Journal of the Korean Statistical Society (2001), 30: 3, pp 387–406

A Note on Adaptive Estimation for Nonlinear Time Series Models

Sahmyeong Kim [1]

ABSTRACT

Adaptive estimators for a class of nonlinear time series models has been proposed by several authors. Koul and Schick (1997) proposed the adaptive estimators without sample splitting for location-type time series models. They also showed by simulation that the adaptive estimators without sample splitting have smaller mean squared errors than those of the adaptive estimators with sample splitting. The present paper generalizes the result in case of location-scale type nonlinear time series models by simulation.

Keywords: Adaptive estimators, Sample splitting, Non sample splitting, Nonlinear time series models, Local Asymptotic Normality.

1. Introduction

The problem of obtaining adaptive estimators in the presence of infinite-dimensional nuisance parameters has been studied by numerous researchers. See the book by Bickel et al. (1993) for mainly the iid case and the regression models. Recent work on adaptive estimation for dependent data includes the following references. Kreiss (1987a,b) has discussed the adaptive estimators in ARMA models. Engle and Gonzalez-Rivera (1991) have discussed an efficient estimator which is based on a nonparametrically estimated density. Linton (1993) suggested a reparameterization of the parameters of interest in the model studied by Engle and Gonzalez-Rivera (1991). Jeganathan (1995) proposed adaptive estimators for more general time series models. Sample splitting technique has been used by many authors including Bickel (1982) to simplify the proofs for the existence of adaptive estimators. Schick (1987) introduced an adaptive estimator in the iid case without sample splitting under slightly stronger conditions than those requiring sample splitting. Koul and Schick (1997) proposed adaptive estimators which are based on non-splitting of the sample and these estimators have better

[1] Assistant Professor, Chung-Ang University, Seoul, Korea

performance (i.e. smaller mean squared errors) than the estimators under the sample splitting method. However, Koul and Schick (1997) have studied the adaptiveness of the estimators in a class of location type nonlinear time series such as threshold autoregressive models.

On the other hand, Shin and So(1999) studied the adaptive maximum likelihood estimation under the nonstationary $AR(p)$ models which do not satisfy the LAN property. In this paper, we present the local asymptotic normality (LAN) property for the time series models under investigation in Section 2. Section 3 presents the adaptive estimator based on sample splitting which was proposed by Drost et al. (1997). In Section 4, we propose adaptive estimators without sample splitting for a class of location-scale nonlinear time series models including ARCH models. Section 5 presents some simulation results which compare the mean squared errors of the adaptive estimators of Drost et al. (1997) and the proposed estimators. Finally, concluding remarks are given in Section 6.

2. Local Asymptotic Normality

In this section we review the local asymptotic normality (LAN) of the nonlinear time series models. In general, the LAN property provides the asymptotic normality of the log-likelihood ratio of two contiguous probability measure sequences. See Akritas and Johnson (1980), Kreiss (1987a), Hwang and Basawa (1993), Jeganathan (1995) and Drost et al. (1997).

Consider the time series model

$$X_t = m_t(\theta) + \sigma_t(\theta)\epsilon_t, \ t = 1, ..., n, \tag{2.1}$$

where $\{\epsilon_t\}$, $t = 1, ..., n$, is a sequence of iid random errors with $E(\epsilon_t) = 0$ and $Var(\epsilon_t) = \sigma^2$, $m_t(\theta) = E(X_t|\mathcal{F}_{t-1})$, $\sigma_t^2(\theta) = Var(X_t|\mathcal{F}_{t-1})$ where $\mathcal{F}_{t-1}$ is the σ-field generated by $X_{t-1}, ..., X_{t-p}$, $p \geq 1$. Let g be a density function of the errors. Define

$$f(X_t;\theta) = log\{g(\epsilon_t(\theta))\} - log\ \sigma_t(\theta). \tag{2.2}$$

Let

$$S_t(\theta) = \frac{df(X_t;\theta)}{d\theta} = H_t(\theta)\psi_t(\theta), \ t = 1, ..., n, \tag{2.3}$$

where

$$H_t(\theta) = -\frac{1}{\sigma_t(\theta)}\{m_t'(\theta), \sigma_t'(\theta)\},$$

$$\psi_t(\theta) = \{\frac{g'}{g}(\epsilon_t(\theta)), 1 + \epsilon_t(\theta)\frac{g'}{g}(\epsilon_t(\theta))\},$$

and prime means the derivative of the function. Consider the ratio of two likelihood functions

$$\Lambda(\theta, \theta_n) = \sum_{t=1}^{n} log\{\frac{g(\epsilon_t(\theta_n))}{\sigma_t(\theta_n)} / \frac{g(\epsilon_t(\theta))}{\sigma_t(\theta)}\}, \tag{2.4}$$

where

$$\theta_n = \theta + h/\sqrt{n} \text{ and } h \varepsilon R^p.$$

The following theorem establishes the LAN property for the location-scale type time series models.

Theorem 2.1.

Assume the following regularity conditions.

(C.1) $\{X_t\}$, $t = 1, ..., n$ *is stationary and ergodic.*

(C.2) $\int \{\frac{g'(x)}{g(x)}\}^2 g(x)dx < \infty$ *and* $\int \{1 + \frac{g'(x)}{g(x)}\}^2 g(x)dx < \infty$

(C.3) $log\{\frac{g(X_0;\theta_n)/\sigma_1(\theta_n)}{g(X_0;\theta)/\sigma_1(\theta)}\} \xrightarrow{p} 0$ *as* $n- > \infty$.

(C.4) $\sum_{t=1}^{n} |||H_t(\theta)|||^2 I_{[\frac{1}{\sqrt{n}}|H_t(\theta)|>\varepsilon]} \xrightarrow{p} 0$ *as* $n \to \infty$.

(C.5) There exists a $p \times p$ *nonsingular matrix* V *such that* $\frac{1}{n}\sum_{t=1}^{n} H_t(\theta)H_t(\theta)^T \xrightarrow{p} V$ *as* $n \to \infty$.

