Adaptive estimators for parameters of the autoregression function of a Markov chain

Wolfgang Wefelmeyer

Department of Mathematics, University of Siegen, 57068 Siegen, Germany

Abstract: Suppose we observe an ergodic Markov chain on the real line, with a parametric model for the autoregression function, i.e. the conditional mean of the transition distribution. If one specifies, in addition, a parametric model for the conditional variance, one can define a simple estimator for the parameter, the maximum quasi-likelihood estimator. It is robust against misspecification of the conditional variance, but not efficient. We construct an estimator which is adaptive in the sense that it is efficient if the conditional variance is misspecified, and asymptotically as good as the maximum quasi-likelihood estimator if the conditional variance is correctly specified. The adaptive estimator is a weighted nonlinear least squares estimator, with weights given by predictors for the conditional variance.

AMS 1991 Subject Classification: Primary 62G20; Secondary 62M05.

Key words and phrases: Autoregression function, Efficient estimator, Markov chain, Misspecification, Quasi-likelihood, Weighted nonlinear least squares estimator.

Running head: Estimating autoregression parameters.

1 Introduction

Consider an ergodic discrete-time Markov chain on the real line, with transition distribution $Q(x, dy)$ and invariant distribution $\pi(dx)$. For the conditional mean, or autoregression function, and the conditional variance write

$$\begin{aligned} m(x) &= \int y Q(x, dy), \\ v(x) &= \int (y - m(x))^2 \, Q(x, dy). \end{aligned}$$

Assume a parametric model for the conditional mean, $m(x) = m_\vartheta(x)$. We want to construct an efficient estimator for ϑ. The approach described below also works for more general time series, and also for continuous-time processes. An outline for general semimartingales is in Wefelmeyer (1993). To keep the model simple and the assumptions specific, we restrict attention to Markov chains. We will take ϑ to be one-dimensional. The generalization to finite-dimensional parameters is straightforward.

Suppose, for the moment, that we choose, in addition to the model $m(x) = m_\vartheta(x)$, a parametric model for the conditional variance, $v(x) = v_\vartheta(x)$, involving the same parameter ϑ. Then the model is a *quasi-likelihood model*. The name was introduced by Wedderburn (1974) for models with a relation between mean and variance. A generalization to discrete-time stochastic processes is due to Godambe (1985). A version for continuous time is considered in Thavaneswaran and Thompson (1986) and Hutton and Nelson (1986). Several surveys are collected in Godambe (1991). The following simple results are essentially known. Suppose we observe $X_0, \ldots, X_n$. The customary estimator for the parameter is the *maximum quasi-likelihood estimator*. It is defined as a solution of the estimating equation

$$\sum_{i=1}^{n} v_\vartheta(X_{i-1})^{-1} m'_\vartheta(X_{i-1}) \left(X_i - m_\vartheta(X_{i-1})\right) = 0, \tag{1.1}$$

with prime denoting differentiation with respect to ϑ. If $v_\vartheta(x)$ does not depend on x, one can omit v_ϑ from the estimating function, and the maximum quasi-likelihood estimator is a conditional least squares estimator in the sense of Klimko and Nelson (1978). The estimating function is a martingale. It remains a martingale if $v(x) = v_\vartheta(x)$ does not hold. Hence the maximum quasi-likelihood estimator is robust against misspecification of the conditional variance: Even if $v(x) = v_\vartheta(x)$ does not hold, the estimator is consistent and asymptotically normal, with variance

$$\pi(m_\vartheta'^2 v / v_\vartheta^2) / \left(\pi(m_\vartheta'^2 / v_\vartheta)\right)^2. \tag{1.2}$$

Here $\pi(f)$ is short for $\int f(x)\pi(dx)$. If $v(x) = v_\vartheta(x)$, then (1.2) equals

$$1/\pi(m_\vartheta'^2 / v_\vartheta). \tag{1.3}$$

The denominator is the *quasi-Fisher information*.

