2m+8)} Z_a \xrightarrow{D} N(0, \Sigma_a) \tag{8.11}$$

where

$$\Sigma_1 = \sigma_K^4 \beta^4 \int \sigma(x)\mu(x) \left[\mathrm{Tr} \left\{ \nabla^2 f(x) \right\} + 2\nabla^T \mu(x) \nabla f(x)/\mu(x) \right]^2 dx +$$

$$4 \|K\|_2^{2m} \beta^{-m} \int \sigma^2(x) dx \tag{8.12}$$

$$\Sigma_2 = \sigma_K^4 \beta^4 \int \sigma(x)\mu(x) \, \mathrm{Tr} \left\{ \nabla^2 f(x) \right\}^2 dx + 4 \|K\|_2^{2m} \beta^{-m} \int \sigma^2(x) dx. \tag{8.13}$$

Proof of Proposition 8.1. Note that the variance of the third term in (8.5) is asymptotically

$$(n - i_m + 1)^{-1} E \int 4 \left\{ f(x) - \widehat{f}_a(x) \right\}^2 \sigma(x)\mu(x) dx$$

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

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

Meanwhile, the variance of the second term is asymptotically smaller than (here we take $a = 2$, the other case is similar)

$$(n - i_m + 1)^{-1} E \int \left\{ f(x) - \widehat{f}_a(x) \right\}^4 \mu(x) dx =$$

$$= \sigma_K^8 h^8 \int \mathrm{Tr} \left\{ \nabla^2 f(x) \right\}^4 \mu(x) dx/(16(n - i_m + 1)) +$$

$$6\sigma_K^4 h^4 \left\{ 1 + o_p(1) \right\} E \int \mathrm{Tr} \left\{ \nabla^2 f(x) \right\}^2 \sum_{i=i_m}^{n} K_h(X_i - x)^2 \sigma(X_i) \xi_i^2 \mu(x)/(4(n - i_m + 1)^3 \mu(x)^2) dx$$

$$+4\sigma_K^2 h^2 \left\{ 1 + o_p(1) \right\} E \int \mathrm{Tr} \left\{ \nabla^2 f(x) \right\} \sum_{i=i_m}^{n} K_h(X_i - x)^3 \sigma^{3/2}(X_i) \xi_i^3 \mu(x)/(2(n - i_m + 1)^4 \mu(x)^3) dx$$

$$+ \left\{ 1 + o_p(1) \right\} E \int \sum_{i=i_m}^{n} K_h(X_i - x)^4 \sigma^2(X_i) \xi_i^4 \mu(x)/(\mu(x)^4 (n - i_m + 1)^5) dx$$

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

$$= O_p \left\{ h^8/(n - i_m + 1) + h^4 (n - i_m + 1)^{-2} h^{-m} + h^2 (n - i_m + 1)^{-3} h^{-2m} + (n - i_m + 1)^{-4} h^{-3m} \right\}$$

$$= O_p \left(h^8 n^{-1} + h^8 n^{-1} + h^{10} n^{-1} + h^{12} n^{-1} \right) = o_p \left(h^4 n^{-1} \right),$$

which has finished the proof of the proposition.

Proof of Theorem 4.1. One similarly defines

$$Z'_a = \widehat{A}'_a - \left\{ \|K\|_2^{2(m+l)} - 2K(0)^{(m+l)} \right\} (n - i^*_{m+l} + 1)^{-1} h'^{-(m+l)} B$$

$$-C'_a \sigma_K^4 h'^4/4 - (n - i^*_{m+l} + 1)^{-1} \sum_{i=i^*_{m+l}}^{n} \sigma(X_i)\xi_i^2$$

then

$$(n - i^*_{m+l} + 1)^{(m+l+8)/(2m+2l+8)} Z'_a \xrightarrow{D} N(0, \Sigma'_a) \tag{8.14}$$

