Estimating the innovation distribution in nonlinear autoregressive models

Anton Schick [*] Wolfgang Wefelmeyer [†]

Abstract

The usual estimator for the expectation of a function under the innovation distribution of a nonlinear autoregressive model is the empirical estimator based on estimated innovations. It can be improved by exploiting that the innovation distribution has mean zero. We show that the resulting estimator is efficient if the innovations are estimated with an efficient estimator for the autoregression parameter. Efficiency of this estimator is necessary except when the expectation of the function can be estimated adaptively. Analogous results hold for heteroscedastic models.

AMS 2000 subject classifications. Primary 62M05, 62M10; secondary 62G20.

Key words and Phrases. Constrained model, empirical estimator, influence function, Markov chain model, semiparametric model.

1 Introduction

Let $X_{1-p}, \dots, X_n$ be observations from a stationary and ergodic (homoscedastic) nonlinear autoregressive model of order p,

$$X_j = r(\vartheta, \mathbf{X}_{j-1}) + \varepsilon_j, \quad j \geq 1, \tag{1.1}$$

where $\mathbf{X}_{j-1} = (X_{j-p}, \dots, X_{j-1})$, and the innovations ε_j are independent and identically distributed with mean 0 and finite second moment. Suppose we want to estimate the

[*] Binghamton University, Department of Mathematical Sciences, Binghamton, NY 13902-6000, USA, anton@math.binghamton.edu, http://math.binghamton.edu/anton/index.html. Research partially supported by NSF Grant DMS 0072174.

[†] Universität Siegen, Fachbereich 6 Mathematik, Walter-Flex-Str. 3, 57068 Siegen, Germany, wefelmeyer@mathematik.uni-siegen.de, http://www.math.uni-siegen.de/statistik/wefelmeyer.html.

expectation $E[h(\varepsilon_1)]$ of some square-integrable function h under the innovation distribution. The usual estimator is the empirical estimator $\frac{1}{n}\sum_{j=1}^n h(\hat{\varepsilon}_j)$ based on the estimated innovations $\hat{\varepsilon}_j = X_j - r(\hat{\vartheta}_n, \mathbf{X}_{j-1})$, where $\hat{\vartheta}_n$ is some estimator of ϑ.

If ϑ is known, we can improve the empirical estimator $\frac{1}{n}\sum_{j=1}^n h(X_j - r(\vartheta, \mathbf{X}_{j-1}))$ using the constraint $E[\varepsilon_1] = 0$. If ϑ is unknown, it suggests itself to replace ϑ by an estimator. In Theorem 1 of Section 2 we give conditions for asymptotic linearity of the resulting estimator for $E[h(\varepsilon_1)]$. In Theorem 4 of Section 3 we characterize efficiency of this estimator. Section 4 states corresponding results for *heteroscedastic* nonlinear regression models. Section 5 contains examples. The proofs of Theorems 1 and 3 are in Section 6.

The results may be viewed as instances of the following general principle, proved for the i.i.d. case by Klaassen and Putter (1999). Suppose a model is parametrized by (ϑ, F). Let $F_{n\vartheta}$ be efficient for F if ϑ is known, and let $\hat{\vartheta}_n$ be efficient for ϑ. Then $F_{n\hat{\vartheta}_n}$ is efficient for F.

2 Asymptotically linear estimators

To begin we recall results on constrained models for independent and identically distributed observations. Let $\varepsilon_1, \dots, \varepsilon_n$ be i.i.d. with distribution function F fulfilling the constraint $E[\psi(\varepsilon_1)] = 0$ for some vector-valued F-square-integrable function ψ such that $E[\psi(\varepsilon_1)\psi^\top(\varepsilon_1)] = \int \psi\psi^\top \, dF$ is invertible. Let h be an F-square-integrable function. Because of the constraint, we can write the expectation $E[h(\varepsilon_1)]$ for each vector a as $H(a, F) = E[h(\varepsilon_1) - a^\top \psi(\varepsilon_1)]$. The obvious estimator for F is the empirical estimator $F_n(x) = \frac{1}{n}\sum_{j=1}^n 1(\varepsilon_j \le x)$. We obtain unbiased estimators of $E[h(\varepsilon_1)]$ by replacing F in $H(a, F)$ by the empirical estimator,

$$H(a, F_n) = \frac{1}{n}\sum_{j=1}^n (h(\varepsilon_j) - a^\top \psi(\varepsilon_j)). \tag{2.1}$$

It is easy to check that the smallest asymptotic variance in this class of estimators is achieved by $a = a_h(F)$ with

$$a_h(F) = (E[\psi(\varepsilon_1)\psi(\varepsilon_1)^\top])^{-1} E[\psi(\varepsilon_1)h(\varepsilon_1)]. \tag{2.2}$$

Under our assumptions, $a_h(F_n)$ is consistent for $a_h(F)$. Hence $H(a_h(F_n), F_n)$ has the same asymptotic variance as the best unbiased "estimator" $H(a_h(F), F_n)$ in the class (2.1). The estimator $H(a_h(F_n), F_n)$ is efficient by Levit (1975).

Consider now the (homoscedastic) nonlinear autoregressive model of order p. By this model we mean a strictly stationary and ergodic time series X_j, $j \geq 1-p$, which satisfies the structural relation (1.1). Here ε_j, $j \geq 1$, are i.i.d. with unknown distribution function F and independent of the initial observations $\mathbf{X}_0$. Let G denote the stationary distribution of $\mathbf{X}_0$. We assume that F has mean 0 and finite second moment. The parameter ϑ is unknown and belongs to some open subset Θ of $\mathbf{R}^k$.

Suppose first that the parameter ϑ is *known*. Write the innovations ε_j as functions of the observations, $\varepsilon_j(\vartheta) = X_j - r(\vartheta, \mathbf{X}_{j-1})$. The constraint on the innovation distribution is $E[\varepsilon_1] = 0$. For each real a we obtain an unbiased estimator for $E[h(\varepsilon_1)]$,

$$H(a, F_{n\vartheta}) = \frac{1}{n} \sum_{j=1}^{n} (h(\varepsilon_j(\vartheta)) - a\varepsilon_j(\vartheta)),$$

where $F_{n\vartheta}(x) = \frac{1}{n} \sum_{j=1}^{n} 1(\varepsilon_j(\vartheta) \leq x)$. The constraint on F can be written $E[\psi(\varepsilon_1)] = 0$ for $\psi(x) = x$. By the above results on constrained models, the asymptotic variance of $H(a, F_{n\vartheta})$ is minimized by $a = a_h(F) = E[\varepsilon_1 h(\varepsilon_1)]/E[\varepsilon_1^2]$, and $H_n(\vartheta) = H(a_h(F_{n\vartheta}), F_{n\vartheta})$ is efficient, with asymptotic variance

$$\sigma_0^2 = E[h(\varepsilon_1)^2] - (E[h(\varepsilon_1)])^2 - (E[\varepsilon_1 h(\varepsilon_1)])^2 \Big/ E[\varepsilon_1^2]. \tag{2.3}$$

We are interested in the model with ϑ *unknown*. Assuming that $\hat{\vartheta}_n$ is asymptotically linear, we show in Theorem 1 that $H_n(\hat{\vartheta}_n)$ is asymptotically linear, and calculate its influence function. Call an estimator $\hat{\kappa}_n$ of an m-dimensional functional $\kappa(\vartheta, F)$ *asymptotically linear* at (ϑ, F) with *influence function* χ (or $\chi(\mathbf{X}_0, X_1)$) if $\chi : \mathbf{R}^{p+1} \to \mathbf{R}^m$ with $E[\chi(\mathbf{X}_0, X_1)|\mathbf{X}_0] = 0$ and $E[\|\chi(\mathbf{X}_0, X_1)\|^2] < \infty$, and if

$$n^{1/2}(\hat{\kappa}_n - \kappa(\vartheta, F)) = n^{-1/2} \sum_{j=1}^{n} \chi(\mathbf{X}_{j-1}, X_j) + o_{P_n}(1). \tag{2.4}$$

By the martingale central limit theorem, an estimator with influence function χ is asymptotically normal with covariance matrix

$$V_\chi = E[\chi(\mathbf{X}_0, X_1)\chi(\mathbf{X}_0, X_1)^\top]. \tag{2.5}$$

For ϑ known, the estimator $H_n(\vartheta)$ is asymptotically linear with influence function

$$h(\varepsilon_1) - a_h(F)\varepsilon_1 - E[h(\varepsilon_1)]. \tag{2.6}$$

Now replace ϑ by an asymptotically linear estimator $\hat\vartheta_n$. The influence function of $H_n(\hat\vartheta_n)$ is given in Theorem 1. We use the following assumptions. They say that $r(\vartheta, x)$ is differentiable in ϑ in an appropriate sense and that the function h has a smooth derivative.

