

Efficient estimation in Markov chain models: an introduction

Wolfgang Wefelmeyer
University of Siegen

Abstract

We outline the theory of efficient estimation for semiparametric Markov chain models, and illustrate in a number of simple cases how the theory can be used to determine lower bounds for the asymptotic variance of estimators and to construct efficient estimators. In particular, we consider estimation of stationary distributions of Markov chains, of autoregression parameters and innovation distributions in AR- and ARCH-models and more general time series, and of parameters in quasi-likelihood models.

AMS 1991 subject classifications. Primary 62M05; secondary 62F12, 62F35, 62G20, 62M10.

Key words and Phrases. Variance bound, empirical estimator, martingale approximation, maximum likelihood estimator, Kullback-Leibler information, estimating equation, misspecified model, weighted least squares, conditional heteroscedasticity, quasi-likelihood, Markov chain Monte Carlo, Gibbs sampler.

1 Introduction

The basic example of a time series is the autoregressive process

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

where the innovations ε_i are independent and identically distributed with mean zero and finite variance. For $|\alpha| < 1$ this is an ergodic Markov chain. We may want to estimate the autoregression parameter α , the distribution of the ε_i , or the stationary distribution of the X_i . The classical estimator for α is the least squares estimator $\hat{\alpha} = \sum X_{i-1}X_i / \sum X_{i-1}^2$. The distribution function of the innovations can be estimated by the empirical distribution function $\frac{1}{n} \sum 1(\hat{\varepsilon}_i \leq r)$ based on the estimated innovations $\hat{\varepsilon}_i = X_i - \hat{\alpha}X_{i-1}$. The obvious estimator for the stationary distribution function is the empirical distribution function $\frac{1}{n} \sum 1(X_i \leq r)$.

None of these three estimators is efficient. How far are they from being efficient? How can one find better estimators? In the following sections we outline how one determines lower bounds for the asymptotic variance of estimators in general semiparametric Markov chain models. Then we characterize efficient estimators and indicate how one uses the characterization to construct such estimators in specific cases.

Our approach requires the model to be ‘locally asymptotically normal’. In Section 2 we introduce this concept for general Markov chain models with possibly infinite-dimensional parameter and illustrate it with a few simple examples. In Section 3 we consider the problem of estimating a one-dimensional function of the parameter and determine an optimal estimator within a simple class of estimators: the ‘asymptotically linear’ and ‘regular’ ones. Section 4 considers martingale estimating equations and indicates when they lead to asymptotically linear and regular estimators. Section 5 shows that the optimal asymptotically linear and regular estimator is already ‘efficient’, i.e., optimal among *all* regular estimators. The presentation is rigorous in Sections 2, 3 and 5, and heuristic in the others.

In Sections 6 to 9 we treat examples: nonparametric models in Section 6, autoregressive models in Section 7, quasi-likelihood models in Section 8. Section 9 briefly describes some other types of Markov chain models that are amenable to our approach. To keep the exposition simple, we restrict attention to *first*-order Markov chains and estimation of *one*-dimensional functions of the parameter. The extension to higher-order Markov chains and higher-dimensional functions is straightforward. Extensions to infinite-dimensional functions and to other types of processes are also possible.

A basic reference for convergence of Markov chains is Meyn and Tweedie (1994). The results on efficient estimation in Sections 2, 3 and 5 generalize those for independent and identically distributed observations, for which we refer to Bickel et al. (1993).

2 Local asymptotic normality for Markov chains

In this section we give conditions under which a general Markov chain model with possibly infinite-dimensional parameter is locally asymptotically normal. This means that if we fix a parameter and rescale the parametrization approximately, the model is approximated by a Gaussian shift model as the sample size increases. The setting will be rather abstract in order to cover all applications in the following sections. Before we describe the general setting, we begin with a simple case, the full nonparametric model, in which the transition distribution itself plays the role of a parameter.

Let X_0, \dots, X_n be realizations of a homogeneous Markov chain on some measurable state space (E, \mathcal{E}) . For notational simplicity we will always assume that the chain starts at a fixed value $X_0 = x_0$, but we continue to write X_0 .

Let us first look at the full nonparametric model, in which no structural assumptions are made on the transition distribution. We ‘parametrize’ the model by the transition distribution and view the ‘parameter space’, the family of all transition distributions, as a manifold.

We fix a transition distribution $Q(x, dy)$ and assume that under it the chain is ergodic with invariant distribution $\pi(dx)$. We write $\pi \otimes Q(dx, dy) = \pi(dy)Q(x, dy)$ and $Q(x, f) = \int Q(x, dy)f(x, y)$ and introduce the Hilbert space

$$H = \{h \in L_2(\pi \otimes Q) : Q(x, h) = 0 \text{ for } x \in E\}.$$

The manifold of transition distributions is smooth with *tangent space* H in the following sense. For $h \in H$ bounded,

$$(2.1) \quad Q_{nh}(x, dy) = Q(x, dy)(1 + n^{-1/2}h(x, y))$$

defines a transition distribution with *derivative* h . We will find it convenient to consider sequences indexed by $n^{-1/2}$.

If h is not bounded, we must modify the definition (2.1) to obtain a true conditional *distribution*. We will see later, however, that it suffices to restrict attention to bounded h because they are dense in H .

Remark. To interpret H as a tangent space, we identify a transition distribution $R(x, dy)$ with $(dR/dQ)(x, dy)^{1/2} - 1$. This function is automatically in $L_2(\pi \otimes Q)$ and is 0 for $R = Q$. The identification implies in particular that we ignore all transition distributions $R(x, dy)$ which are not absolutely continuous with respect to $Q(x, dy)$. We will see later that our ‘tangent space’ is still big enough for our purposes. \square

Let P denote the joint law of X_0, \dots, X_n if the chain is generated by Q , and P_{nh} the corresponding law if Q_{nh} is true. By ergodicity,

$$(2.2) \quad \frac{1}{n} \sum_{i=1}^n h(X_{i-1}, X_i)^2 \rightarrow \pi \otimes Qh^2.$$

From this result and a Taylor expansion we obtain a stochastic approximation of the log-likelihood,

$$(2.3) \quad \begin{aligned} \log \frac{dP_{nh}}{dP} &= \sum_{i=1}^n \log \frac{dQ_{nh}}{dQ}(X_{i-1}, X_i) \\ &= n^{-1/2} \sum_{i=1}^n h(X_{i-1}, X_i) - \frac{1}{2} \pi \otimes Qh^2 + o_P(1). \end{aligned}$$

By a martingale central limit theorem,

$$(2.4) \quad n^{-1/2} \sum_{i=1}^n h(X_{i-1}, X_i) \Rightarrow (\pi \otimes Qh^2)^{1/2} N \quad \text{under } P.$$

Here N denotes a random variable with standard normal distribution. Relations (2.3) and (2.4) constitute a nonparametric version of *local asymptotic normality*. The reason for choosing the rate $n^{-1/2}$ in (2.1) is now apparent: with this choice, the likelihood ratio has a nondegenerate limit.

Let us now consider *submodels* of the full nonparametric model. Such submodels may be described in different ways, three of which will appear in the following sections:

- There is a restriction on the transition distribution, say the chain is reversible (Example 5 in Section 6), or $Q(x, dy)$ is symmetric about x .
- There is a parametric family of restrictions on the transition distribution, say on the conditional mean,

$$\int Q(x, dy)y = \alpha x,$$

and we may want to estimate that parameter. See Section 8, and also Example 10 in Section 9.

- The transition distribution is parametrized (with a possibly infinite-dimensional parameter), as in the AR(1)-model,

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

with α unknown and p a mean 0 density which is possibly also unknown. See Example 3 below and Section 7, and also Example 11 in Section 9.

For the first two types of model we will continue using the transition distribution as a parameter. This can also be done for the third type, but is usually more convenient, and certainly more common, to parametrize by the obvious parameters.

Let Θ be a possibly infinite-dimensional set, the *parameter space*, and Q_ϑ , $\vartheta \in \Theta$, a family of transition distributions on the state space (E, \mathcal{E}) . We view Θ as a manifold, fix $\vartheta \in \Theta$ so that $Q = Q_\vartheta$ is ergodic with invariant distribution $\pi = \pi_\vartheta$, and assume that Θ is smooth in the following sense. There is a linear space K , the *tangent space*, and a linear map $D : K \rightarrow H$, and for each $k \in K$ there is a sequence ϑ_{nk} such that $Q_{nk} = Q_{\vartheta_{nk}}$ is *Hellinger differentiable* with *derivative* Dk ,

$$(2.5) \quad \int Q(x, dy) \left(\left(\frac{dQ_{nk}}{dQ}(x, y) \right)^{1/2} - 1 - \frac{1}{2}n^{-1/2}(Dk)(x, y) \right)^2 \leq n^{-1}r_n(x),$$

where r_n decreases to 0 pointwise and is π -integrable for large n . This version of Hellinger differentiability is due to Höpfner et al. (1990).