where

$$\Sigma'_1 = \sigma_K^4 \beta^4 \int \sigma(x)\mu(x') \left[\mathrm{Tr} \left\{ \nabla^2 f(x) \right\} + 2\nabla^T \mu(x') \nabla f(x)/\mu(x') \right]^2 dx' +$$

$$4 \|K\|_2^{2m+2l} \beta^{-(m+l)} \int \sigma^2(x) dx \tag{8.15}$$

$$\Sigma'_2 = \sigma_K^4 \beta^4 \int \sigma(x)\mu(x) \, \mathrm{Tr} \left\{ \nabla^2 f(x) \right\}^2 dx + 4 \|K\|_2^{2m+2l} \beta^{-(m+l)} \int \sigma^2(x) dx. \tag{8.16}$$

We then show that the difference between $(n - i_m + 1)^{-1} \sum_{i=i_m}^{n} \sigma(X_i)\xi_i^2$ and $(n - i^*_{m+l} + 1)^{-1} \sum_{i=i^*_{m+l}}^{n} \sigma(X_i)\xi_i^2$ is negligible. This difference is

$$(n-i_m+1)^{-1} \sum_{i=i_m}^{i^*_{m+l}} \sigma(X_i)\xi_i^2 + (i_m - i^*_{m+l})(n-i_m+1)^{-1}(n-i^*_{m+l}+1)^{-1} \sum_{i=i^*_{m+l}}^{n} \sigma(X_i)\xi_i^2 = O_p\left(\frac{1}{n}\right).$$

Thus

$$P\left[AFPE'_a < AFPE_a\right] =$$

$$P\left[Z'_a - Z_a < \text{constant}(n - i^*_{m+l} + 1)^{-4/(m+l+4)} - \text{constant } (n - i_m + 1)^{-4/(m+4)}\right].$$

Note that

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

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

which give

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

where

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

Proof of Theorem 4.4. Using arguments like before, one needs only to show that if x'' is a proper subvector of $x' = (x_{i_1}, ..., x_{i_{m'}})$, then

$$C'' > C'$$

where C' is as in (8.7) and

$$C'' = \int f(x)^2 \mu(x) dx - \int E^2 \left\{ f(x) \mid x'' \right\} \mu(x'') dx''$$

which yields

$$C'' - C' = \int \left[E\left\{ f(x) \mid x'' \right\} - E\left\{ f(x) \mid x' \right\} \right]^2 \mu(x') dx' > 0$$

as we assume that the true model includes all the lags $i_1, ..., i_m$.

References

[1] Akaike, H. (1969), Fitting Autoregressive Models for Prediction, *Annals of the Institute of Statistical Mathematics*, 21, 243-247.

[2] Akaike, H. (1971), Autoregressive Model Fitting for Control, *Annals of the Institute of Statistical Mathematics*, 23, 163-180.

[3] Ango Nze, P. (1992), Critères d'ergodicité de Quelques Modèles à Représentation Markovienne, *C.R. Acad. Sci. Paris*, sér I, 315, 1301-1304.

[4] Auestad, B. and Tjøstheim, D. (1990), Identification of Nonlinear Time Series: First Order Characterization and Order Determination, *Biometrika*, 77, 669-687.

[5] Chen, R. and Tsay, R. S. (1993), Nonlinear Additive ARX Models, *Journal of the American Statistical Association*, 88, 955-967.

[6] Cheng, B. and Tong, H. (1992), On Consist Non-parametric Order Determination and Chaos (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 54, 427-474.

[7] Davydov, Yu. A. (1973), Mixing Conditions for Markov Chains, *Theory of Probability and its Applications*, 18, 312-328.

[8] 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.

[9] Engle, R. (1982), Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation, *Econometrica*, 50, 987-1008.

[10] Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman and Hall.

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

[12] Härdle, W., Tsybakov, A. B. and Yang, L. (1997), Nonparametric Vector Autoregression, *Journal of Statistical Planning and Inference*, to appear.

[13] Lütkepohl, H. (1991), *Introduction to Multiple Time Series Analysis*, Springer-Verlag, Heidelberg.

