CHAPTER 1

EFFICIENT ESTIMATORS FOR TIME SERIES

Anton Schick
Department of Mathematical Sciences, Binghamtom University
Binghamton, NY 13902-6000, USA
E-mail: anton@math.binghamton.edu

Wolfgang Wefelmeyer
Mathematisches Institut, Universität zu Köln
Weyertal 86–90, 50931 Köln, Germany
E-mail: wefelm@math.uni-koeln.de

We illustrate several recent results on efficient estimation for semiparametric time series models with a simple class of models: first-order nonlinear autoregression with independent innovations. We consider in particular estimation of the autoregression parameter, the innovation distribution, conditional expectations, the stationary distribution, the stationary density, and higher-order transition densities.

1. Introduction

Inference for semiparametric time series is well-studied. Two recent monographs are Taniguchi and Kakizawa (2000) and Fan and Yao (2003). The classical nonparametric estimators are however inefficient. In the last twenty years, efficient estimators for various functionals of such models have been constructed. The main effort was on estimators for the Euclidean parameters, but recently other functionals of time series have also been treated. We describe some of these results in a simple situation, observations $X_0, \ldots, X_n$ from a stationary nonlinear autoregressive model $X_i = r_\vartheta(X_{i-1}) + \varepsilon_i$ with independent innovations ε_i. For notational simplicity we restrict attention to the first-order case and assume that ϑ is one-dimensional. The innovations are assumed to have mean zero, finite variance, and a positive density f. The model is semiparametric, with "parameter of interest" ϑ and "nui-

sance parameter" f. We will also be interested in f, in which case ϑ would be the nuisance parameter. In the time series and regression literature, our model would be called "parametric" because the (auto-)regression function depends on a finite-dimensional parameter.

In Section 2 we recall a characterization of regular and efficient estimators in the context of our model. Section 3 describes an efficient estimator for ϑ as a one-step improvement of a $n^{1/2}$-consistent initial estimator. Section 4 shows that appropriately weighted residual-based empirical estimators are efficient for linear functionals $E[h(\varepsilon)]$ of the innovation distribution. Section 5 introduces similarly weighted residual-based kernel estimators $\hat{f}_w$ for the innovation density and shows that plug-in estimators $\int h(y)\hat{f}_w(y)\,dy$ are also efficient for linear functionals $E[h(\varepsilon)]$. Section 6 uses the representation $E(h(X_{n+1} \mid X_n) = E[h(\varepsilon + r_\vartheta(x))]$ for a conditional expectation to construct $n^{1/2}$-consistent and efficient estimators for it. The results extend to higher-order lags. As m tends to infinity, the conditional expectation $E(h(X_{n+m} \mid X_n)$ of lag m converges to the expectation $E[h(X)]$ under the stationary law. This gives rise to efficient estimators for such expectations, as shown in Section 7. The stationary density g has the representation $g(y) = E[f(y - r_\vartheta(X))]$. In Section 8 we use this representation to construct $n^{1/2}$-consistent and efficient estimators for g. The two-step transition density q_2 has the representation $q_2(x,z) = E[f(z - r_\vartheta(\varepsilon + r_\vartheta(x)))]$. Section 9 suggests $n^{1/2}$-consistent and efficient estimators for q_2. This extends to higher-order lags.

Our estimators for the autoregression parameter, the innovation distribution and the stationary distribution in Sections 3, 4 and 7 have the same parametric convergence rate as the usual nonparametric estimators, but smaller asymptotic variances. On the other hand, our estimators for conditional expectations, the stationary density and higher-order transition densities in Sections 5, 8 and 9 have better, parametric, convergence rates than the nonparametric estimators. The parametric rates in Sections 5 and 9 require a parametric form of the autoregression function, because the representations of the functionals there are functions of the value of the autoregression function at a point. The parametric rates in Sections 4, 7 and 8 would extend to models with semiparametric or nonparametric autoregression functions, because the functionals considered there are smooth functionals of the autoregression function, and the plug-in principle works.

2. Characterization of efficient estimators

In this section we recall a characterization of efficient estimators in the context of our semiparametric time series model. The standard reference in the case of i.i.d. data is Bickel, Klaassen, Ritov and Wellner (1998). The theory for time series is similar, especially for processes driven by independent innovations, as in our model.

Let $X_0, \ldots, X_n$ be observations from the stationary nonlinear time series $X_i = r_\vartheta(X_{i-1}) + \varepsilon_i$. Assume that the innovations ε_i are i.i.d. with mean zero, finite variance σ^2, and density f. Conditions for geometric ergodicity in terms of the growth of r_ϑ are in Mokkadem (1987), Bhattacharya and Lee (1995a, b) and An and Huang (1996). The model is described by the transition density $f(y - r_\vartheta(x))$ from $X_0 = x$ to $X_1 = y$, and parametrized by ϑ and f. Write F for the distribution function of f. Let g denote the stationary density. We will write (X, ε) for (X_0, ε_1) and (X, Y) for (X_0, X_1). In order to characterize efficient estimators of smooth functionals of (ϑ, f), we show that the model is locally asymptotically normal. For this, fix ϑ and f and introduce perturbations $\vartheta_{nu} = \vartheta + n^{-1/2}u$ and $f_{nv}(x) \doteq f(x)(1 + n^{-1/2}v(x))$ (in the sense of Hellinger differentiability). Here the local parameter u runs through $\mathbb{R}$. For f_{nv} to be again a probability density with mean zero, the local parameter v must lie in the linear space

$$V = \{v \in L_2(F) : E[v(\varepsilon)] = E[\varepsilon v(\varepsilon)] = 0\}.$$

In other words: $v(\varepsilon)$ must be orthogonal to 1 and ε. If r_ϑ is appropriately differentiable in ϑ with derivative $\dot{r}_\vartheta$, then the transition density is perturbed as

$$f_{nv}(y - r_{\vartheta_{nu}}(x)) \doteq f(y - r_\vartheta(x))\big(1 + n^{-1/2}\big(v(y - r_\vartheta(x)) + u\dot{r}_\vartheta(x)\ell(y - r_\vartheta(x))\big)\big),$$

with $\ell = -f'/f$ the score function for location of the innovation distribution. The perturbation of the transition density is the *tangent*; it is convenient to write it as a random variable

$$t_{uv}(X, Y) = v(\varepsilon) + u\dot{r}_\vartheta(X)\ell(\varepsilon).$$

The *tangent space* of the model is

$$T = \{t_{uv}(X, Y) : u \in \mathbb{R},\ v \in V\}.$$

Let P_{n+1} denote the joint law of $(X_0, \ldots, X_n)$, with density

$$g(X_0) \prod_{i=1}^{n} f(X_i - r_\vartheta(X_{i-1})),$$

and write $P_{n+1,uv}$ for the joint law under ϑ_{nu} and f_{nv}. Koul and Schick (1997) prove *local asymptotic normality*, i.e. a quadratic approximation of the local log-likelihood of the form

$$\log \frac{dP_{n+1,uv}}{dP_{n+1}}(X_0,\dots,X_n) = n^{-1/2}\sum_{i=1}^{n} t_{uv}(X_{i-1},X_i) - \frac{1}{2}E[t_{uv}^2(X,Y)] + o_p(1),$$

where the linear term $n^{-1/2}\sum_{i=1}^{n} t_{uv}(X_{i-1},X_i)$ is asymptotically normal with variance $E[t_{uv}^2(X,Y)]$ by a martingale central limit theorem.

