An introduction to efficient estimation for semiparametric time series

Priscilla E. Greenwood
Arizona State University

Ursula U. Müller
Universität Bremen

Wolfgang Wefelmeyer
Universität Siegen

Abstract

We illustrate several recent results on efficient estimation for semiparametric time series models with two types of AR(1) models: having independent and centered innovations, and having general and conditionally centered innovations. We consider in particular estimation of the autoregression parameter, the stationary distribution, the innovation distribution, and the stationary density.

1 Introduction

The purpose of this paper is to illustrate a number of recent results on efficient estimation for semiparametric time series models in the context of a linear autoregressive process of order one, $X_i = \vartheta X_{i-1} + \varepsilon_i$. In addition, we sketch the construction of efficient estimators in this context. Historically, it was first assumed that the innovations ε_i are i.i.d. with zero mean. We call this model I. For many applications, especially in the recent econometrics literature, independence of the innovations is considered too strong an assumption and is replaced by the weaker condition that the ε_i may depend on the previous observation, with $E(\varepsilon_i \mid X_{i-1}) = 0$. We call this model II. It can be described as a nonparametric Markov chain model fulfilling the constraint $E(X_i \mid X_{i-1}) = \vartheta X_{i-1}$. Structurally, the two models are quite different. Model I is very close to an i.i.d. model, and efficient estimation of ϑ in particular has a long history. Efficient estimation of the stationary law and of the law of the innovations is more recent. Model II is closer to the full nonparametric Markov chain model, for which efficient estimation was considered in the last decade only. Nevertheless, in certain respects model II is simpler than model I because it has less structure. The full nonparametric Markov chain model has just one (infinite-dimensional) parameter, the transition kernel. Model II is described by the transition kernel and the parameter ϑ in the model for the conditional mean. This is not a parametrization, because the transition kernel and the parameter in the constraint on the transition kernel do not vary independently. Model I has two parameters: the innovation law and the autoregression parameter.

This paper does not follow the order of the historical development. Instead, we follow the order of nesting of these models: Sections 2 to 4 treat first the full nonparametric Markov chain model, then model II, and finally model I. We will only sketch the arguments and refer the reader to the literature for details.

2 Nonparametric efficiency of the least squares estimator

We begin by recalling some results on efficient estimation for general Markov chain models: local asymptotic normality, characterizations of efficient and regular estimators, and efficiency of empirical estimators. Let $X_0, \dots, X_n$ be observations from a real-valued stationary Markov chain with transition kernel $Q(x, dy)$. Let π, $P = \pi \otimes Q$ and $P^{(n)}$ denote the laws of X_0, (X_0, X_1) and $(X_0, \dots, X_n)$, respectively. Write $Ph = \int h(x,y)\, P(dx, dy)$ for the expectation of $h(X_0, X_1)$, and $Q_x h = \int h(x,y) Q(x, dy)$ for the conditional expectation of $h(X_0, X_1)$ given $X_0 = x$.

Local asymptotic normality. To describe asymptotic variance bounds and characterize efficient estimators, we introduce a *local model* through (Hellinger differentiable) perturbations $Q_{nv}(x, dy) \doteq Q(x, dy)(1 + n^{-1/2} v(x,y))$. For Q_{nv} to be a transition kernel, the *local parameter* v must be in the space

$$V = \{v \in L_2(P) : Qv = 0\}.$$

This is the *tangent space* of the full nonparametric model. Write π_{nv}, P_{nv} and $P_{nv}^{(n)}$ for the corresponding laws if Q_{nv} is true. We have *local asymptotic normality*

$$\begin{aligned} \log \frac{dP_{nv}^{(n)}}{dP^{(n)}}(X_0, \dots, X_n) &= n^{-1/2} \sum_{i=1}^{n} v(X_{i-1}, X_i) - \frac{1}{2} P v^2 + o_{P^{(n)}}(1), \\ n^{-1/2} \sum_{i=1}^{n} v(X_{i-1}, X_i) &\Rightarrow (Pv^2)^{1/2} N \quad \text{under } P^{(n)}, \end{aligned}$$

with N a standard normal random variable. Proofs under increasingly weaker conditions are given by Roussas (1965), Höpfner, Jacod and Ladelli (1990), Penev (1991) and Höpfner (1993a, 1993b).

Characterization of efficient estimators. Consider now a submodel, described by a subset of the transition kernels. Its tangent space is obtained by perturbing Q within the submodel. The tangent space is a subset of V, say V_*, which we take to be (closed and) linear. A real-valued functional $s(Q)$ is *differentiable* at Q with *gradient* g if $g \in V$ and

$$n^{1/2}(s(Q_{nv}) - s(Q)) \to P[gv] \quad \text{for all } v \in V_*.$$

The *canonical gradient* g_* is the projection of g onto V_*. An estimator $\hat{s}$ of s is *regular* at Q with *limit* L if L is a random variable such that

$$n^{1/2}(\hat{s} - s(Q_{nv})) \Rightarrow L \quad \text{under } P_{nv}^{(n)} \text{ for all } v \in V_*.$$

The convolution theorem says that $L = (Pg_*^2)^{1/2}N + M$ in distribution, with random variable M independent of N. This justifies calling a regular estimator $\hat{s}$ *efficient* for s at Q if $L = (Pg_*^2)^{1/2}N$ in distribution.