[14] Nummelin, E. and Tuominen, P. (1982), Geometric Ergodicity of Harris–recurrent Markov Chains with Application to Renewal Theory, *Stochastic Processes and their Applications*, 12, 187-202.

[15] Ruppert, D. and Wand, M. P. (1994), Multivariate Locally Weighted Least Squares Regression, *Annals of Statistics*, 22, 1346-1370.

[16] Saulis, L. and Statulevičius, V. A. (1991), *Limit Theorems for Large Deviations*, Kluwer.

[17] Tjøstheim, D. (1994), Non-linear Time Series Analysis: A Selective Review, *Scandinavian Journal of Statistics*, 21, 97-130.

[18] Tjøstheim, D. and Auestad, B. (1994), Nonparametric Identification of Nonlinear Time Series: Selecting Significant Lags, *Journal of the American Statistical Association*, 89, 1410-1419.

[19] 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.

[20] Vieu, P. (1994), Order Choice in Nonlinear Autoregressive Models, *Statistics*, 24, 1-22.

[21] Wand, M. P. and Jones, M. C. (1995), *Kernel Smoothing*, Chapman and Hall, London.

[22] Yang, L. and Härdle, W. (1997), Nonparametric Autoregression with Multiplicative Volatility and Additive Mean, revised for *Journal of Time Series Analysis*.

[23] Yang, L. and Tschernig, R. (1997), Fast Estimation of Multivariate Second Derivatives, unpublished manuscript.

[24] Yao, Q. and Tong, H. (1994), On Subset Selection in Non-parametric Stochastic Regression, *Statistica Sinica*, 4, 51-70.