Assumption 1. There is a G-square-integrable function $\dot r$ such that, for each constant C,

$$(2.7)\qquad \sup_{\|\Delta\|\le Cn^{-1/2}} \sum_{j=1}^n \Big(r(\vartheta+\Delta, \mathbf{X}_{j-1}) - r(\vartheta, \mathbf{X}_{j-1}) - \Delta^\top \dot r(\mathbf{X}_{j-1})\Big)^2 = o_{P_n}(1).$$

Assumption 2. The function h is absolutely continuous and F-square-integrable, and its (almost everywhere) derivative h' is F-square-integrable and satisfies

$$\int \sup_{|a|\le\eta} (h'(x-a)-h'(x))^2 dF(x) \to 0 \quad \text{as } \eta\to 0.$$

Theorem 1. *Suppose Assumptions 1 and 2 hold. Let $\hat\vartheta_n$ be asymptotically linear for ϑ, with influence function $\chi(\mathbf{X}_0, X_1)$. Then $H_n(\hat\vartheta_n)$ is asymptotically linear for $E[h(\varepsilon_1)]$, with influence function*

$$h(\varepsilon_1) - a_h(F)\varepsilon_1 - E[h(\varepsilon_1)] - (E[h'(\varepsilon_1)] - a_h(F))E[\dot r(\mathbf{X}_0)^\top]\chi(\mathbf{X}_0, X_1).$$

If $E[h'(\varepsilon_1)] = a_h(F)$ or $E[\dot r(\mathbf{X}_0)] = 0$, the influence function reduces to the influence function (2.6) of $H_n(\vartheta)$. In general, none of these two conditions holds, and the asymptotic variance of $H_n(\hat\vartheta_n)$ depends on $\hat\vartheta_n$ through its influence function $\chi(\mathbf{X}_0, X_1)$.

We have assumed that h is differentiable. This excludes the interesting case $h(x) = 1[x\le t]$, for which $E[h(\varepsilon_1)] = F(t)$. To treat this case, one can rely on expansions of $F_{n\hat\vartheta_n}$ available in the literature; see Koul (1996) for general nonlinear models, and Boldin (1982), Koul (1992, Chapter 7) and Koul and Leventhal (1989) for linear autoregressive models. General empirical processes involving "pseudo-observations" $\hat\varepsilon_j = \hat f(X_j)$ are studied in Ghoudi and Rémillard (1998). For completeness we describe such an expansion. For this we strengthen Assumption 1 and require smoothness of F. More precisely, we require the following.

(a1) There is a G-square-integrable function $\dot r$ such that, for each constant C,

$$\max_{1\le j\le n} \sup_{\|\Delta\|\le Cn^{-1/2}} \Big| r(\vartheta+\Delta, \mathbf{X}_{j-1}) - r(\vartheta, \mathbf{X}_{j-1}) - \Delta^\top \dot r(\mathbf{X}_{j-1})\Big| = o_{P_n}(n^{-1/2}).$$

(a2) The distribution function F has a positive and uniformly continuous density f.

(a3) The estimator $\hat{\vartheta}_n$ is $n^{1/2}$-consistent for ϑ.

Under these assumptions we have the expansion

$$\sup_{t\in\mathbf{R}}\left|F_{n\hat{\vartheta}_n}(t)-F_{n\vartheta}(t)-f(t)E[\dot{r}(\mathbf{X}_0)^\top](\hat{\vartheta}_n-\vartheta)\right|=o_{P_n}(n^{-1/2}).$$

This is essentially Corollary 1.6 in Koul (1996), except that we have replaced his condition (h1) by the weaker condition (a1). Inspection of his proof shows that our (a1) is sufficient to guarantee the critical requirements (3.8) and (3.11) needed in his proof. Koul (1996) also constructs $n^{1/2}$-consistent estimators for ϑ.

For $h(x)=1[x\le t]$ we have $a_h(F)=E[\varepsilon_1 1(\varepsilon_1\le t)]/E[\varepsilon_1^2]=a_t(F)$, say. The empirical estimator $a_t(F_{n\hat{\vartheta}_n})=\sum_{j=1}^n\hat{\varepsilon}_j 1(\hat{\varepsilon}_j\le t)\Big/\sum_{j=1}^n\hat{\varepsilon}_j^2$ is consistent for $a_t(F)$ uniformly in t in the sense that $\sup_{t\in\mathbf{R}}|a_t(F_{n\hat{\vartheta}_n})-a_t(F)|=o_{P_n}(1)$. Thus we obtain that the improved empirical distribution function $F^*_{n\hat{\vartheta}_n}=F_{n\hat{\vartheta}_n}-a_t(F_{n\hat{\vartheta}_n})\frac{1}{n}\sum_{j=1}^n\hat{\varepsilon}_j$ admits the expansion

$$(2.8)\quad \sup_{t\in\mathbf{R}}\left|F^*_{n\hat{\vartheta}_n}(t)-F_{n\vartheta}(t)-(f(t)-a_t(F))E[\dot{r}(\mathbf{X}_0)^\top](\hat{\vartheta}_n-\vartheta)\right|=o_{P_n}(n^{-1/2}).$$

For asymptotically linear $\hat{\vartheta}_n$ this gives the desired result corresponding to Theorem 1.

3 Efficient and adaptive estimators

In this section we characterize efficient and adaptive estimators among those of the form $H_n(\hat{\vartheta}_n)$. This requires the nonlinear autoregressive model (1.1) to be locally asymptotically normal. We need an additional assumption, finiteness of the Fisher information for location. As in Section 2, fix a distribution function F with mean 0 and finite second moment.

Assumption 3. The distribution function F has positive and absolutely continuous density f with finite Fisher information for location: $E[\ell_1^2(\varepsilon_1)]=\int\ell_1^2\,dF<\infty$, with $\ell_1(x)=-f'(x)/f(x)$.

Remark 1. Under Assumptions 2 and 3 we have $E[h'(\varepsilon_1)]=E[\ell_1(\varepsilon_1)h(\varepsilon_1)]$.

Local asymptotic normality requires local perturbations of the model around the true parameter (ϑ,F). The perturbed parameters must still be in the parameter space.

Introduce local parameters $w = (u, v)$, with $u \in \mathbf{R}^k$ and v in the space $\mathcal{V}$ of F-square-integrable functions fulfilling

$$\int E[\xi(\varepsilon_1)v(\varepsilon_1)] = 0, \quad \text{with } \xi(x) = (1, x)^\top.$$

Set $\vartheta_{nu} = \vartheta + n^{-1/2}u$, and let F_{nv} be the distribution function with density

$$f_{nv}(x) = f(x)(1 + n^{-1/2}v_n(x)).$$

Here v_n is defined as follows: With $a_n = n^{1/8}$ and φ the standard normal density, let $\xi_n(x) = (1, (-a_n) \vee x \wedge a_n)^\top$ be a trimmed version of $\xi(x)$, let $\tilde{v}_n(x) = (-a_n) \vee v(x) \wedge a_n$ be a trimmed version of $v(x)$, let $\bar{v}_n(x) = \int \tilde{v}_n(x - a_n^{-1}y)\varphi(y)dy$ be a smoothed and trimmed version of $v(x)$, and then define

$$v_n(x) = \bar{v}_n(x) - E[\bar{v}_n(\varepsilon_1)\xi(\varepsilon_1)^\top](E[\xi_n(\varepsilon_1)\xi(\varepsilon_1)^\top])^{-1}\xi_n(x).$$