An estimator $\hat{s}$ is *asymptotically linear* for s at Q with *influence function* w if $w \in W$ and

$$n^{1/2}(\hat{s} - s(Q)) = n^{-1/2}\sum_{i=1}^{n} w(X_{i-1}, X_i) + o_{P^{(n)}}(1).$$

An asymptotically linear estimator is regular if and only if its influence function is a gradient, and a (regular) estimator is efficient if and only if it is asymptotically linear with influence function equal to the canonical gradient. The convolution theorem is due to Hájek (1970). For the characterizations of regular and efficient estimators we refer to Bickel, Klaassen, Ritov and Wellner (1998, Section 3.3).

Efficiency of empirical estimators. We return to the full nonparametric Markov chain model. Suppose the chain is geometrically ergodic. Let $h(x,y)$ be P-square-integrable. We want to estimate the linear functional $s(Q) = Ph = Eh(X_0, X_1)$. By a perturbation expansion, see Kartashov (1985a, 1985b, 1996),

$$n^{1/2}(P_{nv}h - Ph) \to P[Ah \cdot v] \quad \text{for all } v \in V,$$

where A is a projection from $L_2(P)$ onto V, defined by

$$Ah(x,y) = h(x,y) - Q_x h + \sum_{j=1}^{\infty}(Q_y^j h - Q_x^{j+1} h). \tag{2.1}$$

Hence Ph is differentiable at Q with canonical gradient Ah. On the other hand, the empirical estimator $\frac{1}{n}\sum_{i=1}^{n} h(X_{i-1}, X_i)$ has the martingale approximation

$$n^{-1/2}\sum_{i=1}^{n}(h(X_{i-1}, X_i) - Ph) = n^{-1/2}\sum_{i=1}^{n} Ah(X_{i-1}, X_i) + o_{P^{(n)}}(1). \tag{2.2}$$

This approximation has been found independently by several authors, in particular Gordin (1969), Maigret (1978), Dürr and Goldstein (1986) and Greenwood and Wefelmeyer (1995). The approximation says that the empirical estimator is asymptotically linear with influence function Ah. Hence the estimator is regular and efficient. This was first proved by Penev (1991) without using the martingale approximation. The above proof is due to Greenwood and Wefelmeyer (1995). A shorter proof, not using the perturbation expansion, is obtained if one parametrizes the chain by the joint law

P rather than the transition kernel Q; see Bickel (1993) and Bickel and Kwon (2001). The approach is convenient for submodels given by constraints on P rather than Q. For a discussion we refer to Greenwood, Schick and Wefelmeyer (2001). We use the above approach because our submodels will be described in terms of the transition kernel of the chain.

Efficiency of the least squares estimator in the full nonparametric model. Write $\tau^2 = \int x^2 \pi(dx) = EX_0^2$ for the stationary second moment. The *least squares functional*

$$\vartheta(Q) = \tau^{-2} \int xy \, P(dx, dy) = E[X_0 X_1]/EX_0^2 \tag{2.3}$$

is the minimizer in ϑ of

$$\int (y - \vartheta x)^2 \, P(dx, dy) = E[(X_1 - \vartheta X_0)^2].$$

A natural estimator is the empirical version of the functional, the *least squares estimator*

$$\hat{\vartheta}_{LS} = \frac{\sum_{i=1}^n X_{i-1} X_i}{\sum_{i=1}^n X_{i-1}^2}. \tag{2.4}$$

This is the minimizer in ϑ of $\sum_{i=1}^n (X_i - \vartheta X_{i-1})^2$, i.e. the solution of the estimating equation

$$\sum_{i=1}^n X_{i-1}(X_i - \vartheta X_{i-1}) = 0. \tag{2.5}$$

The least squares estimator is the ratio of two empirical estimators. Since continuously differentiable functions of efficient estimators are efficient, $\hat{\vartheta}_{LS}$ is efficient in the full nonparametric model. We have

$$\begin{aligned} n^{1/2}(\hat{\vartheta}_{LS} - \vartheta) &= \frac{n^{-1/2} \sum_{i=1}^n X_{i-1}(X_i - \vartheta X_{i-1})}{\frac{1}{n} \sum_{i=1}^n X_{i-1}^2} \\ &= \tau^{-2} n^{-1/2} \sum_{i=1}^n X_{i-1}(X_i - \vartheta X_{i-1}) + o_{P^{(n)}}(1). \end{aligned} \tag{2.6}$$

Hence, by the martingale approximation (2.2), $\hat{\vartheta}_{LS}$ is asymptotically linear with influence function

$$w = \tau^{-2} A h_{LS}, \quad \text{where } h_{LS}(x, y) = x(y - \vartheta x). \tag{2.7}$$

The martingale central limit theorem implies that its asymptotic variance is

$$\tau^{-4} E[X_0^2 \varepsilon_1^2] + 2\tau^{-4} \sum_{j=2}^\infty E[X_0 \varepsilon_1 X_{j-1} \varepsilon_j] \tag{2.8}$$

with $\varepsilon_j = X_j - \vartheta X_{j-1}$. We note that since $\hat{\vartheta}_{LS}$ is efficient and regular, its influence function $\tau^{-2} A h_{LS}$ must be the canonical gradient of ϑ in the full nonparametric model.