The tangent space T is a subspace of

$$S = \{s(X,Y) \in L_2(P_2) : E(s(X,Y) \mid X) = 0\}.$$

Consider a real-valued functional κ of (ϑ, f). Call κ *differentiable* at (ϑ, f) with *gradient* $s \in S$ if

$$n^{1/2}(\kappa(\vartheta_{nu}, f_{nv}) - \kappa(\vartheta, f)) \to E[s(X,Y)t_{uv}(X,Y)], \quad (u,v) \in \mathbb{R} \times V.$$

The gradient is not uniquely determined, but its projection t_* onto T, the *canonical gradient*, is. Let $\hat\kappa$ be an estimator of κ. Call $\hat\kappa$ *regular* at (ϑ, f) with *limit* L if

$$n^{1/2}(\hat\kappa - \kappa(\vartheta_{nu}, f_{nv})) \Rightarrow L \quad \text{under } P_{n+1,uv}, \quad (u,v) \in \mathbb{R} \times V.$$

Call $\hat\kappa$ *asymptotically linear* at (ϑ, f) with *influence function* $s \in S$ if

$$n^{1/2}(\hat\kappa - \kappa(\vartheta, f)) = n^{-1/2}\sum_{i=1}^{n} s(X_{i-1},X_i) + o_p(1).$$

By a martingale central limit theorem, such an estimator is asymptotically normal with variance $E[s^2(X,Y)]$. The convolution theorem of Hájek (1970) and Le Cam (1971) in the version of Bickel *et al.* (1998, Section 2.3) implies the following three results:

(1) The distribution of L is a convolution, $L = N + M$ in distribution, where N is normal with variance $E[t_*^2(X,Y)]$ and M is independent of N.
(2) A regular estimator has limit $L = N$ if and only if it is asymptotically linear with influence function t_*.
(3) An asymptotically linear estimator is regular if and only if its influence function is a gradient.

A regular estimator with limit $L = N$ is least dispersed among all regular estimators. Such an estimator is called *efficient*. It follows from

(1)–(3) that $\hat{\kappa}$ is regular and efficient if and only if it is asymptotically linear with influence function equal to the canonical gradient,

$$n^{1/2}(\hat{\kappa} - \kappa(\vartheta, f)) = n^{-1/2} \sum_{i=1}^{n} t_*(X_{i-1}, X_i) + o_p(1). \tag{1}$$

Remark 1. Note that S is the tangent space of the nonparametric model of *all* first-order Markov chains on the state space $\mathbb{R}$. Such a chain is described by its transition distribution $Q(x, dy)$, which is perturbed as $Q_{ns}(x, dy) \doteq Q(x, dy)(1 + n^{-1/2}s(x, y))$ with $s \in S$. The reason for embedding T into S, i.e. the autoregressive model into a nonparametric Markov chain model, is the following. Often the functional of interest has a natural extension to a larger model. In such a larger model it is typically easier to determine a gradient, for example as the influence function of some nonparametric estimator. The canonical gradient is then found by projecting the given gradient onto T. Also, in some cases an efficient estimator is found by correcting the given nonparametric estimator.

The choice of the larger space S determines how many (regular) asymptotically linear estimators exist. For the choice $S = T$, any functional would have a unique gradient, and all (regular) asymptotically linear estimators would be asymptotically equivalent. We could also pick a larger S than above, for example the tangent space of all Markov chains of *arbitrary* order, which would give more "asymptotically linear" estimators, but for our purposes the chosen S turns out to be large enough. □

Remark 2. We have introduced gradients and influence functions as elements of the tangent space S of *transition* distributions $Q(x, dy)$. Bickel (1993) and Bickel and Kwon (2001) describe Markov chain models by the *joint* law $P_2(dx, dy)$ of two successive observations. This is particularly convenient when the model and the functional of interest are naturally described in terms of P_2. Results for Markov chains can then be obtained from results for bivariate i.i.d. models, and vice versa. See also the discussion in Greenwood, Schick and Wefelmeyer (2001). □

To calculate canonical gradients, it is convenient to decompose the tangents t_{uv} into orthogonal components. We have $E[\varepsilon\ell(\varepsilon)] = 1$. Hence the projection of $\ell(\varepsilon)$ onto V is $\ell_V(\varepsilon) = \ell(\varepsilon) - \sigma^{-2}\varepsilon$. Write $\mu = E[\dot{r}_\vartheta(X)]$. Then

$$t_{uv}(X, Y) = v(\varepsilon) + u\mu\ell_V(\varepsilon) + us_0(X, Y)$$

with $s_0(X,Y) = \dot r_\vartheta(X)\ell(\varepsilon) - \mu\ell_V(\varepsilon)$ orthogonal to V. The variance of $s_0(X,Y)$ is

$$\Lambda = E[s_0^2(X,Y)] = RJ - \mu^2 J_V$$

with $R = E[\dot r_\vartheta^2(X)]$, $J = E[\ell^2(\varepsilon)]$ and $J_V = E[\ell_V^2(\varepsilon)] = J - \sigma^{-2}$.

In the following we describe the canonical gradients for some of the functionals considered in Sections 3–9 below.

Autoregression parameter. By Schick and Wefelmeyer (2002a, Section 2), the canonical gradient of $\kappa(\vartheta, f) = \vartheta$ is

$$t_*(X,Y) = \Lambda^{-1} s_0(X,Y). \tag{2}$$

If f happens to be normal, then $\ell(\varepsilon) = \sigma^{-2}\varepsilon$, so $J = \sigma^{-2}$, $J_V = 0$, $\Lambda = R\sigma^{-2}$ and hence $s_0(X,Y) = \dot r_\vartheta(X)\sigma^{-2}\varepsilon$ and $t_*(X,Y) = R^{-1}\dot r_\vartheta(X)\varepsilon$.