Remark. Our description of the tangent space omits one important feature: The space K should contain *all* directions k from which we can approximate ϑ within Θ . We have already mentioned in the previous Remark that this is not always possible. In the applications we simply try to make K as large as we can. In Section 3 we will determine a lower bound for the asymptotic variance of estimators of functions $t(\vartheta)$. The bound depends on K . It is the larger the larger K . We will know that K was chosen large enough whenever we can construct an estimator which attains the bound. \square

The operator induces an inner product on K ,

$$(2.6) \quad \langle k, k' \rangle = \pi \otimes Q(Dk \cdot Dk'),$$

and a corresponding norm $\|k\| = \langle k, k \rangle^{1/2}$. Write P_{nk} for the joint law of X_0, \dots, X_n if Q_{nk} is true. A Taylor expansion now gives local asymptotic normality of the form

$$(2.7) \quad \log \frac{dP_{nk}}{dP} = n^{-1/2} \sum_{i=1}^n (Dk)(X_{i-1}, X_i) - \frac{1}{2} \|k\|^2 + o_P(1).$$

As in (2.4),

$$(2.8) \quad n^{-1/2} \sum_{i=1}^n (Dk)(X_{i-1}, X_i) \Rightarrow \|k\|N \quad \text{under } P.$$

Local asymptotic normality for Markov chains is basically due to Roussas (1965). The nonparametric version (2.3), (2.4) is in Penev (1991). The parametric version may be obtained by modifying Höpfner (1993a) who treats Markov step processes; see also Höpfner (1993b).

The nonparametric model may be considered as a special case of the (infinite-dimensional) parametric model:

Example 1. (*Full nonparametric model.*) If the transition distribution Q is completely unknown, parametrize Q by itself, set $K = H$ and define, for bounded $h \in H$, the perturbed transition distribution Q_{nh} through (2.1). Then Hellinger differentiability (2.5) holds with $Dh = h$, and the inner product (2.6) is

$$\langle h, h' \rangle = \pi \otimes Q(hh'),$$

the natural inner product on $L_2(\pi \otimes Q)$. □

At the other end of the spectrum are the models with a one-dimensional parameter.

Example 2. (*One-dimensional parameter.*) If $\Theta \subset \mathbf{R}$, set $K = \mathbf{R}$ and $\vartheta_{nr} = \vartheta + n^{-1/2}r$. Write ℓ' for the Hellinger derivative of Q_τ at $\tau = \vartheta$,

$$(2.9) \quad \int Q(x, dy) \left(\left(\frac{dQ_\tau}{dQ}(x, y) \right)^{1/2} - 1 - \frac{1}{2}(\tau - \vartheta)\ell'(x, y) \right)^2 \leq (\tau - \vartheta)^2 r_\tau(x)$$

with r_τ decreasing to 0 pointwise for $\tau \rightarrow \vartheta$ and π -integrable for τ close to ϑ . Then $Dr = r\ell'$, and the inner product (2.6) is

$$\langle r, r' \rangle = rr'I$$

with $I = \pi \otimes Q\ell'$ the *Fisher information*. The inner product is not the natural one. We have local asymptotic normality of the form

$$(2.10) \quad \log \frac{dP_{nr}}{dP} = rn^{-1/2} \sum_{i=1}^n \ell'(X_{i-1}, X_i) - \frac{1}{2} r^2 I + o_P(1).$$

The multivariate version of this example is straightforward. \square

Example 2 describes a whole class of models, in particular the following autoregressive model.

Example 3. (*AR(1) with normal innovations.*) Let $X_i = X_{i-1} + \varepsilon_i$ with ε_i i.i.d. and standard normal. The X_i form a Markov chain with transition distribution

$$Q(x, dy) = \varphi(y - \vartheta x)dx,$$

where φ is the standard normal density. If $\vartheta < 1$, then the chain is ergodic, and Hellinger differentiability holds with $\ell'(x, y) = x(y - \vartheta x)$. Hence $(Dr)(x, y) = rx(y - \vartheta x)$, the Fisher information is

$$I = \pi \otimes Q\ell'^2 = \int \pi(dx)x^2 \int \varphi(y - \vartheta x)dy(y - \vartheta x)^2 = \mathbf{E} X^2,$$

and the inner product (2.6) is

$$\langle r, r' \rangle = rr' \mathbf{E} X^2.$$

We have local asymptotic normality of the form

$$\log \frac{dP_{nr}}{dP} = rn^{-1/2} \sum_{i=1}^n X_{i-1}(X_i - \vartheta X_{i-1}) - \frac{1}{2}r^2 \mathbf{E} X^2 + o_P(1).$$

\square

3 Asymptotically linear and regular estimators

As in the previous section, we consider a general Markov chain model with possibly infinite-dimensional parameter. We now want to estimate a one-dimensional function of the parameter. We introduce a class of estimators, the ‘asymptotically linear’ estimators, whose properties are particularly easy to study. Our aim is to find an optimal estimator in this class. We show that this problem has a meaningful answer if we further restrict attention to ‘regular’ estimators. The restriction to asymptotically linear estimators is justified by the characterization (5.2) of efficient estimators.

Let $Q_\vartheta, \vartheta \in \Theta$, be a family of transition distributions as in Section 2. In addition, we consider a function $t : \Theta \rightarrow \mathbf{R}$ which we want to estimate. Fix $\vartheta \in \Theta$. We call an estimator T_n *asymptotically linear* for t at ϑ with *influence function* f if $f \in H$ and

$$(3.1) \quad n^{1/2}(T_n - t(\vartheta)) = n^{-1/2} \sum_{i=1}^n f(X_{i-1}, X_i) + o_P(1).$$

From the martingale central limit theorem used in (2.4),

$$n^{1/2}(T_n - t(\vartheta)) \Rightarrow (\pi \otimes Qf^2)^{1/2}N \quad \text{under } P.$$

Hence the asymptotic variance of an asymptotically linear estimator with influence function f is $\pi \otimes Qf^2$

What is a good asymptotically linear estimator? Without further restricting the class of estimators, this question has no meaningful answer: the asymptotic variance can be made 0 at ϑ by taking $f = 0$. To exclude estimators which have small variance for certain ϑ at the expense of other parameters, we will restrict attention to ‘regular’ estimators, in the sense that their distribution converges *continuously* in the parameter to a limit distribution.

For a proper definition, pick a tangent space K and sequences ϑ_{nk} as in Section 2. An estimator T_n is called *regular* for t at ϑ with limit L if

$$(3.2) \quad n^{1/2}(T_n - t(\vartheta_{nk})) \Rightarrow L \quad \text{under } P_{nk} \text{ for } k \in K.$$

If the function t is smooth, regularity implies a restriction on the influence function of an asymptotically linear estimator. To describe what we mean by smooth, we recall that $Q_{nk} = Q_{\vartheta_{nk}}$ has derivative Dk in the sense of (2.5). The function t is called *differentiable* at ϑ with *gradient* h^t if $h^t \in H$ and

$$(3.3) \quad n^{1/2}(t(\vartheta_{nk}) - t(\vartheta)) \rightarrow \pi \otimes Q(Dk \cdot h^t) \quad \text{for } k \in K.$$

The *canonical gradient* is the projection Dk^t of an arbitrary gradient h^t into DK . For the canonical gradient, (3.3) reads

$$(3.4) \quad n^{1/2}(t(\vartheta_{nk}) - t(\vartheta)) \rightarrow \langle k, k^t \rangle \quad \text{for } k \in K.$$

Conversely, any gradient h^t has the property that $h^t - Dk^t$ is orthogonal to DK ,

$$(3.5) \quad \pi \otimes Q(Dk \cdot (h^t - Dk^t)) = 0 \quad \text{for } k \in K.$$

In particular, the canonical gradient is the shortest gradient,

$$(3.6) \quad \|k^t\| = \pi \otimes Q(Dk^t)^2 \leq \pi \otimes Q(h^t)^2.$$

Remark. The canonical gradient is not always easy to calculate. Sometimes it is easier to find another gradient first (which may, in turn, be canonical in a larger model). One can then try to project that gradient into the tangent space.

Another approach is possible when the tangent space K comes equipped with some inner product $\langle k, k' \rangle_0$, say, as is usually the case. It may then be possible to find the gradient k_0^t of t with respect to this inner product,

$$n^{1/2}(t(\vartheta_{nk}) - t(\vartheta)) \rightarrow \langle k, k_0^t \rangle_0 \quad \text{for } k \in K.$$

Comparing with (3.3) and (3.5), we see that the canonical gradient Dk^t is now determined by

$$\langle k, k_0^t \rangle_0 = \langle Dk, Dk^t \rangle \quad \text{for } k \in K.$$

If D has an adjoint $D^* : H \rightarrow K$, we have

$$\langle Dk, Dk^t \rangle = \langle k, D^* Dk^t \rangle_0.$$

Hence, if $D^* D$ has an inverse, the canonical gradient is Dk^t with

$$k^t = (D^* D)^{-1} k_0^t.$$

It suffices to find an operator C such that

$$\langle Dk, Dk^t \rangle = \langle k, Ck^t \rangle_0.$$

Then the canonical gradient is $k^t = C^{-1} k_0^t$. It may happen that C^{-1} is difficult to determine but that C can be written as a perturbation of the identity operator, say $C = I - B$. If B is not too large, C^{-1} may then be written as the von Neumann series $C^{-1} = \sum_{j=0}^{\infty} B^j$. For an application of this approach see Greenwood et al. (1997). \square

We can now characterize the regular among the asymptotically linear estimators:

If T_n is asymptotically linear for t at ϑ with influence function f , then T_n is regular if and only if f is a gradient for t at ϑ , and then the limit of T_n is $L = (\pi \otimes Qf^2)^{1/2} N$.