Then under the conditions (C.1)-(C.5), we have

$$\Lambda(\theta, \theta_n) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} h^T S_t(\theta) - \frac{1}{2} h^T W h + o_p(1), \tag{2.5}$$

and

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} S_t(\theta) \xrightarrow{d} N(0, W), \tag{2.6}$$

where

$$W = E_\theta\{H_t(\theta)\psi_t(\theta)\psi_t(\theta)^T H_t(\theta)^T\}.$$

Proof. See Koul and Schick (1997) or Drost et al. (1997).

Basically Koul and Schick (1997) and Drost et al. (1997) have almost the same regularity conditions for the LAN property. However Drost et al. (1997) considered the LAN property for the location-scale time series models whereas Koul and Schick (1997) only considered a class of location type nonlinear time series models. Koul and Schick (1997) established efficient estimation in general to guarantee the existence of the adaptiveness while Jeganathan (1995) assumed the symmetry of the error density and Drost et al. (1997) estimated only the

component of the parameter of interest which is adaptively estimable. In this paper, we also assume that the error densities are symmetric about 0 because adaptive estimation is not always possible.

3. Adaptive Estimation With Sample Splitting

In this section we will present adaptive estimators using discretized and $\sqrt{n}$-consistent estimators for initial estimators and sample splitting method. The concept of discretization was developed by Le Cam (1960) and the technique has been used to construct adaptive estimators in the semiparameter set-up.

Definition 3.1.

For any sequence of estimators $\{\theta_n\}$, $n \geq 1$, $\{\hat{\theta}_n\}$ *is said to be a sequence of discretized estimators if* $\hat{\theta}_n$ *is the nearest point of* $\{\theta; \theta = \frac{c}{\sqrt{n}}(N^p)\}$ *from* θ_n *where* N *is the set of positive integers and* θ *is a* $p \times 1$ *vector of parameters.*

The technique of discretization enables us to treat $\hat{\theta}_n$ as non random and to establish the validity of the one-step type estimators without further additional conditions.

The sample splitting method was developed by Bickel (1982) and Schick (1986) in the iid cases. Klaassen (1987) showed that discretization may be avoided by additional sample splitting. We will follow Schick's (1986) method of sample splitting. Suppose that $\{\hat{\theta}_n\}$ is a sequence of discretized estimators. Let $\{\hat{\epsilon}_{nt}\}$ be a sequence of residuals for a given model $\hat{\epsilon}_{nt} = \epsilon_t(\hat{\theta}_n)$, $t = 1, ..., n$. Then we divide the set of residuals into two samples. We then estimate $\frac{g'}{g}(\epsilon_t(\theta))$, where g is a density of the errors, using only the second part of the residuals, i.e., $\hat{e}_2 = (\hat{\epsilon}_{n[\frac{n}{2}]+1}, ..., \hat{\epsilon}_{nn})$ if we want to calculate these ratios at $\hat{\epsilon}_{nt}$ with $t \leq [\frac{n}{2}]$. Use only the residuals $\hat{e}_1 = (\hat{\epsilon}_{n1}, ..., \hat{\epsilon}_{n[\frac{n}{2}]})$ if we want to calculate these ratios at $\hat{\epsilon}_{nt}$ with $t > [\frac{n}{2}]$. In order to derive the adaptive estimators with sample splitting, we need to impose some regularity conditions which were proposed by Schick (1986) for the iid case and by Drost et al. (1997) for a general class of time series models. The next lemma shows the existence of the consistent estimators of the information matrix $W(\theta)$ in (2.4).

Lemma 3.1. *(Drost et al., 1997) Under the conditions A-G in Drost et al. (1997), suppose* $\{\tilde{\theta}_n\}$ *is a sequence of discretized and* $\sqrt{n}$*-consistent estimators of* θ*. Then we have*

$$\begin{aligned}\hat{W}_n &= \frac{1}{n}\sum_{t=1}^{\alpha} H_t(\tilde{\theta}_n)\hat{\psi}_{2t}(\tilde{\theta}_n)\hat{\psi}_{2t}(\tilde{\theta}_n)^T H_t(\tilde{\theta}_n)^T \\ &+ \frac{1}{n}\sum_{t=\alpha+1}^{n} H_t(\tilde{\theta}_n)\hat{\psi}_{1t}(\tilde{\theta}_n)\hat{\psi}_{1t}(\tilde{\theta}_n)^T H_t(\tilde{\theta}_n)^T \qquad (3.1) \\ &\xrightarrow{p} W(\theta),\end{aligned}$$

where $H_t(\tilde{\theta}_n)$ *is defined in (2.3),*

$$\begin{aligned}\hat{\psi}_{1t}(\tilde{\theta}_n) &= \hat{\psi}_t(\tilde{\epsilon}_{n1}, \tilde{\epsilon}_{n2}, ..., \tilde{\epsilon}_{\alpha}), \\ \hat{\psi}_{2t}(\tilde{\theta}_n) &= \hat{\psi}_t(\tilde{\epsilon}_{\alpha+1}, \tilde{\epsilon}_{\alpha+2}, ..., \tilde{\epsilon}_n), \tilde{\epsilon}_{nt} = \epsilon_t(\tilde{\theta}_n),\end{aligned}$$

$\hat{\psi}_t$ *is a kernel density estimator of* ψ *in (2.3) and* α *is* $[\frac{n}{2}]$, $t = 1, ..., n$.

The next theorem guarantees the existence of adaptive estimators for a class of general time series models.