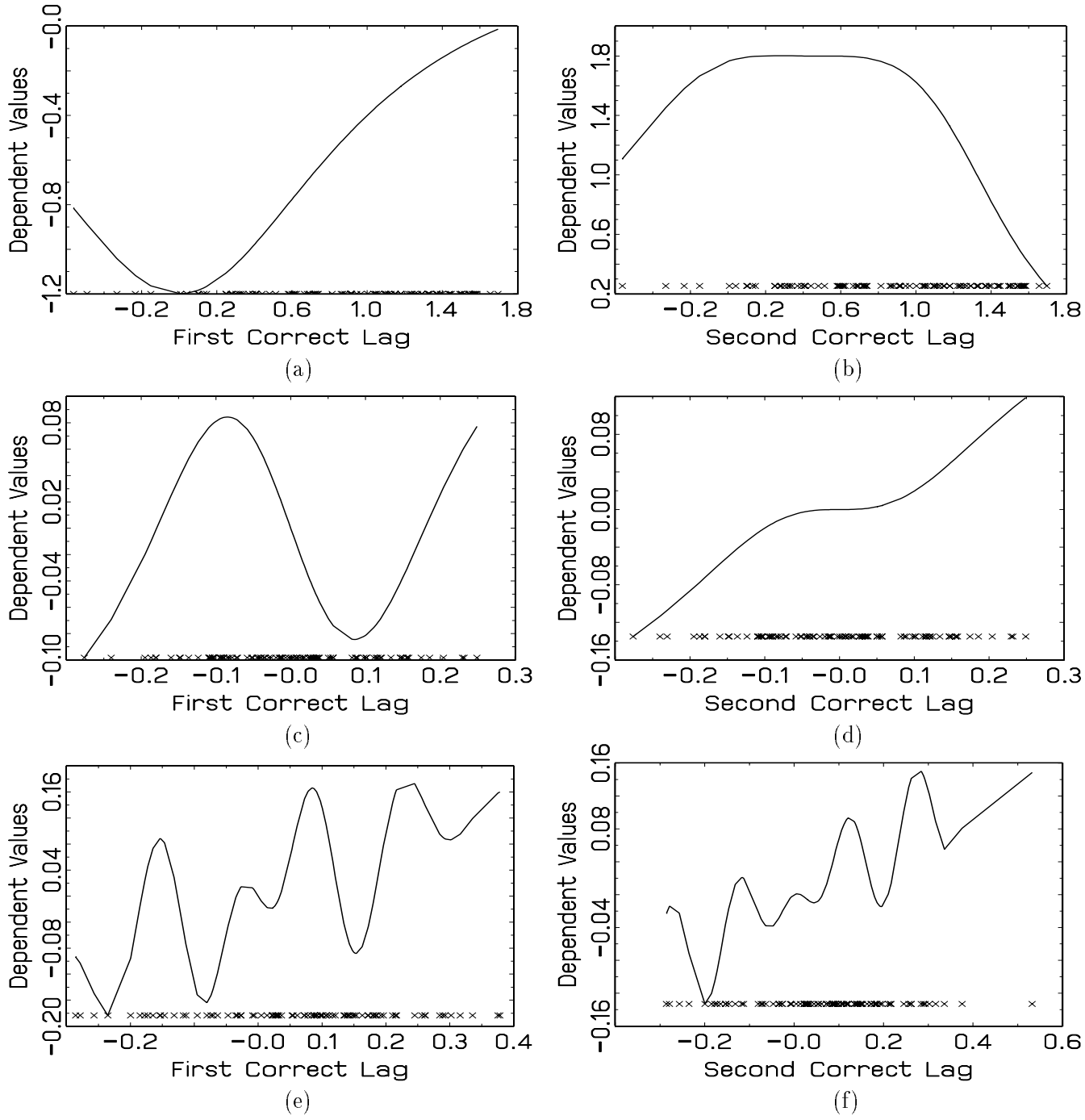

Figure 1: Additive nonlinear functions used in the Monte-Carlo experiments. The stars indicate one realization of the empirical distribution of 100 observations: (a) Lag 1 in the NLAR1 process; (b) Lag 2 in the NLAR1 process; (c) Lag 6 in the NLAR2 process; (d) Lag 10 in the NLAR2 process; (e) Lag 6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