The proof is simple and instructive: By the martingale central limit theorem,

$$n^{-1/2} \sum Dk(X_{i-1}, X_i) \quad \text{and} \quad n^{-1/2} \sum f(X_{i-1}, X_i)$$

are jointly asymptotically normal under P . The means are 0 and the covariance is $\pi \otimes Q(Dk \cdot f)$. Because of asymptotic linearity (3.1) and local asymptotic normality (2.3), (2.4), LeCam's third lemma (see, e.g., Bickel et al. 1993, p. 503, Lemma 3) implies

$$n^{1/2}(T_n - t(\vartheta)) \Rightarrow (\pi \otimes Qf^2)^{1/2} N + \pi \otimes Q(Dk \cdot f) \quad \text{under } P_{nk}.$$

With differentiability (3.3) of t ,

$$n^{1/2}(T_n - t(\vartheta_{nk})) \Rightarrow (\pi \otimes Qf^2)^{1/2} N + \pi \otimes Q(Dk \cdot (f - Dk^t)) \quad \text{under } P_{nk}.$$

Hence T_n is regular for t at ϑ if and only if

$$\pi \otimes Q(Dk \cdot (f - Dk^t)) = 0 \quad \text{for } k \in K.$$

By the characterization (3.5), the last relation says that f is a gradient for t at ϑ . \square

By the characterization of regular estimators, we now have a meaningful answer to the question of which influence function is optimal. Call an influence function *regular* if the corresponding estimator is regular. By (3.6):

The optimal regular influence function is the canonical gradient Dk^t , and the corresponding estimator has asymptotic variance $\|k^t\|^2$.

Remark. The canonical gradient is defined as the projection of an arbitrary gradient into the tangent space. When the tangent space is not closed, it suffices to take the projection into its closure. In particular, it suffices to check regularity of an estimator for a dense subset of K . We have already seen in the full nonparametric model that this is convenient: In (2.1) it suffices to construct Q_{nh} for *bounded* h . \square

Example 2 (cont'd). (*One-dimensional parameter.*) Let Q_ϑ , $\vartheta \in \Theta$, be a one-dimensional parametric model for the transition distributions. We want to estimate the parameter ϑ . By (3.6) and Example 2 in Section 2, the canonical gradient is of the form $r^t \ell'$, and r^t solves

$$n^{1/2}(\vartheta_{nr} - \vartheta) = r \stackrel{!}{=} \langle r, r^t \rangle = r r^t I;$$

hence $r^t = I^{-1}$. Here $I = \pi \otimes Q \ell'^2$ is the Fisher information. \square

Example 3 (cont'd). (*AR(1) with normal innovations.*) Let $X_i = X_{i-1} + \varepsilon_i$ with ε_i i.i.d. and standard normal, and $\vartheta < 1$. We want to estimate the parameter ϑ . The canonical gradient is $(Dr^t)(x, y) = r^t x(y - \vartheta x)$ with r^t determined by (3.4),

$$n^{1/2}(\vartheta_{nr} - \vartheta) = r \stackrel{!}{=} \langle r, r^t \rangle = r r^t \mathbb{E} X^2 \quad \text{for } r \in \mathbf{R}.$$

We obtain $r^t = (\mathbb{E} X^2)^{-1}$ and

$$(Dr^t)(x, y) = (\mathbb{E} X^2)^{-1} x(y - \vartheta x).$$

The optimal regular influence function equals the gradient. Hence the least squares estimator

$$(3.7) \quad T_n = \frac{\sum X_{i-1} X_i}{\sum X_{i-1}^2}$$

is an optimal regular and asymptotically linear estimator. This follows immediately by ergodicity,

$$(3.8) \quad \begin{aligned} n^{1/2}(T_n - t(\vartheta)) &= \frac{n^{-1/2} \sum X_{i-1} (X_i - \vartheta X_{i-1})}{\frac{1}{n} \sum X_{i-1}^2} \\ &= (\mathbb{E} X^2)^{-1} n^{-1/2} \sum X_{i-1} (X_i - \vartheta X_{i-1}) + o_P(1). \end{aligned}$$

Of course, we expect that the least squares estimator is even efficient in this model because it is the maximum likelihood estimator. See Example 2 (cont'd) in Section 4. \square

4 Martingale estimating equations

In the setting of the previous section, we consider again estimation of a real-valued function $t(\vartheta)$ of a possibly infinite-dimensional parameter. We indicate how asymptotically linear and regular estimators can be obtained as solutions of estimating equations. We refer to Andrews (1994) for a study of estimating equations in general semiparametric time series models. Estimating equations usually cannot lead to efficient estimators unless t is one-one. However, in certain models, efficient estimators may be obtained from ‘adaptive’ versions of estimating equations; see Section 8.

Let Q_ϑ , $\vartheta \in \Theta$, be a family of transition distributions and $t : \Theta \rightarrow \mathbf{R}$ a function which we want to estimate.

A *martingale estimating equation* is based on a function $e : \mathbf{R} \times E \times E \rightarrow \mathbf{R}$ such that $e_{t(\vartheta)} \in H$ for $\vartheta \in \Theta$. Then $\sum e_{t(\vartheta)}(X_{i-1}, X_i)$ is a martingale, and an estimator T_n of $t(\vartheta)$ is obtained as a solution $t = T_n$ of the equation

$$(4.1) \quad \sum_{i=1}^n e_t(X_{i-1}, X_i) = 0.$$

From the point of view of the theory of estimating functions, martingale estimating equations have been studied in very general settings. We refer to Godambe (1985), Godambe and Heyde (1987), Godambe and Thompson (1989), and to the book Godambe (1991). Under appropriate conditions on the function e , a Taylor expansion gives

$$(4.2) \quad \begin{aligned} 0 &= n^{-1/2} \sum_{i=1}^n e_{T_n}(X_{i-1}, X_i) \\ &= n^{-1/2} \sum_{i=1}^n e_{t(\vartheta)}(X_{i-1}, X_i) + n^{1/2} (T_n - t(\vartheta)) \frac{1}{n} \sum_{i=1}^n e'_{t(\vartheta)}(X_{i-1}, X_i) + o_P(1). \end{aligned}$$

By ergodicity,

$$\frac{1}{n} \sum_{i=1}^n e'_{t(\vartheta)}(X_{i-1}, X_i) \rightarrow \pi \otimes Q e'_{t(\vartheta)} \quad (P).$$

We obtain

$$n^{1/2} (T_n - t(\vartheta)) = -(\pi \otimes Q e'_{t(\vartheta)})^{-1} n^{-1/2} \sum_{i=1}^n e_{t(\vartheta)}(X_{i-1}, X_i) + o_P(1).$$

Hence T_n is asymptotically linear for t at ϑ with influence function

$$(4.3) \quad f(x, y) = -(\pi \otimes Q e'_{t(\vartheta)})^{-1} e_{t(\vartheta)}(x, y).$$

In particular, the asymptotic variance of T_n is

$$(4.4) \quad (\pi \otimes Q e'_{t(\vartheta)})^{-2} \pi \otimes Q e_{t(\vartheta)}^2.$$

The influence function (4.3) is not automatically regular or, equivalently, a gradient. Note, however, that so far we have used the condition $e_{t(\vartheta)} \in H$ only for a fixed parameter ϑ . We indicate now that under appropriate differentiability conditions, the assumption

$$(4.5) \quad Q_{nk}(x, e_{t(\vartheta_{nk})}) = 0 \quad \text{for } k \in K$$

implies that the influence function is regular. Since t is differentiable (3.4) with canonical gradient Dk^t , we obtain

$$e_{t(\vartheta_{nk})} \approx e_{t(\vartheta)} + n^{-1/2} \langle k, k^t \rangle e'_{t(\vartheta)}.$$

Since Q_{nk} is differentiable (2.1) with derivative Dk , relation (4.5) implies

$$\begin{aligned} 0 &= Q_{nk}(x, e_{t(\vartheta_{nk})}) \\ &\approx Q(x, e_{t(\vartheta)}) + n^{-1/2} \left(Q(x, Dk \cdot e_{t(\vartheta)}) + \langle k, k^t \rangle Q(x, e'_{t(\vartheta)}) \right). \end{aligned}$$

Taking expectations with respect to π ,

$$(4.6) \quad \pi \otimes Q(Dk \cdot e_{t(\vartheta)}) = -\langle k, k^t \rangle \pi \otimes Qe'_{t(\vartheta)} \quad \text{for } k \in K.$$

The characterization (3.5) now implies that the influence function (4.3) is a gradient.