The following properties are easy to check. The function v_n is absolutely continuous, $|v_n| \leq Cn^{1/8}$ and $|v_n'| \leq Cn^{1/4}$ for some finite constant C, and $\int v_n \xi^\top dF = 0$ and $\int (v_n - v)^2 dF \to 0$. In particular, F_{nv} fulfills Assumption 3, has zero mean and finite variance. Moreover,

$$n^{1/2} \int h(x)(f_{nv}(x) - f(x))dx = E[h(\varepsilon_1)v_n(\varepsilon_1)] \to E[h(\varepsilon_1)v(\varepsilon_1)].$$

Since F has a density, the stationary distribution G of $\mathbf{X}_j$ has a density, say $g = g_{\vartheta,F}$. Write g_{nw} for the density $g_{\vartheta_{nu},F_{nv}}$. Write P_{nw} for the joint law of $X_{1-p}, \dots, X_n$ if ϑ_{nu} and F_{nv} are the underlying parameters. Consider the local log-likelihood ratio

$$\log \frac{dP_{nw}}{dP_n} = \log \frac{g_{nw}(\mathbf{X}_0)}{g(\mathbf{X}_0)} + \sum_{j=1}^n \log \frac{f_{nv}(\varepsilon_j(\vartheta_{nu}))}{f(\varepsilon_j(\vartheta))}.$$

Here the random variables $\varepsilon_j(\vartheta) = X_j - r(\vartheta, \mathbf{X}_{j-1})$ are the innovations, written as functions of the observations. We have local asymptotic normality as follows. For the proof see Koul and Schick (1997). They also give conditions for smoothness (3.1) of the stationary density.

Theorem 2. *Let $w = (u, v) \in \mathbf{R}^k \times \mathcal{V}$. Suppose Assumptions 1 and 3 hold and the stationary density depends smoothly on the parameter,*

$$\int |g_{nw}(\mathbf{x}) - g(\mathbf{x})| d\mathbf{x} \to 0. \tag{3.1}$$

Then

$$\log \frac{dP_{nw}}{dP_n} = n^{-1/2} \sum_{j=1}^n S_w(\mathbf{X}_{j-1}, \varepsilon_j) - \frac{1}{2} R_w + o_{P_n}(1), \tag{3.2}$$

$$n^{-1/2} \sum_{j=1}^n S_w(\mathbf{X}_{j-1}, \varepsilon_j) \Rightarrow N(0, R_w) \quad \textit{under } P_n, \tag{3.3}$$

where $N(0, R_w)$ is normal with mean 0 and variance R_w, and where

$$\begin{aligned} S_w(\mathbf{X}_0, \varepsilon_1) &= v(\varepsilon_1) + u^\top \dot r(\mathbf{X}_0)\ell_1(\varepsilon_1), \\ R_w &= E[S_w(\mathbf{X}_0, \varepsilon_1)^2]. \end{aligned}$$

From now on we view $S_w(\mathbf{X}_0, \varepsilon_1)$ as an element of the Hilbert space $L_2(G \times F)$, and $\mathcal{V}$ as a subset of the Hilbert space $L_2(F)$. The *tangent space*

$$T = \{S_w(\mathbf{X}_0, \varepsilon_1) : w = (u, v) \in \mathbf{R}^k \times \mathcal{V}\}$$

is a closed linear subspace of $L_2(G \times F)$.

Consider now the problem of estimating an m-dimensional functional $\kappa(\vartheta, F)$. We say that κ is *differentiable* at (ϑ, F) with *gradient* $\dot\kappa$ if $E[\|\dot\kappa(\mathbf{X}_0, X_1)\|^2] < \infty$ and for all $w = (u, v) \in \mathbf{R}^k \times \mathcal{V}$,

$$n^{1/2}(\kappa(\vartheta_{nu}, F_{nv}) - \kappa(\vartheta, F)) \to E[\dot\kappa(\mathbf{X}_0, X_1) S_w(\mathbf{X}_0, \varepsilon_1)]. \tag{3.4}$$

The function $\dot\kappa$ is not uniquely determined, but its projection $\dot\kappa_0$ onto T^m is. We call $\dot\kappa_0$ the *canonical gradient*, and assume that $V_{\dot\kappa_0} = E[\dot\kappa_0(\mathbf{X}_0, X_1)\dot\kappa_0(\mathbf{X}_0, X_1)^\top]$ is positive definite.

Now let $\hat\kappa_n$ be an estimator of κ. We say that $\hat\kappa_n$ is *regular* at (ϑ, F) with *limit* L if

$$n^{1/2}(\hat\kappa_n - \kappa(\vartheta, F)) \Rightarrow L \quad \text{under } P_{nw} \text{ for all } w = (u, v) \in \mathbf{R}^k \times \mathcal{V}.$$

The convolution theorem of Hájek (1970) in the version of Bickel et al. (1998, Section 2.3) implies the following three results:

1. The distribution of L is a convolution,

$$L = N(0, V_{\dot\kappa_0}) + K \quad \text{in distribution},$$

with K independent of $N(0, V_{\dot\kappa_0})$.

2. A regular estimator $\hat{\kappa}_n$ has limit $L = N(0, V_{\dot{\kappa}_0})$ if and only if $\hat{\kappa}_n$ is asymptotically linear (2.4) with influence function χ equal to the canonical gradient $\dot{\kappa}_0$.

3. An asymptotically linear estimator is regular if and only if its influence function is a gradient.

An estimator with limit $L = N(0, V_{\dot{\kappa}_0})$ is least dispersed among all regular estimators. We call such an estimator *efficient*. It follows from 1. to 3. that an estimator is regular and efficient if and only if it is asymptotically linear with influence function equal to the canonical gradient,

$$n^{1/2}(\hat{\kappa}_n - \kappa(\vartheta, F)) = n^{-1/2} \sum_{j=1}^{n} \dot{\kappa}_0(\mathbf{X}_{j-1}, X_j) + o_{P_n}(1). \tag{3.5}$$

We apply the characterization (3.5) to estimators of ϑ and of $E[h(\varepsilon_1)]$. To calculate the corresponding canonical gradients, it is convenient to decompose the elements $S_w(\mathbf{X}_0, \varepsilon_1)$ in the tangent space T into orthogonal components,

$$S_w(\mathbf{X}_0, \varepsilon_1) = v(\varepsilon_1) + u^\top E[\dot{r}(\mathbf{X}_0)](\ell_1(\varepsilon_1) - a_{\ell_1}(F)\varepsilon_1) + u^\top S(\mathbf{X}_0, \varepsilon_1) \tag{3.6}$$

with

$$S(\mathbf{X}_0, \varepsilon_1) = (\dot{r}(\mathbf{X}_0) - E[\dot{r}(\mathbf{X}_0)])\ell_1(\varepsilon_1) + E[\dot{r}(\mathbf{X}_0)]a_{\ell_1}(F)\varepsilon_1 \tag{3.7}$$

and $a_{\ell_1}(F) = E[\varepsilon_1 \ell_1(\varepsilon_1)]/E[\varepsilon_1^2]$. From Remark 1 we obtain

$$a_{\ell_1}(F) = (E[\varepsilon_1^2])^{-1}. \tag{3.8}$$

By construction, $\ell_1(\varepsilon_1) - a_{\ell_1}(F)\varepsilon_1$ is in $\mathcal{V}$, and $S(\mathbf{X}_0, \varepsilon_1)$ is orthogonal to $\mathcal{V}$,

$$E[S(\mathbf{X}_0, \varepsilon_1)v(\varepsilon_1)] = 0 \quad \text{for } v \in \mathcal{V}. \tag{3.9}$$

We assume from now on that the dispersion matrix $\Lambda = E[S(\mathbf{X}_0, \varepsilon_1)S(\mathbf{X}_0, \varepsilon_1)^\top]$ is positive definite.