Theorem 3.1. *(Drost et al., 1997) Under the assumptions of Lemma 3.1, let* $\tilde{\theta}_n$ *be a discretized and* $\sqrt{n}$*-consistent estimator of* θ*. Define* $\hat{\theta}_n$ *by*

$$\hat{\theta}_n = \tilde{\theta}_n + \hat{W}_n^{-1}\{\sum_{t=1}^{\alpha} H_t(\tilde{\theta}_n)\hat{\psi}_{2t}(\tilde{\theta}_n) + \sum_{t=\alpha+1}^{n} H_t(\tilde{\theta}_n)\hat{\psi}_{1t}(\tilde{\theta}_n)\}.$$

Then

$$\sqrt{n}(\hat{\theta}_n - \theta) - \frac{1}{\sqrt{n}} W(\theta)^{-1}\sum_{t=1}^{n} H_t(\theta)\psi_t(\theta) \xrightarrow{p} 0. \qquad (3.2)$$

Gonzalez-Rivera (1991) proposed an efficient estimator and compared it with the quasi-maximum likelihood estimator (QMLE) for the GARCH models. However, they also showed by simulation that their estimator was not actually efficient compared to QMLE. Drost and Klaassen (1997) showed that the adaptive estimators under sample splitting are efficient over the QMLE in the GARCH models. Their simulation study illustrated that the adaptive estimators worked quite well when the sample size is large, i.e., 1000 or more.

4. Adaptive Estimation Without Sample Splitting

In this section we will propose the adaptive estimators of location-scale time series models without sample splitting when the error distribution is symmetric

about 0. Jeganathan (1995) and Koul and Schick (1997) established the adaptive estimators for the location-type time series models with symmetry of the error distribution. We expect that the adaptive estimators without sample splitting will have better performance for small sample sizes. Koul and Schick (1997) pointed out that Jeganathan (1995) had an error in assuming $H_t(\tilde{\theta}_n)\psi_t(\tilde{\theta}_n)$ to be a martingale difference. Koul and Schick (1997) assumed instead that a truncation of $H_t(\tilde{\theta}_n)$ has a martingale property. We will adapt the truncation form for the next theorem. First, we define a function $G_n(x)$ from R^p to R^p as follow.

$$G_n(x) = xI_{[\|x\| \le c_n]} + c_n I_{[x > c_n]} - c_n I_{[x < -c_n]}, \tag{4.1}$$

where $x \in R^p$ and $\{c_n\}$ is a sequence of real numbers converging to infinity.

Next we define the kernel density $\hat{\psi}$ to estimate

$$\psi(x) = (\frac{g'}{g}(x), 1 + x\frac{g'}{g}(x)) = (\psi_1(x), \psi_2(x)), \tag{4.2}$$

where g is an error density. $\hat{\psi}_1$ is defined as

$$\hat{\psi}_1(x) = \frac{g_n'(x, \epsilon_{n1}, ..., \epsilon_{nn}) - g_n'(-x, \epsilon_{n1}, ..., \epsilon_{nn})}{a_n + g_n(x, \epsilon_{n1}, ..., \epsilon_{nn}) + g_n(-x, \epsilon_{n1}, ..., \epsilon_{nn})},$$

where

$$g_n(x, \epsilon_{n1}, ..., \epsilon_{nn}) = \frac{1}{na_n}\sum_{t=1}^{n} K(\frac{x - \epsilon_{nt}}{a_n}),$$

$$g_n'(x, \epsilon_{n1}, ..., \epsilon_{nn}) = \frac{1}{na_n^2}\sum_{t=1}^{n} K'(\frac{x - \epsilon_{nt}}{a_n}), \tag{4.3}$$

K is a kernel, i.e., $\int K = 1$, a_n is a bandwidth and

$$\epsilon_{nt} = \epsilon_t(\tilde{\theta}_n), \ t = 1, ..., n.$$

Similarly,

$$\hat{\psi}_2(x) = 1 + x\hat{\psi}_1(x).$$

Define an estimator as follow.

$$\hat{\theta}_n = \tilde{\theta}_n + \frac{1}{n}\hat{W}^{-1}\sum_{t=1}^{n} \bar{H}_{nt}\hat{\psi}(\epsilon_{nt}), \tag{4.4}$$

where $\tilde{\theta}_n$ is a discretized and $\sqrt{n}$-consistent estimator of θ, $\bar{H}_{nt} = G_n(H_t(\tilde{\theta}_n))$, and

$$\hat{W} = \frac{1}{n}\sum_{t=1}^{n} \bar{H}_{nt}\hat{\psi}(\epsilon_{nt})\hat{\psi}(\epsilon_{nt})^T\bar{H}_{nt}^T.$$

Theorem 4.1.

Assume the following regularity conditions.

(C.6) $C^T\sqrt{n}\int\hat{\psi}(\varepsilon_1,...,\varepsilon_n,x)g(x)dx \xrightarrow{p} 0$, *where* C *is an orthogonal matrix of full rank and* g *is a density of* $\varepsilon_1,...,\varepsilon_n$.

(C.7) $\frac{1}{n}\sum_{t=1}^{n} H_t(\theta) \xrightarrow{p} H$, *where* H *is square integrable.*

(C.8) $\int\{\hat{\psi}(\varepsilon_1,...,\varepsilon_n,x)-\psi(x)\}^2 g(x)dx \xrightarrow{p} 0$ as $n \xrightarrow{p} \infty$.

(C.9) For every θ_n *such that* $\theta_n-\theta = O(n^{-\frac{1}{2}})$ *and every sequence* $\{C_n\}$ *where* $Cn \xrightarrow{p} \infty$, $\frac{1}{n}\sum_{t=1}^{n}\|H_t(\theta)\|^2 I[H_t(\theta) > C_n] \xrightarrow{p} 0$ as $n \to \infty$.

(C.10) $n^{-1}a_n^{-3}b_n^{-1}C_n^{-2} \to 0$ as $n \to \infty$.

Then under the conditions (C.1) - (C.10), for a sequence of discretized $\sqrt{n}$*-consistent estimator* $\{\theta_n\}$ *of* θ, *we have*

$$\sqrt{n}(\hat{\theta}_n-\theta)-\frac{1}{\sqrt{n}}W^{-1}\sum_{t=1}^{n}H_t(\theta)\psi(\varepsilon_t(\theta)) \xrightarrow{p} 0 \tag{4.5}$$

Proof. The proof is referred to the appendix.