The model for the conditional variance is often less reliable than the model for the conditional mean. In fact, in many cases it is fairly arbitrarily chosen for the purpose of defining a maximum quasi-likelihood estimator. This point is emphasized, e.g., by Zeger and Liang (1986). If the conditional variance is misspecified, the maximum quasi-likelihood estimator is inefficient. This is obvious because its asymptotic variance (1.2) involves the misspecified function v_ϑ. Crowder (1987) gives an example in which the maximum quasi-likelihood estimator is very inefficient under misspecification. Two questions arise: Can one quantify the efficiency loss? Can one determine an efficient estimator?

To answer these questions, we return to the situation in which we have specified a parametric model $m(x) = m_\vartheta(x)$ for the autoregression function, but consider the conditional variance $v(x)$ as unknown. We show in Section 2 that then an asymptotic variance bound for regular estimators of ϑ is given by

$$1/\pi(m_\vartheta'^2/v). \tag{1.4}$$

By the Schwarz inequality, this is strictly smaller than the asymptotic variance (1.2) of the maximum quasi-likelihood estimator unless $v(x) = v_\vartheta(x)$ holds. We show in Theorem 2 that an efficient estimator, with asymptotic variance (1.4), is obtained if we replace $v_\vartheta(X_{i-1})$ in the estimating equation (1.1) by a predictor v_{i-1} for the conditional variance $v(X_{i-1})$. In particular, whatever the model for the conditional variance, the estimator is asymptotically as good as the maximum quasi-likelihood estimator when the conditional variance is correctly specified, and strictly better when it is not.

If $m_\vartheta(x) = \vartheta x$, the efficient estimator can be written in closed form. The model contains the first-order autoregressive model. We compare our estimator with the best estimator in the latter model.

We have restricted attention to models in which the parametric specification of the conditional variance does not hold. Then the maximum quasi-likelihood estimator is inefficient. If the conditional variance is correctly specified, is the maximum quasi-likelihood estimator then efficient? In other words, is the asymptotic variance bound then equal to (1.3), the inverse of the quasi-Fisher information? The answer is, again, negative. This is shown in Wefelmeyer (1992).

2 Results

Let $X_0, \ldots, X_n$ be observations from an ergodic Markov chain on the real line, with transition distribution $Q(x, dy)$ and invariant distribution $\pi(dx)$. To fix things, set $X_0 = 0$. The results remain true for other initial distributions. Suppose a parametric model for the autoregression function,

$$\int yQ(x, dy) = m_\vartheta(x). \tag{2.1}$$

Let ϑ vary in an open subset of the real line. Write $\mathcal{Q}_\vartheta$ for the set of transition distributions Q for which (2.1) holds. In the following we fix ϑ and $Q \in \mathcal{Q}_\vartheta$.

Assumptions. Let $m_\tau(x)$ be twice differentiable near $\tau = \vartheta$. Assume that the second derivative $m_\tau''(x)$ is continuous at $\tau = \vartheta$ uniformly in x, and that m_ϑ' and m_ϑ'' are π-square integrable. Let y be $Q(x, dy)$-square integrable uniformly in x,

$$\sup_x \int y^2 I(|y| > c) Q(x, dy) \to 0, \qquad c \to \infty.$$

Assume that v is bounded away from zero.

First we prove that the model is locally asymptotically normal. Let H denote the set of all bounded functions $h(x,y)$ such that

$$\int h(x,y)Q(x,dy) = 0 \qquad \text{for all } x, \tag{2.2}$$

$$\int yh(x,y)Q(x,dy) = m'_\vartheta(x) \qquad \text{for all } x. \tag{2.3}$$

We construct a local model in which these functions appear as *score functions.* Let $L_2(\pi \otimes Q)$ denote the space of functions $f(x,y)$ with finite second moment, written as

$$\pi \otimes Q(f^2) = \int\int f(x,y)^2 \pi(dx) Q(x,dy).$$

Write P_n for the joint distribution of $X_0, \ldots, X_n$ if Q is true.

Theorem 1 *For each $h \in H$ and $u \in \mathbb{R}$ there exists a transition distribution $Q^{nuh} \in \mathcal{Q}_{\vartheta+n^{-1/2}u}$ such that*

$$\log dP_n^{nuh}/dP_n = un^{-1/2}\sum_{i=1}^n h(X_{i-1},X_i) - \frac{1}{2}u^2\pi \otimes Q(h^2) + o_{P_n}(1),$$

$$n^{-1/2}\sum_{i=1}^n h(X_{i-1},X_i) \Rightarrow N_h \quad \textit{under } P_n,$$

where N_h is normal with mean zero and variance $\pi \otimes Q(h^2)$.