3 Linear autoregression

Let $X_0, \dots, X_n$ be observations from a Markov chain fulfilling the following constraint: There is a number ϑ such that

$$E(X_1 \mid X_0 = x) = \int y\, Q(x, dy) = \vartheta x. \tag{3.1}$$

This is model II. It is a submodel of the full nonparametric Markov chain model of Section 2 and can be written as

$$X_i = \vartheta X_{i-1} + \varepsilon_i,$$

where the ε_i are martingale increments, i.e.

$$E(\varepsilon_1 \mid X_0) = E(X_1 - \vartheta X_0 \mid X_0) = 0.$$

Asymptotic variance of the least squares estimator. The estimating equation (2.5) for $\hat\vartheta_{LS}$ is now a *martingale estimating equation* since, in model II, $X_i - \vartheta X_{i-1}$ are martingale increments. This simplifies the influence function and the asymptotic variance of $\hat\vartheta_{LS}$. The expansion (2.6) now says that $\hat\vartheta_{LS}$ has influence function $h_{LS}(x,y) = x(y - \vartheta x)$ and asymptotic variance

$$\tau^{-4} P h_{LS} = \tau^{-4} E[X_0^2 \varepsilon_1^2] = \tau^{-4} E[X_0^2 \rho^2(X_0)],$$

where

$$\rho^2(x) = E(\varepsilon_1^2 \mid X_0 = x) = \int (y - \vartheta x)^2\, Q(x, dy)$$

is the conditional variance of the innovation given that the previous observation is x. By the characterization of regular estimators, h_{LS} is a gradient of ϑ in model II.

Weighted least squares estimators. We will now see that the least squares estimator is not efficient in model II. A large class of alternative estimators in model II is obtained as solutions $\hat\vartheta_W$ of martingale estimating equations

$$\sum_{i=1}^n W_\vartheta(X_{i-1})(X_i - \vartheta X_{i-1}) = 0 \tag{3.2}$$

with predictable weights $W_\vartheta(X_{i-1})$. A Taylor expansion argument gives

$$n^{1/2}(\hat\vartheta_W - \vartheta) = -E[W_\vartheta(X_0)X_0]^{-1} n^{-1/2} \sum_{i=1}^n W_\vartheta(X_{i-1})(X_i - \vartheta X_{i-1}) + o_{P^{(n)}}(1).$$

Hence $\hat\vartheta_W$ has asymptotic variance

$$E[W_\vartheta(X_0)X_0]^{-2} E[W_\vartheta(X_0)^2 \rho^2(X_0)].$$

By the Cauchy–Schwarz inequality, the asymptotic variance is minimized for

$$W_*(x) = \rho^{-2}(x)x.$$

Since ρ depends on the unknown transition kernel, the weight function W_* cannot be used in the estimating equation (3.2). Suppose we replace ρ by a consistent (kernel) estimator $\hat{\rho}$. The resulting estimating equation

$$\sum_{i=1}^{n} \hat{\rho}^{-2}(X_{i-1})X_{i-1}(X_i - \vartheta X_{i-1}) = 0 \tag{3.3}$$

leads to a weighted least squares estimator

$$\hat{\vartheta}_{II} = \frac{\sum_{i=1}^{n} \hat{\rho}^{-2}(X_{i-1})X_{i-1}X_i}{\sum_{i=1}^{n} \hat{\rho}^{-2}(X_{i-1})X_{i-1}^2} \tag{3.4}$$

with influence function $M^{-1}\mu$ and asymptotic variance M^{-1}, where

$$\mu(x,y) = \rho^{-2}(x)x(y - \vartheta x) \quad \text{and} \quad M = P\mu^2 = E[\rho^{-2}(X_0)X_0^2]$$

play the roles of *score function* and *Fisher information* for ϑ. The variance M^{-1} is smaller than the asymptotic variance of the least squares estimator $\hat{\vartheta}_{LS}$ unless ρ happens to be constant. Hence $\hat{\vartheta}_{LS}$ is not efficient in model II, in general.

Efficiency of the best weighted least squares estimator. We show that the weighted least squares estimator $\hat{\vartheta}_{II}$ defined in (3.4) is efficient in model II. To calculate the canonical gradient for ϑ in model II, we determine the tangent space of the model. A perturbed transition kernel Q_{nv} must also fulfill the constraint on $E(X_1 \mid X_0)$, possibly with perturbed parameter $\vartheta_{nt} = \vartheta + n^{-1/2}t$:

$$\int y \, Q_{nv}(x, dy) = \vartheta_{nt}x \quad \text{for some } x \in \mathbf{R}.$$

This means that the tangent space V_{II} of model II is the union of the affine spaces

$$V_t = \{v \in V : \int yv(x,y)\, Q(x,dy) = tv\}.$$

For $t = 0$ this is the tangent space when ϑ is known. Since $v(X_0, X_1)$ is conditionally centered, we can write V_0 as

$$V_0 = \{v \in V : \int (y - \vartheta x)v(x,y)\, Q(x,dy) = 0\}.$$

The orthogonal complement of V_0 in V_{II} is spanned by the function $\ell(x,y) = \rho^{-2}(x)x(y - \vartheta x)$.