Innovation distribution. By Schick and Wefelmeyer (2002a, Section 2), the canonical gradient of a linear functional $\kappa(\vartheta, f) = E[h(\varepsilon)]$ of the innovation distribution has canonical gradient

$$t_*(X,Y) = h_V(\varepsilon) - \mu E[h_V(\varepsilon)\ell(\varepsilon)]\Lambda^{-1} s_0(X,Y), \tag{3}$$

where $h_V(\varepsilon) = h(\varepsilon) - E[h(\varepsilon)] - \sigma^{-2}E[\varepsilon h(\varepsilon)]\varepsilon$ is the projection of $h(\varepsilon)$ onto V. In the submodel with ϑ *known*, the canonical gradient of $E[h(\varepsilon)]$ is $h_V(\varepsilon)$.

Conditional expectation. The conditional expectation with lag one of a function h can be written

$$E(h(Y) \mid X = x) = \int h(y) f(y - r_\vartheta(x))\, dy = \int h(y + r_\vartheta(x)) f(y)\, dy.$$

This is the (unconditional) expectation $E[h(\varepsilon, \vartheta)]$ of a function $h(y, \vartheta) = h(y + r_\vartheta(x))$ depending on ϑ, and the gradient is similar to (3), with additional terms from this dependence on ϑ,

$$t_*(X,Y) = h_V(\varepsilon, \vartheta) + (\dot r_\vartheta(x) - \mu)E[h_V(\varepsilon, \vartheta)\ell(\varepsilon)]\Lambda^{-1} s_0(X,Y). \tag{4}$$

3. Autoregression parameter

A simple estimator of ϑ is the *least squares estimator*. It is defined as the minimizer in ϑ of $\sum_{i=1}^n (X_i - r_\vartheta(X_{i-1}))^2$ and is the solution of the martingale estimating equation

$$\sum_{i=1}^n \dot r_\vartheta(X_{i-1})(X_i - r_\vartheta(X_{i-1})) = 0.$$

By Taylor expansion, its influence function is seen to be $s(X,Y) = R^{-1}\dot{r}_\vartheta(X)\varepsilon$. We have seen in Section 2 that this equals the canonical gradient of ϑ only if the innovations happen to be normally distributed.

An efficient estimator of ϑ is obtained in Koul and Schick (1997) as a one-step improvement of an initial $n^{1/2}$-consistent estimator $\hat{\vartheta}$, for example the least squares estimator. Rewrite the canonical gradient of ϑ as

$$t_*(X,Y) = \Lambda^{-1}\big((\dot{r}_\vartheta(X) - \mu)\ell(\varepsilon) - \mu\sigma^{-2}\varepsilon\big).$$

Estimate $\ell = -f'/f$ by $\hat{\ell} = -\hat{f}'/\hat{f}$ with $\hat{f}$ an appropriate kernel estimator. Then estimate J by $\hat{J} = (1/n)\sum_{i=1}^n \hat{\ell}^2(\hat{\varepsilon}_i)$, and μ, R and σ^2 by empirical estimators $\hat{\mu} = (1/n)\sum_{i=1}^n \dot{r}_{\hat{\vartheta}}(X_i)$, $\hat{R} = (1/n)\sum_{i=1}^n \dot{r}^2_{\hat{\vartheta}}(X_i)$ and $\hat{\sigma}^2 = (1/n)\sum_{i=1}^n \hat{\varepsilon}_i^2$, where $\hat{\varepsilon}_i = X_i - r_{\hat{\vartheta}}(X_{i-1})$ are the residuals. For $\Lambda = RJ - \mu^2 J_V = (R-\mu^2)J - \mu^2\sigma^{-2}$ we obtain the estimator $\hat{\Lambda} = (\hat{R} - \hat{\mu}^2)J - \hat{\mu}^2\hat{\sigma}^{-2}$, and for $t_*(X_{i-1}, X_i)$ we obtain the estimator

$$\hat{t}_*(X_{i-1}, X_i) = \hat{\Lambda}^{-1}\big((\dot{r}_{\hat{\vartheta}}(X_i) - \hat{\mu})\hat{\ell}(\hat{\varepsilon}_i) - \hat{\mu}\hat{\sigma}^{-2}\hat{\varepsilon}_i\big).$$

The efficient one-step improvement of $\hat{\vartheta}$ is then

$$\hat{\vartheta} + \frac{1}{n}\sum_{i=1}^n \hat{t}_*(X_{i-1}, X_i).$$

It does not require sample splitting. A related result for parameters of the moving average coefficients in invertible linear processes is in Schick and Wefelmeyer (2002b).

Remark 3. The *linear* autoregressive model $X_i = \vartheta X_{i-1} + \varepsilon_i$ is a degenerate case. Here $r_\vartheta(X) = \vartheta X$, $\dot{r}_\vartheta(X) = X$, and $\mu = 0$. Hence $s_0(X,Y) = X\ell(\varepsilon)$. Furthermore, $R = E[\dot{r}^2_\vartheta(X)] = E[X^2]$, which is the stationary variance,

$$R = \tau^2 = \sigma^2 \sum_{j=0}^\infty \vartheta^{2j} = \frac{\sigma^2}{1-\vartheta^2}.$$

Hence $\Lambda = \tau^2 J$, and the canonical gradient (2) reduces to $t_*(X,Y) = \tau^{-2}J^{-1}X\ell(\varepsilon)$. The least squares estimator is $\hat{\vartheta} = \sum_{i=1}^n X_{i-1}X_i / \sum_{i=1}^n X_{i-1}^2$. An efficient estimator for ϑ is the one-step improvement

$$\hat{\vartheta} + \hat{\tau}^{-2}\hat{J}^{-1}\frac{1}{n}\sum_{i=1}^n X_i\hat{\ell}(\hat{\varepsilon}_i)$$

with $\hat{\varepsilon}_i = X_i - \hat{\vartheta}X_{i-1}$ and $\hat{\tau}^2 = (1-\hat{\vartheta}^2)^{-1}(1/n)\sum_{i=1}^n \hat{\varepsilon}_i^2$, and with $\hat{J}$ and $\hat{\ell}$ as before.