Conditions for local asymptotic normality for Markov chains are well known, starting with Roussas (1965). For a nonparametric model see Penev (1991). We will use a version of the conditions given by Höpfner (1993) for Markov step processes. The main point of Theorem 1 is that we can find Q^{nuh} which belong to $\mathcal{Q}_{\vartheta+n^{-1/2}u}$.

The model P_n^{nuh}, $h \in H$, $u \in \mathbb{R}$, is the *local model* at Q. We comment briefly on the choice of local model. The smaller the local model, the smaller the asymptotic variance bound for estimators, and the larger the class of competing estimators. Our aim here is a local model which is not too large and hence does not exclude reasonable estimators from competing, but still large enough for the corresponding variance bound to be attained globally, i.e. for all Q, by some estimator. The last requirement holds by Theorem 2 below. It describes an estimator which attains the variance bound. The first requirement is fulfilled because of Theorem 1. It says that the local model lies within the given model. We need (2.2) for $Q^{nuh}(x,\cdot)$ to be a probability measure, and (2.3) for Q^{nuh} to be in $\mathcal{Q}_{\vartheta+n^{-1/2}u}$, i.e.

$$\int yQ^{nuh}(x,dy) = m_{\vartheta+n^{-1/2}u}(x).$$

The boundedness of h simplifies the construction of Q^{nuh}. The price: H is not closed. Let $\overline{H}$ denote the closure of H in $L_2(\pi \otimes Q)$.

We recall some well-known consequences of local asymptotic normality. A convenient reference is Greenwood and Wefelmeyer (1990).

First we describe an asymptotic variance bound for estimators of ϑ. Call an estimator $\hat{\vartheta}_n$ *regular* at Q with *limit* L if, for all $h \in H$ and $u \in \mathbb{R}$,

$$n^{1/2}(\hat{\vartheta}_n - \vartheta - n^{-1/2}u) \Rightarrow L \qquad \text{under } P_n^{nuh}.$$

By the convolution theorem,

$$L = M + N \qquad \text{in distribution,}$$

where M is independent of N, the random variable N is normal with mean zero and variance I^{-1}, and

$$I = \pi \otimes Q(s^2).$$

Here s is the *efficient score function* in $\overline{H}$. It minimizes $\pi \otimes Q(h^2)$ over $h \in \overline{H}$. Hence it is characterized by

$$\pi \otimes Q(s^2) = \pi \otimes Q(sh) \qquad \text{for } h \in H. \tag{2.4}$$

We prove that the efficient score function is given by

$$s(x, y) = v(x)^{-1} m'_\vartheta(x)\,(y - m_\vartheta(x))\,, \tag{2.5}$$

and that

$$I = \pi(m'^2_\vartheta / v). \tag{2.6}$$

First we show that the function s defined in (2.5) fulfills (2.4). Use (2.1) to write

$$\begin{aligned} \pi \otimes Q(s^2) &= \int v(x)^{-2} m'_\vartheta(x)^2 \int (y - m_\vartheta(x))^2\, Q(x, dy)\pi(dx) \\ &= \pi(m'^2_\vartheta / v). \end{aligned}$$

This is (2.6). Use (2.1) to (2.3) to write

$$\begin{aligned} \pi \otimes Q(sh) &= \int v(x)^{-1} m'_\vartheta(x) \int (y - m_\vartheta(x))\, h(x, y) Q(x, dy)\pi(dx) \\ &= \pi(m'^2_\vartheta / v). \end{aligned}$$

Comparing the two equations, we obtain (2.4). To prove $s \in \overline{H}$, it remains to check that (2.2) and (2.3) hold for $h = s$. This follows with relation (2.1), and we are done.

If the true transition distribution fulfills the condition $v = v_\vartheta$ on the conditional variance, then the efficient score function equals the *quasi-score function*

$$v_\vartheta(x)^{-1} m'_\vartheta(x)\,(y - m_\vartheta(x))$$

on which the estimating equation (1.1) of the maximum quasi-likelihood estimator is based.