Using (4.6) again, now with $k = k^t$, we can write the asymptotic variance (4.4) of the estimator as

$$(4.7) \quad \frac{\|k^t\|^2 \pi \otimes Qe_{t(\vartheta)}^2}{\pi \otimes Q(Dk \cdot e_{t(\vartheta)})^2}.$$

Remark. By the Schwarz inequality or by (3.6), the asymptotic variance (4.7) is minimal if e is proportional to the canonical gradient Dk^t , and the minimal asymptotic variance is $\|k^t\|^2$. Of course, such a function e cannot, in general, be used in an estimating equation because it will depend on ϑ not only through $t(\vartheta)$. \square

Example 2 (cont'd). (*One-dimensional parameter.*) Let Q_ϑ , $\vartheta \in \Theta$, be a one-dimensional parametric model for the transition distributions. We want to estimate the parameter ϑ . By Example 2 in Section 2, the tangent space is $r\ell'$, $r \in \mathbf{R}$. By Example 2 (cont'd) in Section 3, the canonical gradient of ϑ is $I^{-1}\ell'_\vartheta$, with $I = \pi \otimes Q\ell''_\vartheta$ the Fisher information. The *maximum likelihood estimator* T_n solves the martingale estimating equation

$$\sum \ell'_\vartheta(X_{i-1}, X_i) = 0.$$

By (4.3), T_n is asymptotically linear with influence function $-(\pi \otimes Q\ell''_\vartheta)^{-1}\ell'_\vartheta$. Differentiating with respect to ϑ under the integral,

$$0 = (Q_\vartheta(x, \ell'_\vartheta))' = Q_\vartheta(x, \ell''_\vartheta) + Q_\vartheta(x, \ell'^2_\vartheta).$$

Hence $-\pi \otimes Q\ell''_\vartheta = I$, and the influence function is seen to equal the canonical gradient, so that the maximum likelihood estimator is efficient. (We note that $-\pi \otimes Q\ell''_\vartheta = I$ is also obtained from (4.6).) Efficiency of the maximum likelihood estimator has been proved for many specific models. We refer in particular to Hwang and Basawa (1994) who assume that $Q_\vartheta(x, dy)$ is an exponential family for each x . \square

5 A characterization of efficient estimators

We have seen in Section 3 that the best regular and asymptotically linear estimator has an influence function which is equal to the canonical gradient. In this section we show that this estimator is already optimal among *all* regular estimators.

This is a consequence of Hájek's (1970) convolution theorem. In our context, it reads as follows. (For a proof see, e.g., Bickel et al. 1993, p. 63, Theorem 2A.)

Let Q_{ϑ} , $\vartheta \in \Theta$, be Hellinger differentiable (2.5), let $t : \Theta \rightarrow \mathbf{R}$ be differentiable (3.3) with canonical gradient Dk^t , and let T_n be regular (3.2) with limit L . Then

$$(5.1) \quad \left(n^{-1/2} \sum_{i=1}^n (Dk^t)(X_{i-1}, X_i), n^{1/2}(T_n - t(\vartheta)) - n^{-1/2} \sum_{i=1}^n (Dk^t)(X_{i-1}, X_i) \right) \\ \Rightarrow (\|k^t\|N, M) \quad \text{under } P,$$

with N standard normal and M independent of N . In particular,

$$L = \|k^t\|N + M \quad \text{in distribution.}$$

By Anderson (1955), L is more spread out than $\|k^t\|N$,

$$P(|L| \leq c) \leq P(\|k^t\|N \leq c) \quad \text{for } c > 0.$$

This justifies calling T_n *efficient* if

$$L = \|k^t\|N \quad \text{in distribution.}$$

It follows from (5.1) that a regular and efficient estimator is asymptotically linear with influence function equal to the canonical gradient. We have seen in Section 3 that the converse is also true:

An estimator is regular and efficient if and only if it is asymptotically linear with influence function equal to the canonical gradient,

$$(5.2) \quad n^{1/2}(T_n - t(\vartheta)) = n^{-1/2} \sum_{i=1}^n (Dk^t)(X_{i-1}, X_i) + o_P(1).$$

Remark. In the full nonparametric model (see Example 1 in Section 2) we have $K = H$, and the gradient h^t is *unique* by (3.5). In particular, there is only *one* regular influence function, namely h^t , and the optimality result for regular influence functions is rather uninteresting. In the i.i.d. case, this is sometimes used as an argument against results on regular estimators in nonparametric models. However, the class of regular estimators

is much larger than the class of regular and asymptotically linear estimators. One might object that reasonable estimators should be asymptotically linear. It should however be noted that the concept of asymptotic linearity can be considerably extended. This is probably not so obvious in the i.i.d. case, but in our definition (3.1) for Markov chains it suggests itself to allow functions f with more than two arguments. This point was made in Wefelmeyer (1991). \square

6 Nonparametric models and martingale approximations

Let X_0, \dots, X_n be observations from an ergodic Markov chain with unknown transition distribution $Q(x, dy)$ and invariant distribution $\pi(dx)$. We are interested in estimating π or, more specifically, the expectation πf of some square-integrable function $f(x)$.

By ergodicity, a consistent estimator is the *empirical estimator*

$$E_n f = \frac{1}{n} \sum_{i=1}^n f(X_i).$$

More generally, the expectation $\pi \otimes Qf$ of $f \in L_2(\pi \otimes Q)$ is estimated by the empirical estimator

$$E_n f = \frac{1}{n} \sum_{i=1}^n f(X_{i-1}, X_i).$$

We expect $E_n f$ to be efficient, but it is not even clear that it is asymptotically linear in the sense of (3.1): We can write

$$n^{1/2}(E_n f - \pi \otimes Qf) = n^{-1/2} \sum_{i=1}^n (f(X_{i-1}, X_i) - \pi \otimes Qf),$$

but $f - \pi \otimes Qf$ is not in H . Note, however, that we can expand $\sum f(X_{i-1}, X_i)$ into a series of martingales, the first step being

$$\sum_{i=1}^n f(X_{i-1}, X_i) = \sum_{i=1}^n (f(X_{i-1}, X_i) - Q(X_{i-1}, f)) + \sum_{i=1}^n Q(X_i, f) + Q(X_n, f) - Q(X_0, f).$$

Under an appropriate geometric ergodicity assumption, if we continue adding and subtracting higher order conditional expectations, we obtain the *martingale approximation*

$$n^{1/2}(E_n f - \pi \otimes Qf) = n^{-1/2} \sum_{i=1}^n (Af)(X_{i-1}, X_i) + o_P(1)$$

with

$$(Af)(x, y) = f(x, y) - Q(x, f) + \sum_{j=1}^{\infty} (Q^j(y, f) - Q^{j+1}(x, f)).$$

Note that $Af \in H$. Hence $E_n f$ has influence function Af . The expansion is due to Gordin (1969); see also Gordin and Lifšic (1978), Maigret (1978), Dürr and Goldstein (1986) and Meyn and Tweedie (1994, Section 17.4).

To prove that the empirical estimator $E_n f$ is efficient, it remains to show that its influence function Af is the canonical gradient of $\pi \otimes Qf$. As in the first part of Section 2, we parametrize the model by the distribution function, take H as tangent space and pick sequences Q_{nh} with derivative h in the sense of (2.1). We view $\pi \otimes Qf$ as a function t of Q . By (3.4) and Example 1 in Section 2, the canonical gradient h^t is defined by

$$n^{1/2}(t(Q_{nh}) - t(Q)) = n^{1/2}(\pi_{nh} \otimes Q_{nh}f - \pi \otimes Qf) \rightarrow \pi \otimes Q(hh^t) \quad \text{for } h \in H.$$

By a perturbation expansion of π_{nh} due to Kartashov (1985a), (1985b), the left side converges to $\pi \otimes Q(hAf)$, and we obtain $h^t = Af$ as expected. The above proof is due to Greenwood and Wefelmeyer (1995). Other proofs, for $f(x)$ with one argument, are in Penev (1991) and Bickel (1993). When f has more than two arguments, the corresponding empirical estimator $E_n f$ is *not* efficient; it is, however, efficient in the appropriate larger model of *higher order* Markov chains.

Example 4. (*Countable state space.*) Suppose the state space E is countable. Then the transition distribution $Q(x, dy)$ is determined by the matrix of transition probabilities q_{xy} , and the stationary distribution $\pi(dx)$ by a vector of probabilities p_x . The above result, applied for $f = \delta_x$ and $f = \delta_{(x,y)}$, shows that

$$N_n^x = \frac{1}{n} \# \{i : X_i = x\} \quad \text{is efficient for } p_x$$

and

$$N_n^{xy} = \frac{1}{n} \# \{i : X_{i-1} = x, X_i = y\} \quad \text{is efficient for } p_x q_{xy}.$$

Hence N_n^{xy}/N_n^x is efficient for q_{xy} . Similar results for Markov step processes and semi-Markov processes are in Greenwood and Wefelmeyer (1994) and (1996a). \square

Example 5. (*Reversible chains.*) Suppose we know that the chain is reversible,

$$\pi(dx)Q(x, dy) = \pi(dy)Q(y, dx).$$

Can we do better than the empirical estimator? It suggests itself to use a symmetrized version,

$$\frac{1}{2n} \sum_{i=1}^n (f(X_{i-1}, X_i) + f(X_i, X_{i-1})).$$

Greenwood and Wefelmeyer (1996b) show that this estimator is efficient. Although the result seems fairly obvious, the proof is not: It is not simple to translate the condition

$$\pi_{nh}(dx)Q_{nh}(x, dy) = \pi_{nh}(dy)Q_{nh}(y, dx)$$

into a restriction on h . □

Remark. Consider the *empirical measure*

$$(6.1) \quad M_n = \frac{1}{n} \sum_{i=1}^n \varepsilon_{(X_{i-1}, X_i)},$$

with ε_a denoting the one-point probability measure in a . The empirical estimator $E_n f$ is the expectation $M_n f$ of f under the empirical measure. Since $E_n f$ is efficient for $\pi \otimes Q f$ for all (square-integrable) f , we expect that sufficiently smooth functions $s(M_n)$ are efficient for $s(\pi \otimes Q)$. The following Example 6 illustrates this point. □