The canonical gradient equals the one in the submodel with f *known*. Hence ϑ can be estimated *adaptively* with respect to f. To prove efficiency in this situation, we need local asymptotic normality only for fixed f. Kreiss (1987a, b) constructs adaptive estimators for parameters in ARMA models with symmetric innovation density and in AR models with mean zero innovation density. Jeganathan (1995) and Drost, Klaassen and Werker (1997) generalize Kreiss (1987a) to nonlinear and heteroscedastic autoregression. See also Koul and Pflug (1990) and Koul and Schick (1996) for adaptive estimation in explosive linear autoregression and in random coefficient autoregression. General results on adaptive estimation in the i.i.d. case are in Klaassen and Putter (2005). □

4. Innovation distribution

In this section we consider estimation of a linear functional $E[h(\varepsilon)]$ of the innovation distribution. Suppose first that ϑ is *known*. Then we know the innovations $\varepsilon_i = X_i - r_\vartheta(X_{i-1})$ and can estimate $E[h(\varepsilon)]$ by the empirical estimator $(1/n)\sum_{i=1}^n h(\varepsilon_i)$. Its influence function is $h(\varepsilon) - E[h(\varepsilon)]$. The canonical gradient for ϑ known is $h_V(\varepsilon) = h(\varepsilon) - E[h(\varepsilon)] - \sigma^{-2}E[\varepsilon h(\varepsilon)]\varepsilon$, so the empirical estimator is not efficient. The reason is that it does not use the information that the innovations have mean zero. There are different ways of using this information.

1. Following Levit (1975) and Haberman (1984), an estimator with influence function $h_V(\varepsilon)$ is obtained by estimating $\sigma^{-2}E[\varepsilon h(\varepsilon)]$ empirically and using the *corrected empirical estimator*

$$\frac{1}{n}\sum_{i=1}^n h(\varepsilon_i) - \frac{\sum_{i=1}^n \varepsilon_i h(\varepsilon_i)}{\sum_{i=1}^n \varepsilon_i^2}\frac{1}{n}\sum_{i=1}^n \varepsilon_i.$$

2. Following Owen (1988, 2001), choose random weights w_i such that the weighted empirical distribution has mean zero, $\sum_{i=1}^n w_i\varepsilon_i = 0$, and use the *weighted empirical estimator*

$$\frac{1}{n}\sum_{i=1}^n w_i h(\varepsilon_i).$$

By the method of Lagrange multipliers, the weights are seen to be of the form $w_i = 1/(1+\lambda\varepsilon_i)$. This implies $\lambda = \sigma^{-2}(1/n)\sum_{i=1}^n \varepsilon_i + o_p(n^{-1/2})$ and therefore

$$\frac{1}{n}\sum_{i=1}^n w_i h(\varepsilon_i) = \frac{1}{n}\sum_{i=1}^n h(\varepsilon_i) - \sigma^{-2}E[\varepsilon h(\varepsilon)]\frac{1}{n}\sum_{i=1}^n \varepsilon_i + o_p(n^{-1/2}).$$

Hence this estimator also has influence function $h_V(\varepsilon)$.

Now return to the autoregressive model of interest, with ϑ *unknown*. The *parametric plug-in principle* says that an efficient estimator for $E[h(\varepsilon)]$ is obtained by replacing ϑ by an efficient estimator $\hat{\vartheta}$. Then the true innovations are replaced by the residuals $\hat{\varepsilon}_i = X_i - r_{\hat{\vartheta}}(X_{i-1})$. Correspondingly, choose weights $\hat{w}_i$ such that $\sum_{i=1}^n \hat{w}_i \hat{\varepsilon}_i = 0$. Efficient estimators for $E[h(\varepsilon)]$ are then obtained as the *residual-based corrected empirical estimator*

$$\frac{1}{n}\sum_{i=1}^n h(\hat{\varepsilon}_i) - \frac{\sum_{i=1}^n \hat{\varepsilon}_i h(\hat{\varepsilon}_i)}{\sum_{i=1}^n \hat{\varepsilon}_i^2} \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i \tag{5}$$

and the *residual-based weighted empirical estimator*

$$\frac{1}{n}\sum_{i=1}^n \hat{w}_i h(\hat{\varepsilon}_i). \tag{6}$$

Depending on h, these improvements can lead to drastic variance reductions. As with the true innovations, we have the expansion

$$\frac{1}{n}\sum_{i=1}^n \hat{w}_i h(\hat{\varepsilon}_i) = \frac{1}{n}\sum_{i=1}^n h(\hat{\varepsilon}_i) - \sigma^{-2} E[\varepsilon h(\varepsilon)] \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i + o_p(n^{-1/2}). \tag{7}$$

The same expansion holds for the estimator (5). For any $n^{1/2}$-consistent estimator $\hat{\vartheta}$,

$$\frac{1}{n}\sum_{i=1}^n h(\hat{\varepsilon}_i) = \frac{1}{n}\sum_{i=1}^n h(\varepsilon_i) - \mu E[h'(\varepsilon)](\hat{\vartheta} - \vartheta) + o_p(n^{-1/2}).$$

By (2), an efficient estimator $\hat{\vartheta}$ has influence function $\Lambda^{-1} s_0(X, Y)$. With $E[h'(\varepsilon)] = E[h(\varepsilon)\ell(\varepsilon)]$, the estimators (5) and (6) are seen to have influence functions equal to the canonical gradient (3). Hence they are efficient.

Efficient estimators of the type (5) were obtained by Wefelmeyer (1994) for linear autoregression, by Schick and Wefelmeyer (2002a) for nonlinear and heteroscedastic autoregression $X_i = r_\vartheta(X_{i-1}) + s_\vartheta(X_{i-1})\varepsilon_i$, and by Schick and Wefelmeyer (2002b) for invertible linear processes with moving average coefficients depending on a finite-dimensional parameter ϑ.

Related results are possible for autoregression $X_i = r(X_{i-1}) + \varepsilon_i$ with a semiparametric or nonparametric model for the autoregression function r. Then $\hat{\varepsilon}_i = X_i - \hat{r}(X_{i-1})$ with $\hat{r}$ a nonparametric estimator of r. Here the (unweighted) residual-based empirical estimator $(1/n)\sum_{i=1}^n h(\hat{\varepsilon}_i)$ is already efficient. The reason is that $\hat{r}$ uses the information in $E[\varepsilon] = 0$. For the corresponding (heteroscesdastic) nonparametric *regression* model, Akritas and Van Keilegom (2001) obtain a functional central limit theorem for the

residual-based empirical distribution function. Different estimators for the regression function are used in Müller, Schick and Wefelmeyer (2004a, b; 2005c). See also Cheng (2002, 2004, 2005).