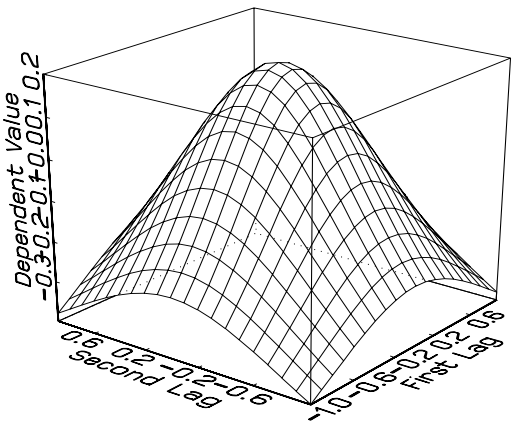

Figure 2: Regression Function of the NLAR4 Process

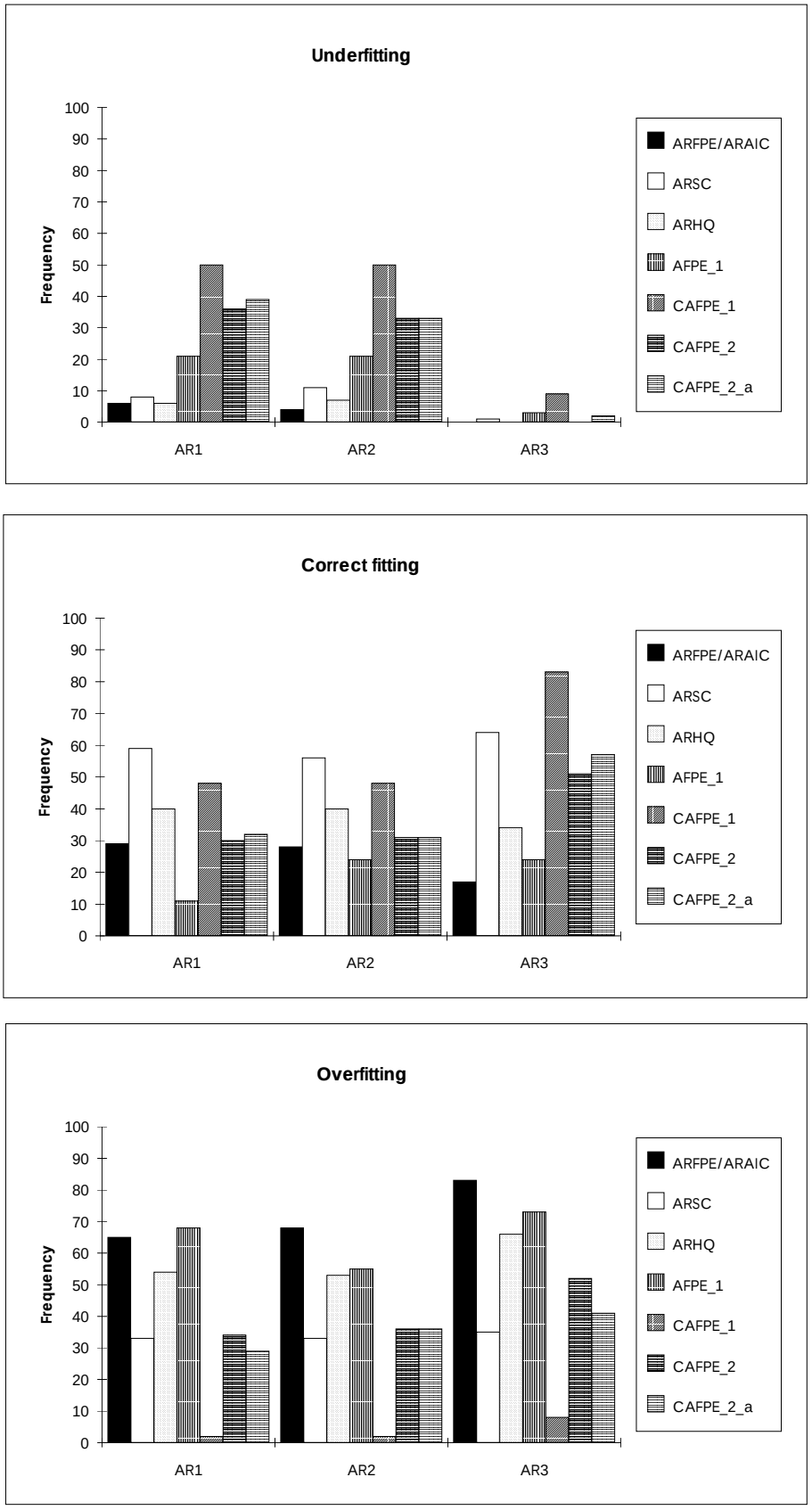

Figure 3: Empirical frequencies of underfitting, correct fitting and overfitting of linear AR models