Consider first the problem of estimating the parameter ϑ. We view the parameter as a functional $\kappa(\vartheta, F) = \vartheta$. We have

$$n^{1/2}(\kappa(\vartheta_{nu}, F_{nv}) - \kappa(\vartheta, F)) = u, \tag{3.10}$$

and by decomposition (3.6) and orthogonality (3.9),

$$\Lambda^{-1}E[S(\mathbf{X}_0, \varepsilon_1)S_w(\mathbf{X}_0, \varepsilon_1)] = \Lambda^{-1}E[S(\mathbf{X}_0, \varepsilon_1)S(\mathbf{X}_0, \varepsilon_1)^\top]u = u. \tag{3.11}$$

Hence $\Lambda^{-1} S(\mathbf{X}_0, \varepsilon_1)$ is a gradient of $\kappa(\vartheta, F) = \vartheta$. The gradient is canonical since

$$a^\top S(\mathbf{X}_0, \varepsilon_1) = S_w(\mathbf{X}_0, \varepsilon_1) \quad \text{for } w = \Big(a,\ -a^\top E[\dot r(\mathbf{X}_0)](\ell_1(\varepsilon_1) - a_{\ell_1}(F)\varepsilon_1)\Big).$$

By characterization (3.5), an estimator $\hat\vartheta_n$ is regular and efficient if and only if it has influence function $\Lambda^{-1} S(\mathbf{X}_0, \varepsilon_1)$,

$$n^{1/2}(\hat\vartheta_n - \vartheta) = \Lambda^{-1} n^{-1/2} \sum_{j=1}^{n} S(\mathbf{X}_{j-1}, \varepsilon_j) + o_{P_n}(1). \tag{3.12}$$

The asymptotic covariance matrix is then Λ^{-1}.

Efficient estimators for ϑ are constructed in Kreiss (1987a), (1987b) for AR(p) and ARMA(p, q) models, and in Drost, Klaassen and Werker (1997) and Koul and Schick (1997) for nonlinear autoregressive models.

Consider now the problem of estimating the functional $\kappa(\vartheta, F) = E[h(\varepsilon_1)]$. The canonical gradient is given in the following theorem.

Theorem 3. *Suppose Assumptions 1 and 3 hold. Then the functional $E[h(\varepsilon_1)]$ is differentiable at (ϑ, F) with canonical gradient*

$$\begin{aligned} \dot\kappa_0(\mathbf{X}_0, X_1) &= h(\varepsilon_1) - a_h(F)\varepsilon_1 - E[h(\varepsilon_1)] \\ &\quad - E\Big[\Big(h(\varepsilon_1) - a_h(F)\varepsilon_1\Big)\ell_1(\varepsilon_1)\Big] E[\dot r(\mathbf{X}_0)^\top]\Lambda^{-1} S(\mathbf{X}_0, \varepsilon_1). \end{aligned} \tag{3.13}$$

For ϑ known we have local asymptotic normality (3.2), (3.3) with $u = 0$, and the functional $E[h(\varepsilon_1)]$ has canonical gradient $h(\varepsilon_1) - a_h(F)\varepsilon_1 - E[h(\varepsilon_1)]$. This canonical gradient equals the canonical gradient for unknown ϑ if and only if $h(\varepsilon_1) - a_h(F)\varepsilon_1$ and $\ell_1(\varepsilon_1)$ are uncorrelated or $E[\dot r(\mathbf{X}_0)^\top] = 0$. In these cases, $E[h(\varepsilon_1)]$ can be estimated as well not knowing ϑ as knowing ϑ. One says then that $E[h(\varepsilon_1)]$ can be estimated *adaptively* with respect to ϑ.

By Remark 1, applied with our h and then with $h(x) = ax$,

$$E[h'(\varepsilon_1)] - a = E\Big[\Big(h(\varepsilon_1) - a\varepsilon_1\Big)\ell_1(\varepsilon_1)\Big]. \tag{3.14}$$

Hence the influence function of $H_n(\hat\vartheta_n)$ in Theorem 1 can be written

$$h(\varepsilon_1) - a_h(F)\varepsilon_1 - E[h(\varepsilon_1)] - E\Big[\Big(h(\varepsilon_1) - a_h(F)\varepsilon_1\Big)\ell_1(\varepsilon_1)\Big] E[\dot r(\mathbf{X}_0)^\top]\, \chi(\mathbf{X}_0, X_1).$$

In particular, if $E[h(\varepsilon_1)]$ can be estimated adaptively, then $H_n(\hat\vartheta_n)$ is efficient for any asymptotically linear $\hat\vartheta_n$. A look at the proof of Theorem 1 shows that $n^{1/2}$-consistency of $\hat\vartheta_n$ would suffice.

In general, $E[h(\varepsilon_1)]$ cannot be estimated adaptively. Let $\hat\vartheta_n$ have influence function χ. Assume also that $\hat\vartheta_n$ is regular. By the characterization of regular estimators, χ must be a gradient of $\kappa(\vartheta, F) = \vartheta$. Since $h(\varepsilon_1) - a_h(F)\varepsilon_1 - E[h(\varepsilon_1)]$ is in $\mathcal{V}$, it follows from relation (3.10) and definition (3.4) that $h(\varepsilon_1) - a_h(F)\varepsilon_1$ and $\chi(\mathbf{X}_0, X_1)$ are uncorrelated. Hence the asymptotic variance of $H_n(\hat\vartheta_n)$ is

$$\sigma^2 = \sigma_0^2 + \Big(E\Big[\Big(h(\varepsilon_1) - a_h(F)\varepsilon_1\Big)\ell_1(\varepsilon_1)\Big]\Big)^2 E[\dot r(\mathbf{X}_0)^\top] V_\chi \, E[\dot r(\mathbf{X}_0)],$$

where σ_0^2 is the asymptotic variance (2.3) of $H_n(\vartheta)$, and V_χ is the asymptotic covariance matrix (2.5) of $\hat\vartheta_n$. The minimal V_χ is Λ^{-1}. We arrive at the following result.

Theorem 4. *Suppose Assumptions 1 to 3 hold.*

1. *The functional $E[h(\varepsilon_1)]$ can be estimated adaptively with respect to ϑ if and only if $E\Big[\Big(h(\varepsilon_1) - a_h(F)\varepsilon_1\Big)\ell_1(\varepsilon_1)\Big] = 0$ or $E[\dot r(\mathbf{X}_0)] = 0$. Then the estimator*

$$H_n(\hat\vartheta_n) = \frac{1}{n}\sum_{j=1}^n \Big(h(\hat\varepsilon_j) - \frac{\sum_{j=1}^n \hat\varepsilon_j h(\hat\varepsilon_j)}{\sum_{j=1}^n \hat\varepsilon_j^2}\hat\varepsilon_j\Big)$$

is efficient whenever $\hat\vartheta_n$ is $n^{1/2}$-consistent.

2. *Suppose $E[h(\varepsilon_1)]$ cannot be estimated adaptively. Let $\hat\vartheta_n$ be regular and asymptotically linear for ϑ, with influence function χ. Then $H_n(\hat\vartheta_n)$ is efficient if and only if the i-th component $\hat\vartheta_{ni}$ of $\hat\vartheta_n$ is efficient whenever $E[\dot r_i(\mathbf{X}_0)] \neq 0$.*

We have mentioned at the end of Section 2 that a version of Theorem 1, namely relation (2.8), holds for $h(x) = 1[x \le t]$ under appropriately modified assumptions. For this function we have $E[h(\varepsilon_1)] = F(t)$. The improved empirical distribution function $F^*_{n\hat\vartheta_n}(t)$ introduced there is efficient for $F(t)$ provided $\hat\vartheta_{ni}$ is also efficient whenever $E[\dot r_i(\mathbf{X}_0)] \neq 0$. This follows immediately from the observation that $f(t) = -E[\ell_1(\varepsilon_1)1(\varepsilon_1 \le t)]$. Efficiency holds even in the functional sense discussed in Bickel et al. (1998, Section 5.2; see also Schick and Susarla, 1990).

4 Heteroscedastic nonlinear autoregression

The results of Sections 2 and 3 generalize to *heteroscedastic* nonlinear autoregression. By a heteroscedastic nonlinear autoregressive model of order p we mean a strictly stationary and ergodic time series X_j, $j \geq 1-p$, which satisfies the structural relation

$$X_j = r(\vartheta, \mathbf{X}_{j-1}) + s(\vartheta, \mathbf{X}_{j-1})\varepsilon_j, \quad j \geq 1. \tag{4.1}$$

Again, ε_j are i.i.d. with unknown distribution function F, and independent of the initial observations $\mathbf{X}_0$. We write again G for the stationary distribution of $\mathbf{X}_0$. We now assume that F has mean 0 and variance 1. We also need that F has finite fourth moment. Assumptions 1 and 2 are replaced by the following two assumptions.