The canonical gradient of ϑ is obtained as the projection of the gradient $h_{LS}(x,y) = x(y-\vartheta x)$ onto V_{II}. It must lie in the orthogonal complement of the tangent space V_0 for known ϑ. Hence it must be of the form $c\mu(x,y)$, with constant c determined by $P[(h_{LS} - c\mu)\mu] = 0$, i.e. $c = M^{-1}$. Therefore, the canonical gradient of ϑ in model II is $M^{-1}\mu$. This equals the influence function of the best weighted least squares estimator $\hat{\vartheta}_{II}$ defined in (3.4), which is therefore efficient. For generalizations of this result we refer to Wefelmeyer (1996, 1997) and Müller and Wefelmeyer (2002a, b).

Improved empirical estimators. We have seen in Section 2 that the empirical estimator $\frac{1}{n}\sum_{i=1}^n h(X_{i-1}, X_i)$ is efficient for $Eh(X_0, X_1)$ in the full nonparametric model. We will now construct better, efficient, estimators in the smaller model II, using the constraint (3.1). We follow a heuristic *plug-in principle*, consisting of two steps. First we consider model II with known ϑ and construct an efficient estimator of $Eh(X_0, X_1)$ for each ϑ. Then we replace ϑ by an efficient estimator $\hat{\vartheta}$. Under appropriate conditions, the resulting estimator for $Eh(X_0, X_1)$ will be efficient in model II with unknown ϑ. This construction principle will resurface later in the paper. For general results on the plug-in principle we refer to Müller, Schick and Wefelmeyer (2001a) and Klaassen and Putter (2002).

Suppose first that ϑ is known. We have seen in (3.3) that the constraint $E(\varepsilon_1 \mid X_0) = 0$ leads to new estimators for ϑ. It also leads to new estimators of $Eh(X_0, X_1)$, constructed by adding a correction term to the empirical estimator,

$$\frac{1}{n}\sum_{i=1}^n \big(h(X_{i-1}, X_i) - c(X_i - \vartheta X_{i-1})\big). \tag{3.5}$$

By the martingale approximation (2.2), they have influence function $Ah(x,y) - c(y-\vartheta x)$ and asymptotic variance $\int (Ah(x,y) - c(y-\vartheta x))^2 \, P(dx,dy)$. By the Cauchy–Schwarz inequality, this variance is minimized for

$$c_*(\vartheta) = \tau^{-2} \int (y-\vartheta x) Ah(x,y) \, P(dx,dy).$$

Since c_* depends on the unknown joint distribution of the Markov chain, it cannot be used in (3.5). We replace c_* by a consistent estimator $\hat{c}_*$, which does not change the asymptotic variance, and obtain the best improved empirical estimator

$$\frac{1}{n}\sum_{i=1}^n \big(h(X_{i-1}, X_i) - \hat{c}_*(\vartheta)(X_i - \vartheta X_{i-1})\big). \tag{3.6}$$

An explicit construction of $\hat{c}_*$ is in Müller, Schick and Wefelmeyer (2001b).

We have seen in Section 2 that $Eh(X_0, X_1)$ has gradient Ah. The canonical gradient in model II with known ϑ is obtained by projecting Ah onto V_0. This amounts to finding c such that $\int (Ah(x,y) - c(y-\vartheta x))^2 \, P(dx,dy)$ is minimized, a problem we already solved when we determined the improved empirical estimator (3.6). It follows that the canonical gradient equals the influence function of the estimator (3.6). Hence this estimator is efficient in model II when ϑ is known.

Now suppose that ϑ is unknown. The plug-in principle says that

$$\hat{E}_{II}h = \frac{1}{n}\sum_{i=1}^{n}\big(h(X_{i-1}, X_i) - \hat{c}_*(\hat{\vartheta}_{II})(X_i - \hat{\vartheta}_{II}X_{i-1})\big) \tag{3.7}$$

is efficient for $Eh(X_0, X_1)$ in model II. A version of this result for nonlinear regression is in Müller and Wefelmeyer (2002a).

4 Independent innovations

Let $X_0, \dots, X_n$ be observations from the AR(1) model $X_i = \vartheta X_{i-1} + \varepsilon_i$, where the innovations ε_i are now i.i.d., with mean zero and finite variance σ^2, and $|\vartheta| < 1$ to ensure geometric ergodicity. This is model I.

