Efficient estimation of invariant distributions of some semiparametric Markov chain models

Anton Schick
SUNY Binghamton

Wolfgang Wefelmeyer
Universität-GH Siegen

Abstract

We characterize efficient estimators for the expectation of a function under the invariant distribution of a Markov chain and outline ways of constructing such estimators. We consider two models. The first is described by a parametric family of constraints on the transition distribution; the second is the heteroscedastic nonlinear autoregressive model. The efficient estimator for the first model adds a correction term to the empirical estimator. In the second model, the suggested efficient estimator is a one-step improvement of an initial estimator which might be obtained by a version of Markov chain Monte Carlo.

AMS 1991 subject classifications. Primary 62F12, 62G20; secondary 62J09, 62M05.

Key words and Phrases. Markov chain, efficient estimator, empirical estimator, sample splitting, nonlinear heteroscedastic autoregression, quasi-likelihood, Markov chain Monte Carlo.

1 Introduction

Let $X_0, \dots, X_n$ be observations from a homogeneous and geometrically ergodic Markov chain with transition distribution $Q(x, dy)$ and invariant distribution $\pi(dx)$. We want to estimate the expectation $\pi(f) = \int \pi(dx) f(x)$ of a function f under π. The usual estimator is the empirical estimator $\frac{1}{n}\sum_{i=1}^n f(X_i)$. It is efficient if nothing is known about the transition distribution; see Penev [25], Bickel [2] and Greenwood and Wefelmeyer [9]. We expect that the empirical estimator can be improved if we have partial knowledge about Q.

Two types of models are considered in the literature. For one type, information about Q is given *indirectly* through restrictions on the *invariant* distribution of the chain. Examples of this type include parametric or semiparametric modeling of π as well as reversibility of the chain, $\pi(dx)Q(x, dy) = \pi(dy)Q(y, dx)$. Efficient estimation in this type of model is considered in Greenwood and Wefelmeyer [10] and Kessler, Schick and Wefelmeyer [18].

For the second type, information is given *directly* about the transition distribution Q. An example is $X_i = \vartheta X_{i-1} + \varepsilon_i$, where the ε_i are martingale increments. The corresponding restriction on Q is

$$\int Q(x, dy)y = \vartheta x. \tag{1.1}$$

Efficient estimators of ϑ in this and related models are constructed in Wefelmeyer [32].

A submodel of (1.1) is the AR(1) model, with ε_i i.i.d. innovations with mean zero density p, in which case

$$Q(x, dy) = p(y - \vartheta x)dy. \tag{1.2}$$

Efficient estimators of ϑ are constructed in Kreiss [20], [21], and of expectations under the innovation distribution in Wefelmeyer [30].

For models of the type (1.1) and (1.2), efficient estimation of $\pi(f)$ has not yet been treated. In this paper we outline ways of characterizing and constructing efficient estimators of $\pi(f)$ for such models.

The paper is organized as follows. In Section 2 we recall, for general Markov chain models, the characterization of estimators which are efficient in the sense of being least dispersed and regular. This characterization says that an estimator is efficient if and only if it is asymptotically linear with influence function equal to the canonical gradient. The canonical gradient is the projection of an arbitrary gradient into the tangent space of the model. In Sections 3 and 4 we compute the canonical gradient of the functional $\pi(f)$ for various models and thus obtain the characterization of efficient estimators in these models.

In Section 3 we consider a family of models which generalize (1.1). They are given by a parametric family of restrictions on the transition distribution,

$$\int Q(x, dy)a_\vartheta(x, y) = 0, \tag{1.3}$$

where a_ϑ may be vector-valued. Besides (1.1), this model includes nonlinear AR(1) models $X_i = m_\vartheta(X_{i-1}) + \varepsilon_i$ with martingale increment innovations ε_i, for which (1.3) reduces to

$$\int Q(x, dy)y = m_\vartheta(x). \tag{1.4}$$

Model (1.3) also includes the ARCH(1) model $X_i = \sigma(1 + \alpha X_{i-1}^2)^{1/2}\varepsilon_i$ with martingale increment innovations ε_i. In this case, (1.3) reduces to

$$
\begin{aligned}
\int Q(x,dy)y &= 0, \\
\int Q(x,dy)y^2 &= \sigma^2(1+\alpha x^2).
\end{aligned} \tag{1.5}
$$

The above models (1.1), (1.4), (1.5) are special cases of *quasi-likelihood* Markov chain models,

$$
\begin{aligned}
\int Q(x,dy)y &= m_\vartheta(x), \\
\int Q(x,dy)(y-m_\vartheta(x))^2 &= v_\vartheta(x).
\end{aligned} \tag{1.6}
$$

While we consider efficient estimation of $\pi(f)$, efficient estimation of ϑ in these models has been treated in Wefelmeyer [31], where additional references to literature on these models may be found.