Assumption 4. The functions $r(\vartheta, \mathbf{X}_0)$ and $s(\vartheta, \mathbf{X}_0)$ are differentiable in the sense of Assumption 1, with derivatives $\dot{r}(\mathbf{X}_0)$ and $\dot{s}(\mathbf{X}_0)$, respectively. Furthermore, $s(\vartheta, \mathbf{x})$ is bounded away from 0 over $\mathbf{x} \in \mathbf{R}^p$ and ϑ in compact subsets of Θ.

Assumption 5. The function h is absolutely continuous and F-square-integrable, and its (almost everywhere) derivative h' satisfies

$$\textstyle\int (1+x^2) h'(x)^2 dF(x) < \infty, \tag{4.2}$$

$$\textstyle\int (1+x^2) \sup_{a^2+b^2 \leq \eta} (h'(x-a-bx) - h'(x))^2 dF(x) \to 0 \quad \text{as } \eta \to 0. \tag{4.3}$$

Suppose first that ϑ is *known*. Write the innovations ε_j as functions of the observations,

$$\varepsilon_j(\vartheta) = \frac{X_j - r(\vartheta, \mathbf{X}_{j-1})}{s(\vartheta, \mathbf{X}_{j-1})}.$$

The constraint on F can be written $E[\psi(\varepsilon_1)] = 0$ for $\psi(x) = (x, x^2-1)^\top$. For each vector a we obtain an unbiased estimator for $E[h(\varepsilon_1)]$,

$$H(a, F_{n\vartheta}) = \frac{1}{n} \sum_{j=1}^n (h(\varepsilon_j(\vartheta)) - a^\top \psi(\varepsilon_j(\vartheta))),$$

where $F_{n\vartheta}(x) = \frac{1}{n}\sum_{j=1}^n 1(\varepsilon_j(\vartheta) \leq x)$. By the results on constrained models in Section 2, an efficient estimator for $E[h(\varepsilon_1)]$ is $H_n(\vartheta) = H(a_h(F_{n\vartheta}), F_{n\vartheta})$ with

$$\begin{aligned} a_h(F) &= \begin{pmatrix} 1 & E[\varepsilon_1^3] \\ E[\varepsilon_1^3] & E[\varepsilon_1^4]-1 \end{pmatrix}^{-1} \begin{pmatrix} E[\varepsilon_1 h(\varepsilon_1)] \\ E[(\varepsilon_1^2-1)h(\varepsilon_1)] \end{pmatrix} \\ &= \frac{1}{E[\varepsilon_1^4] - 1 - (E[\varepsilon_1^3])^2} \begin{pmatrix} (E[\varepsilon_1^4]-1)E[\varepsilon_1 h(\varepsilon_1)] - E[\varepsilon_1^3]E[(\varepsilon_1^2-1)h(\varepsilon_1)] \\ E[(\varepsilon_1^2-1)h(\varepsilon_1)] - E[\varepsilon_1^3]E[\varepsilon_1 h(\varepsilon_1)] \end{pmatrix}. \end{aligned} \tag{4.4}$$

Suppose now that ϑ is unknown. Replace ϑ in $H_n(\vartheta)$ by an asymptotically linear estimator $\hat{\vartheta}_n$. The influence function of $H_n(\hat{\vartheta}_n)$ is given in Theorem 5.

Theorem 5. *Suppose Assumptions 4 and 5 hold. Let $\hat{\vartheta}_n$ be asymptotically linear for ϑ, with influence function $\chi(\mathbf{X}_0, X_1)$. Then $H_n(\hat{\vartheta}_n)$ is asymptotically linear for $E[h(\varepsilon_1)]$, with influence function*

$$\begin{aligned} &h(\varepsilon_1) - a_h(F)^\top \psi(\varepsilon_1) - E[h(\varepsilon_1)] \\ &\qquad -E\Big[\Big(h'(\varepsilon_1) - a_h(F)^\top \psi'(\varepsilon_1)\Big)(1, \varepsilon_1)\Big] E[M(\mathbf{X}_0)^\top]\chi(\mathbf{X}_0, X_1), \end{aligned}$$

where $M(\mathbf{X}_0)$ is the $k \times 2$ matrix

$$M(\mathbf{X}_0) = \frac{1}{s(\vartheta, \mathbf{X}_0)}(\dot{r}(\mathbf{X}_0), \dot{s}(\mathbf{X}_0)).$$

To discuss efficiency of $H_n(\hat{\vartheta}_n)$, we need local asymptotic normality of the model, and the following heteroscedastic generalization of Assumption 3.

Assumption 6. The distribution function F has positive and absolutely continuous density f satisfying

$$\int (1 + x^2)\Big(\frac{f'(x)}{f(x)}\Big)^2 dF(x) < \infty.$$

Assumption 6 implies that the Fisher informations for location and scale are finite: $\int \ell_1^2 dF < \infty$ and $\int \ell_2^2 dF < \infty$, with $\ell_2(x) = -1 + x\ell_1(x)$. We set $\ell = (\ell_1, \ell_2)^\top$.

A version of Remark 1 holds for ℓ_2.

Remark 2. Under Assumptions 5 and 6 we have $E[\varepsilon_1 h'(\varepsilon_1)] = E[\ell_2(\varepsilon_1)h(\varepsilon_1)]$.

Local perturbations (ϑ_{nu}, F_{nv}) around the true parameter (ϑ, F) are introduced as in Section 3. Write again $\mathcal{V}$ for the local parameter space, which now consists of F-square-integrable functions v fulfilling

$$E[\xi(\varepsilon_1)v(\varepsilon_1)] = 0, \quad \text{with } \xi(x) = (1, x, x^2 - 1)^\top.$$

The local log-likelihood ratio

$$\log \frac{dP_{nw}}{dP_n} = \log \frac{g_{nw}(\mathbf{X}_0)}{g(\mathbf{X}_0)} + \sum_{j=1}^n \log \frac{f_{nv}(\varepsilon_j(\vartheta_{nu}))/s(\vartheta_{nu}, \mathbf{X}_{j-1})}{f(\varepsilon_j(\vartheta))/s(\vartheta, \mathbf{X}_{j-1})}$$

is asymptotically normal as follows.

Theorem 6. *Let $w = (u, v) \in \mathbf{R}^k \times \mathcal{V}$. Suppose Assumptions 4 and 6 hold and the stationary density g_{nw} fulfills (3.1). Then local asymptotic normality (3.2), (3.3) holds with $S_w(\mathbf{X}_0, \varepsilon_1) = v(\varepsilon_1) + u^\top M(\mathbf{X}_0)\ell(\varepsilon_1)$.*

The proof is similar to the proof of Theorem 2. For fixed nuisance parameter F, i.e., for $v = 0$, the theorem is proved in Jeganathan (1995) and in Drost, Klaassen and Werker (1997).

We decompose $S_w(\mathbf{X}_0, \varepsilon_1)$ into orthogonal components,

$$S_w(\mathbf{X}_0, \varepsilon_1) = v(\varepsilon_1) + u^\top E[M(\mathbf{X}_0)](\ell(\varepsilon_1) - a_\ell(F)^\top \psi(\varepsilon_1)) + u^\top S(\mathbf{X}_0, \varepsilon_1)$$

with

$$S(\mathbf{X}_0, \varepsilon_1) = (M(\mathbf{X}_0) - E[M(\mathbf{X}_0)])\ell(\varepsilon_1) + E[M(\mathbf{X}_0)]a_\ell(F)^\top \psi(\varepsilon_1) \tag{4.5}$$

and $a_\ell(F)$ a 2×2 matrix with columns defined as in (4.4) for $h = \ell_1$ and $h = \ell_2$. From Remarks 1 and 2 we obtain

$$\begin{aligned} a_\ell(F) &= (E[\psi(\varepsilon_1)\psi(\varepsilon_1)^\top])^{-1} E[\psi(\varepsilon_1)\ell(\varepsilon_1)] \\ &= \begin{pmatrix} 1 & E[\varepsilon_1^3] \\ E[\varepsilon_1^3] & E[\varepsilon_1^4] - 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \\ &= \frac{1}{E[\varepsilon_1^4] - 1 - (E[\varepsilon_1^3])^2} \begin{pmatrix} E[\varepsilon_1^4] - 1 & -2E[\varepsilon_1^3] \\ -E[\varepsilon_1^3] & 2 \end{pmatrix}. \end{aligned} \tag{4.6}$$

By construction, the components of $\ell(\varepsilon_1) - a_\ell(F)^\top \psi(\varepsilon_1)$ are in $\mathcal{V}$, and $S(\mathbf{X}_0, \varepsilon_1)$ is orthogonal to $\mathcal{V}$. We assume that the dispersion matrix $\Lambda = E[S(\mathbf{X}_0, \varepsilon_1)S(\mathbf{X}_0, \varepsilon_1)^\top]$ is positive definite. As in Section 3, the influence function for efficient estimators of ϑ is $\Lambda^{-1}S(\mathbf{X}_0, \varepsilon_1)$. The influence function for efficient estimators of $E[h(\varepsilon_1)]$ is given in the following theorem.