Adaptivity. The transition kernel of model I is parametrized by ϑ and by the density, say f, of the innovations:

$$Q(x, dy) = f(y - \vartheta x)\, dy.$$

The tangent space of model I will therefore be expressed in terms of perturbations $\vartheta_{nt} = \vartheta + n^{-1/2}t$ and (Hellinger differentiable) perturbations $f_{nu}(z) \doteq f(z)(1+n^{-1/2}u(z))$. Since f has mean zero, the functions u vary in

$$U = \{u \in L_2(f) : Eu(\varepsilon) = E[\varepsilon u(\varepsilon)] = 0\}.$$

The transition density is then perturbed as

$$f_{nu}(y - \vartheta_{nt}x) \doteq f(y - \vartheta x)\big(1 + n^{-1/2}\big(u(y - \vartheta x) + tx\ell(y - \vartheta x)\big)\big)$$

with $\ell = -f'/f$. The function ℓ is the score function for location of the innovation distribution and very different from the score function μ in Section 3. We obtain that the tangent space of model I is

$$V_I = \{u(y - \vartheta x) + tx\ell(y - \vartheta x) : u \in U, t \in \mathbf{R}\}.$$

Since X_i and $\varepsilon_i = X_i - \vartheta X_{i-1}$ are independent in this model, the decomposition of V_I into functions $u(y - \vartheta x)$ and the linear span of the function $x\ell(y - \vartheta x)$ is orthogonal. Note that the functions $u(y - \vartheta x)$ form the tangent space for fixed ϑ, while the multiples of $x\ell(y - \vartheta x)$ are the tangent space for fixed f. Because they are orthogonal, ϑ can be estimated *adaptively* with respect to f in the sense that the variance bound does not increase if we do not know f.

Estimating the autoregression parameter. Since ϑ is adaptive with respect to f, its canonical gradient can be calculated in the model with fixed f. Hence it is of the form $t_* x\ell(y-\vartheta x)$ with t_* determined by

$$n^{1/2}(\vartheta_{nt}-\vartheta)=t\overset{!}{=}t_*t\int x^2\ell(y-\vartheta x)^2\,P(dx,dy)=t_*t\tau^2J,$$

where $J=E[\ell(\varepsilon)^2]$ is the Fisher information for location of the innovation distribution. The solution of the above equation is $t_*=\tau^{-2}J^{-1}$. Hence a regular and efficient estimator for ϑ in model I has influence function $\tau^{-2}J^{-1}x\ell(y-\vartheta x)$. It can be constructed as *one-step improvement* of a $n^{-1/2}$-consistent *initial estimator*, for example the least squares estimator $\hat\vartheta_{LS}$:

$$\hat\vartheta_I=\hat\vartheta_{LS}+\hat\tau^{-2}\hat J^{-1}\frac{1}{n}\sum_{i=1}^n X_{i-1}\hat\ell(X_i-\hat\vartheta_{LS}X_{i-1}), \tag{4.1}$$

where $\hat\tau$, $\hat J$ and $\hat\ell$ are appropriate estimators of τ, J and ℓ, respectively. For constructions in more general autoregressive models see Kreiss (1987a, b), Jeganathan (1995), Drost, Klaassen and Werker (1997), Koul and Schick (1997) and Schick and Wefelmeyer (2002b).

Of course, in model I the asymptotic variance of the best weighted least squares estimator $\hat\vartheta_{II}$ cannot be smaller than the asymptotic variance of $\hat\vartheta_I$. Indeed, since $\rho(x)=\sigma$ in model I, the asymptotic variance of $\hat\vartheta_{II}$ and of the ordinary least squares estimator $\hat\vartheta_{LS}$ both become $\tau^{-2}\sigma^2$, while the asymptotic variance of $\hat\vartheta_I$ is $\tau^{-2}J^{-1}$. Since $E[\varepsilon\ell(\varepsilon)]=1$, the Cauchy–Schwarz inequality gives $\sigma^2\geq J^{-1}$, with equality only if ε is proportional to $\ell(\varepsilon)$, i.e. for normal errors.

Estimating the innovation distribution. Estimators of functionals of the innovation distribution can be based on residuals, i.e. on estimated innovations $\hat\varepsilon_i=X_i-\hat\vartheta X_{i-1}$. Consider for example a linear functional $Eh(\varepsilon)=\int h(z)f(z)\,dz$, where h is f-square-integrable. A simple estimator is the empirical estimator $\frac{1}{n}\sum_{i=1}^n h(\hat\varepsilon_i)$ based on the residuals. It will not be efficient in model I since it uses neither independence nor centeredness of the innovations. To obtain an efficient estimator for $Eh(\varepsilon)$, we use again the plug-in principle employed in Section 3 for efficient estimation of $Eh(X_0,X_1)$. If ϑ is known, we can observe the innovations. They are independent with mean zero density f. Similarly as in Section 3, we can use the constraint $E\varepsilon=0$ to introduce modified empirical estimators

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

Their asymptotic variance is $E[(h(\varepsilon)-Eh(\varepsilon)-c\varepsilon)^2]$. The variance is minimized for $c_*=\sigma^{-2}E[\varepsilon h(\varepsilon)]$. The minimal variance is $E[h(\varepsilon)^2]-E[h(\varepsilon)]^2-\sigma^{-2}E[\varepsilon h(\varepsilon)]^2$. It is not changed if we replace c_* by a consistent estimator $\hat{c}_*$, for example a ratio of empirical estimators. With this choice, the best modified empirical estimator for $Eh(\varepsilon)$ is

$$\frac{1}{n}\sum_{i=1}^{n}\Big(h(\varepsilon_i) - \frac{\sum_{i=1}^{n}\varepsilon_i h(\varepsilon_i)}{\sum_{i=1}^{n}\varepsilon_i^2}\varepsilon_i\Big). \tag{4.2}$$