5. Simulation Results

We present some small sample results by simulation. We simulated TAR(1) model and ARCH (1) model with sample sizes n=100 and 200. We considered two symmetric error densities, the double exponential ($g(x) = \frac{1}{2}\exp(-|x|)$ and the mixture normal ($g(x) = \frac{1}{2}N(-2,1) + \frac{1}{2}N(2,1)$). For our simulation we use the kernel $K(x) = \frac{3}{4\sqrt{5}}(1-\frac{x^2}{5})I_{[|x|\le\sqrt{5}]}$ and $a_n = 0.3, 0.4, 0.5, 0.6, 0.7$. Drost and Klaassen(1997) recommended that reasonable choices of the Bandwidth may be between 0.25 and 0.75. They also pointed out that the values between 0.25 and 0.75 for the kernel density do not affect the conclusions below. We use the conditional least-squares estimators as the initial estimators of the parameter of interest.

For the TAR(1) model, the adaptive estimators of θ_1 and θ_2 are defined as follows.

$$\begin{aligned}
\hat{\theta}_1 &= \tilde{\theta}_{1n} + \frac{\sum_{t=1}^{n} X_{t-1}^{+}\hat{\psi}_1(\epsilon_{nt})}{\sum_{t=1}^{n}(X_{t-1}^{+})^2\hat{\psi}_1^2(\epsilon_{nt})}, \\
\hat{\theta}_2 &= \tilde{\theta}_{2n} + \frac{\sum_{t=1}^{n} X_{t-1}^{-}\hat{\psi}_1(\epsilon_{nt})}{\sum_{t=1}^{n}(X_{t-1}^{-})^2\hat{\psi}_1^2(\epsilon_{nt})},
\end{aligned} \tag{5.1}$$

where $\tilde{\theta}_n = (\tilde{\theta}_{1n}, \tilde{\theta}_{2n})^T$ is the conditional least-squares estimator of $\theta = (\theta_1, \theta_2)^T$, $X_{t-1}^+ = X_{t-1} I_{(X_{t-1} \geq 0)}$, $\epsilon_{nt} = \epsilon_t(\tilde{\theta}_n)$, $X_{t-1}^- = X_{t-1} I_{(X_{t-1} < 0)}$ and $\hat{\psi}_1$ is defined in (4.4). Then we calculate two types of the adaptive estimators, i.e., using sample splitting and non-sample splitting. In order to calculate the adaptive estimators with sample splitting in the TAR(1) model, we estimate $\hat{\psi}_1$ of (5.1) by the sample splitting method in Section 3. Whereas the adaptive estimators without sample splitting for the TAR(1) model can be calculated by estimating $\hat{\psi}_1$ using the whole sample.

For the ARCH(1) model, when the model is defined as $X_t = \theta X_{t-1} + \sigma_t(\theta)\epsilon_t(\theta)$, where $\sigma_t^2(\theta) = \alpha_0 + \alpha_1(X_{t-1} - \theta X_{t-2})^2$, the adaptive estimator is defined as

$$\hat{\theta} = \tilde{\theta}_n + \frac{\sum_{t=1}^{n} [\{X_{t-1}\hat{\psi}_1(\tilde{\theta}_n) + X_{t-1}^2(1+\epsilon_{nt})\hat{\psi}_1(\tilde{\theta}_n)\}/\sigma_t(\tilde{\theta}_n)]}{\sum_{t=1}^{n} [\{X_{t_1}\hat{\psi}_1(\tilde{\theta}_n) + X_{t-1}^2(1+\epsilon_{nt})\hat{\psi}_1(\tilde{\theta}_n)\}/\sigma_t(\tilde{\theta}_n)]^2}, \tag{5.2}$$

where $\tilde{\theta}_n$ is the conditional least-squares estimator of θ, and

$$\sigma_t^2(\tilde{\theta}_n) = \hat{\alpha}_0 + \hat{\alpha}_1(X_{t-1} - \tilde{\theta}_n X_{t-2})^2, \ t = 1, ..., n,$$

$\hat{\alpha}_0 = \alpha_0(\tilde{\theta}_n)$, $\hat{\alpha}_1 = \alpha_1(\tilde{\theta}_n)$ are the least-squares estimators of α_0 and α_1.

We calculate the adaptive estimators of θ where ψ_1 is estimated by the sample splitting technique and also estimated by the whole sample.

The number of simulations is 1000 and we have $\theta = (0.3, -.3)^T$ for the TAR(1) model. We also have $\theta = 0.7$ with $\alpha_0 = 0.7$ and $\alpha_1 = 0.1$ for the ARCH(1) model. For the mixture normal error distribution, we can see that the sample variance and sample mean squared error (MSE) of the adaptive estimators without sample splitting are smaller than those of the adaptive estimators with sample splitting for both the TAR(1) and ARCH(1) models for the sample sizes (n=100 and 200). Similar results can be seen for the double exponential error distribution. We also note that for the TAR(1) models, the values of sample variance, sample MSE and bias do not vary significantly when the value of bandwidth changes in the kernel density estimator. Also the choice of the kernel density does not seem to affect these performance characteristics. In the ARCH(1) model, however, we note that the sample MSE values get larger when the bandwidth is increased, the error distribution is mixed normal and non-sample splitting method is used.

6. Concluding Remarks

The adaptive estimator proposed in this chapter is an extension of that of Koul and Schick (1997). Simulation results for AR(1) model comparing the two types of the adaptive estimators were presented by Koul and Schick (1997). Their simulation results showed that the adaptive estimators without sample splitting have better performance (smaller MSEs) than the adaptive estimators with sample splitting in the small sample sizes. Our simulation results also show that the adaptive estimators without sample splitting are better than the other adaptive esitmators with sample splitting for the TAR(1) and the ARCH(1) model. Therefore our simulation results are an extension of Koul and Schick (1997) for more general time series models.

Drost and Klaassen (1997) showed that the adaptive estimators with sample splitting performed very well when the sample size is 1000 or more. We expect that the performance of the two type of the estimators (with or without sample splitting) would be the same when the sample size is considerably large. Although we do not report the simulation results, we found that the MSEs of the adaptive estimators with sample splitting are almost the same as those of the adaptive estimators without sample splitting when the sample size is 900 under the double exponential error density for both the TAR(1) and ARCH(1) models.

Table 1: Bias, Sample Variance and Sample MSE in TAR(1) model

$n = 100$, mixed normal, $\theta_1 = 0.3$ and $\theta_2 = -.3$.