Theorem 7. *Suppose Assumptions 4 to 6 hold. Then the functional $E[h(\varepsilon_1)]$ is differentiable at (ϑ, F) with canonical gradient*

$$\dot{\kappa}_0(\mathbf{X}_0, X_1) = h(\varepsilon_1) - a_h(F)^\top \psi(\varepsilon_1) - E[h(\varepsilon_1)] - D^\top \Lambda^{-1} S(\mathbf{X}_0, \varepsilon_1)$$

with $D = E[M(\mathbf{X}_0)]E\Big[\Big(h(\varepsilon_1) - a_h(F)^\top \psi(\varepsilon_1)\Big)\ell(\varepsilon_1)\Big]$.

By Remarks 1 and 2,

$$E\Big[\Big(h'(\varepsilon_1) - a^\top \psi'(\varepsilon_1)\Big)(1, \varepsilon_1)^\top\Big] = E\Big[\Big(h(\varepsilon_1) - a^\top \psi(\varepsilon_1)\Big)\ell(\varepsilon_1)\Big]. \tag{4.7}$$

Hence, by Theorem 5, the influence function of $H_n(\hat\vartheta_n)$ can be written

$$h(\varepsilon_1) - a_h(F)^\top \psi(\varepsilon_1) - E[h(\varepsilon_1)] - D^\top \chi(\mathbf{X}_0, X_1).$$

We arrive at the following result.

Theorem 8. *Suppose Assumptions 4 to 6 hold.*

1. *The functional $E[h(\varepsilon_1)]$ can be estimated adaptively with respect to ϑ if and only if $D = 0$. Then the estimator $H_n(\hat\vartheta_n) = \frac{1}{n}\sum_{j=1}^n \Big(h(\hat\varepsilon_j) - \hat d_n^{-1}\Big(\hat c_{n1}\hat\varepsilon_j + \hat c_{n2}(\hat\varepsilon_j^2 - 1)\Big)\Big)$, with*

$$\begin{aligned}
\hat d_n &= \frac{1}{n}\sum_{j=1}^n \hat\varepsilon_j^4 - 1 - \Big(\frac{1}{n}\sum_{j=1}^n \hat\varepsilon_j^3\Big)^2,\\
\hat c_{n1} &= \Big(\frac{1}{n}\sum_{j=1}^n \hat\varepsilon_j^4 - 1\Big)\frac{1}{n}\sum_{j=1}^n \hat\varepsilon_j h(\hat\varepsilon_j) - \frac{1}{n}\sum_{j=1}^n \hat\varepsilon_j^3 \frac{1}{n}\sum_{j=1}^n (\hat\varepsilon_j^2 - 1) h(\hat\varepsilon_j),\\
\hat c_{n2} &= \frac{1}{n}\sum_{j=1}^n (\hat\varepsilon_j^2 - 1) h(\hat\varepsilon_j) - \frac{1}{n}\sum_{j=1}^n \hat\varepsilon_j^3 \frac{1}{n}\sum_{j=1}^n \hat\varepsilon_j h(\hat\varepsilon_j),
\end{aligned}$$

is efficient whenever $\hat\vartheta_n$ is $n^{1/2}$-consistent.

2. *Suppose $E[h(\varepsilon_1)]$ cannot be estimated adaptively. Let $\hat\vartheta_n$ be regular and asymptotically linear for ϑ, with influence function χ. Then $H_n(\hat\vartheta_n)$ is efficient if and only if $D^\top(V_\chi - \Lambda^{-1})D = 0$. This condition holds in particular if $\hat\vartheta_n$ is efficient, i.e., $V_\chi = \Lambda^{-1}$.*

In the non-adaptive case, a sufficient condition for efficiency of $H_n(\hat\vartheta_n)$ is that $\hat\vartheta_{ni}$ is efficient whenever $E[\dot r_i(\mathbf{X}_0)/s(\vartheta, \mathbf{X}_0)] \neq 0$ or $E[\dot s_i(\mathbf{X}_0)/s(\vartheta, \mathbf{X}_0)] \neq 0$. This is also necessary if none of the two components of $E\Big[\Big(h(\varepsilon_1) - a_h(F)^\top \psi(\varepsilon_1)\Big)\ell(\varepsilon_1)\Big]$ vanishes, which will be the case for most parameters (ϑ, F).

Efficient estimators of ϑ can be constructed along the lines of Drost, Klaassen and Werker (1997).

5 Examples

Example 1. The simplest example is linear autoregression of order p,

$$X_j = \vartheta^\top \mathbf{X}_{j-1} + \varepsilon_j,$$

where ε_j are i.i.d. with mean 0. The structural relation (1.1) is satisfied with $r(\vartheta, \mathbf{X}_0) = \vartheta^\top \mathbf{X}_0$. We have $\dot r(\mathbf{X}_0) = \mathbf{X}_0$ and $E[\dot r(\mathbf{X}_0)] = 0$. Hence the model is adaptive. Let $\hat\vartheta_n$

be $n^{1/2}$-consistent and $\hat{\varepsilon}_j = X_j - \hat{\vartheta}_n^\top \mathbf{X}_{j-1}$. Then the estimator $H_n(\hat{\vartheta}_n)$ in Theorem 4 is efficient. For $p = 1$ this was shown in Wefelmeyer (1994). An analogous result holds for linear *regression*; see Klaassen and Putter (1999, Example 5.3).

Example 2. A non-adaptive generalization of Example 1 is

$$X_j - \mu = \rho^\top (\mathbf{X}_{j-1} - \mu \mathbf{1}) + \varepsilon_j,$$

where ε_j are i.i.d. with mean 0, and $\mathbf{1} = (1, \dots, 1)^\top$. Here $\vartheta = (\rho^\top, \mu)^\top$ and $r(\mathbf{X}_0, \vartheta) = \mu + \rho^\top (\mathbf{X}_0 - \mu \mathbf{1})$. Assumption 1 holds with $\dot{r}(\mathbf{X}_0) = (\mathbf{X}_0 - \mu \mathbf{1}, 1 - \rho^\top \mathbf{1})^\top$. We have $E[\dot{r}(\mathbf{X}_0)] = (0, 1 - \rho^\top \mathbf{1})^\top$. Hence, by Theorem 4, an efficient estimator for $E[h(\varepsilon_1)]$ requires an efficient estimator for μ, but not for ρ.

Write $g = (g_\rho^\top, g_\mu)^\top$ for the canonical gradient of $\vartheta = (\rho^\top, \mu)^\top$. To calculate the canonical gradient g_μ of μ, note first that by (3.7) and (3.8),

$$S(\mathbf{X}_0, \varepsilon_1) = \begin{pmatrix} (\mathbf{X}_0 - \mu \mathbf{1}) \ell_1(\varepsilon_1) \\ (1 - \rho^\top \mathbf{1})(E[\varepsilon_1^2])^{-1} \varepsilon_1 \end{pmatrix}.$$

The covariance matrix Λ of $S(\mathbf{X}_0, \varepsilon_1)$ is diagonal, with $(1 - \rho^\top \mathbf{1})^2 (E[\varepsilon_1^2])^{-1}$ as lower right entry. Hence $g_\mu = (1 - \rho^\top \mathbf{1})^{-1} \varepsilon_1$. It is easy to check that this is the influence function of the empirical estimator $\overline{X}_n = \frac{1}{n} \sum_{j=1}^n X_j$, which is therefore efficient. Now estimate $\rho = (\rho_1, \dots, \rho_p)^\top$ by the empirical autocorrelation coefficients

$$\hat{\rho}_{ni} = \frac{\sum_{j=1}^n (X_j - \overline{X}_n)(X_{j-i} - \overline{X}_n)}{\sum_{j=1}^n (X_j - \overline{X}_n)^2}.$$

Estimate the innovations ε_j by $\hat{\varepsilon}_j = X_j - \overline{X}_n - \hat{\rho}_n^\top (\mathbf{X}_{j-1} - \overline{X}_n \mathbf{1})$. Set $\hat{\vartheta}_n = (\hat{\rho}_n^\top, \overline{X}_n)^\top$. Then the estimator $H_n(\hat{\vartheta}_n)$ in Theorem 4 is efficient for $E[h(\varepsilon_1)]$.