The *heteroscedastic nonlinear autoregression model*

$$X_i = m_\vartheta(X_{i-1}) + s_\vartheta(X_{i-1})\varepsilon_i$$

with *independent* innovations is a submodel of the quasi-likelihood model with $v_\vartheta(x) = s_\vartheta(x)^2$. The transition distribution has the form

$$Q(x,dy) = s_\vartheta(x)^{-1}p\big(s_\vartheta(x)^{-1}(y-m_\vartheta(x))\big)dy,$$

where p is the common density of the innovations. The model includes the ARCH(1) model $X_i = \sigma(1+\alpha X_{i-1}^2)^{1/2}\varepsilon_i$ with independent innovations ε_i. In this case, $m_\vartheta(x) = 0$ and $s_\vartheta(x)^2 = \sigma^2(1+\alpha x^2)$. We treat efficient estimation of $\pi(f)$ in Section 4. The construction of an efficient estimator is outlined only for the case of a *known* innovation density p. Efficient estimation of ϑ is treated in Linton [23], Hwang and Basawa [15], Drost, Klaassen and Werker [4], [5], Jeganathan [12], Koul and Schick [19].

The model with independent innovations turns out to be considerably less tractable than the model with martingale innovations, which is close to nonparametric. The two models differ in several aspects. We comment on the differences in Section 5.

In this paper we focus on the main ideas and suppress the necessary regularity conditions, geometric ergodicity of the Markov chain and appropriate differentiability properties of $m_\vartheta(x)$ and $s_\vartheta(x)$ as functions of ϑ.

2 Characterization of efficient estimators

Consider a family $\mathcal{Q}$ of transition distributions on some measurable state space. Fix $Q \in \mathcal{Q}$ such that the corresponding Markov chain is ergodic with invariant distribution

π. A *local model* at $\mathcal{Q}$ is obtained by perturbing Q as

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

in such a way that Q_{nh} lies within $\mathcal{Q}$. In regular cases the *local parameter* h will run through a *linear* subspace H_0 of

$$H = \{h \in L_2(\pi \otimes Q) : \int Q(x,dy)h(x,y) = 0\}.$$

The space H_0 is called the *tangent space*. Suppose we observe $X_0, \dots, X_n$ driven by Q, with fixed initial distribution $\mu(dx)$. Write P_n and P_{nh} for the joint distribution of $X_0, \dots, X_n$ if Q and Q_{nh}, respectively, are the transition distributions. Then the log-likelihood admits a stochastic expansion

$$\begin{aligned}\log \frac{dP_{nh}}{dP_n}(X_0,\dots,X_n) &= n^{-1/2}\sum_{i=1}^n \log \frac{dQ_{nh}}{dQ}(X_{i-1},X_i) \\ &= n^{-1/2}\sum_{i=1}^n h(X_{i-1},X_i) - \frac{1}{2}\pi \otimes Q(h^2) + o_P(1),\end{aligned}$$

and by a martingale central limit theorem,

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

with N standard normal. This is *local asymptotic normality* in the sense of LeCam [22]. Proofs for Markov chains under increasingly weaker conditions are given by Roussas [27], Penev [25] and Höpfner [13], [14].

Consider a real-valued functional t on $\mathcal{Q}$. It is called *differentiable* at Q with *gradient* $g \in H$ if

$$n^{1/2}(t(Q_{nh}) - t(Q)) \to \pi \otimes Q(hg) \quad \text{for } h \in H_0.$$

The *canonical gradient* is the projection g_0 of g onto H_0.

By the convolution theorem in the version of Pfanzagl and Wefelmeyer [26], Theorem 9.3.1, an estimator T_n of $t(Q)$ is efficient if and only if it is asymptotically linear with influence function equal to the canonical gradient,

$$n^{1/2}(T_n - t(Q)) = n^{-1/2}\sum_{i=1}^n g_0(X_{i-1},X_i) + o_P(1). \tag{2.1}$$

We are interested in estimating the expectation $\pi(f)$ of a function f under the invariant distribution. The usual estimator is the *empirical estimator* $\frac{1}{n}\sum_{i=1}^n f(X_i)$. This estimator is asymptotically linear,

$$n^{1/2}\Big(\frac{1}{n}\sum_{i=1}^n f(X_i) - \pi(f)\Big) = n^{-1/2}\sum_{i=1}^n Af(X_{i-1},X_i) + o_P(1), \tag{2.2}$$

with influence function

$$Af(x,y) = \sum_{j=0}^{\infty} \Big(\int Q^j(y,dz)f(z) - \int Q^{j+1}(x,dz)f(z) \Big).$$

This follows from the martingale approximation of Gordin [7]; see also Gordin and Lifšic [8] and Meyn and Tweedie [24], Section 17.4.