Example 6. (*Misspecified parametric models.*) Suppose we have a one-dimensional parametric model Q_α , $\alpha \in A \subset \mathbf{R}$, for the transition distribution. Let ℓ'_α denote the Hellinger derivative of Q_α in the sense of (2.9). Let $q_\alpha(x, y)$ be a density of $Q_\alpha(x, dy)$. The maximum likelihood estimator maximizes

$$\frac{1}{n} \sum_{i=1}^n \log q_\alpha(X_{i-1}, X_i) = E_n \log q_\alpha.$$

Suppose now that the model is misspecified, and that the true transition distribution is Q . Let $s(\pi \otimes Q)$ denote the parameter α which maximizes the *Kullback-Leibler information* $\pi \otimes Q \log q_\alpha$. Then the maximum likelihood estimator is $s(M_n)$, where M_n is the empirical measure (6.1). By the preceding Remark, we expect $s(M_n)$ to be efficient. It is, however, not clear how to check that s is sufficiently smooth. A proof avoiding this problem is due to Beran (1977) in the i.i.d. case, and to Greenwood and Wefelmeyer (1997) in the Markov chain setting. Related results for stationary Gaussian time series are in Dahlhaus and Wefelmeyer (1996). The asymptotic distribution of the maximum likelihood estimator under misspecification was derived by Huber (1967) in the i.i.d. case, by Ogata (1980) for Markov chains, and by Hosoya (1989) for general stationary linear processes. □

7 Autoregression

Consider the linear first-order autoregressive model

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

where the ε_i are i.i.d. with a density p which has mean 0 and finite variance, and where $|\alpha| < 1$. This is an ergodic Markov chain with transition distribution

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

In Example 3 we have treated the case that p is known (and standard normal). Now we treat p as an unknown nuisance parameter.

To prove local asymptotic normality, fix α and write π for the invariant distribution of $Q = Q_\alpha$. The parameter of the model is $\vartheta = (\alpha, p)$. Perturb α and p separately: $\alpha_{na} = \alpha + n^{-1/2}a$ and, for a bounded function $b(x)$,

$$p_{nb}(x) = p(x)(1 + n^{-1/2}b(x)).$$

Since p_{nb} must be a probability density, we must have $E b(\varepsilon) = 0$. Since p_{nb} must have mean 0, we must have $E \varepsilon b(\varepsilon) = 0$. We obtain the tangent space $K = \mathbf{R} \times B$ with

$$B = \{b \in L_2(p) : E b(\varepsilon) = 0 \text{ and } E \varepsilon b(\varepsilon) = 0\}.$$

Write $\vartheta_{nk} = (\alpha_{na}, p_{nb})$. With $\ell' = p'/p$, the corresponding transition distribution $Q_{nk} = Q_{\vartheta_{nk}}$ is

$$\begin{aligned} Q_{nk}(x, dy) &= p_{nb}(y - \alpha_{na}x)dy \\ &\approx Q(x, dy) \left(1 + n^{-1/2}(-ax\ell'(y - \alpha x) + b(y - \alpha x))\right). \end{aligned}$$

Hence the derivative of Q_{nk} is

$$(Dk)(x, y) = -ax\ell'(y - \alpha x) + b(y - \alpha x).$$

Since $E \varepsilon = 0$, the two functions on the right are orthogonal, and the inner product (2.6) induced on K is

$$(7.1) \quad \langle k, k' \rangle = aa'E X^2 E \ell'(\varepsilon)^2 + E b(\varepsilon)b'(\varepsilon),$$

where the expectation of X is taken with respect to the stationary distribution. We obtain local asymptotic normality

$$\begin{aligned} \log \frac{dP_{nk}}{dP} &= -an^{-1/2} \sum_{i=1}^n X_{i-1} \ell'(X_i - \alpha X_{i-1}) + n^{-1/2} \sum_{i=1}^n b(X_i - \alpha X_{i-1}) \\ &\quad - \frac{1}{2}a^2 E X^2 E \ell'(\varepsilon)^2 - \frac{1}{2}E b(\varepsilon)^2 + o_P(1). \end{aligned}$$

For a proof see Huang (1986) or Kreiss (1987b). For *known* innovation distribution see Akahira (1976) and Akritas and Johnson (1982).

We consider first the problem of estimating the autoregression parameter α . The classical estimator for α is the *least squares estimator*

$$\hat{\alpha}_n = \frac{\sum X_{i-1} X_i}{X_{i-1}^2}.$$

As in (3.8) we see that $\hat{\alpha}_n$ is asymptotically linear with influence function

$$f(x, y) = (E X^2)^{-1}x(y - \alpha x).$$

Hence its asymptotic variance is $(\mathbb{E} X^2)^{-1} \mathbb{E} \varepsilon^2$.

To calculate a variance bound, we view the parameter α of interest as a function t of the parameter $\vartheta = (\alpha, p)$ of the model. By (3.4), the canonical gradient Dk^t , with $k^t = (a^t, b^t)$, solves

$$\begin{aligned} n^{1/2}(t(\vartheta_{nk}) - t(\vartheta)) &= n^{1/2}(\alpha_{na} - \alpha) = \alpha \\ \stackrel{!}{=} \langle k, k^t \rangle &= aa^t \mathbb{E} X^2 \mathbb{E} \ell'(\varepsilon)^2 + \mathbb{E} b(\varepsilon) b^t(\varepsilon) \quad \text{for } a \in \mathbf{R}, b \in B. \end{aligned}$$

The solution is $a^t = (\mathbb{E} X^2 \mathbb{E} \ell'(\varepsilon)^2)^{-1}$, $b^t = 0$. Hence by (5.2) a regular and efficient estimator T_n for α is characterized by

$$n^{1/2}(T_n - \alpha) = (\mathbb{E} X^2 \mathbb{E} \ell'(\varepsilon)^2)^{-1} n^{-1/2} \sum_{i=1}^n X_{i-1} \ell'(X_i - \alpha X_{i-1}) + o_P(1).$$

Such an estimator is constructed in Kreiss (1987b). For an efficient estimator under the stronger model assumption that p is symmetric see Kreiss (1987a). The asymptotic variance of the efficient estimator is $(\mathbb{E} X^2 \mathbb{E} \ell'(\varepsilon)^2)^{-1}$. Hence the relative efficiency of the least squares estimator is $(\mathbb{E} \varepsilon^2 \mathbb{E} \ell'(\varepsilon)^2)^{-1}$. This equals the relative efficiency of the empirical estimator in the i.i.d. location model generated by the density p .

Consider now the problem of estimating the distribution of the innovations ε_i . To be specific, we estimate the expectation $\mathbb{E} f(\varepsilon)$ of some square-integrable function f . We do not observe the ε_i and cannot use the empirical estimator $\frac{1}{n} \sum_{i=1}^n f(\varepsilon_i)$. With $\hat{\alpha}_n = \sum X_{i-1} X_i / \sum X_{i-1}^2$ denoting the least squares estimator, we can replace the ε_i by *estimated* innovations $\hat{\varepsilon}_{ni} = X_i - \hat{\alpha}_n X_{i-1}$ and use the corresponding empirical estimator

$$E_n f = \frac{1}{n} \sum_{i=1}^n f(\hat{\varepsilon}_{ni}).$$

For $f(x) = 1(x \leq r)$ this is the empirical distribution function. It is well studied. Functional central limit theorems are obtained by Boldin (1982), (1983) for autoregressive processes, by Boldin (1989) for moving average processes and by Kreiss (1991) for general linear processes. Explosive autoregressive processes, with $|\alpha| > 1$, are treated in Koul and Leventhal (1989). See also the monograph by Koul (1992). It can be shown that the asymptotic variance $\mathbb{E} f(\varepsilon)^2$ of the estimator is not changed if we replace ε_i by $\hat{\varepsilon}_{ni}$. Nevertheless, $E_n f$ is not efficient. The reason is that we have not used the information that the innovations have mean 0. To see how we might improve the empirical estimator, we calculate the canonical gradient of $\mathbb{E} f(\varepsilon)$, viewed as a function t of (α, p) . By (3.4), the canonical gradient Dk^t , with $k^t = (a^t, b^t)$, solves

$$\begin{aligned} n^{1/2}(t(\vartheta_{nk}) - t(\vartheta)) &= n^{1/2} \left(\int p_{nb}(x) dx f(x) - \int p(x) dx f(x) \right) = \mathbb{E} b(\varepsilon) f(\varepsilon) \\ \stackrel{!}{=} \langle k, k^t \rangle &= aa^t \mathbb{E} X^2 \mathbb{E} \ell'(\varepsilon)^2 + \mathbb{E} b(\varepsilon) b^t(\varepsilon) \quad \text{for } a \in \mathbf{R}, b \in B. \end{aligned}$$

We obtain $a^t = 0$ and must solve

$$E b(\varepsilon) f(\varepsilon) = E b(\varepsilon) b^t(\varepsilon) \quad \text{for } b \in B.$$

Using both restrictions $E b(\varepsilon) = 0$ and $E \varepsilon b(\varepsilon) = 0$, one checks that $b^t(x) = f(x) - E f(x) - cx$ with $c = E \varepsilon f(\varepsilon) / E \varepsilon^2$. Hence the canonical gradient is

$$(Dk^t)(x, y) = f(y - \alpha x) - E f(\varepsilon) - c(y - \alpha x).$$

It is easy to see that an efficient estimator for $E f(\varepsilon)$, with influence function equal to the canonical gradient, is

$$E_n f - \hat{c}_n \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_{ni} \quad \text{with } \hat{c}_n = \frac{\sum \hat{\varepsilon}_{ni} f(\hat{\varepsilon}_{ni})}{\sum \hat{\varepsilon}_{ni}^2}.$$

For a proof see Wefelmeyer (1994). The tangent space $K = \mathbf{R} \times B$ is also calculated in Huang (1986) and Kreiss (1987b). These authors forget the restriction $E \varepsilon b(\varepsilon) = 0$. One would conclude from their result that the usual empirical estimator is efficient.