Remark 4. The *linear* autoregressive model $X_i = \vartheta X_{i-1} + \varepsilon_i$ is a degenerate case. Then $r_\vartheta(X) = \vartheta X$, $\dot r_\vartheta(X) = X$, and $\mu = 0$. By Section 2, the canonical gradient of $E[h(\varepsilon)]$ is $h_V(\varepsilon)$ and equals the canonical gradient in the submodel with ϑ *known*. Hence $E[h(\varepsilon)]$ can be estimated *adaptively* with respect to ϑ. It follows in particular that the estimators (5) and (6) are efficient even if an inefficient estimator for ϑ is used. Also, to prove efficiency, we need local asymptotic normality only for fixed ϑ. □

5. Innovation density

In this section we describe weighted residual-based kernel estimators $\hat f_w$ for the innovation density f. They will be efficient in the (weak) sense that they lead to efficient plug-in estimators $\int h(x)\hat f_w(x)\,dx$ for linear functionals $E[h(\varepsilon)]$. This will be used in Sections 6–9 to construct efficient estimators for conditional expectations, the stationary distribution and density, and higher-order transition densities.

As in Section 4, let $\hat\vartheta$ be $n^{1/2}$-consistent, introduce residuals $\hat\varepsilon_i = X_i - r_{\hat\vartheta}(X_{i-1})$ and choose weights $\hat w_i$ such that $\sum_{i=1}^n \hat w_i\hat\varepsilon_i = 0$. The innovation density f can be estimated by unweighted and weighted residual-based kernel estimators

$$\hat f(y) = \frac{1}{n}\sum_{i=1}^n k_b(y-\hat\varepsilon_i) \quad\text{and}\quad \hat f_w(y) = \frac{1}{n}\sum_{i=1}^n \hat w_i k_b(y-\hat\varepsilon_i),$$

where $k_b(x) = k(x/b)/b$ with k a kernel and b a bandwidth. For an appropriate choice of bandwidth, a Taylor expansion gives

$$\hat f_w(y) = \hat f(y) - \sigma^{-2} y f(y)\frac{1}{n}\sum_{i=1}^n \hat\varepsilon_i + o_p(n^{-1/2}).$$

This is analogous to expansion (7) for the weighted residual-based empirical estimator $(1/n)\sum_{i=1}^n \hat w_i h(\hat\varepsilon_i)$. We see that $\hat f_w$ differs from $\hat f$ by a term of order $n^{-1/2}$. Since $\hat f_w$ and $\hat f$ converge to f at a slower rate, weighting has a negligible effect if we are interested in estimating f itself.

Now we compare the residual-based kernel estimator $\hat f$ with the kernel estimator f^* based on the true innovations,

$$f^*(y) = \frac{1}{n}\sum_{i=1}^n k_b(y-\varepsilon_i).$$

We obtain

$$\hat{f}(y) = f^*(y) + f'(y)\mu(\hat{\vartheta} - \vartheta) + o_p(n^{-1/2}).$$

For the weighted residual-based kernel estimator we therefore have

$$\hat{f}_w(y) = f^*(y) + \sigma^{-2} y f(y) \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i - (\sigma^{-2} y f(y) + f'(y))\mu(\hat{\vartheta} - \vartheta) + o_p(n^{-1/2}).$$

Müller, Schick and Wefelmeyer (2005a) give conditions under which these stochastic expansions hold for some norms such as the supremum norm and the V-norm $\|f\|_V = \int |f(y)| V(y)\, dy$, where typically $V(y) = (1 + |y|)^m$ for some non-negative integer m. For $m = 0$ this is the L_1-norm with respect to Lebesgue measure.

The stronger norms, with $m > 0$, are useful when we want to estimate e.g. a moment $E[\varepsilon^m]$ with a plug-in estimator $\int y^m \hat{f}_w(y)\, dy$. This is an example of the *nonparametric plug-in principle*: Even though $\hat{f}_w$ converges to f at a rate slower than $n^{-1/2}$, the smooth functional $\int y^m \hat{f}_w(y)\, dy$ of $\hat{f}_w$ converges to $\int y^m f(y)\, dy = E[\varepsilon^m]$ at the parametric rate $n^{-1/2}$.

The estimator $\int y^m \hat{f}_w(y)\, dy$ is even efficient if an efficient estimator $\hat{\vartheta}$ is used. More generally, for all sufficiently regular h bounded by (a multiple of) V, the plug-in estimators $\int h(y) \hat{f}_w(y)\, dy$ are efficient for $\int h(y) f(y)\, dy = E[h(\varepsilon)]$. We may therefore call $\hat{f}_w$ *efficient for plug-in*. This (weak) efficiency concept for function estimators was introduced by Klaassen, Lee and Ruymgaart (2001).

Weighting can lead to considerable variance reduction. For example, in the *linear* autoregression model $X_i = \vartheta X_{i-1} + \varepsilon_i$ we have $\mu = 0$ and

$$\int y^m \hat{f}(y)\, dy = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^m + o_p(n^{-1/2}),$$

$$\int y^m \hat{f}_w(y)\, dy = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^m - \sigma^{-2} E[\varepsilon^{m+1}] \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i + o_p(n^{-1/2}).$$

The asymptotic variances are $E[\varepsilon^{2m}]$ and $E[\varepsilon^{2m}] - \sigma^{-2}(E[\varepsilon^{m+1}])^2$, respectively. For $m = 3$ and f normal these variances are $15\sigma^6$ and $6\sigma^6$, respectively, a variance reduction of nearly two thirds.

6. Conditional expectation

The conditional expectation $E(h(X_{n+1}) \mid X_n = x)$ with lag one of a known function h can be estimated by a nonparametric estimator

$$\frac{\sum_{i=1}^{n} k_b(x - X_{i-1}) h(X_i)}{\sum_{i=1}^{n} k_b(x - X_{i-1})},$$

where $k_b(x) = k(x/b)/b$ with k a kernel and b a bandwidth. If the time series is known to be first-order Markov with transition density $q(x, y)$ from x to y, then we have

$$E(h(X_{n+1}) \mid X_n = x) = \int h(y)q(x, y)\, dy,$$

and an estimator is obtained by plugging in a (kernel) estimator for q. In the nonlinear autoregressive model $X_i = r_\vartheta(X_{i-1}) + \varepsilon_i$, the transition density is $q(x, y) = f(y - r_\vartheta(x))$, and we can write

$$E(h(X_{n+1}) \mid X_n = x) = \int h(y + r_\vartheta(x))f(y)\, dy = E[h(\varepsilon + r_\vartheta(x))].$$