Example 3. A heteroscedastic and non-adaptive example is the SETAR(1,2) process, defined by

$$X_j = (\alpha_1 + \beta_1 X_{j-1} + \sigma_1 \varepsilon_j) 1[X_{j-1} < 0] + (\alpha_2 + \beta_2 X_{j-1} + \sigma_2 \varepsilon_j) 1[X_{j-1} \ge 0], \quad j \ge 1,$$

where the ε_j are i.i.d. with mean 0 and variance 1. The structural relation (4.1) holds for $\vartheta = (\alpha_1, \beta_1, \sigma_1, \alpha_2, \beta_2, \sigma_2)^\top$ and

$$\begin{aligned} r(\vartheta, x) &= (\alpha_1 + \beta_1 x) 1[x < 0] + (\alpha_2 + \beta_2 x) 1[x \ge 0], \\ s(\vartheta, x) &= \sigma_1 1[x < 0] + \sigma_2 1[x \ge 0]. \end{aligned}$$

In this model, Assumption 4 holds with

$$\begin{aligned}\dot r(x) &= (1[x<0], x1[x<0], 0, 1[x\ge 0], x1[x>0], 0)^\top,\\ \dot s(x) &= (0,0,1[x<0],0,0,1[x\ge 0])^\top,\end{aligned}$$

and the vector D is given by

$$-\Big(\frac{pI_0}{\sigma_1}, \frac{\mu_1 I_0}{\sigma_1}, \frac{I_1}{\sigma_1}, \frac{(1-p)I_0}{\sigma_2}, \frac{\mu_2 I_0}{\sigma_2}, \frac{I_1}{\sigma_2}\Big)^\top,$$

where $I_k = E[\varepsilon_1^k(h'(\varepsilon) - a_h(F)^\top \psi'(\varepsilon))]$, $p = P(X_0 < 0)$, $\mu_1 = E[X_0 1[X_0 < 0]]$, $\mu_2 = E[X_0 1[X_0 > 0]]$. Hence the estimator $H_n(\hat\vartheta_n)$ of Theorem 8 is, in general, efficient only if all components of $\hat\vartheta_n$ are efficient. Such estimators for ϑ can be constructed as one-step improvements of $n^{1/2}$-consistent estimators, using the approach of Drost, Klaassen and Werker (1997).

An initial estimator for (α_1, β_1) is the minimizer $(\hat\alpha_1, \hat\beta_1)$ of

$$\sum_{j=1}^n (X_j - \alpha_1 - \beta_1 X_{j-1})^2 1[X_{j-1} < 0];$$

and σ_1 can be estimated by the square root $\hat\sigma_1$ of

$$\hat\sigma_1^2 = \frac{\sum_{j=1}^n (X_j - \hat\alpha_1 - \hat\beta_1 X_{j-1})^2 1[X_{j-1} < 0]}{\sum_{j=1}^n 1[X_{j-1} < 0]}.$$

For α_2, β_2 and σ_2 we have corresponding estimators.

Example 4. Another heteroscedastic and non-adaptive example is

$$X_j = \alpha X_{j-1} + \sqrt{\beta + \gamma X_{j-1}^2}\,\varepsilon_j, \quad j \ge 1,$$

where the ε_j are i.i.d. with mean 0 and variance 1. The structural relation (4.1) holds for $\vartheta = (\alpha, \beta, \gamma)^\top$ and $r(\vartheta, x) = \alpha x$, $s(\vartheta, x) = \sqrt{\beta + \gamma x^2}$. The time series is ergodic if β and γ are positive and $\alpha^2 + \gamma < 1$. Assumption 4 holds with

$$\dot r(x) = (x, 0, 0)^\top, \quad \dot s(x) = \frac{1}{2\sqrt{\beta + \gamma x^2}}(0, 1, x^2)^\top.$$

We obtain

$$M^\top(x) = \begin{bmatrix} \frac{x}{\sqrt{\beta+\gamma x^2}} & 0 & 0 \\ 0 & \frac{1}{2(\beta+\gamma x^2)} & \frac{x^2}{2(\beta+\gamma x^2)} \end{bmatrix}.$$

Hence the estimator $H_n(\hat\vartheta_n)$ of Theorem 8 is, in general, efficient only if all components of $\hat\vartheta_n$ are efficient. Such estimators for ϑ can be constructed as one-step improvements of $n^{1/2}$-consistent estimators, using the approach of Drost, Klaassen and Werker (1997). Here one should be able to avoid sample splitting using ideas of Koul and Schick (1997).

6 Proofs

Lemma 1. *Suppose Assumption 1 holds and $\hat{\vartheta}_n$ is $n^{1/2}$-consistent. Then*

$$\sum_{j=1}^{n}\Big(r(\hat{\vartheta}_n,\mathbf{X}_{j-1})-r(\vartheta,\mathbf{X}_{j-1})-\dot{r}(\mathbf{X}_{j-1})^{\top}(\hat{\vartheta}_n-\vartheta)\Big)^2=o_{P_n}(1), \tag{6.1}$$

$$\sum_{j=1}^{n}\Big(r(\hat{\vartheta}_n,\mathbf{X}_{j-1})-r(\vartheta,\mathbf{X}_{j-1})\Big)^2=O_{P_n}(1), \tag{6.2}$$

$$\max_{1\le j\le n}|r(\hat{\vartheta}_n,\mathbf{X}_{j-1})-r(\vartheta,\mathbf{X}_{j-1})|=o_{P_n}(1). \tag{6.3}$$

Proof. Relation (6.1) follows immediately from Assumption 1 and $n^{1/2}$-consistency of $\hat{\vartheta}_n$. For notational convenience, write $r_{nj}=r(\hat{\vartheta}_n,\mathbf{X}_{j-1})-r(\vartheta,\mathbf{X}_{j-1})$. Relation (6.1) implies $\max_{1\le j\le n}|r_{nj}-\dot{r}(\mathbf{X}_{j-1})^{\top}(\hat{\vartheta}_n-\vartheta)|=o_{P_n}(1)$. Hence (6.2) follows if we show $\frac{1}{n}\sum_{j=1}^{n}\|\dot{r}(\mathbf{X}_{j-1})\|^2=O_{P_n}(1)$, and (6.3) follows if we show $n^{-1/2}\max_{1\le j\le n}\|\dot{r}(\mathbf{X}_{j-1})\|=o_{P_n}(1)$. Both statements follow from stationarity and square-integrability of $\dot{r}$.