Remark. The tangent space is $K = \mathbf{R} \times B$, and by (7.1) the spaces \mathbf{R} and B are *orthogonal*. As a consequence, α has a canonical gradient Dk^t with k^t of the form $(a^t, 0)$, while k^t is of the form $(0, b^t)$ for $E f(\varepsilon)$, a function of the second parameter, p . This means that, asymptotically, each of the two parameters can be estimated just as well knowing as not knowing the other parameter. Such a model is called *adaptive*, and efficient estimators for these parameters are also called adaptive.

Suppose we have proved, for each fixed p , local asymptotic normality of the one-dimensional model with transition distribution $p(y - \alpha x)dy$. Suppose we have constructed an estimator not depending on p which is efficient in all of these one-dimensional models. Then the parameter α is adaptive, and the estimator is efficient in the model with both α and p unknown. It is not necessary to prove local asymptotic normality of this model. Kreiss (1987a) uses this approach for ARMA-models, and Gassiat (1993) for noncausal AR-models. \square

Example 7. (*ARCH*.) The first-order ARCH-model is

$$X_i = \sigma(1 + \alpha X_{i-1}^2)^{1/2} \varepsilon_i,$$

where the ε_i are i.i.d. with a density p which has mean 0 and variance 1. This is a Markov chain with transition distribution

$$Q(x, dy) = \frac{1}{\sigma(1 + \alpha x^2)^{1/2}} p\left(\frac{y}{\sigma(1 + \alpha x^2)^{1/2}}\right) dy.$$

It can be treated in a similar way as the AR(1)-model above. A review of ARCH-models is Bollerslev et al. (1992). Efficient estimators in this model and generalized ARCH-models are constructed in Engle and González-Rivera (1991), Linton (1993) and Drost and Klaassen (1996) under increasingly weaker assumptions. \square

Example 8. (*Nonlinear autoregression.*) A generalization of AR- and ARCH-models are nonlinear and heteroscedastic autoregressive models

$$X_i = m_\alpha(X_{i-1}) + v_\alpha(X_{i-1})^{1/2}\varepsilon_i,$$

where the ε_i are i.i.d. with a density p which has mean 0 and variance 1 as in Example 7. Efficient estimation in nonlinear autoregressive models is studied by Hwang and Basawa (1993), Drost et al. (1994a), (1994b), Jeganathan (1995) and Koul and Schick (1996b). \square

8 Modeling conditional moments

Let X_0, \dots, X_n be observations from a real-valued ergodic Markov chain with transition distribution $Q(x, dy)$ and invariant distribution $\pi(dx)$. Suppose we have a parametric model for the conditional mean, or autoregression function, but that the transition distribution is unspecified otherwise. A simple such specification is

$$(8.1) \quad \int Q(x, dy)y = \alpha x.$$

(The linear autoregressive model of Section 7 is a *submodel*, with a transition distribution of the form $Q(x, dy) = p(y - \alpha x)dy$, where p is a mean 0 density.)

Such a model suggests a special class of martingale estimating equations. Note that the model can be described by saying that the innovations $\varepsilon_i = X_i - \alpha X_{i-1}$ are martingale increments. In particular, ‘stochastic integrals’ of the form

$$(8.2) \quad \sum_{i=1}^n w_\alpha(X_{i-1})(X_i - \alpha X_{i-1})$$

are martingales. We obtain martingale estimating equations of the form

$$\sum e_\alpha(X_{i-1}, X_i) = 0$$

as in Section 4, now with $t(\alpha) = \alpha$ and $e_\alpha(x, y) = w_\alpha(x)(y - \alpha x)$. As before, we write T_n for the corresponding estimator. The choice $w_\alpha(x) = x$ leads to the least squares estimator, $T_n = \sum X_{i-1}X_i / X_{i-1}^2$. In general, to solve the estimating equation, we expand it as in (4.2),

$$\begin{aligned} 0 &= n^{-1/2} \sum_{i=1}^n e_{T_n}(X_{i-1}, X_i) \\ &= n^{-1/2} \sum_{i=1}^n e_\alpha(X_{i-1}, X_i) + n^{1/2}(T_n - \alpha) \frac{1}{n} \sum_{i=1}^n e'_\alpha(X_{i-1}, X_i) + o_P(1). \end{aligned}$$

Here $e'_\alpha(x, y) = -xw'_\alpha(x) + w'_\alpha(x)(y - \alpha x)$, and we observe a special feature of the estimating function: The terms $w'_\alpha(X_{i-1})(X_i - \alpha X_{i-1})$ are martingale increments; hence

$$\frac{1}{n} \sum_{i=1}^n w'_\alpha(X_{i-1})(X_i - \alpha X_{i-1}) = o_P(1)$$

and the influence function of T_n does not contain the derivative of w_α :

$$(8.3) \quad n^{1/2}(T_n - \alpha) = (\mathbb{E} X w_\alpha(X))^{-1} n^{-1/2} \sum_{i=1}^n w_\alpha(X_{i-1})(X_i - \alpha X_{i-1}) + o_P(1).$$

The asymptotic variance of T_n is

$$(8.4) \quad (\mathbb{E} X w_\alpha(X))^{-2} \pi(v w_\alpha^2),$$

where v denotes the conditional variance

$$v(x) = \int Q(x, dy)(y - \alpha x)^2.$$

Are there better estimators than the least squares estimator? By the Schwarz inequality, the variance (8.4) is minimized by $w_\alpha(x) = v(x)^{-1}x$, but this function cannot be used in an estimating equation (8.2) because $v(x)$ is unknown. However, as shown in Wefelmeyer (1997a), the conditional variance $v(x)$ can be replaced by an estimator $\hat{v}_n(x)$ without changing the asymptotic variance of the estimator which solves the corresponding estimating equation. In this sense, the estimator is ‘adaptive’. The solution of this estimating equation is the *weighted* least squares estimator

$$T_n = \frac{\sum \hat{v}_n(X_{i-1})^{-1} X_{i-1} X_i}{\sum \hat{v}_n(X_{i-1})^{-1} X_{i-1}^2}.$$

By (8.4), its asymptotic variance is $(\mathbb{E} v(X)^{-1} X^2)^{-1}$.

Are there better estimators than the optimal weighted least squares estimator? To decide this, we calculate the canonical gradient of α . As in the first part of Section 2, we parametrize the model by the transition distribution, fix Q and consider perturbations (2.1)

$$Q_{nh}(x, dy) = Q(x, dy)(1 + n^{-1/2}h(x, y))$$

with $h \in H$. The transition distribution Q_{nh} must fulfill the model assumption (8.1), with a possibly perturbed $\alpha_{nh} = \alpha + n^{-1/2}a$,

$$\int Q_{nh}(x, dy)y = \alpha_{nh}x,$$

i.e.,

$$\int Q(x, dy)y + n^{-1/2} \int Q(x, dy)h(x, y)y = \alpha x + n^{-1/2}ax.$$

It follows that the tangent space is the union of the affine spaces

$$H_a = \{h \in H : \int Q(x, dy)h(x, y)y = ax\}, \quad a \in \mathbf{R}.$$

By (3.4), the canonical gradient h^t of α , viewed as a function t of Q , is in one of these affine spaces and solves

$$(8.5) \quad \begin{aligned} n^{1/2}(t(Q_{nh}) - t(Q)) &= n^{1/2}(\alpha_{nh} - \alpha) = a \\ &\stackrel{!}{=} \langle h, h^t \rangle = \pi \otimes Q(hh^t) \quad \text{for } a \in \mathbf{R}, h \in H_a. \end{aligned}$$

Our candidate is the influence function of the optimal weighted least squares estimator,

$$g(x, y) = (\mathbb{E} v(X)^{-1} X^2)^{-1} v(x)^{-1} x (y - \alpha x).$$

This function is clearly in H , and since

$$\int Q(x, dy) v(x)^{-1} x (y - \alpha x) y = v(x)^{-1} x \int Q(x, dy) (y - \alpha x)^2 = x,$$

we also have $g \in H_a$ for $a = (\mathbb{E} v(X)^{-1} X^2)^{-1}$. It remains to check whether $h^t = g$ solves (8.5). But

$$\pi \otimes Q(hg) = (\mathbb{E} v(X)^{-1} X^2)^{-1} \int \pi(dx) v(x)^{-1} x \int Q(x, dy) h(x, y) (y - \alpha x) = a.$$

Hence the optimal weighted least squares estimator is efficient.