(Sample Splitting)

	Band Width	Bias	Sample Variance	Sample MSE
$\theta_1 = 0.3$	0.3	0.0194	0.0369	0.0373
	0.4	0.0195	0.0361	0.0364
	0.5	0.0197	0.0350	0.0353
	0.6	0.0199	0.0336	0.0340
	0.7	0.0202	0.0322	0.0325
$\theta_2 = -.3$	0.3	0.0159	0.0330	0.0333
	0.4	0.0162	0.0323	0.0325
	0.5	0.01643	0.0313	0.0315
	0.6	0.0168	0.0301	0.0304
	0.7	0.0172	0.0289	0.0291

Table 2: Bias, Sample Variance and Sample MSE in TAR(1) model

$n = 100$, mixed normal, $\theta_1 = 0.3$ and $\theta_2 = -.3$.

(Non Sample Splitting)

	Band Width	Bias	Sample Variance	Sample MSE
$\theta_1 = 0.3$	0.3	0.0182	0.0199	0.0202
	0.4	0.0181	0.0199	0.0202
	0.5	0.0231	0.0215	0.0220
	0.6	0.0176	0.0212	0.0215
	0.7	0.0203	0.023	0.0218
$\theta_2 = -.3$	0.3	0.0168	0.0211	0.0214
	0.4	0.0167	0.0211	0.0213
	0.5	0.0212	0.0198	0.0203
	0.6	0.0246	0.0208	0.0215
	0.7	0.0219	0.0196	0.0201

Table 3: Bias, Sample Variance and Sample MSE in TAR(1) model

$n = 200$, mixed normal, $\theta_1 = 0.3$ and $\theta_2 = -.3$.

(Sample Splitting)

	Band Width	Bias	Sample Variance	Sample MSE
$\theta_1 = 0.3$	0.3	0.0083	0.0220	0.0220
	0.4	0.0084	0.0211	0.0211
	0.5	0.0114	0.0177	0.0178
	0.6	0.0088	0.0190	0.0191
	0.7	0.0091	0.0178	0.0179
$\theta_2 = -.3$	0.3	0.0110	0.0220	0.0220
	0.4	0.0102	0.0212	0.0213
	0.5	0.0007	0.0182	0.0182
	0.6	0.0103	0.0192	0.0193
	0.7	0.0104	0.0179	0.0181

Table 4: Bias, Sample Variance and Sample MSE in TAR(1) model

$n = 200$, mixed normal, $\theta_1 = 0.3$ and $\theta_2 = -.3$.

(Non Sample Splitting)

	Band Width	Bias	Sample Variance	Sample MSE
$\theta_1 = 0.3$	0.3	0.0101	0.0099	0.0100
	0.4	0.0124	0.0098	0.0099
	0.5	0.0124	0.0098	0.0099
	0.6	0.0100	0.0099	0.0100
	0.7	0.0100	0.0099	0.0101
$\theta_2 = -.3$	0.3	0.0066	0.0098	0.0098
	0.4	0.0110	0.0092	0.0094
	0.5	0.0108	0.0092	0.0094
	0.6	0.0066	0.0097	0.0098
	0.7	0.066	0.0097	0.0098

Table 5: Bias, Sample Variance and Sample MSE in TAR(1) model

$n = 100$, double exponential, $\theta_1 = 0.3$ and $\theta_2 = -.3$.

(Sample Splitting)

	Band Width	Bias	Sample Variance	Sample MSE
$\theta_1 = 0.3$	0.3	0.0153	0.0321	0.0323
	0.4	0.0204	0.0319	0.0323
	0.5	0.0164	0.0273	0.0276
	0.6	0.0213	0.0255	0.0259
	0.7	0.0220	0.0224	0.0229
$\theta_2 = -.3$	0.3	0.0290	0.0326	0.0334
	0.4	0.0292	0.0353	0.0361
	0.5	0.0290	0.0282	0.0291
	0.6	0.0299	0.0285	0.0293
	0.7	0.0304	0.0252	0.0261

Table 6: Bias, Sample Variance and Sample MSE in TAR(1) model

$n = 100$, double exponential, $\theta_1 = 0.3$ and $\theta_2 = -.3$.

(Non Sample Splitting)

	Band Width	Bias	Sample Variance	Sample MSE
$\theta_1 = 0.3$	0.3	0.0207	0.0190	0.0200
	0.4	0.0244	0.0187	0.0192
	0.5	0.0244	0.0185	0.0191
	0.6	0.0206	0.0187	0.0191
	0.7	0.0206	0.0187	0.0191
$\theta_2 = -.3$	0.3	0.2820	0.0207	0.0215
	0.4	0.0303	0.0213	0.0222
	0.5	0.0302	0.0211	0.0220
	0.6	0.0281	0.0203	0.0211
	0.7	0.0280	0.0203	0.0211

Table 7: Bias, Sample Variance and Sample MSE in TAR(1) model

$n = 200$, double exponential, $\theta_1 = 0.3$ and $\theta_2 = -.3$.

(Sample Splitting)

	Band Width	Bias	Sample Variance	Sample MSE
$\theta_1 = 0.3$	0.3	0.0110	0.0137	0.0138
	0.4	0.0149	0.0130	0.0133
	0.5	0.0141	0.0139	0.0141
	0.6	0.0139	0.0123	0.0124
	0.7	0.0137	0.0107	0.0109
$\theta_2 = -.3$	0.3	0.0134	0.0151	0.0153
	0.4	0.0134	0.0133	0.0135
	0.5	0.0140	0.0158	0.0160
	0.6	0.0146	0.0142	0.0144
	0.7	0.0153	0.0126	0.0128

Table 8: Bias, Sample Variance and Sample MSE in TAR(1) model

$n = 200$, double exponential, $\theta_1 = 0.3$ and $\theta = -.3$.