Because of the convolution theorem, we may call I^{-1} an asymptotic variance bound for regular estimators of ϑ. An estimator is *efficient* at Q if its limit distribution under P_n is N. It is well known that an estimator $\hat{\vartheta}_n$ is regular and efficient at Q if and only if it is asymptotically linear with influence function equal to $I^{-1}s$,

$$n^{1/2}(\hat{\vartheta}_n - \vartheta) = I^{-1}n^{-1/2}\sum_{i=1}^{n} s(X_{i-1}, X_i) + o_{P_n}(1). \tag{2.7}$$

Again, a convenient reference is Greenwood and Wefelmeyer (1990). We use this characterization to prove Theorem 2. To describe an efficient estimator, we need an estimator for the conditional variance $v(x)$ appearing in the efficient score function (2.5). Let v_n be a strongly consistent predictor for $v(X_n)$. For example, take $v_n = \hat{v}_n(X_n)$ with $\hat{v}_n$ a uniformly strongly consistent kernel estimator for the conditional variance function v. Conditions for uniform strong consistency of $\hat{v}_n$ are given in Collomb (1984) and Truong and Stone (1992). We do not repeat them here.

Theorem 2 *Any $n^{1/2}$-consistent solution $\vartheta = \hat{\vartheta}_n$ of*

$$n^{-1/2}\sum_{i=1}^{n} v_{i-1}^{-1} m_\vartheta'(X_{i-1})\left(X_i - m_\vartheta(X_{i-1})\right) = o_{P_n}(1)$$

is regular and efficient at Q.

If P_n is true, $X_i - m_\vartheta(X_{i-1})$ is a martingale increment, and $v_{i-1}^{-1}m_\vartheta'(X_{i-1})$ is predictable. Hence the estimating function in Theorem 2 is a martingale. The proof of Theorem 2 makes use of this property. This is why we estimate $v(X_{i-1})$ by a predictor v_{i-1} in place of an estimator making full use of the observations $X_0, \ldots, X_n$, like $\hat{v}_n(X_{i-1})$.

Corollary. *Assume $m_\vartheta(x) = \vartheta x$. Then the weighted least squares estimator*

$$\hat{\vartheta}_n = \sum_{i=1}^{n} v_{i-1}^{-1} X_{i-1} X_i \Big/ \sum_{i=1}^{n} v_{i-1}^{-1} X_{i-1}^2 \tag{2.8}$$

is regular and efficient at Q.

The model considered in the Corollary, with $m_\vartheta(x) = \vartheta x$, contains the autoregressive model

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

where ϑ_i are i.i.d. with unknown mean zero density p. This submodel is described by transition distributions of the form

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

Then $v(x) = \int y^2 p(y)dy$ does not depend on x, and the estimator (2.8) defined in the Corollary has asymptotic variance $1-\vartheta^2$. On the other hand, as shown by Huang (1986), the variance bound for regular estimators in the autoregressive model is

$$(1-\vartheta^2)/\sigma^2 I, \tag{2.9}$$

where $\sigma^2 = \int y^2 p(y)dy$ and $I = \int \ell'(y)^2 p(y)dy$, with ℓ' the logarithmic derivative of p. An estimator attaining the variance bound (2.9) is constructed in Kreiss (1987). By the Schwarz inequality, (2.9) is strictly smaller than $1-\vartheta^2$ unless $\ell'(y)$ is proportional to y, which would be the case when p is a normal density. Except in this case, the weighted least squares estimator (2.8) is not efficient if the autoregressive model is true. On the other hand, it is robust in that it remains consistent as long as the process is Markov with conditional mean ϑx, and it is efficient in this larger model.