By Kartashov [16], [17], the transition distribution Q_{nh} has an invariant distribution, say π_{nh}, and the functional $t(Q) = \pi(f)$ is differentiable at Q with gradient Af,

$$n^{1/2}(\pi_{nh}(f) - \pi(f)) \to \pi \otimes Q(hAf) \quad \text{for } h \in H_0.$$

The canonical gradient g_0 is the projection of Af onto H_0. The empirical estimator is thus efficient if and only if Af is in H_0. If nothing is known about the transition distribution, then H_0 equals H, and the empirical estimator is efficient.

Both the martingale approximation (2.2) and differentiability of $\pi(f)$ require that the chain is geometrically ergodic. Sufficient conditions for model (1.1) are in Schick [29], and sufficient conditions for nonlinear autoregression models are in Guegan and Diebolt [11], Bhattacharya and Lee [3] and An and Huang [1].

3 Constrained transition distributions

In this section we consider a Markov chain model which is given by the following parametric family of restrictions on the transition distribution $Q(x,dy)$:

$$\int Q(x,dy)a_\vartheta(x,y) = 0. \tag{3.1}$$

For simplicity, we first restrict attention to *one*-dimensional ϑ and *real*-valued a_ϑ. Extensions to higher dimensions are indicated in Remark 1. The tangent space is obtained by perturbing Q as, say, $Q_{nh}(x,dy) \doteq Q(x,dy)(1 + n^{-1/2}h(x,y))$ subject to the restriction (3.1) with a perturbed ϑ, say $\vartheta_{nu} = \vartheta + n^{-1/2}u$:

$$\int Q_{nh}(x,dy)a_{\vartheta_{nu}}(x,y) = o(n^{-1/2}).$$

This implies the following restriction on the local parameter h:

$$\int Q(x,dy)a_\vartheta(x,y)h(x,y) = uc(x) \tag{3.2}$$

with

$$c(x) = -\int Q(x,dy)a'_\vartheta(x,y), \tag{3.3}$$

where the prime denotes a derivative with respect to the parameter ϑ. Here and in the following, we suppress the dependence on ϑ whenever the function is not just a function of ϑ but depends also on the unknown transition distribution Q. Let H_1 denote the set of all $h \in H$ satisfying (3.2) with $u = 1$. Then the tangent space H_0 is the linear span of H_1,

$$H_0 = [H_1] = \{uh : h \in H_1, u \in \mathbf{R}\}.$$

Since the difference of any two solutions solves the corresponding *homogeneous* equation, H_0 can be decomposed as

$$H_0 = [h] + K, \tag{3.4}$$

where h is any solution of (3.2) with $u \neq 0$, and K is the set of all solutions of the corresponding homogeneous equation,

$$K = \{h \in H : \int Q(x, dy) a_\vartheta(x, y) h(x, y) = 0\}.$$

Let

$$\psi(x, y) = v(x)^{-1/2} a_\vartheta(x, y),$$

where v is the conditional variance of a_ϑ,

$$v(x) = \int Q(x, dy) a_\vartheta(x, y)^2.$$

Note that in the definition of K the function $a_\vartheta(x, y)$ can be replaced by $\psi(x, y)$ or $e(x)\psi(x, y)$. Hence the orthogonal complement $K^\perp$ of K in H is the set of all functions $h \in H$ of the form $h(x, y) = e(x)\psi(x, y)$ with $e \in L_2(\pi)$ arbitrary. There is only one solution of (3.2) with $u = 1$ in $K^\perp$, namely $h = \varphi$ with

$$\varphi(x, y) = v(x)^{-1/2} c(x) \psi(x, y). \tag{3.5}$$

Thus we can decompose H_0 and H as sums of *orthogonal* subspaces,

$$H_0 = [\varphi] \oplus K, \quad H = H_0 \oplus L \tag{3.6}$$

with $L = \{h \in K^\perp : h \perp \varphi\}$. The elements h of L are of the form

$$h(x, y) = e(x)\psi(x, y) \quad \text{with } \pi(v^{-1/2} ce) = 0.$$

Consider now the problem of estimating the expectation $\pi(f)$ of a function f under the invariant distribution π. According to Section 2, the canonical gradient g_0 is the projection of Af onto H_0. It is of the form $u_0\varphi + k_0$, where $u_0\varphi$ is the projection of Af

onto $[\varphi]$, and k_0 the projection of Af onto K. By the definition of $K^\perp$, the projection of Af onto $K^\perp$ is $e_0(x)\psi(x,y)$ with

$$e_0(x) = \int Q(x,dy)\psi(x,y)Af(x,y).$$

Hence the projection of Af onto K is

$$k_0(x,y) = Af(x,y) - e_0(x)\psi(x,y).$$

Furthermore,

$$u_0 = \frac{\pi \otimes Q(\varphi Af)}{\pi \otimes Q(\varphi^2)} = \frac{\pi(v^{-1/2}ce_0)}{\pi(v^{-1}c^2)}.$$