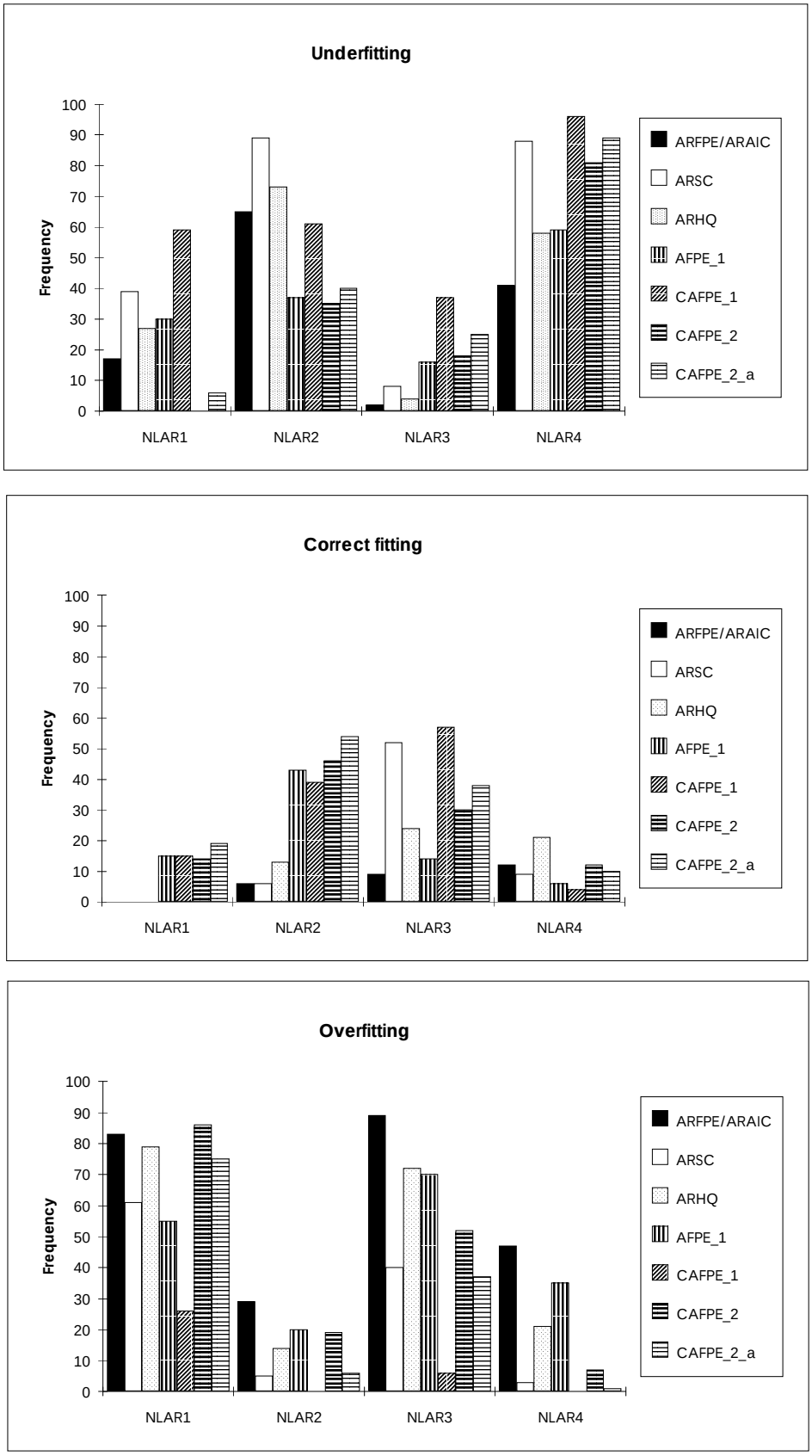

Figure 4: Empirical frequencies of underfitting, correct fitting and overfitting of nonlinear AR models

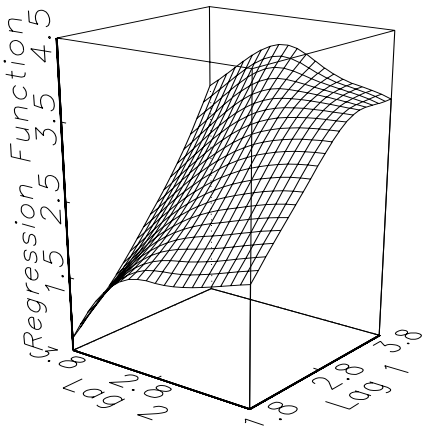

Figure 5: Regression Function for logged lynx data obtained with the local linear estimator and bandwidth 0.335