3 Proofs

Proof of Theorem 1. Fix $h \in H$ and $u \neq 0$. To construct Q^{nuh}, consider first

$$Q_0^{nuh}(x,dy) = \left(1 + n^{-1/2} u h(x,y)\right) Q(x,dy).$$

Since h is bounded and (2.2) holds, this is a transition distribution for n sufficiently large. By (2.1) and (2.3),

$$\int y Q_0^{nuh}(x,dy) = m_\vartheta(x) + n^{-1/2} u m'_\vartheta(x).$$

On the other hand, there exists $\vartheta_{nu}(x)$ between ϑ and $\vartheta + n^{-1/2}u$ such that

$$m_{\vartheta+n^{-1/2}u}(x) = m_\vartheta(x) + n^{-1/2} u m'_{\vartheta_{nu}(x)}(x).$$

Since $m_\tau(x)$ is twice differentiable, relation (2.1) holds up to $O(n^{-1})$ for $Q = Q_0^{nuh}$ and $\vartheta = \vartheta + n^{-1/2}u$. We want a transition distribution Q^{nuh} such that this holds exactly. Then $Q^{nuh} \in \mathcal{Q}_{\vartheta+n^{-1/2}u}$. To this end, we add to h in the definition of Q_0^{nuh} a function proportional to y, properly truncated so that Q^{nuh} is nonnegative for large n,

$$Q^{nuh}(x,dy) = \left(1 + n^{-1/2} u \left(h(x,y) + r_n(x,y)\right)\right) Q(x,dy),$$

where

$$r_n(x,y) = v_n(x)^{-1} \left(m'_{\vartheta_{nu}(x)}(x) - m'_\vartheta(x)\right) \left(b_n(y) - \int b_n(y) Q(x,dy)\right)$$

with

$$\begin{aligned} b_n(y) &= y I(|y| \leq n^{1/4}), \\ v_n(x) &= \int b_n(y)^2 Q(x,dy) - m_\vartheta(x) \int b_n(y) Q(x,dy). \end{aligned}$$

Since y is $Q(x,dy)$-square integrable uniformly in x, the function v_n is uniformly close to v, hence also bounded away from zero. Hence $r_n(x,y)$ is bounded. By construction,

$$\int r_n(x,y)Q(x,dy) = 0.$$

Hence Q^{nuh} is a transition distribution. Also,

$$\int y r_n(x,y)Q(x,dy) = m'_{\vartheta_{nu}(x)}(x) - m'_\vartheta(x),$$

so that

$$\int y Q^{nuh}(x,dy) = m_{\vartheta+n^{-1/2}u}(x).$$

In other words, $Q^{nuh} \in \mathcal{Q}_{\vartheta+n^{-1/2}u}$.

It remains to prove local asymptotic normality. This is basically due to Roussas (1965). Höpfner (1993) proves local asymptotic normality for Markov step processes under weaker assumptions. His argument is easily modified to cover discrete-time Markov processes. To apply it, we need only check an appropriate version of Hellinger differentiability for Q^{nuh}, condition $H1''$ in Höpfner *et al.* (1990). Hellinger differentiability is implied by a corresponding differentiability in quadratic mean,

$$\int r_n(x,y)^2 Q(x,dy) \le R_n(x) \tag{3.1}$$

with $R_n \downarrow 0$ pointwise and R_n π-integrable for large n. To prove (3.1), recall that $m''_\tau(x)$ is continuous at $\tau = \vartheta$ uniformly in x. Hence there exists $\varepsilon > 0$ such that for n sufficiently large,

$$\int r_n(x,y)^2 Q(x,dy) \le n^{-1/2}u^2 v_n(x)^{-2}\left(m''_\vartheta(x)+\varepsilon\right)^2.$$

As seen above, v_n is bounded away from zero. The function m''_ϑ is π-square integrable, and (3.1) follows.

Proof of Theorem 2. Let $\varepsilon > 0$ and $c > 0$. Choose $c > 0$ and N such that

$$P_n\{n^{1/2}|\hat\vartheta_n - \vartheta| > c\} \le \varepsilon \qquad \text{for } n \ge N,$$
$$P\{\sup_{n\ge N}\sup_x |v_n(x) - v(x)| > \varepsilon\} \le \varepsilon.$$

Restrict attention to $X_0, X_1, \ldots$ such that for $n \ge N$ and $x \in \mathbb{R}$,

$$n^{1/2}|\hat\vartheta_n - \vartheta| \le c,$$
$$|v_n(x) - v(x)| \le \varepsilon.$$