In conclusion, the canonical gradient is

$$g_0(x,y) = Af(x,y) - \big(e_0(x) - u_0 v(x)^{-1/2}c(x)\big)\psi(x,y).$$

Alternatively, we can write

$$g_0(x,y) = Af(x,y) - w(x)a_\vartheta(x,y) \tag{3.7}$$

with

$$w(x) = v(x)^{-1}(d_0(x) - u_0 c(x))$$

and

$$d_0(x) = v(x)^{1/2}e_0(x) = \int Q(x,dy)a_\vartheta(x,y)Af(x,y).$$

Hence, by (2.1), an efficient estimator T_n of $\pi(f)$ is characterized by

$$n^{1/2}(T_n - \pi(f)) = n^{-1/2}\sum_{i=1}^n \big(Af(X_{i-1},X_i) - w(X_{i-1})a_\vartheta(X_{i-1},X_i)\big) + o_P(1). \tag{3.8}$$

Its asymptotic variance is

$$\pi \otimes Q(Af)^2 - \pi(vw^2). \tag{3.9}$$

This shows that in this model the variance bound is reduced by

$$\pi(vw^2) = \pi(v^{-1}d_0^2) - \frac{(\pi(v^{-1}cd_0))^2}{\pi(v^{-1}c^2)}.$$

By the Schwarz inequality, this is strictly positive unless d_0 is proportional to c.

In view of the characterization (3.8) and the martingale approximation (2.2) of the empirical estimator, we obtain an efficient estimator

$$T_n = \frac{1}{n}\sum_{i=1}^n f(X_i) - W_n$$

if we can construct W_n such that

$$n^{1/2}W_n = n^{-1/2}\sum_{i=1}^n w(X_{i-1})a_\vartheta(X_{i-1},X_i) + o_P(1). \tag{3.10}$$

Let us now sketch a possible construction of W_n. Write

$$\begin{aligned} n^{-1/2}&\sum_{i=1}^n w_n(X_{i-1})a_{\vartheta_n}(X_{i-1},X_i) \\ &= n^{-1/2}\sum_{i=1}^n w_n(X_{i-1})\Big(a_{\vartheta_n}(X_{i-1},X_i) - \int Q(X_{i-1},dy)a_{\vartheta_n}(X_{i-1},y)\Big) \\ &\quad + n^{-1/2}\sum_{i=1}^n w_n(X_{i-1})\int Q(X_{i-1},dy)a_{\vartheta_n}(X_{i-1},y). \end{aligned}$$

If w_n and ϑ_n are *deterministic* sequences, the first right-hand term is a martingale. It approximates the right side of (3.10) if

$$\iint \pi(dx)Q(x,dy)\big(w_n(x)a_{\vartheta_n}(x,y) - w(x)a_\vartheta(x,y)\big)^2 \to 0.$$

If $n^{1/2}(\vartheta_n - \vartheta)$ is bounded, the second term is approximately

$$n^{1/2}(\vartheta_n - \vartheta)\frac{1}{n}\sum_{i=1}^n w_n(X_{i-1})c(X_{i-1}).$$

The average converges to $\pi(wc) = 0$ if $\pi(|(w_n - w)c|) \to 0$.

These arguments remain valid for *estimators* $\hat{w}$ and $\hat{\vartheta}$ in place of w_n and ϑ_n if $\hat{w}$ is based on *independent* copies of our sample and $\hat{\vartheta}$ is a *discretized* $n^{1/2}$-consistent estimator. Since independent samples are not available to us, we use the sample splitting technique of Schick [29]. We pick three subsamples so that we can use separate subsamples for estimating certain terms. This amounts to using two independent copies $Y_0,\dots,Y_n$ and $Z_0,\dots,Z_n$ of the observations $X_0,\dots,X_n$.

To estimate w, we must estimate v, d_0 and c. Note that $Af(x,y) = Uf(y) - \int Q(x,dy)Uf(y)$ with

$$Uf(y) = \sum_{j=0}^\infty \int Q^j(y,dz)(f(z) - \pi(f)).$$

Hence $d_0(x)$ can be approximated by

$$\int Q(x,dy)a_\vartheta(x,y)\sum_{j=0}^m \int Q^j(y,dz)f(z),$$

where m tends to infinity sufficiently slowly. Here we have replaced the infinite series by a finite sum and omitted the centering. We estimate $\int Q^j(x,dy)f(y)$ by a j-step kernel estimator,

$$\hat{Q}^j f(x) = \frac{\sum_{k=1}^{n-m} f(Z_{k+j})K_n(Z_k - x)}{\sum_{k=1}^{n-m} K_n(Z_k - x)},$$

where $K_n(x) = h_n^{-1}K(h_n^{-1}x)$ for some density K and bandwidth h_n tending to zero. Hence an estimator for $d_0(x)$ is

$$\hat{d}_0(x) = \frac{\sum_{i=1}^n a_{\hat{\vartheta}}(Y_{i-1},Y_i)K_n(Y_{i-1}-x)}{\sum_{i=1}^n K_n(Y_{i-1}-x)} \sum_{j=0}^m \hat{Q}^j f(Y_i).$$