(Non Sample Splitting)

	Band Width	Bias	Sample Variance	Sample MSE
$\theta_1 = 0.3$	0.3	0.0112	0.0092	0.0093
	0.4	0.0170	0.0103	0.0106
	0.5	0.0112	0.0091	0.0092
	0.6	0.0111	0.0091	0.0092
	0.7	0.0126	0.0086	0.0087
$\theta_2 = -.3$	0.3	0.0185	0.0093	0.0097
	0.4	0.0110	0.0089	0.0090
	0.5	0.0185	0.0092	0.0096
	0.6	0.0184	0.0092	0.0096
	0.7	0.0175	0.0105	0.0107

Table 9: Bias, Sample Variance and Sample MSE in ARCH(1) model

$n = 100$, mixed normal, $\theta = 0.7$, $\alpha_0 = 0.7$ and $\alpha_1 = 0.1$.

	Band Width	Bias	Sample Variance	Sample MSE
Sample Splitting	0.3	0.0610	0.0086	0.0123
	0.4	0.0661	0.0089	0.0133
	0.5	0.0628	0.0083	0.0122
	0.6	0.0644	0.0083	0.0124
	0.7	0.0647	0.0082	0.0124
Non Sample Splitting	0.3	0.0088	0.0059	0.0060
	0.4	-0.0061	0.0057	0.0058
	0.5	-0.0286	0.0056	0.0064
	0.6	-0.0521	0.0054	0.0081
	0.7	-0.0431	0.0073	0.0092

Table 10: Bias, Sample Variance and Sample MSE in ARCH(1) model

$n = 200$, mixed normal, $\theta = 0.7$, $\alpha_0 = 0.7$ and $\alpha_1 = 0.1$.

	Band Width	Bias	Sample Variance	Sample MSE
Sample Splitting	0.3	0.0481	0.0043	0.0067
	0.4	0.0473	0.0043	0.0065
	0.5	0.0494	0.0040	0.0065
	0.6	0.0475	0.0039	0.0062
	0.7	0.0497	0.0041	0.0066
Nonsample Splitting	0.3	0.0068	0.0027	0.0028
	0.4	-0.0075	0.0026	0.0027
	0.5	-0.0292	0.0025	0.0034
	0.6	-0.0600	0.0029	0.0059
	0.7	-0.0431	0.0039	0.0059

Table 11: Bias, Sample Variance and Sample MSE in ARCH(1) model

$n = 100$, double exponential, $\theta = 0.7$, $\alpha_0 = 0.7$ and $\alpha_1 = 0.1$.

	Band Width	Bias	Sample Variance	Sample MSE
Sample Splitting	0.3	0.0028	0.0070	0.0078
	0.4	0.0029	0.0077	0.0086
	0.5	0.0279	0.0070	0.0078
	0.6	0.0274	0.0070	0.0077
	0.7	0.0267	0.0108	0.0115
Non Sample Splitting	0.3	-0.0267	0.0054	0.0061
	0.4	-0.0181	0.0060	0.0063
	0.5	0.0159	0.0068	0.0070
	0.6	0.0223	0.0068	0.0073
	0.7	0.0314	0.0069	0.0079

Table 12: Bias, Sample Variance and Sample MSE in ARCH(1) model

$n = 200$, double exponential, $\theta = 0.7$, $\alpha_0 = 0.7$ and $\alpha_1 = 0.1$.

	Band Width	Bias	Sample Variance	Sample MSE
Sample Splitting	0.3	0.0133	0.0039	0.0041
	0.4	0.0168	0.0039	0.0042
	0.5	0.0133	0.0040	0.0042
	0.6	0.0132	0.0040	0.0041
	0.7	0.0164	0.0040	0.0042
Nonsample Splitting	0.3	-0.0270	0.0028	0.0035
	0.4	0.0115	0.0036	0.0037
	0.5	0.0078	0.0038	0.0039
	0.6	0.0108	0.0039	0.0040
	0.7	0.0143	0.0036	0.0038

Appendix A. Proof of Theorem 4.1

Observe that

$$\begin{aligned}
&\sqrt{n}(\hat{\theta}_n - \theta) - \tfrac{1}{\sqrt{n}} W^{-1} \sum_{t=1}^{n} H_t(\theta)\psi_t(\theta) \\
&= \sqrt{n}(\hat{\theta}_n - \theta) + \tfrac{1}{\sqrt{n}} W^{-1}(\tilde{\theta}_n) \sum_{t=1}^{n} H_t(\tilde{\theta}_n)\psi(\epsilon_{nt}) - \tfrac{1}{\sqrt{n}} W^{-1} \sum_{t=1}^{n} H_t(\theta)\psi_t(\theta) \\
&+ \tfrac{1}{\sqrt{n}} \hat{W}^{-1}(\tilde{\theta}_n) \sum_{t=1}^{n} \{\bar{H}_t(\tilde{\theta}_n)\bar{\psi}(\epsilon_{nt}) - H_t(\tilde{\theta}_n)\psi(\epsilon_{nt})\} \\
&+ \tfrac{1}{\sqrt{n}} \hat{W}^{-1}(\tilde{\theta}_n) \sum_{t=1}^{n} \{\bar{H}_t(\tilde{\theta}_n)\hat{\psi}(\epsilon_{nt}) - \bar{H}_t(\tilde{\theta}_n)\bar{\psi}(\epsilon_{nt})\} \\
&= A_1 + A_2 + A_3.
\end{aligned} \tag{A.1}$$

We need to prove $A_1 + A_2 + A_3 \xrightarrow{p} 0$. First since $\tilde{\theta}_n$ is a discretized $\sqrt{n}$-estimator, $A_1 \xrightarrow{P} 0$ if $\hat{W}^{-1}(\hat{\theta}_n)$ under $P_{\tilde{\theta}_n}$. See the theorem 2.5.2 in Bickel et al. (1993). Therefore it suffices to show that

$$\hat{W}^{-1}(\tilde{\theta}_n) \to \hat{W}^{-1} \quad \text{under} \quad P_{\tilde{\theta}_n} \tag{A.2}$$

$$\tfrac{1}{\sqrt{n}}(\tilde{\theta}_n) \sum_{t=1}^{n} \{\bar{H}_t(\tilde{\theta}_n)\bar{\psi}(\epsilon_{nt}) - H_t(\tilde{\theta}_n)\psi(\epsilon_{nt})\} \xrightarrow{p} 0, \tag{A.3}$$

and

$$\tfrac{1}{\sqrt{n}}(\tilde{\theta}_n) \sum_{t=1}^{n} \{\bar{H}_t(\tilde{\theta}_n)\hat{\psi}(\epsilon_{nt}) - \bar{H}_t(\tilde{\theta}_n)\bar{\psi}(\epsilon_{nt})\} \xrightarrow{p} 0, \tag{A.4}$$

under $P_{\hat{\theta}_n}$.