Rewrite the estimating equation as

$$
\begin{aligned}
o_{P_n}(1) &= n^{-1/2}\sum v_{i-1}^{-1} m'_{\hat\vartheta_n}(X_{i-1})\left(X_i - m_{\hat\vartheta_n}(X_{i-1})\right)\\
&= n^{-1/2}\sum v_{i-1}^{-1} m'_{\hat\vartheta_n}(X_{i-1})\left(X_i - m_{\vartheta}(X_{i-1})\right)\\
&\quad - n^{-1/2}\sum v_{i-1}^{-1} m'_{\hat\vartheta_n}(X_{i-1})\left(m_{\hat\vartheta_n}(X_{i-1}) - m_{\vartheta}(X_{i-1})\right)\\
&= n^{-1/2}\sum v(X_{i-1})^{-1} m'_{\vartheta}(X_{i-1})\left(X_i - m_{\vartheta}(X_{i-1})\right)\\
&\quad - n^{1/2}(\hat\vartheta_n - \vartheta)\pi(m'^2_\vartheta/v)\\
&\quad + n^{-1/2}\sum v_{i-1}^{-1}\left(m'_{\hat\vartheta_n}(X_{i-1}) - m'_{\vartheta}(X_{i-1})\right)\left(X_i - m_{\vartheta}(X_{i-1})\right)\\
&\quad + n^{-1/2}\sum \left(v_{i-1}^{-1} - v(X_{i-1})^{-1}\right) m'_{\vartheta}(X_{i-1})\left(X_i - m_{\vartheta}(X_{i-1})\right)\\
&\quad - n^{1/2}(\hat\vartheta_n - \vartheta)n^{-1}\sum\left(v(X_{i-1})^{-1} m'_\vartheta(X_{i-1})^2 - \pi(m'^2_\vartheta/v)\right)\\
&\quad - n^{-1/2}\sum\left(v_{i-1}^{-1} m'_{\hat\vartheta_n}(X_{i-1}) - v(X_{i-1})^{-1} m'_\vartheta(X_{i-1})\right)\left(m_{\hat\vartheta_n}(X_{i-1}) - m_\vartheta(X_{i-1})\right)\\
&\quad - n^{-1/2}\sum v(X_{i-1})^{-1} m'_\vartheta(X_{i-1})\left(m_{\hat\vartheta_n}(X_{i-1}) - m_\vartheta(X_{i-1}) - (\hat\vartheta_n - \vartheta)m'_\vartheta(X_{i-1})\right).
\end{aligned}
$$

We have to show that the third to seventh right hand terms are of order $o_{P_n}(1)$. Then Theorem 2 follows by solving the estimating equation for $\hat\vartheta_n - \vartheta$ and applying a martingale central limit theorem. Split the sums into sums from 1 to N and from $N+1$ to n. The terms involving sums from 1 to N go to zero pointwise because of the factors $n^{-1/2}$ or n^{-1}. Hence we may assume w.l.g. that $N = 0$. The assumptions imply that v is bounded and bounded away from zero. We assume w.l.g. that $1 \le v \le c$. In the following we show that the third to seventh terms are small in probability if ε is small. They are negligible if we take ε to zero. The third term is the crucial one.

Third term. Choose ϑ_{ni} between ϑ and $\hat\vartheta_n$ such that

$$
m'_{\hat\vartheta_n}(X_{i-1}) = m'_\vartheta(X_{i-1}) + (\hat\vartheta_n - \vartheta)m''_{\vartheta_{ni}}(X_{i-1}).
$$

Write the third term as

$$
(\hat\vartheta_n - \vartheta)A_n + n^{1/2}(\hat\vartheta - \vartheta)B_n
$$

with

$$
\begin{aligned}
A_n &= n^{-1/2}\sum v_{i-1}^{-1} m''_\vartheta(X_{i-1})\left(X_i - m_\vartheta(X_{i-1})\right),\\
B_n &= n^{-1}\sum v_{i-1}^{-1}\left(m''_{\vartheta_{ni}}(X_{i-1}) - m''_\vartheta(X_{i-1})\right)\left(X_i - m_\vartheta(X_{i-1})\right).
\end{aligned}
$$