Similarly, estimate $v(x)$, $c(x)$, u_0 by

$$\begin{aligned}
\hat{v}(x) &= \frac{\sum_{i=1}^n a_{\hat{\vartheta}}(Y_{i-1},Y_i)^2 K_n(Y_{i-1}-x)}{\sum_{i=1}^n K_n(Y_{i-1}-x)}, \\
\hat{c}(x) &= -\frac{\sum_{i=1}^n a'_{\hat{\vartheta}}(Y_{i-1},Y_i)K_n(Y_{i-1}-x)}{\sum_{i=1}^n K_n(Y_{i-1}-x)}, \\
\hat{u}_0 &= \frac{\sum_{i=1}^n \hat{v}(X_i)^{-1}\hat{c}(X_i)\hat{d}_0(X_i)}{\sum_{i=1}^n \hat{v}(X_i)^{-1}\hat{c}(X_i)^2}.
\end{aligned}$$

In $\hat{u}_0$ we have used the original observations since the sample splitting techniques of Schick [29] allow *multiplicative* constants to be estimated by the full data. Then $w(x)$ is estimated by

$$\hat{w}(x) = \hat{v}(x)^{-1}(\hat{d}_0(x) - \hat{u}_0\hat{c}(x)),$$

and our outline of the construction of W_n is finished.

For technical reasons it may be necessary to set the estimators $\hat{d}_0(x)$, $\hat{v}(x)$ and $\hat{c}(x)$ equal to zero when $|x|$ is large or the denominators are relatively small; see Schick [29].

Remark 1. If a_ϑ is r-dimensional and ϑ is s-dimensional, then a'_ϑ and c are $r \times s$-matrices, v is the conditional $r \times r$-dispersion matrix of a_ϑ, d_0 is an r-vector, and u_0 is an s-vector solving $\pi(c^\top v^{-1}c)u_0 = \pi(c^\top v^{-1}d_0)$. Hence the gradient is

$$g_0(x,y) = Af(x,y) - (d_0(x) - c(x)u_0)^\top v(x)^{-1}a_\vartheta(x,y).$$

Example 1. Consider the nonlinear AR(1) model $X_i = m_\vartheta(X_{i-1}) + \varepsilon_i$ with martingale increment innovations ε_i and ϑ s-dimensional. Here $r = 1$, and (3.1) holds with

$a_\vartheta(x,y) = y - m_\vartheta(x)$, so that $\int Q(x,dy)y = m_\vartheta(x)$. We have

$$\begin{aligned} c &= m'_\vartheta, \\ v(x) &= \int Q(x,dy)(y - m_\vartheta(x))^2, \\ d_0(x) &= \int Q(x,dy)(y - m_\vartheta(x))Af(x,y), \\ u_0 &= \big(\pi(v^{-1}m'^\top_\vartheta m'_\vartheta)\big)^{-1}\pi(v^{-1}d_0 m'^\top_\vartheta). \end{aligned}$$

Hence the canonical gradient for $\pi(f)$ is

$$g_0(x,y) = Af(x,y) - (d_0(x) - m'_\vartheta(x)u_0)v(x)^{-1}(y - m_\vartheta(x)).$$

For the linear AR(1) model we have $m_\vartheta(x) = \vartheta x$, so that $m'_\vartheta(x) = x$. For the SETAR model we have $m_\vartheta(x) = \vartheta_1 x 1_{(x<0)} + \vartheta_2 x 1_{(x>0)}$, so that $m'_\vartheta(x) = (x1_{(x<0)}, x1_{(x>0)})$.

4 Heteroscedastic nonlinear autoregression

In this section we consider the heteroscedastic nonlinear autoregression model of order one,

$$X_i = m_\vartheta(X_{i-1}) + s_\vartheta(X_{i-1})\varepsilon_i,$$

with independent innovations ε_i which have mean 0, variance 1, finite fourth moment and unknown positive density p with finite Fisher information for location and scale. The transition distribution has the form $Q(x,dy) = s_\vartheta(x)^{-1}p(\varepsilon_\vartheta(x,y))dy$ with $\varepsilon_\vartheta(x,y) = s_\vartheta(x)^{-1}(y - m_\vartheta(x))$. The tangent space is obtained by perturbing p as $p_{nk}(x) \doteq p(x)(1 + n^{-1/2}k(x))$ and ϑ as $\vartheta_{nu} = \vartheta + n^{-1/2}u$. Because the innovations have mean 0 and variance 1, we must have

$$\int p(x)dx\, k(x) = 0, \qquad \int p(x)dx\, xk(x) = 0, \qquad \int p(x)dx\, x^2k(x) = 1.$$