Lemma A.1. *Assume conditions of (C.1)-(C.8). Then we have* $\hat{W}(\hat{\theta}_n) \to W$ *under* $P_{\tilde{\theta}_n}$.

Proof.

$\hat{W}(\hat{\theta}_n) = \frac{1}{n} \sum_{t=1}^{n} \bar{H}_t(\tilde{\theta}_n)(\epsilon_{nt})\psi(\epsilon_{nt})^T \bar{H}_t(\tilde{\theta}_n)^T$, then for any $a \in R^p$, $a \neq 0$,

$$\begin{aligned}
|\tfrac{1}{n} \sum_{t=1}^{n} a^T \bar{H}_t(\tilde{\theta}_n)\hat{\psi}(\epsilon_{nt})|^2 \quad & -a^T W a| \\
&\leq \tfrac{1}{n} | \sum_{t=1}^{n} \{|a^T \bar{H}_t(\tilde{\theta}_n)\hat{\psi}(\epsilon_{nt})|^2 \\
&\quad -|a^T \bar{H}_t(\tilde{\theta}_n)\psi(\epsilon_{nt})|^2\}| \\
&\quad + \tfrac{1}{n} | \sum_{t=1}^{n} |a^T \bar{H}_t(\tilde{\theta}_n)\psi(\epsilon_{nt})|^2 - a^T W a| \\
&= B_1 + B_2
\end{aligned}$$

Write $\frac{1}{n}\sum_{t=1}^{n}\{|a^T\bar{H}_t(\tilde{\theta}_n)\hat{\psi}(\epsilon_{nt})|^2 - |a^T\bar{H}_t(\tilde{\theta}_n)\psi(\epsilon_{nt})|^2\}$
$= \frac{1}{n}\sum_{t=1}^{n} a^T\bar{H}_t(\tilde{\theta}_n)\bar{H}_t(\tilde{\theta}_n)a\int(1+x^2)\{\hat{\psi}_1(x)-\psi_1(x)\}^2g(x)dx$

By the condition (C.1) and (C.7), B_1 is finite. Since

$$\int(1+x^2)\{\hat{\psi}_1(x)-\psi_1(x)\}^2g(x)dx$$
$$= \int\{\hat{\psi}_1(x)-\psi_1(x)\}^2g(x)dx + \int x^2\{\hat{\psi}_1(x)-\psi_1(x)\}^2g(x)dx$$

Consider

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} a^T\bar{H}_t(\tilde{\theta}_n)\{\hat{\psi}(\epsilon_{nt})-\psi(\epsilon_{nt})\} = a^TD_1$$

Then it is easily seen that a^TD_1 is a matingale. Therefore,

$$\frac{1}{n}\sum_{t=1}^{n}[a^T\bar{H}_t(\tilde{\theta}_n)\{\hat{\psi}(\epsilon_{nt})-\psi(\epsilon_{nt})\}\{\hat{\psi}(\epsilon_{nt})-\psi(\epsilon_{nt})\}^TH_t(\tilde{\theta}_n)^Ta]$$
$$= \frac{1}{n}\sum_{t=1}^{n}\int\{a^T\bar{H}_t(\tilde{\theta}_n)\hat{\psi}(x) - a^TH_t(\tilde{\theta}_n)\psi(x)\}^2g(X)dx$$
$$= \frac{1}{n}\sum_{t=1}^{n} a^T\bar{H}_t(\tilde{\theta}_n)\bar{H}_t(\tilde{\theta}_n)a[\int(1+x^2)\{\hat{\psi}_1(x)-\psi(x)\}^2g(x)dx]$$

By using condition (C.7) and condition (C.2), we know that $(a^TD_1)^2$ is finite.

Write

$$\int(1+x^2)\{\hat{\psi}_1(x)-\psi_1(x)\}^2g(x)dx$$
$$= \int\{\hat{\psi}_1(x)-\psi_1(x)\}^2g(x)dx + \int x^2\{\hat{\psi}_1(x)-\psi_1(x)\}^2g(x)dx$$

Schick (1987) and Koul and Schick (1997) showed that

$$\int\{\hat{\psi}_1(x)-\psi_1(x)\}^2g(x)dx = o(n^{-1}a_n^{-4})$$

and Park (1990) also showed that

$$\int x^2(\hat{\psi}_1(x)-\psi_1(x))^2g(x)dx = o(n^{-1}a_n^{-4})$$

and

$$\int x(\hat{\psi}_1(x)-\psi_1(x))^2g(x)dx = o(n^{-1}a_n^{-4})$$

Therefore

$$B_1 \xrightarrow{p} 0 \text{ under } P_{\tilde{\theta}_n}.$$

Next, we have

$$\begin{aligned} B_2 &= \tfrac{1}{n}\sum_{t=1}^{n} |a^T \bar{H}_t(\tilde{\theta}_n)\psi(\epsilon_{nt})|^2 - a^T W a \\ &= \tfrac{1}{n}\sum_{t=1}^{n} \{|a^T \bar{H}_t(\tilde{\theta}_n)\hat{\psi}(\epsilon_{nt})|^2 - |a^T \bar{H}_t(\tilde{\theta}_n)\psi(\epsilon_{nt})|^2\} \\ &= \tfrac{1}{n}\sum_{t=1}^{n} \{|a^T H_t(\tilde{\theta}_n)\psi(\epsilon_{nt})|^2 - a^T W a\} \end{aligned}$$

The first term goes to 0 if $C_n \to \infty$ because of conditions (C.1), (C.5) and (C.7). It is easily checked that the second term also goes to 0.