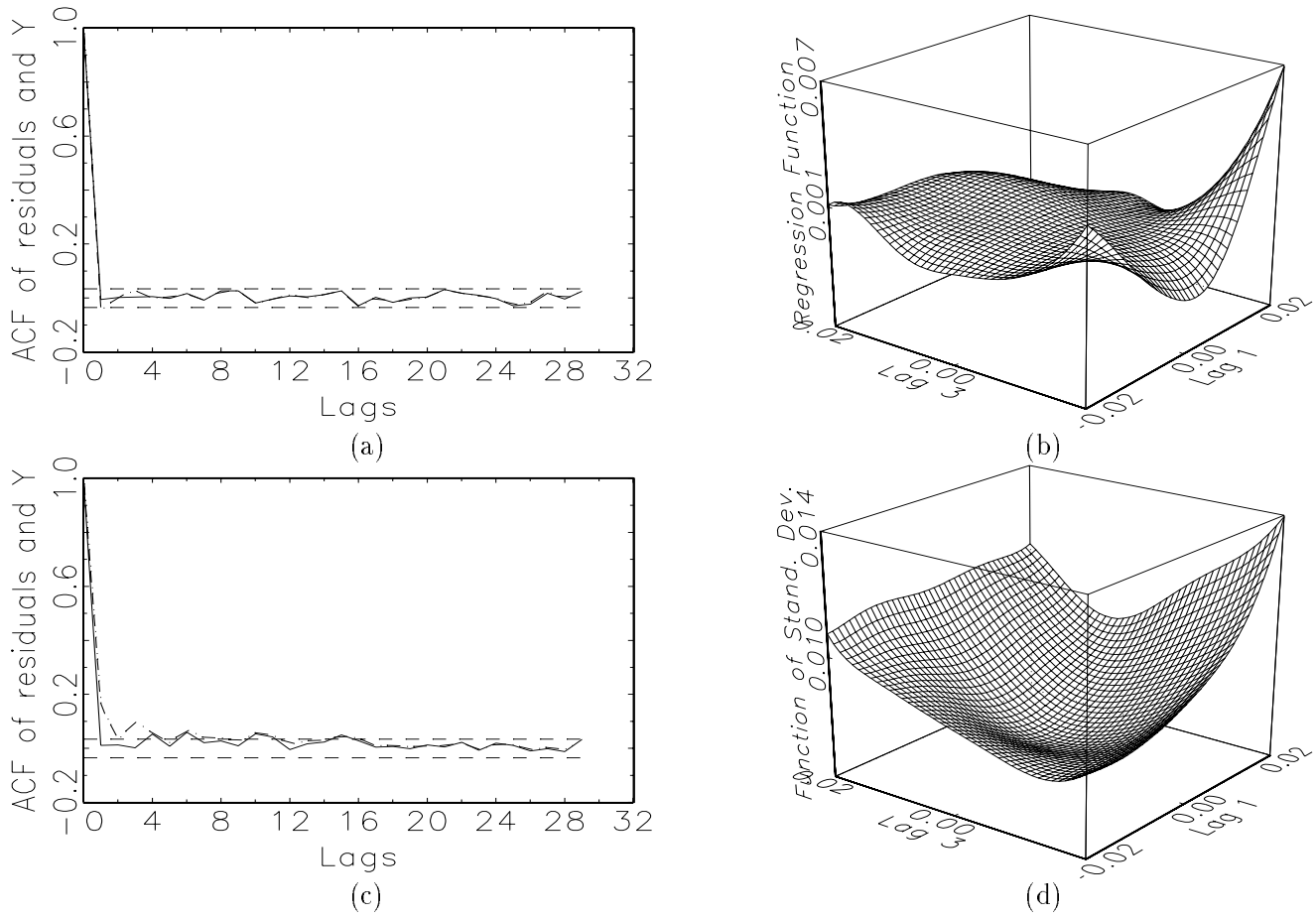

Figure 6: Local linear estimates for daily DM/US-Dollar series: (a) ACF of estimated residuals (solid line) and of observations (dashed line); (b) Regression Function; (c) ACF of squared estimated residuals (solid line) and of squared observations (dashed line); (d) Conditional Standard Deviation