Let K denote the set of all such k. Let ℓ_1, ℓ_2 denote the score functions for location and scale, $\ell_1(x) = -p'(x)/p(x)$, $\ell_2(x) = x\ell_1(x) - 1$. The perturbed Q is

$$Q_{nuk}(x,dy) = s_{\vartheta_{nu}}(x)^{-1}p_{nk}(\varepsilon_{\vartheta_{nu}}(x,y))dy \doteq Q(x,dy)\big(1 + n^{-1/2}\big(L(x,y)u + k(\varepsilon_\vartheta(x,y))\big)\big)$$

with

$$L(x,y) = s_\vartheta(x)^{-1}\big(\ell_1(\varepsilon_\vartheta(x,y))m'_\vartheta(x) + \ell_2(\varepsilon_\vartheta(x,y))s'_\vartheta(x)\big).$$

Here $m'_\vartheta(x)$ and $s'_\vartheta(x)$ are *row* vectors of the same dimension as ϑ. Thus the tangent space is

$$H_0 = \{Lu + k(\varepsilon_\vartheta) : u \in \mathbf{R}, k \in K\}.$$

With L_* denoting the projection of L onto $K(\varepsilon_\vartheta) = \{k(\varepsilon_\vartheta) : k \in K\}$ and $L_0 = L - L_*$, we can write H_0 as the orthogonal sum $H_0 = [L_0] \oplus K(\varepsilon_\vartheta)$. Note that $K(\varepsilon_\vartheta)$ is the orthogonal complement of $[1, \varepsilon_\vartheta, \varepsilon_\vartheta^2]$ in the set of all functions of ε_ϑ. Hence the projection of a function $h \in H$ onto $K(\varepsilon_\vartheta)$ is obtained by first taking the conditional expectation $\mathrm{E}\,(h|\varepsilon_\vartheta)$ given ε_ϑ and then subtracting from $\mathrm{E}\,(h|\varepsilon_\vartheta)$ its projection onto $[1, \varepsilon_\vartheta, \varepsilon_\vartheta^2]$. An orthogonal basis of this space is $[1, \varepsilon_\vartheta, \varepsilon_\vartheta^2 - 1 - \mu_3\varepsilon_\vartheta]$, where μ_k denotes the k-th moment of p. Since $\mathrm{E}\,(h|\varepsilon_\vartheta)$ has expectation 0, the projection of h onto $K(\varepsilon_\vartheta)$ is

$$Bh(\varepsilon_\vartheta) = \mathrm{E}\,(h|\varepsilon_\vartheta) - \varepsilon_\vartheta \pi \otimes Q(h\varepsilon_\vartheta) - (\varepsilon_\vartheta^2 - 1 - \mu_3\varepsilon_\vartheta)\frac{\pi \otimes Q(h \cdot (\varepsilon_\vartheta^2 - 1 - \mu_3\varepsilon_\vartheta))}{\mu_4 - 1 - \mu_3^2}.$$

To calculate L_*, we recall that

$$\pi \otimes Q(\ell(\varepsilon_\vartheta)(\varepsilon_\vartheta, \varepsilon_\vartheta^2 - 1)) = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix},$$

where $\ell = (\ell_1, \ell_2)^\top$. See, e.g., Schick [28], Remark 3.10. Hence

$$\begin{aligned} L_* &= \Big(\ell_1(\varepsilon_\vartheta) - \varepsilon_\vartheta + (\varepsilon_\vartheta^2 - 1 - \mu_3\varepsilon_\vartheta)\frac{\mu_3}{\mu_4 - 1 - \mu_3^2}\Big)\pi(s_\vartheta^{-1}m_\vartheta') \\ &\quad + \Big(\ell_2(\varepsilon_\vartheta) - (\varepsilon_\vartheta^2 - 1 - \mu_3\varepsilon_\vartheta)\frac{2}{\mu_4 - 1 - \mu_3^2}\Big)\pi(s_\vartheta^{-1}s_\vartheta'), \end{aligned}$$

and $L_0(x, y) = \Lambda(x, \varepsilon_\vartheta(x, y))$ with

$$\begin{aligned} \Lambda(x, \varepsilon) &= \ell_1(\varepsilon)\big(s_\vartheta(x)^{-1}m_\vartheta'(x) - \pi(s_\vartheta^{-1}m_\vartheta')\big) + \ell_2(\varepsilon)\big(s_\vartheta(x)^{-1}s_\vartheta'(x) - \pi(s_\vartheta^{-1}s_\vartheta')\big) \\ &\quad + \varepsilon\pi(s_\vartheta^{-1}m_\vartheta') + \frac{\varepsilon^2 - 1 - \mu_3\varepsilon}{\mu_4 - 1 - \mu_3^2}\big(2\pi(s_\vartheta^{-1}s_\vartheta') - \mu_3\pi(s_\vartheta^{-1}m_\vartheta')\big). \end{aligned}$$