Remark. We have just checked that the influence function of the optimal weighted least squares estimator is a gradient. By the characterization in Section 3, this would be unnecessary if we knew that the estimator is regular. Indeed, every estimating equation based on a martingale (8.2) gives a regular estimator, provided w_α is so smooth that the estimator is asymptotically linear of the form (8.3). In other words:

Influence functions of the form $f(x, y) = (\mathbb{E} X w_\alpha(X))^{-1} w_\alpha(x) (y - \alpha x)$ are regular.

By (3.5), we must check that

$$\pi \otimes Q(hf) = \pi \otimes Q(hg) \quad \text{for } a \in \mathbf{R}, h \in H_a.$$

We calculate both sides for $h \in H_a$, ignoring constants:

$$\int \pi(dx) w_\alpha(x) \int Q(x, dy) h(x, y) (y - \alpha x) = a \mathbb{E} X w_\alpha(X)$$

and

$$\int \pi(dx) v(x)^{-1} \int Q(x, dy) h(x, y) (y - \alpha x) = a \mathbb{E} v(X)^{-1} X^2.$$

Taking into account the constants in f and g , the check is finished. \square

Remark. If the innovations $\varepsilon_i = X_i - \alpha X_{i-1}$ happen to be independent, the model reduces to the AR(1)-model of Section 7, and the conditional variance $v(x)$ reduces to

$$v(x) = \int p(y - \alpha x) (y - \alpha x)^2 = \mathbb{E} \varepsilon^2$$

and does not depend on x . Then the optimal weighted least squares estimator is asymptotically equivalent to the ordinary least squares estimator. \square

Example 9. (*Quasi-likelihood models.*) A quasi-likelihood model for Markov chains is described by parametric models for both the conditional mean and the conditional variance,

$$\begin{aligned}\int Q(x, dy)y &= m_\alpha(x), \\ \int Q(x, dy)(y - m_\alpha(x))^2 &= v_\alpha(x),\end{aligned}$$

with a common parameter α . For examples see Zeger and Qaqish (1988) and Huh-tala (1992). A submodel is the nonlinear and heteroscedastic autoregression model of Example 8 in Section 7, with transition distribution

$$Q(x, dy) = \frac{1}{v_\alpha(x)^{1/2}} p\left(\frac{y - m_\alpha(x)}{v_\alpha(x)^{1/2}}\right) dy.$$

The approach of the present section works for quasi-likelihood models; see Wefelmeyer (1996). Wefelmeyer (1997b) treats a generalization in which m_α and v_α also depend on covariates. \square

9 Other types of Markov chain models

In this section we briefly describe further models which also fall under our Markov chain setting.

Example 10. (*Discretely observed diffusions.*) Suppose we observe an ergodic diffusion process

$$dX_t = b_\alpha(X_t) + \sigma_\alpha(X_t)dW_t$$

at discrete time points $0, \Delta, 2\Delta, \dots$. The observations form a Markov chain X_0, \dots, X_n . The model is in principle parametrized by α . However, while the stationary distribution of X_i is the same as for the diffusion process, the transition distribution may be difficult to calculate. Recent references on estimators for α are Bibby and Sørensen (1995), Pedersen (1995), Kessler and Sørensen (1995) and Sørensen (1996). In the context of the present review, we mention the approach of Kessler and Wefelmeyer (1996): Consider the transition distribution as unknown. We have a parametric family of restrictions on the transition distributions, namely a parametric model for the stationary distribution of X_i . A simple estimator for α is the estimator that would be the maximum likelihood estimator if the observations were independent, but this estimator turns out to be inefficient. \square

Example 11. (*Markov chain Monte Carlo.*) Suppose we want to calculate the expectation πf of a function f on a product space $E = E_1 \times \dots \times E_k$. This may be difficult directly or numerically, or even by ordinary Monte Carlo integration. If we can simulate

the conditional distributions $p_j(x_{-j}, dx_j)$ of the j -th component of π given the other components x_{-j} , we may generate a k -dimensional Markov chain based on updates of a single component by

$$Q_j(x, dy) = p_j(x_{-j}, dy_j)\varepsilon_{x_{-j}}(dy_{-j}).$$

An introduction to such *Markov chain Monte Carlo* procedures is the monograph Gilks et al. (1996). The *Gibbs sampler with deterministic sweep* uses the transition distribution $Q = Q_1 \cdots Q_k$, the one with *random sweep* uses $Q = \frac{1}{k} \sum Q_j$. If we denote the simulations from the corresponding Markov chain by X^0, \dots, X^n , the expectation πf can be estimated by the empirical estimator $\frac{1}{n} \sum_{i=1}^n f(X^i)$. Its asymptotic variance is studied, e.g., by Peskun (1973), Frigessi, Hwang and Younes (1992), Green and Han (1992), Liu, Wong and Kong (1994) and (1995) and Clifford and Nicholls (1995). Does the empirical estimator make effective use of the simulations? Greenwood et al. (1997) view π as the unknown parameter and calculate variance bounds for estimators of πf . It turns out that it is best not to use the usual empirical estimator but to use the empirical estimator for deterministic sweep which considers the update of a single component as a new observation:

$$\frac{1}{k} \sum_{j=1}^k \frac{1}{n} \sum_{i=1}^n f(X_1^i, \dots, X_j^i, X_{j+1}^{i-1}, \dots, X_k^{i-1}).$$

This estimator is close to efficient. □

Example 12. (*Random coefficient autoregression.*) We observe X_0, \dots, X_n with

$$X_i = (\alpha + Z_i)X_{i-1} + \varepsilon_i,$$

where the ε_i are i.i.d. with a density p which has mean 0 and finite variance, and the Z_i are i.i.d. with mean 0 and distribution function G such that $\alpha + \text{var } G < 1$. The X_i are a Markov chain with transition distribution

$$Q(x, dy) = \int p(y - (\alpha + z)x) dG(z) dy.$$

The model is studied in the monographs by Nicholls and Quinn (1982) and Tong (1990). Efficient estimators for α are constructed in Koul and Schick (1996a). Weighted least squares estimators and local asymptotic normality for generalized random coefficient autoregressive models are studied in Hwang and Basawa (1996a), (1996b). □

Example 13. (*Regression with autoregressive errors.*) We observe (U_i, Y_i) , $i = 1, \dots, n$, where the Y_i follow a linear regression model

$$Y_i = \beta U_i + X_i,$$

the covariates U_i are i.i.d., with density g , and the errors X_i are AR(1),

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

where the ε_i are i.i.d. with a density p which has mean 0 and finite variance, and where $|\alpha| < 1$. Writing

$$Y_i = \beta U_i + \alpha(Y_{i-1} - \beta U_{i-1}) + \varepsilon_i,$$

we see that the observations (U_i, Y_i) are a Markov chain with transition distribution

$$Q(u_{i-1}, y_{i-1}, du_i, dy_i) = g(u_i)p(y_i - \beta u_i - \alpha(y_{i-1} - \beta u_{i-1}))du_i dy_i$$

and four parameters, α , β , p , g . Local asymptotic normality for *deterministic* U_i is proved in Swensen (1985), for *moving average* X_i in Garel (1989). For the additive regression model $Y_i = \beta U_i + \gamma(V_i) + X_i$ of Engle et al. (1986), efficient estimators for α and β , respectively, are constructed in Schick (1993) and (1996). Quite different efficient estimators are obtained for long range stationary Gaussian errors X_i : see Dahlhaus (1995). \square

Acknowledgment. I thank the referee for a number of suggestions which improved the presentation.

References

- Akahira, M. (1976). On the asymptotic efficiency of estimators in an autoregressive process. *Ann. Inst. Statist. Math.* 28, 35–48.
- Akritis, M. G. and Johnson, R. A. (1982). Efficiencies of tests and estimators for p-order autoregressive processes when the error distribution is nonnormal. *Ann. Inst. Statist. Math.* 34, 579–589.
- Anderson, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proc. Amer. Math. Soc.* 6, 170–176.
- Andrews, D. W. K. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica* 62, 43–72.
- Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* 5, 445–463.
- Bibby, B. M. and Sørensen, M. (1995). Martingale estimating functions for discretely observed diffusion processes. *Bernoulli* 1, 17–39.
- 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. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