Therefore

$$\begin{aligned} &\tfrac{1}{n}\sum_{t=1}^{n} H_t(\tilde{\theta}_n)\hat{\psi}(\epsilon_{nt})\hat{\psi}(\epsilon_{nt})^T H_t(\tilde{\theta}_n)^T \\ &= W + o_{\tilde{\theta}_n(1)}. \end{aligned}$$

Lemma A.2.

$$A_2 = \tfrac{1}{\sqrt{n}}\hat{W}^{-1}(\tilde{\theta}_n)\sum_{t=1}^{n}\{H_t(\tilde{\theta}_n)\bar{\psi}(\epsilon_{nt}) - H_t(\tilde{\theta}_n)\psi(\epsilon_{nt})\} \xrightarrow{p} 0 \text{ as } n \to \infty.$$

Proof. If suffices to show that

$$\frac{1}{\sqrt{n}}\{H_t(\tilde{\theta}_n)\bar{\psi}(\epsilon_{nt}) - H_t(\tilde{\theta}_n)\psi(\epsilon_{nt})\} \xrightarrow{p} 0$$

Since

$$\hat{W}^{-1}(\tilde{\theta}_n) \xrightarrow{p} W^{-1}.$$

Take $C \in R^P$, then

$$\begin{aligned} &\frac{1}{\sqrt{n}} E_{\tilde{\theta}_n}[C^T H_t(\tilde{\theta}_n)\bar{\psi}(\epsilon_{nt}) - C^T H_t(\tilde{\theta}_n)\psi(\epsilon_{nt})|F_{t-1}] \\ &= \tfrac{1}{\sqrt{n}} E_{\tilde{\theta}_n}[\bar{\psi}(\epsilon_{nt}) - \psi(\epsilon_{nt})]C^T H_t(\tilde{\theta}_n) \\ &= 0. \end{aligned}$$

Since $E_{\tilde{\theta}_n}(\bar{\psi}(\epsilon_{nt})) = 0$ and $E_{\tilde{\theta}_n}(\psi(\epsilon_{nt})) = 0$.

Therefore $\frac{1}{\sqrt{n}}C^T H_t(\tilde{\theta}_n)\bar{\psi}(\epsilon_{nt}) - C^T H_t(\tilde{\theta}_n)\psi(\epsilon_{nt})$ is a zero-mean martingale. Then, we have

$$\frac{1}{n}\sum_{t=1}^{n} E_{\tilde{\theta}_n}[C^T H_t(\tilde{\theta}_n)\bar{\psi}(\epsilon_{nt}) - C^T H_t(\tilde{\theta}_n)\psi(\epsilon_{nt})]^2$$

$$= [C^T H_t(\tilde{\theta}_n)\{\bar{\psi}(\epsilon_{nt}) - \psi(\epsilon_{nt})\}^T H_t^T(\theta_n)C^T]$$

$$\xrightarrow{p} 0.$$

Lemma A.3.

$$A_3 = \frac{1}{\sqrt{n}}\hat{W}^{-1}(\tilde{\theta}_n)\sum_{t=1}^{n}\bar{H}_t(\tilde{\theta}_n)\{\hat{\psi}(\epsilon_{nt}) - \bar{\psi}(\epsilon_{nt})\} \xrightarrow{p} 0 \text{ as } n\to\infty.$$

Proof. See Theorem 5.2 of Koul and Schick (1997). By using Lemma A.1-A.3, we complete the proof of Theorem 4.1.

REFERENCES

Akritas, M.G and Johnson, R. A. (1982). Efficiencies of tests and estimators for pth-order autoregressive processes when the error distribution is nonnormal. *Ann. Inst. Statist. Math.* 34, 579-589.

Bickel, P. J. (1982). On adaptive estimation. *The Annals of Statistics* 647-671.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Weller, J. A. (1993). Efficient and adaptive estimation for semiparametric models. Johns Hopkins University Press.

Drost, F.C. and Klaassen, C. A. J. (1997). Efficient estimation in semiparametric GARCH models. *journal of Econometrics* 81, 193-221.

Drost, F.C. and Klaassen, C. A. J. and Werker, B. J. M. (1997). Adaptive estimation in time-series models. *The Annals of Statistics* 25, 786-817.

Hwang, S. Y. and Basawa, I. V. (1993). Asymptotic optimal inference for a class of nonlinear time series models. *Stochastic Processes and Their Applications* 46, 91-113.

Jeganathan, P. (1995). Some aspects of asymptotic theory with applications to time series models. *Econometric Theory* 11, 818-887.

Klaassen, C. A. J. (1987). Consistent estimation of the influence functions of locally asymptotically linear estimates. *The Annals of Statistics* 15, 1548-1562.

Koul, H. L. and Schick, A. (1997). Efficient estimation in nonlinear autoregressive time-series models. *Bernoulli* 3, 247-277.

Kreiss, J. P. (1985). A note on M-estimation in stationary ARMA processes. *Statistics and Decisions* 3, 317-336.

Kreiss, J. P. (1987a). On adaptive estimation in stationary ARMA processes. *The Annals of Statistics* 15, 112-133.

Kreiss, J. P. (1987b). On adaptive estimation in autoregressive models when there are nuisance functions. *Statistics Decisions* 5, 59-76.

Le Cam, L. (1960). Locally asymptotically normal families of distributions. *University of California Publ. Statistics.* 3, 37-98.

Le Cam, L. and Yang, G. L. (1990) *Asymptotics in Statistics* : Some Basic Concepts. New York. Springer.

Linton, O. (1993). Adaptive estimation in ARCH models. *Econometric Theory* 9, 539-569.

Park, B. U. (1990). Efficient estimation in the two-sample semiparametric location-scale models. *Probability Theory and Related Fields* 86, 21-39.

Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics* 14, 1139-1151.

Schick, A. (1987). A note on the construction of asymptotically linear estimates. *Journal of Statistical Planning and Inference* 16, 89-105.

Shin, D. W. and So, B. S. (1999). Unit root test based on adaptive maximum likelihood estimation, *Econometric Theory* 15, 1-23.