Consider now the problem of estimating the expectation $\pi(f)$ of a function f under the invariant distribution π. According to Section 2, the canonical gradient g_0 is the projection of Af onto $H_0 = [L_0] \oplus K(\varepsilon_\vartheta)$. It is of the form $g_0 = L_0 u_0 + BAf(\varepsilon_\vartheta)$, where $u_0 = \pi \otimes Q(L_0^\top L_0)^{-1}\pi \otimes Q(L_0^\top Af)$. Hence by (2.1), an efficient estimator T_n of $\pi(f)$ is characterized by

$$n^{1/2}(T_n - \pi(f)) = n^{-1/2}\sum_{i=1}^n \big(\Lambda(X_{i-1}, \varepsilon_i)u_0 + BAf(\varepsilon_i)\big) + o_P(1)$$

with $\varepsilon_i = s_\vartheta(X_{i-1})^{-1}\big(X_i - m_\vartheta(X_{i-1})\big)$. Its asymptotic variance is

$$\pi \otimes Q(Af \cdot L_0)\big(\pi \otimes Q(L_0^\top L_0)\big)^{-1}\pi \otimes Q(L_0^\top Af) + \int p(x)dx\,(BAf(x))^2.$$

The construction of an efficient estimator for $\pi(f)$ in this model is considerably more involved than the construction for the model in the previous section. For this reason, we will not treat it here in generality. Instead, we outline the construction in the special case of a *known* innovation density p, location function $m_\vartheta(x) = \vartheta x$, and scale function $s_\vartheta(x) = 1$. This model is parametric, and we write π_ϑ and Q_ϑ for π and Q. It is easy to check that in this case the tangent space H_0 is the linear span of $h(x, y) = x\ell_1(y - \vartheta x)$, and the canonical gradient is

$$g_0 = \frac{\pi_\vartheta \otimes Q_\vartheta(hAf)}{\pi_\vartheta \otimes Q_\vartheta(h^2)} h.$$

Note that $\pi_\vartheta \otimes Q_\vartheta(h^2) = (1 - \vartheta^2)^{-1} J_1$ with J_1 the Fisher information for location. Note also that here $d_\vartheta = \pi_\vartheta \otimes Q_\vartheta(hAf)$ is the derivative of $\pi_\vartheta(f)$ with respect to ϑ.

Let $\hat\vartheta$ be an efficient estimator of ϑ. Then its influence function is $(1-\vartheta^2)J_1^{-1}h$. Thus $\pi_{\hat\vartheta}(f)$ has influence function g_0 and is efficient. Now the problem is that we usually do not know π_ϑ explicitly. However, we can approximate $\pi_{\hat\vartheta}(f)$ by the empirical estimator $\frac{1}{m}\sum_{j=1}^m f(Y_j)$ based on realizations $Y_0, \dots, Y_m$ from the Markov chain with transition distribution $Q_{\hat\vartheta}$. This is a version of a Markov chain Monte Carlo method. For an introduction to such methods see Gilks, Richardson and Spiegelhalter [6]. To guarantee that the error in the approximation is negligible compared to that of the estimator $\pi_{\hat\vartheta}(f)$, the simulation sample size must be considerably larger than n.

Another efficient estimator of $\pi_\vartheta(f)$ is

$$\pi_{\hat\vartheta}(f) + \hat d(1 - \hat\vartheta^2)\frac{1}{nJ_1}\sum_{i=1}^n X_{i-1}\ell_1(X_i - \hat\vartheta X_{i-1}),$$

where $\hat\vartheta$ is a discretized version of a $n^{1/2}$-consistent estimator such as the least squares estimator, and $\hat d$ is a consistent estimator of d_ϑ such as $\frac{1}{2}n^{1/2}(\pi_{\hat\vartheta+n^{-1/2}}(f) - \pi_{\hat\vartheta-n^{-1/2}}(f))$. To see that this works, note that

$$n^{1/2}(\pi_{\hat\vartheta}(f) - \pi_\vartheta(f)) = n^{1/2}(\hat\vartheta - \vartheta)d_\vartheta + o_P(1)$$

and

$$n^{-1/2}\sum_{i=1}^n X_{i-1}\big(\ell_1(X_i - \hat\vartheta X_{i-1}) - \ell_1(X_i - \vartheta X_{i-1})\big) = -(1 - \vartheta^2)^{-1}J_1 n^{1/2}(\hat\vartheta - \vartheta) + o_P(1).$$

The last expansion follows from Koul and Schick [19], (2.11). The expectations, in turn, can be estimated using Markov chain Monte Carlo as before.

The last approach to constructing efficient estimators extends to the case of general m_ϑ and s_ϑ in an obvious way. The case of *unknown* p requires estimating the gradient g_0 by methods similar to those in Section 3.