This estimator is efficient in model I with known ϑ. The result goes back to Levit (1975). The plug-in principle now says that

$$\frac{1}{n}\sum_{i=1}^{n}\Big(h(\hat{\varepsilon}_i) - \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i h(\hat{\varepsilon}_i)}{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}\hat{\varepsilon}_i\Big) \tag{4.3}$$

is efficient for $Eh(\varepsilon)$ in model I if an efficient estimator $\hat{\vartheta}_I$ for ϑ is used in the residuals $\hat{\varepsilon}_i = X_i - \hat{\vartheta}_I X_{i-1}$. This result is due to Wefelmeyer (1994). For generalizations to nonlinear autoregression and to invertible linear time series see Schick and Wefelmeyer (2002a, b). Generalizations to *nonparametric* autoregression models $X_i = r(X_{i-1}) + \varepsilon_i$, with unknown regression function r, are also possible, even though r cannot be estimated at the parametric rate $n^{-1/2}$. See Akritas and Van Keilegom (2001) and Müller, Schick and Wefelmeyer (2002) for corresponding results in nonparametric regression.

Estimating the stationary distribution. The simplest estimator for the expectation $Eh(X_0)$ of a π-square-integrable function h is the empirical estimator discussed in Section 2. It will not be efficient in model I since it uses neither independence nor centeredness of the innovations. The condition $EX_0 = 0$ could be used just as for $Eh(\varepsilon)$ in (4.2). To make use of the independence of the innovations, we observe that for $|\vartheta| < 1$ the AR(1) process is invertible and has a moving average representation

$$X_0 = \sum_{j=0}^{\infty}\vartheta^j\varepsilon_{-j} = \sum_{j=1}^{\infty}\vartheta^{j-1}\varepsilon_j$$

in distribution. Hence

$$Eh(X_0) = Eh(S) \quad \text{with } S = \sum_{j=1}^{\infty}\vartheta^{j-1}\varepsilon_j.$$

This is approximated by

$$Eh(S^{(m)}) \quad \text{with } S^{(m)} = \sum_{j=1}^{m}\vartheta^{j-1}\varepsilon_j$$

if m increases with n. To estimate $Eh(X_0)$, we will again use the plug-in principle. Assume first that ϑ is known. Then we can observe the innovations $\varepsilon_i = X_i - \vartheta X_{i-1}$, and we can estimate $Eh(S^{(m)})$ by a U-statistic defined as follows. Let Φ denote the set of one-one functions φ from $\{1,\dots,m\}$ into $\{1,\dots,n\}$. For $\varphi\in\Phi$ set

$$S_\varphi(\vartheta)=\sum_{j=1}^{m}\vartheta^{j-1}\varepsilon_{\varphi(j)}=\sum_{j=1}^{m}\vartheta^{j-1}(X_{\varphi(j)}-\vartheta X_{\varphi(j)-1}).$$

We estimate $Eh(S^{(m)})$ by the average over these sums, the U-statistic

$$U(\vartheta)=\frac{(n-m)!}{n!}\sum_{\varphi\in\Phi}h(S_\varphi(\vartheta)).$$

If m is fixed, and if the stationary distribution is unrestricted, $U(\vartheta)$ is efficient for $Eh(S^{(m)})$. Schick and Wefelmeyer (2002c) show that $U(\vartheta)$ is also efficient for $Eh(X_0)$ if m increases with n at an appropriate rate.

The only constraint on the stationary distribution is $EX_0=0$. This constraint can be used to improve the empirical estimator, similarly as in (3.5) and (4.2): Consider

$$U(\vartheta,c)=U(\vartheta)-c\frac{1}{n}\sum_{i=1}^{n}(X_i-\vartheta X_{i-1}).$$

Now note that by the Hoeffding decomposition, $U(\vartheta)$ is asymptotically linear with influence function $w(y-\vartheta x)$, where

$$w=\sum_{j=1}^{\infty}w_j\quad\text{with}\quad w_j(z)=E(h(S)\mid\varepsilon_j=z)-Eh(S).$$

The asymptotic variance of $U(\vartheta,c)$ is therefore minimized for

$$c_*(\vartheta)=\sigma^{-2}E[\varepsilon_1 w(\varepsilon_1)].$$

The asymptotic variance is not changed if we replace c_* by a consistent estimator, e.g. by

$$\hat c_*(\vartheta)=\frac{\sum_{i=1}^{n}(X_i-\vartheta X_{i-1})\sum_{j=1}^{m}H_{ji}(\vartheta)}{\sum_{i=1}^{n}(X_i-\vartheta X_{i-1})^2}$$

with

$$H_{ji}(\vartheta)=\frac{(n-m)!}{(n-1)!}\sum_{\varphi\in\Phi,\varphi(j)=i}h(S_\varphi(\vartheta)).$$

If m increases at an appropriate rate, $U(\vartheta,\hat c_*(\vartheta))$ is therefore efficient for $Eh(X_0)$ in model I with ϑ known. The plug-in principle now says that $U(\hat\vartheta_I,\hat c_*(\hat\vartheta_I))$ is efficient in model I. This result is proved in Schick and Wefelmeyer (2002c) for general causal and invertible linear processes.