This is an (unconditional) expectation under the innovation distribution as in Section 4, but now of a function $h(y, \vartheta) = h(y + r_\vartheta(x))$ depending on ϑ. This suggests estimating the conditional expectation by

$$\frac{1}{n}\sum_{i=1}^{n} h(\hat\varepsilon_i + r_{\hat\vartheta}(x)) = \int h(\cdot, \hat\vartheta)\, d\hat{\mathbb{F}}$$

with $\hat{\mathbb{F}}(y) = (1/n)\sum_{i=1}^n \mathbf{1}(\hat\varepsilon_i \le y)$ the empirical distribution function of the residuals $\hat\varepsilon_i = X_i - r_{\hat\vartheta}(X_{i-1})$. If $\hat\vartheta$ is $n^{1/2}$-consistent, then $\int h(\cdot, \hat\vartheta)\, d\hat{\mathbb{F}}$ will be $n^{1/2}$-consistent. For efficiency we need to use an efficient estimator $\hat\vartheta$ and to replace $\hat{\mathbb{F}}$ by a version that uses the information $E[\varepsilon] = 0$. As seen in Section 4, one way of doing this is by taking a weighted version $\hat{\mathbb{F}}_w(y) = (1/n)\sum_{i=1}^n w_i\mathbf{1}(\hat\varepsilon_i \le y)$. The resulting estimator for the conditional expectation is $\int h(\cdot, \hat\vartheta)\, d\hat{\mathbb{F}}_w$. Similar as in Section 5, a Taylor expansion gives

$$\int h(\cdot, \hat\vartheta)\, d\hat{\mathbb{F}}_w = \frac{1}{n}\sum_{i=1}^{n} h_V(\varepsilon_i, \vartheta) + (\dot r_\vartheta(x) - \mu)E[h'_V(\varepsilon, \vartheta)](\hat\vartheta - \vartheta) + o_p(n^{-1/2}).$$

We have $E[h'_V(\varepsilon, \vartheta)] = E[h_V(\varepsilon, \vartheta)\ell(\varepsilon)]$. By (2), an efficient estimator has influence function $\Lambda^{-1}s_0(X, Y)$; so $\int h(\cdot, \hat\vartheta)\, d\hat{\mathbb{F}}_w$ has influence function equal to the canonical gradient (4) and is therefore efficient.

These results extend to higher lags. For example, for lag two the conditional expecation $E(h(X_{n+2}) \mid X_n = x)$ becomes $E[h(\varepsilon_2 + r_\vartheta(\varepsilon_1 + r_\vartheta(x)))]$ and is estimated $n^{1/2}$-consistently by the residual-based von Mises statistic

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h(\hat\varepsilon_j + r_{\hat\vartheta}(\hat\varepsilon_i + r_{\hat\vartheta}(x))) = \iint h(z + r_{\hat\vartheta}(y + r_{\hat\vartheta}(x)))\, d\hat{\mathbb{F}}(y)d\hat{\mathbb{F}}(z).$$

An efficient estimator is obtained if we replace $\hat{\mathbb{F}}$ by the weighted version $\hat{\mathbb{F}}_w$ and use an efficient estimator for ϑ.

To treat lags higher than two, set $\varrho_{1\vartheta}(x,y) = y + r_\vartheta(x)$ and define recursively $\varrho_{m\vartheta}(x,y_1,\dots,y_m) = y_m + r_\vartheta(\varrho_{m-1,\vartheta}(x,y_1,\dots,y_{m-1}))$. Then an m-step conditional expectation can be written

$$\nu(h) = E(h(X_{n+m}) \mid X_n = x) = E[h(\varrho_{m\vartheta}(x,\varepsilon_1,\dots,\varepsilon_m))].$$

With $\hat\vartheta$ efficient, an efficient estimator for $E(h(X_{n+m}) \mid X_n = x)$ is the weighted residual-based von Mises statistic

$$\int\cdots\int h(\varrho_{m\hat\vartheta}(x,y_1,\dots,y_m))\, d\hat{\mathbb{F}}_w(y_1)\dots d\hat{\mathbb{F}}_w(y_m).$$

To prove $n^{1/2}$-consistency of such von Mises statistics, we need an appropriate balance of smoothness assumptions on h and on f. For discontinuous h we must assume that f is smooth. Then we can replace $\hat{\mathbb{F}}$ or $\hat{\mathbb{F}}_w$ by smoothed versions $d\hat{\mathbb{F}}(y) = \hat f(y)\,dy$ and $d\hat{\mathbb{F}}_w(y) = \hat f_w(y)\,dy$ with residual-based unweighted or weighted kernel estimators $\hat f$ and $\hat f_w$ as in Section 5. Write

$$\hat\nu(h) = \int\cdots\int h(\varrho_{m\hat\vartheta}(x,y_1,\dots,y_m))\hat f_w(y_1)\dots\hat f_w(y_m)\,dy_1\cdots dy_m.$$

Müller, Schick and Wefelmeyer (2005b) prove functional central limit theorems for processes $\{n^{1/2}(\nu(h)-\hat\nu(h)) : h\in\mathcal{H}\}$ and appropriate function classes $\mathcal{H}$.

Simulations show that smoothing also improves the small-sample behavior of the von Mises statistics, especially for discontinuous h.

Remark 5. The parametric rates for estimators of conditional expectations do not extend to autoregression $X_i = r(X_{i-1})+\varepsilon_i$ with semiparametric or nonparametric autoregression function r. A conditional expectation of lag one has representation $E(h(X_{n+1}) \mid X_n = x) = E[h(\varepsilon + r(x))]$. This is a (smooth) function of $r(x)$, and the convergence rate of an estimator for the conditional expectation is in general determined by the rate at which we can estimate $r(x)$. Nevertheless, estimators based on this representation may still be better than nonparametric estimators. For results in nonparametric (censored) *regression* we refer to Van Keilegom, Akritas and Veraverbeke (2001) and Van Keilegom and Veraverbeke (2001, 2002). □

7. Stationary distribution

A simple estimator for the expectation $E[h(X)]$ of a known function h under the stationary distribution is the empirical estimator $(1/n)\sum_{i=1}^n h(X_i)$.

The estimator does not make use of the autoregressive structure of our model. A better estimator can be obtained as follows. As m tends to infinity, the m-step conditional expectation $E(h(X_m) \mid X_0 = x)$ converges to $E[h(X)]$ at an exponential rate. By Section 6, for fixed m, an efficient estimator of $E(h(X_m) \mid X_0 = x)$ is the weighted residual-based von Mises statistic

$$\hat{\kappa}_m(x) = \int \cdots \int h(\varrho_{m\hat{\vartheta}}(x, y_1, \ldots, y_m))\, d\hat{\mathbb{F}}_w(y_1) \ldots d\hat{\mathbb{F}}_w(y_m)$$

or its smoothed version. We expect that $\hat{\kappa}_{m(n)}(x)$ is efficient for $E[h(X)]$ if $m(n)$ increases with n at an appropriate (logarithmic) rate.