Lemma 2. *Suppose Assumptions 1 and 2 hold, and $\hat{\vartheta}_n$ is $n^{1/2}$-consistent. Then*

$$n^{-1/2}\sum_{j=1}^{n}h(\hat{\varepsilon}_j)=n^{-1/2}\sum_{j=1}^{n}h(\varepsilon_j)-E[h'(\varepsilon_1)]E[\dot{r}(\mathbf{X}_0)^{\top}]n^{1/2}(\hat{\vartheta}_n-\vartheta)+o_{P_n}(1).$$

Proof. By Taylor expansion,

$$h(\hat{\varepsilon}_j)=h(\varepsilon_j)+(\hat{\varepsilon}_j-\varepsilon_j)h'(\varepsilon_j)+(\hat{\varepsilon}_j-\varepsilon_j)\int_0^1\Big(h'(\varepsilon_j+t(\hat{\varepsilon}_j-\varepsilon_j))-h'(\varepsilon_j)\Big)dt.$$

We have

$$\hat{\varepsilon}_j-\varepsilon_j=-(r(\hat{\vartheta}_n,\mathbf{X}_{j-1})-r(\vartheta,\mathbf{X}_{j-1})). \tag{6.4}$$

Assumption 2 implies $\frac{1}{n}\sum_{j=1}^{n}(h'(\varepsilon_j))^2=O_{P_n}(1)$. Hence relation (6.1), the Cauchy–Schwarz inequality and the ergodic theorem give

$$\begin{aligned} n^{-1/2}\sum_{j=1}^{n}(\hat{\varepsilon}_j-\varepsilon_j)h'(\varepsilon_j) &= -n^{-1/2}\sum_{j=1}^{n}h'(\varepsilon_j)\dot{r}(\mathbf{X}_{j-1})^{\top}(\hat{\vartheta}_n-\vartheta)+o_{P_n}(1)\\ &= -E[h'(\varepsilon_1)]E[\dot{r}(\mathbf{X}_0)^{\top}]n^{1/2}(\hat{\vartheta}_n-\vartheta)+o_{P_n}(1). \end{aligned}$$

It remains to show

$$(6.5)\qquad n^{-1/2}\sum_{j=1}^{n}(\hat{\varepsilon}_j-\varepsilon_j)\int_0^1\Big(h'(\varepsilon_j+t(\hat{\varepsilon}_j-\varepsilon_j))-h'(\varepsilon_j)\Big)dt=o_{P_n}(1).$$

Relations (6.4) and (6.2) imply $\sum_{j=1}^{n}(\hat{\varepsilon}_j-\varepsilon_j)^2=O_{P_n}(1)$. Hence, by the Cauchy–Schwarz inequality, relation (6.5) holds if

$$\frac{1}{n}\sum_{j=1}^{n}\int_0^1\Big(h'(\varepsilon_j+t(\hat{\varepsilon}_j-\varepsilon_j))-h'(\varepsilon_j)\Big)^2dt=o_{P_n}(1).$$

This, in turn, follows from Assumption 2, since by relation (6.3) and (6.4) we have $\max_{1\le j\le n}|\hat{\varepsilon}_j-\varepsilon_j|=o_{P_n}(1)$.

Proof of Theorem 1. It is easy to check that $a_h(F_{n\hat{\vartheta}_n})$ is a consistent estimate of $a_h(F)$. It follows from Lemma 2 that $\frac{1}{n}\sum_{j=1}^{n}h(\hat{\varepsilon}_j)$ has influence function

$$h(\varepsilon_1)-E[h(\varepsilon_1)]-E[h'(\varepsilon_1)]E[\dot{r}(\mathbf{X}_0)^\top]\chi(\mathbf{X}_0,X_1).$$

Since the choice $h(x)=x$ fulfills Assumption 2, we obtain as special case that $\frac{1}{n}\sum_{j=1}^{n}\hat{\varepsilon}_j$ has influence function $\varepsilon_1-E[\dot{r}(\mathbf{X}_0)^\top]\chi(\mathbf{X}_0,X_1)$. Combining the above shows that $H_n(\hat{\vartheta}_n)=\frac{1}{n}\sum_{j=1}^{n}h(\hat{\varepsilon}_j)-a_h(F_{n\hat{\vartheta}_n})\frac{1}{n}\sum_{j=1}^{n}\hat{\varepsilon}_j$ has the desired influence function.

Proof of Remark 1. Consider the location model generated by the density f. The function $\int h(x)f(x-a)dx=\int h(x+a)f(x)dx$ is differentiable at $a=0$; its derivative can be written in two ways. By Assumption 2 and dominated convergence,

$$\int h(x+a)f(x)dx-\int h(x)f(x)dx=a\int h'(x)f(x)dx+o(a).$$

By Assumption 3 and Lemma 7.2 in Ibragimov and Has'minskii (1981),

$$\int h(x)f(x-a)dx-\int h(x)f(x)dx=a\int h(x)\ell_1(x)f(x)dx+o(a).$$

Hence $\int h'(x)f(x)dx=\int h(x)\ell_1(x)f(x)dx$.

Proof of Theorem 3. Note first that $h(\varepsilon_1)-a_h(F)\varepsilon_1-E[h(\varepsilon_1)]$ is in $\mathcal{V}$. Hence the function $\dot{\kappa}_0(\mathbf{X}_0,X_1)$ defined in (3.13) is in the tangent space T. By definition of v_n,

$$n^{1/2}(\kappa(\vartheta_{nu},F_{nv})-\kappa(\vartheta,F))=n^{1/2}\Big(\int hdF_{nv}-\int hdF\Big)\to E[h(\varepsilon_1)v(\varepsilon_1)]$$

for all $(u,v) \in \mathbf{R}^k \times \mathcal{V}$. To prove that $\dot{\kappa}_0(\mathbf{X}_0, X_1)$ is a gradient, we must show that

$$E[\dot{\kappa}_0(\mathbf{X}_0, X_1) S_w(\mathbf{X}_0, \varepsilon_1)] = E[h(\varepsilon_1)v(\varepsilon_1)] \quad \text{for all } w = (u,v) \in \mathbf{R}^k \times \mathcal{V}.$$

But this follows from straightforward calculations using the representation (3.6), the orthogonality condition (3.9), the facts that $h(\varepsilon_1) - a_h(F)\varepsilon_1 - E[h(\varepsilon_1)]$ and $\ell_1(\varepsilon_1) - a_{\ell_1}(F)\varepsilon_1$ belong to $\mathcal{V}$ and the identities

$$E\Big[\Big(h(\varepsilon_1) - a_h(F)\varepsilon_1 - E[h(\varepsilon_1)]\Big)v(\varepsilon_1)\Big] = E[h(\varepsilon_1)v(\varepsilon_1)], \quad v \in \mathcal{V},$$

$$E\Big[\Big(h(\varepsilon_1) - a_h(F) - E[h(\varepsilon_1)]\Big)(\ell_1(\varepsilon_1) - a_{\ell_1}(F)\varepsilon_1))\Big] = E\Big[\Big(h(\varepsilon_1) - a_h(F)\varepsilon_1\Big)\ell_1(\varepsilon_1)\Big].$$

This proves that $\dot{\kappa}_0(\mathbf{X}_0, X_1)$ is a gradient. Since it is in the tangent space T, it is canonical.

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models.* Springer, New York.

Boldin, M. V. (1982). Estimation of the distribution of noise in an autoregression scheme. *Theory Probab. Appl.* **27**, 866–871.

Drost, F. C., Klaassen, C. A. J. and Werker, B. J. M. (1997). Adaptive estimation in time-series models. *Ann. Statist.* **25**, 786–817.

Ghoudi, K. and Rémillard, B. (1998). Empirical processes based on pseudo-observations. In: *Asymptotic Methods in Probability and Statistics* (B. Szyszkowicz, ed.), 171-197, North-Holland, Amsterdam.

Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. Verw. Gebiete* **14**, 323–330.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation. Asymptotic Theory.* Applications of Mathematics 16, Springer, New York.

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

Klaassen, C. A. J. and Putter, H. (1999). Efficient estimation of Banach parameters in semiparametric models. Technical Report, Department of Mathematics, University of Amsterdam.

Koul, H. L. (1992). *Weighted Empiricals and Linear Models.* IMS Lectures Notes-Monograph Series 21, Institut of Mathematical Statistics, Hayward, California.

Koul, H. L. (1996). Asymptotics of some estimators and sequential residual empiricals in nonlinear time series. *Ann. Statist.* **24**, 380–404.

Koul, H. L. and Levental, S. (1989). Weak convergence of the residual empirical process in explosive autoregression. *Ann. Statist.* **17**, 1784–1794.

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

Kreiss, J.-P. (1987a). On adaptive estimation in stationary ARMA processes. *Ann. Statist.* **15**, 112–133.

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

Levit, B. Y. (1975). Conditional estimation of linear functionals. *Problems Inform. Transmission* **11**, 39–54.

Schick, A. and Susarla, V. (1990). An infinite dimensional convolution theorem with applications to random censoring and missing data models. *J. Statist. Plann. Inference* **24**, 13–23.

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.