Note that A_n has predictable quadratic variation

$$
n^{-1}\sum v_{i-1}^{-2} m''_\vartheta(X_{i-1})^2 v(X_{i-1}) \le (1-\varepsilon)^{-2} c n^{-1}\sum m''_\vartheta(X_{i-1})^2.
$$

Since m''_ϑ is π-square integrable, it follows that $A_n = O_{P_n}(1)$. It is only here that we need a predictor v_{i-1} based on $X_0,\dots,X_{i-1}$ rather than an estimator $\hat v_n(X_{i-1})$ based on $X_0,\dots,X_n$. The average B_n is small in probability since $m''_\tau - m''_\vartheta$ is small for τ near ϑ

and $y - m_\vartheta(x)$ has variance $\pi(v) \le c$. Hence the third term is small by $n^{1/2}$-consistency of $\hat\vartheta_n$.

Fourth term. We have $|v_{i-1}^{-1} - v(X_{i-1})^{-1}| \le \varepsilon$. The function $m'_\vartheta(x)\,(y - m_\vartheta(x))$ has variance

$$\pi(m'^2_\vartheta v) \le c\pi(m'^2_\vartheta).$$

Hence the fourth term is small by Chebyshev's inequality.

Fifth term. The average is of order $o_{P_n}(1)$ by the law of large numbers.

Last two terms. Choose ϑ_{ni} between ϑ and $\hat\vartheta_n$ such that

$$m_{\hat\vartheta_n}(X_{i-1}) = m_\vartheta(X_{i-1}) + (\hat\vartheta_n - \vartheta)m'_{\vartheta_{ni}}(X_{i-1}).$$

Then the sixth term can be written as

$$n^{1/2}(\hat\vartheta_n - \vartheta)n^{-1}\sum \left(v_{i-1}^{-1}m'_{\hat\vartheta_n}(X_{i-1}) - v(X_{i-1})^{-1}m'_\vartheta(X_{i-1})\right) m'_{\vartheta_{ni}}(X_{i-1}).$$

Similarly, the seventh term can be written as

$$n^{1/2}(\hat\vartheta_n - \vartheta)n^{-1}\sum v(X_{i-1})^{-1}m'_\vartheta(X_{i-1})\left(m'_{\vartheta_{ni}}(X_{i-1}) - m'_\vartheta(X_{i-1})\right).$$

As before one checks that the averages are small in probability.

References

Collomb, G. (1984). Propriétés de convergence presque complète du prédicteur à noyeau. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 66, 441–460.

Crowder, M. (1987). On linear and quadratic estimating functions. *Biometrika* 74, 591–597.

Godambe, V.P. (1985). The foundations of finite sample inference in stochastic processes. *Biometrika* 72, 419–428.

Godambe, V.P., ed. (1991). *Estimating Functions.* Oxford University Press.

Greenwood, P.E. and Wefelmeyer, W. (1990). Efficiency of estimators for partially specified filtered models. *Stochastic Process. Appl.* 36, 353–370.

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

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

Huang, W.-M. (1986). A characterization of limiting distributions of estimators in an autoregressive process. *Ann. Inst. Statist. Math.* 38, 137–144.

Hutton, J.E. and Nelson, P.I. (1986). Quasi-likelihood estimation for semimartingales. *Stochastic Process. Appl.* 22, 245–257.

Klimko, L.A. and Nelson, P.I. (1978). On conditional least squares estimation for stochastic processes. *Ann. Statist.* 6, 629–642.

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

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, 978–992.

Thavaneswaran, A. and Thompson, M.E. (1986). Optimal estimation for semimartingales. *J. Appl. Probab.* 23, 409–427.

Truong, Y.K. and Stone, C.J. (1992). Nonparametric function estimation involving time series. *Ann. Statist.* 20, 77–97.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 61, 439–447.

Wefelmeyer, W. (1992). Quasi-likelihood models and optimal inference. To appear in: *Ann. Statist.*

Wefelmeyer, W. (1993). Estimating functions and efficiency in a filtered model. In: H. Niemi, G. Högnäs, A.N. Shiryaev and A.V. Melnikov, Eds., *Frontiers in Pure and Applied Probability*, VSP, Utrecht, 287–295.

Zeger, S.L. and Ljang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42, 121–130.