The bias induced by the choice of starting point x can be removed by averaging, i.e. by using instead of $\hat{\kappa}_{m(n)}(x)$ the estimator

$$\frac{1}{n} \sum_{i=1}^{n} \hat{\kappa}_{m(n)}(X_i).$$

For invertible linear processes with moving average coefficients depending on a finite-dimensional parameter, a corresponding result is proved in Schick and Wefelmeyer (2004b).

8. Stationary density

The usual estimator for the stationary density g of a time series is the kernel estimator $(1/n) \sum_{i=1}^{n} k_b(x - X_i)$, where $k_b(x) = k(x/b)/b$ with kernel k and bandwidth b. For our nonlinear autoregressive model, the stationary density can be written

$$g(y) = \int f(y - r_\vartheta(x)) g(x)\, dx = E[f(y - r_\vartheta(X))].$$

This is an expectation under the stationary distribution as in Section 7, but now for a function $h(x, \vartheta, f) = f(y - r_\vartheta(x))$ depending on ϑ and f. By the plug-in principle mentioned in Section 5, we expect to obtain a $n^{1/2}$-consistent estimator if we plug appropriate kernel estimators $\hat{f}$ for f and $\hat{g}$ for g and a $n^{1/2}$-consistent estimator $\hat{\vartheta}$ for ϑ into this representation,

$$g^*(y) = \int \hat{f}(y - r_{\hat{\vartheta}}(x)) \hat{g}(x)\, dx.$$

Note however that g is the density of the convolution of ε and $r_\vartheta(X)$. Even if the density g of X is nice, the distribution of $r_\vartheta(X)$ may be unpleasant. A degenerate case would be a constant autoregression function, say $r_\vartheta =$

0. Then we observe independent $X_i = \varepsilon_i$ with density $g = f$, and $n^{1/2}$-consistent estimation of g is not possible. There may be a problem even if r_ϑ is smooth and strictly increasing but with derivative vanishing at some point. However, if r_ϑ has a derivative that is bounded away from zero, $g^*(y)$ will be $n^{1/2}$-consistent. For efficiency we need an efficient estimator $\hat{\vartheta}$ as in Section 3 and a weighted residual-based kernel estimator $\hat{f}_w$ for the innovation density as in Section 5, and we must replace the estimator $\hat{g}(x)\,dx$ by an efficient estimator as in Section 7.

In the i.i.d. case, $n^{1/2}$-consistent estimators for convolution densities are studied by Frees (1994), Saavedra and Cao (2000) and Schick and Wefelmeyer (2004c). A $n^{1/2}$-consistent estimator for the stationary density of a first-order moving average process is obtained in Saavedra and Cao (1999). Schick and Wefelmeyer (2004a) introduce an efficient version, and Schick and Wefelmeyer (2004d) prove functional central limit theorems for higher-order moving average processes and density estimators viewed as elements of function spaces. For general invertible linear processes, $n^{1/2}$-consistent estimators of the stationary density are constructed in Schick and Wefelmeyer (2005).

9. Transition density

The one-step transition density $q(x, y)$ from $X_0 = x$ to $X_1 = y$ of a first-order Markov chain can be estimated by the Nadaraya–Watson estimator

$$\hat{q}(x, y) = \frac{\sum_{i=1}^n k_b(x - X_{i-1})k_b(y - X_i)}{\sum_{i=1}^n k_b(x - X_{i-1})},$$

where $k_b(x) = k(x/b)/b$ with kernel k and bandwidth b. The two-step transition density $q_2(x, z)$ from $X_0 = x$ to $X_2 = z$ has the representation $q_2(x, z) = \int q(x, y)q(y, z)\,dy$ and can be estimated by $\int \hat{q}(x, y)\hat{q}(y, z)\,dy$. For our nonlinear autoregressive model, the two-step transition density can be written

$$q_2(x, z) = \int f(z - r_\vartheta(y))f(y - r_\vartheta(x))\,dy = E[f(z - r_\vartheta(\varepsilon + r_\vartheta(x)))].$$

This is an expectation under the innovation distribution as in Section 4, but now for a function $h(y, \vartheta, f) = f(z - r_\vartheta(y + r_\vartheta(x)))$ depending on ϑ and f. There is some formal similarity with the representation of the stationary density in Section 8, but also an essential difference: There we had an expectation under the *stationary* distribution; here the expectation is taken with respect to the *innovation* distribution. This makes efficient

estimation of the transition density easier, because it is easier to estimate the innovation distribution efficiently. As efficient estimator of $q_2(x,z)$ we suggest

$$\hat{q}_2(x,z) = \int \hat{f}_w(z - r_{\hat{\vartheta}}(y))\hat{f}_w(y - r_{\hat{\vartheta}}(x))\,dy$$

with $\hat{\vartheta}$ efficient. The result extends to higher-order lags.

Acknowledgments

Anton Schick was supported by NSF grant DMS 0405791.

References

1. Akritas, M. G. and Van Keilegom, I. (2001). Non-parametric estimation of the residual distribution. *Scand. J. Statist.* **28**, 549–567.
2. An, H. Z. and Huang, F. C. (1996). The geometrical ergodicity of nonlinear autoregressive models. *Statist. Sinica* **6**, 943–956.
3. Bhattacharya, R. N. and Lee, C. (1995a). Ergodicity of nonlinear first order autoregressive models. *J. Theoret. Probab.* **8**, 207–219.
4. Bhattacharya, R. and Lee, C. (1995b). On geometric ergodicity of nonlinear autoregressive models. *Statist. Probab. Lett.* **22**, 311–315. Erratum: **41** (1999), 439–440.
5. Bickel, P. J. (1993). Estimation in semiparametric models. In: *Multivariate Analysis: Future Directions* (C. R. Rao, ed.), pp. 55–73, North-Holland, Amsterdam.
6. Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York.
7. Bickel, P. J. and Kwon, J. (2001). Inference for semiparametric models: Some questions and an answer (with discussion). *Statist. Sinica* **11**, 863–960.
8. Cheng, F. (2002). Consistency of error density and distribution function estimators in nonparametric regression. *Statist. Probab. Lett.* **59**, 257–270.
9. Cheng, F. (2004). Weak and strong uniform consistency of a kernel error density estimator in nonparametric regression. *J. Statist. Plann. Inference* **119**, 95–107.
10. Cheng, F. (2005). Asymptotic distributions of error density and distribution function estimators in nonparametric regression. *J. Statist. Plann. Inference* **128**, 327–349.
11. Drost, F. C., Klaassen, C. A. J. and Werker, B. J. M. (1997). Adaptive estimation in time-series models, *Ann. Statist.* **25**, 786–817.
12. Fan, J. and Yao, Q. (2003). *Nonlinear Time Series. Nonparametric and Parametric Methods*. Springer Series in Statistics, Springer, New York.
13. Frees, E. W. (1994). Estimating densities of functions of observations. *J. Amer. Statist. Assoc.* **89**, 517–525.