Estimating the stationary density. There is a rich literature on kernel estimators for the stationary density of time series: see e.g. Chanda (1983), Yakowitz (1989) and Tran (1992). Such estimators converge more slowly than the parametric rate $n^{-1/2}$, depending on the smoothness of the density. For linear processes one can however use the independence of the innovations to obtain density estimators with convergence rate $n^{-1/2}$. This was first observed by Saavedra and Cao (1999, 2000). They write the stationary density d of a moving average process $X_i = \varepsilon_i + \vartheta\varepsilon_{i-1}$ with innovation density f in the convolution representation $d(x) = \int f(x - \vartheta y)f(y)\,dy$ and estimate $d(x)$ by plugging in a kernel estimator $\hat{f}$ for f. The resulting plug-in estimator $\hat{d}(x) = \int \hat{f}(x - \hat{\vartheta}y)\hat{f}(y)\,dy$ converges at the parametric rate $n^{-1/2}$. Schick and Wefelmeyer (2002d) show that $\hat{d}(x)$ is efficient if an efficient estimator for ϑ is used, and Schick and Wefelmeyer (2002e) prove functional convergence of such estimators for general MA(q) processes. In the i.i.d. case, related plug-in estimators for other smooth functionals of densities and regression functions have been studied before: see e.g. Hall and Marron (1987), Bickel and Ritov (1988) and Birgé and Massart (1995) for nonlinear integral functionals of densities and their derivatives, and Goldstein and Messer (1992) and Samarov (1993) for analogous results in regression models.

By definition, the stationary density, say k, of our model I also has a convolution representation:

$$k(x) = \int f(y - \vartheta x)k(y)\,dy.$$

We therefore obtain a $n^{1/2}$-consistent density estimator

$$\bar{k}(x) = \int \hat{f}(y - \hat{\vartheta}x)\hat{k}(y)\,dy$$

for appropriate choices of $\hat{k}$, $\hat{f}$ and $\hat{\vartheta}$. Such an estimator will however not be efficient, since it does not use the assumption of independent innovations to its full extent. To get an efficient estimator, one would again need to exploit the moving average representation $X_0 = \sum_{j=1}^{\infty} \vartheta^{j-1}\varepsilon_j$ (in distribution) and approximate the density k by the density of $S^{(m)} = \sum_{j=1}^{m} \vartheta^{j-1}\varepsilon_j$ for m increasing with n at an appropriate rate.

References

Akritas, M. G. and Van Keilegom, I. (2001). Non-parametric estimation of the residual distribution. *Scand. J. Statist.* **28**, 549–567.

Bickel, P. J. (1993). Estimation in semiparametric models. In: *Multivariate Analysis: Future Directions* (C. R. Rao, ed.) 55–73, North-Holland, Amsterdam.

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

Bickel, P. J. and Kwon, J. (2001). Inference for semiparametric models: Some questions and an answer (with discussion). *Statist. Sinica* **11**, 863–960.

Bickel, P. J. and Ritov, Y. (1998). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā Ser. A* **50**, 381–393.

Birgé, L. and Massart, P. (1995). Estimation of integral functionals of a density. *Ann. Statist.* **23**, 11–29.

Chanda, K. C. (1983). Density estimation for linear processes. *Ann. Inst. Statist. Math.* **35**, 439–446.

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

Dürr, D. and Goldstein, S. (1986). Remarks on the central limit theorem for weakly dependent random variables. In: *Stochastic Processes — Mathematics and Physics* (S. Albeverio, P. Blanchard and L. Streit, eds.), 104–118, Lecture Notes in Mathematics 1158, Springer, Berlin.

Goldstein, L. and Messer, K. (1992). Optimal plug-in estimators for nonparametric functional estimation. *Ann. Statist.* **20**, 1306–1328.

Gordin, M. I. (1969). The central limit theorem for stationary processes. *Soviet Math. Dokl.* **10**, 1174-1176.

Greenwood, P. E., Schick, A. and Wefelmeyer, W. (2001). Comment [on Bickel and Kwon, 2001]. *Statist. Sinica* **11**, 892–906.

Greenwood, P. E. and Wefelmeyer, W. (1995). Efficiency of empirical estimators for Markov chains. *Ann. Statist.* **23**, 132–143.

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

Hall, P. and Marron, J. S. (1990). On variance estimation in nonparametric regression. *Biometrika* **77**, 415–419.

Höpfner, R. (1993a). On statistics of Markov step processes: Representation of log-likelihood ratio processes in filtered local models. *Probab. Theory Related Fields* **94**, 375–398.

Höpfner, R. (1993b). Asymptotic inference for Markov step processes: Observation up to a random time. *Stochastic Process. Appl.* **48**, 295–310.