- Boldin, M. V. (1982). Estimation of the distribution of noise in an autoregressive scheme. *Theory Probab. Appl.* 27, 866–871.
- Boldin, M. V. (1983). Testing hypotheses in autoregressive schemes by the Kolmogorov and ω^2 tests. *Soviet Math. Dokl.* 28, 550–553.
- Boldin, M. V. (1989). On testing hypotheses in the sliding average scheme by the Kolmogorov-Smirnov and ω^2 tests. *Theory Probab. Appl.* 34, 758–746.
- Bollerslev, T., Chou, R. Y. and Kroner, K. (1992). ARCH modeling in finance. A review of the theory and empirical evidence. *J. Econometrics* 52, 5–59.
- Clifford, P. and Nicholls, G. (1995). A Metropolis sampler for polygonal image reconstruction. Technical report, Department of Statistics, Oxford University.
- Dahlhaus, R. (1995). Efficient location and regression estimation for long range dependent regression models. *Ann. Statist.* 23, 1029–1047.
- Dahlhaus, R. and Wefelmeyer, W. (1996). Asymptotically optimal estimation in misspecified time series models. *Ann. Statist.* 24, 952–974.
- Drost, F. C. and Klaassen, C. A. J. (1996). Efficient estimation in semiparametric GARCH models. CentER discussion paper 9638, Tilburg University. To appear in: *J. Econometrics*.
- Drost, F. C., Klaassen, C. A. J. and Werker, B. J. M. (1994a). Adaptiveness in time series models. In: *Asymptotic Statistics* (P. Mandl and M. Hušková, eds.), 467–474. Physica-Verlag, Heidelberg.
- Drost, F. C., Klaassen, C. A. J. and Werker, B. J. M. (1994b). Adaptive estimation in time-series models. CentER discussion paper 9488, Tilburg University. To appear in: *Ann. Statist.* 25 (1997).
- 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.
- Engle, R. F. and González-Rivera (1991). Semiparametric ARCH models. *J. Business Economic Statist.* 9, 345–359.
- Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* 81, 310–320.
- Frigessi, A., Hwang, C.-R. and Younes, L. (1992). Optimal spectral structure of reversible stochastic matrices, Monte Carlo methods and the simulation of Markov random fields. *Ann. Appl. Probab.* 2, 610–628.

- Garel, B. B. (1989). The asymptotic distribution of the likelihood ratio for M. A. processes with a regression trend. *Statist. Decisions* 7, 167–184.
- Gassiat, E. (1993). Adaptive estimation in noncausal stationary AR processes. *Ann. Statist.* 21, 2022–2042.
- Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., eds. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.
- Godambe, V. P. (1985). The foundations of finite sample estimation in stochastic processes. *Biometrika* 72, 419–428.
- Godambe, V. P., ed. (1991). *Estimating Functions*. Oxford University Press.
- Godambe, V. P. and Heyde, C. C. (1987). Quasi-likelihood and optimal estimation. *Internat. Statist. Rev.* 55, 231–244.
- Godambe, V. P. and Thompson, M. E. (1989). An extension of quasi-likelihood estimation. *J. Statist. Plann. Inference* 22, 137–152.
- Gordin, M. I. (1969). The central limit theorem for stationary processes. *Soviet Math. Dokl.* 10, 1174–1176.
- Gordin, M. I. and Lifšic, B. A. (1978). The central limit theorem for stationary Markov processes. *Soviet Math. Dokl.* 19, 392–394.
- Green, P. J. and Han, X.-l. (1992). Metropolis methods, Gaussian proposals and antithetic variables. In: *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis* (P. Barone, A. Frigessi and M. Piccioni, eds.), 142–164, *Lecture Notes in Statistics* 74, Springer-Verlag, Berlin.
- Greenwood, P. E., McKeague, I. W. and Wefelmeyer, W. (1997). Information bounds for Gibbs samplers. Submitted.
- Greenwood, P. E. and Wefelmeyer, W. (1994). Nonparametric estimators for Markov step processes. *Stochastic Process. Appl.* 52, 1–16.
- Greenwood, P. E. and Wefelmeyer, W. (1995). Efficiency of empirical estimators for Markov chains. *Ann. Statist.* 23, 132–143.
- Greenwood, P. E. and Wefelmeyer, W. (1996a). Empirical estimators for semi-Markov processes. *Math. Meth. Statist.* 5, 299–315.
- Greenwood, P. E. and Wefelmeyer, W. (1996b). Reversible Markov chains and optimality of symmetrized empirical estimators. Under revision for: *Bernoulli*.
- Greenwood, P. E. and Wefelmeyer, W. (1997). Maximum likelihood estimator and Kullback–Leibler information in misspecified Markov chain models. To appear in: *Teor. Veroyatnost. i Primenen.*

- Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. Verw. Gebiete* 14, 323–330.
- 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.
- Hosoya, Y. (1989). The bracketing condition for limit theorems on stationary linear processes. *Ann. Statist.* 17, 401–418.
- Huang, W.-M. (1986). A characterization of limiting distributions of estimators in an autoregressive process. *Ann. Inst. Statist. Math.* 38, 137–144.
- Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* 1, 221–233.
- Huhtala, K. (1992). A quasi-likelihood Markov model for the hardenability of steel. In: *Proceedings of the Sixth European Conference on Mathematics in Industry* (F. Hodnett, ed.), 191–194. Teubner, Stuttgart.
- Hwang, S. Y. and Basawa, I. V. (1993). Asymptotic optimal inference for a class of nonlinear time series models. *Stochastic Process. Appl.* 46, 91–113.
- Hwang, S. Y. and Basawa, I. V. (1994). Large sample inference for conditional exponential families with applications to nonlinear time series. *J. Statist. Plann. Inference* 38, 141–158.
- Hwang, S. Y. and Basawa, I. V. (1996a). Parameter estimation for generalized random coefficient autoregressive processes. To appear in: *Proceedings of the Franco-Belgian Conference on Nonlinear Time Series*.
- Hwang, S. Y. and Basawa, I. V. (1996b). The local asymptotic normality of a class of generalized random coefficient autoregressive processes. To appear in: *Statist. Probab. Lett.*
- 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.

- Kessler, M. and Sørensen, M. (1995). Estimating equations based on eigenfunctions for a discretely observed diffusion process. Research Report 332, Department of Theoretical Statistics, University of Aarhus.
- Kessler, M. and Wefelmeyer, W. (1996). The information in the marginal law of a Markov chain. Unpublished manuscript.
- Koul H. L. (1992). Weighted Empiricals and Linear Models. IMS Lecture Notes—Monograph Series 21, Hayward, California.
- Koul, H. L. and Leventhal, S. (1989). Weak convergence of the residual empirical process in explosive autoregression. *Ann. Statist.* 17, 1784–1784.
- Koul, H. L. and Schick, A. (1996a). Adaptive estimation in a random coefficient autoregressive model. *Ann. Statist.* 24, 1025–1052.
- Koul, H. L. and Schick, A. (1996b). Efficient estimation in nonlinear autoregressive time series models. Unpublished manuscript. To appear in: Bernoulli.
- 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.
- Kreiss, J.-P. (1991). Estimation of the distribution function of noise in stationary processes. *Metrika* 38, 285–297.
- Linton, O. (1993). Adaptive estimation in ARCH models. *Econometric Theory* 9, 539–569.
- Liu, J. S., Wong, W. H. and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* 81, 27–40.
- Liu, J. S., Wong, W. H. and Kong, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *J. Roy. Statist. Soc. Ser. B* 57, 157–169.
- 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.
- Meyn, S. P. and Tweedie, R. L. (1994). *Markov Chains and Stochastic Stability*. Second Printing. Springer, London.
- Nicholls, D. F. and Quinn, B. G. (1982). *Random Coefficient Autoregressive Models: An Introduction*. Lecture Notes in Statistics 11. Springer, New York.
- Ogata, Y. (1980). Maximum likelihood estimates of incorrect Markov models for time series and the derivation of AIC. *J. Appl. Probab.* 17, 59–72.

- Pedersen, A. R. (1995). Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. *Bernoulli* 1, 257–279.
- Penev, S. (1991). Efficient estimation of the stationary distribution for exponentially ergodic Markov chains. *J. Statist. Plann. Inference* 27, 105–123.
- Peskun, P. H. (1973). Optimum Monte Carlo sampling using Markov chains. *Biometrika* 60, 607–612.
- Roussas, G. G. (1965). Asymptotic inference in Markov processes. *Ann. Math. Statist.* 36, 987–992.
- Schick, A. (1993). An adaptive estimator of the autocorrelation coefficient in a semi-parametric regression model with autoregressive errors. Unpublished manuscript.
- Schick, A. (1996). Efficient estimation in a semiparametric additive regression model with autoregressive errors. *Stochastic Process. Appl.* 61, 339–361.
- Sørensen, M. (1996). On estimation for discretely observed diffusions: a review. Research Report 348, Department of Theoretical Statistics, University of Aarhus.
- Swensen, A. R. (1985). The asymptotic distribution of the likelihood ratio for autoregressive time series with a regression trend. *J. Multivariate Anal.* 16, 54–70.
- Tong, H. (1990). *Nonlinear Time Series: A Dynamical Approach*. Oxford Statistical Science Series 6. Oxford University Press.
- Wefelmeyer, W. (1991). A generalization of asymptotically linear estimators. *Statist. Probab. Lett.* 11, 195–199.
- 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. (1997a). Adaptive estimators for parameters of the autoregression function of a Markov chain. To appear in: *J. Statist. Plann. Inference*.
- Wefelmeyer, W. (1997b). Quasi-likelihood regression models for Markov chains. To appear.
- Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics* 44, 1019–1031.

Wolfgang Wefelmeyer
Universität-GH Siegen
Fachbereich 6 Mathematik
Hölderlinstr. 3
57068 Siegen, Germany
wefelmeyer@mathematik.uni-siegen.d400.de
<http://www.math.uni-siegen.de/statistik/wefelmeyer.html>