14. Greenwood, P. E., Schick, A. and Wefelmeyer, W. (2001). Comment [on Bickel and Kwon (2001)]. *Statist. Sinica* **11**, 892–906.
15. Haberman, S. J. (1984). Adjustment by minimum discriminant information. *Ann. Statist.* **12**, 971–988. Erratum: **14** (1986), 358.
16. Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. Verw. Gebiete* **14**, 323–330.
17. Jeganathan, P. (1995). Some aspects of asymptotic theory with applications to time series models. *Econometric Theory* **11**, 818–887.
18. Klaassen, C. A. J., Lee, E.-J. and Ruymgaart, F. H. (2001). On efficiency of indirect estimation of nonparametric regression functions. In: *Algebraic Methods in Statistics and Probability* (M. A. G. Viana and D. S. P. Richards, eds.), 173–184, Contemporary Mathematics 287, American Mathematical Society, Providence, RI.
19. Klaassen, C. A. J. and Putter, H. (2005). Efficient estimation of Banach parameters in semiparametric models. *Ann. Statist.* **33**, 307–346.
20. Koul, H. L. and Pflug, G. C. (1990). Weakly adaptive estimators in explosive autoregression. *Ann. Statist.* **18**, 939–960.
21. Koul, H. L. and Schick, A. (1996). Adaptive estimation in a random coefficient autoregressive model. *Ann. Statist.* **24**, 1025–1052.
22. Koul, H. L. and Schick, A. (1997). Efficient estimation in nonlinear autoregressive time-series models. *Bernoulli* **3**, 247–277.
23. Kreiss, J.-P. (1987a). On adaptive estimation in stationary ARMA processes. *Ann. Statist.* **15**, 112–133.
24. Kreiss, J.-P. (1987b). On adaptive estimation in autoregressive models when there are nuisance functions. *Statist. Decisions* **5**, 59–76.
25. Le Cam, L. (1971). Limits of experiments. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **l**, 245–261.
26. Levit, B. Y. (1975). Conditional estimation of linear functionals. *Problems Inform. Transmission* **11**, 39–54.
27. Mokkadem, A. (1987). Sur un modèle autorégressif non linéaire: ergodicité et ergodicité géométrique. *J. Time Ser. Anal.* **8**, 195–204.
28. Müller, U. U., Schick, A. and Wefelmeyer, W. (2004a). Estimating linear functionals of the error distribution in nonparametric regression. *J. Statist. Plann. Inference* **119**, 75–93.
29. Müller, U. U., Schick, A. and Wefelmeyer, W. (2004b). Estimating functionals of the error distribution in parametric and nonparametric regression. *J. Nonparametr. Statist.* **16**, 525–548.
30. Müller, U. U., Schick, A. and Wefelmeyer, W. (2005a). Weighted residual-based density estimators for nonlinear autoregressive models. *Statist. Sinica* **15**, 177–195.
31. Müller, U. U., Schick, A. and Wefelmeyer, W. (2005b). Efficient prediction for linear and nonlinear autoregressive models. Technical Report, Department of Mathematical Sciences, Binghamton University.
32. Müller, U. U., Schick, A. and Wefelmeyer, W. (2005c). Estimating the error distribution function in nonparametric regression. Technical Report, Department of Mathematical Sciences, Binghamton University.

33. Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.
34. Owen, A. B. (2001). *Empirical Likelihood*. Monographs on Statistics and Applied Probability **92**, Chapman & Hall / CRC, London.
35. Saavedra, A. and Cao, R. (1999). Rate of convergence of a convolution-type estimator of the marginal density of an MA(1) process. *Stochastic Process. Appl.* **80**, 129–155.
36. Saavedra, A. and Cao, R. (2000). On the estimation of the marginal density of a moving average process. *Canad. J. Statist.* **28**, 799–815.
37. Schick, A. and Wefelmeyer, W. (2002a). Estimating the innovation distribution in nonlinear autoregressive models. *Ann. Inst. Statist. Math.* **54**, 245–260.
38. Schick, A. and Wefelmeyer, W. (2002b). Efficient estimation in invertible linear processes. *Math. Methods Statist.* **11**, 358–379.
39. Schick, A. and Wefelmeyer, W. (2004a). Root n consistent and optimal density estimators for moving average processes. *Scand. J. Statist.* **31**, 63–78.
40. Schick, A. and Wefelmeyer, W. (2004b). Estimating invariant laws of linear processes by U-statistics. *Ann. Statist.* **32**, 603-632.
41. Schick, A. and Wefelmeyer, W. (2004c). Root n consistent density estimators for sums of independent random variables. *J. Nonparametr. Statist.* **16**, 925–935.
42. Schick, A. and Wefelmeyer, W. (2004d). Functional convergence and optimality of plug-in estimators for stationary densities of moving average processes. *Bernoulli* **10**, 889–917.
43. Schick, A. and Wefelmeyer, W. (2005). Simple root-n consistent density estimators for invertible linear processes: The sup-norm. Technical Report, Department of Mathematical Sciences, Binghamton University.
44. Taniguchi, M. and Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series*. Springer Series in Statistics, Springer, New York.
45. Van Keilegom, I., Akritas, M. G. and Veraverbeke, N. (2001). Estimation of the conditional distribution in regression with censored data: a comparative study. *Comput. Statist. Data Anal.* **35**, 487–500.
46. Van Keilegom, I. and Veraverbeke, N. (2001). Hazard rate estimation in nonparametric regression with censored data. *Ann. Inst. Statist. Math.* **53**, 730–745.
47. Van Keilegom, I. and Veraverbeke, N. (2002). Density and hazard estimation in censored regression models. *Bernoulli* **8**, 607–625.
48. Wefelmeyer, W. (1994). An efficient estimator for the expectation of a bounded function under the residual distribution of an autoregressive process. *Ann. Inst. Statist. Math.* **46**, 309–315.