Höpfner, R., Jacod, J. and Ladelli, L. (1990). Local asymptotic normality and mixed normality for Markov statistical models. *Probab. Theory Related Fields* **86**, 105–129.

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

Kartashov, N. V. (1985a). Criteria for uniform ergodicity and strong stability of Markov chains with a common phase space. *Theory Probab. Math. Statist.* **30**, 71–89.

Kartashov, N. V. (1985b). Inequalities in theorems of ergodicity and stability for Markov chains with common phase space. I. *Theory Probab. Appl.* **30**, 247–259.

Kartashov, N. V. (1996). *Strong Stable Markov Chains.* VSP, Utrecht.

Klaassen, C. A. J. and Putter, H. (2002). Efficient estimation of Banach parameters in semiparamatric models. Technical Report, Korteweg-de Vries Institute for Mathematics, University of Amsterdam, http://preprint.beta.uva.nl/

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.

Maigret, N. (1978). Théorème de limite centrale fonctionnel pour une chaîne de Markov récurrente au sens de Harris et positive. *Ann. Inst. H. Poincaré Probab. Statist.* **14**, 425–440.

Müller, U. U., Schick, A. and Wefelmeyer, W. (2001a). Plug-in estimators in semiparametric stochastic process models. In: *Selected Proceedings of the Symposium on Inference for Stochastic Processes* (I. V. Basawa, C. C. Heyde and R. L. Taylor, eds.), 213-234, IMS Lecture Notes-Monograph Series 37, Institute of Mathematical Statistics, Beachwood, Ohio.

Müller, U. U., Schick, A. and Wefelmeyer, W. (2001b). Improved estimators for constrained Markov chain models. *Statist. Probab. Lett.* **54**, 427-435.

Müller, U. U., Schick, A. and Wefelmeyer, W. (2002). Estimating linear functionals of the error distribution in nonparametric regression. To appear in: *J. Statist. Plann. Inference.*

Müller, U. U. and Wefelmeyer, W. (2002a). Estimators for models with constraints involving unknown parameters. To appear in: *Math. Methods Statist.*

Müller, U. U. and Wefelmeyer, W. (2002b). Autoregression, estimating functions, and optimality criteria. To appear in: *Proceedings of the International Conference on Statistics, Combinatorics and Related Areas.*

Penev, S. (1991). Efficient estimation of the stationary distribution for exponentially ergodic Markov chains. *J. Statist. Plann. Inference* **27**, 105–123.

Roussas, G. G. (1965). Asymptotic inference in Markov processes. *Ann. Math. Statist.* **36**, 987–992.

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.

Saavedra, A. and Cao, R. (2000). On the estimation of the marginal density of a moving average process. *Canad. J. Statist.* **28**, 799–815.

Samarov, A. (1993). Exploring regression structure using nonparametric functional estimation. *J. Amer. Statist. Assoc.* **88**, 836–847.

Schick, A. and Wefelmeyer, W. (1999). Efficient estimation of invariant distributions of some semiparametric Markov chain models. *Math. Meth. Statist.* **8**, 426-440.

Schick, A. and Wefelmeyer, W. (2002a). Estimating the innovation distribution in nonlinear autoregressive models. *Ann. Inst. Statist. Math.* **54**, 245–260.

Schick, A. and Wefelmeyer, W. (2002b). Efficient estimation in invertible linear processes. To appear in: *Math. Methods Statist.*

Schick, A. and Wefelmeyer, W. (2002c). Estimating invariant laws of linear processes by U-statistics. Technical Report, Department of Mathematical Sciences, Binghamton University, http://math.binghamton.edu/anton/preprint.html.

Schick, A. and Wefelmeyer, W. (2002d). Root n consistent and optimal density estimators for moving average processes. Technical Report, Department of Mathematical Sciences, Binghamton University, http://math.binghamton.edu/anton/preprint.html.

Schick, A. and Wefelmeyer, W. (2002e). Functional convergence and optimality of plug-in estimators for stationary densities of moving average processes. Technical Report, Department of Mathematical Sciences, Binghamton University, http://math.binghamton.edu/anton/preprint.html.

Tran, L. T. (1992). Kernel density estimation for linear processes. *Stochastic Process. Appl.* **41**, 281–296.

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.

Wefelmeyer, W. (1996). Quasi-likelihood models and optimal inference. *Ann. Statist.* **24**, 405–422.

Wefelmeyer, W. (1997). Adaptive estimators for parameters of the autoregression function of a Markov chain. *J. Statist. Plann. Inference* **58**, 389–398.

Yakowitz, S. (1989). Nonparametric density and regression estimation for Markov sequences without mixing assumptions. *J. Multivariate Anal.* **30**, 124–136.

Priscilla E. Greenwood
Department of Mathematics and Statistics
Arizona State University
Tempe, AZ 85287-1804, USA
pgreenw@matematik.su.se
http://www.math.ubc.ca/people/faculty/pgreenw/pgreenw.html

Ursula U. Müller
Fachbereich 3: Mathematik und Informatik
Universität Bremen
Postfach 330 440
28334 Bremen, Germany
uschi@math.uni-bremen.de
http://www.math.uni-bremen.de/∼uschi/

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