Remark 2. How much information can be gained from knowing that the innovations ε_i are i.i.d. rather than martingale increments? Suppose that the true model is the AR(1) model $X_i = \rho X_{i-1} + \eta_i$ with independent innovations η_i which have mean 0, variance σ^2, finite fourth moment μ_4 and unknown positive density p. This is the model of Section 4, with $m_\vartheta(x) = \rho x$ and $s_\vartheta(x) = \sigma$. Suppose we want to estimate the second moment of the invariant distribution, $\pi(f)$ for $f(x) = x^2$. It is easy to check that

$$Af(x,y) = \frac{y^2 - \rho^2 x^2 - \sigma^2}{1-\rho^2}.$$

Hence the asymptotic variance of the empirical estimator $\frac{1}{n}\sum_{i=1}^n X_i^2$ for $\pi(f)$ is

$$\frac{\mu_4 - \sigma^4 + 4\sigma^4\rho^2(1-\rho^2)^{-1}}{(1-\rho^2)^2}.$$

To calculate the optimal variance for estimators of $\pi(f)$, we determine the tangent space for the AR(1) model with independent innovations. We do this directly rather than by specializing Section 4. Let $\ell(x) = -p'(x)/p(x)$ denote the score function for location. The tangent space consists of functions $ax\ell(y-\rho x) + \varphi(y-\rho x)$ with $a \in \mathbf{R}$ and $\varphi \in L_2(p)$ with $\int p(x)dx\,\varphi(x) = \int p(x)dx\,x\varphi(x) = 0$. Note that Af is orthogonal to $x\ell(y-\rho x)$. The projection of Af onto the tangent space is therefore

$$\frac{(y-\rho x)^2 - \mu_3\sigma^{-2}(y-\rho x) - \sigma^2}{1-\rho^2}.$$

Hence the optimal variance for estimators of $\pi(f)$ in the AR(1) model with independent innovations is

$$\frac{\mu_4 - \sigma^4 - \mu_3^2\sigma^{-2}}{(1-\rho^2)^2}.$$

The variance reduction over the empirical estimator is

$$\frac{\mu_3^2}{\sigma^2(1-\rho^2)^2} + \frac{4\sigma^4\rho^2}{(1-\rho^2)^3}.$$

Consider now the larger AR(1) model in which the innovations are arbitrary martingale increments. The optimal variance for estimators of $\pi(f)$ is then obtained from (3.9) as

$$\frac{\mu_4 - \sigma^4 + 4\sigma^4\rho^2(1-\rho^2)^{-1}}{(1-\rho^2)^2} - \frac{\mu_3^2}{\sigma^2(1-\rho^2)^2},$$

since now $v(x) = \sigma^2$ and $w(x) = \mu_3\sigma^{-2}(1-\rho^2)^{-1}$. The variance reduction over the empirical estimator is

$$\frac{\mu_3^2}{\sigma^2(1-\rho^2)^2}.$$

Hence if we know that the innovations are independent, we can reduce the variance by

$$\frac{4\sigma^4\rho^2}{(1-\rho^2)^3}.$$

5 Conclusion

The main example of the Markov chain model of Section 3 is the heteroscedastic nonlinear autoregressive model of order one,

$$X_i = m_\vartheta(X_{i-1}) + s_\vartheta(X_{i-1})\varepsilon_i,$$

with *martingale increment* innovations ε_i. Section 4 treats the submodel with *independent* innovations. At first sight, these two autoregression models are quite similar. One purpose of our paper is to point out that both the characterization and the construction of efficient estimators are, in fact, rather different for the two models. Quite generally, efficient estimators are the more complicated the further the model is from the two extreme ends of the spectrum: parametric and fully nonparametric.

The autoregressive model with *martingale increment* innovations is close to the nonparametric end. By (3.6), the orthogonal complement L of the tangent space H_0 consists of functions of the form $e(x)\psi(x, y)$ with known $\psi(x, y)$. This space L, although infinite-dimensional, has a very explicit description. Hence the canonical gradient g_0 of $\pi(f)$, the projection of Af onto H_0, is most easily obtained via the projection m_0 of Af onto the orthogonal complement of H_0, leading to $g_0 = Af - m_0$ with m_0 defined implicitly through (3.7). Since Af is the influence function of the empirical estimator, the form of the gradient suggests constructing an efficient estimator by correcting the empirical estimator as

$$T_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i) - W_n$$

with $n^{1/2}W_n = n^{-1/2}\sum_{i=1}^{n} w(X_{i-1}, X_i) + o_P(1)$ as in (3.10).

The autoregressive model with *independent* innovations is far from both the parametric and the nonparametric end. We do not have an explicit description of the orthogonal complement of the tangent space H_0 and calculate the canonical gradient of $\pi(f)$ by projecting Af directly onto $H_0 = [L_0] \oplus K(\varepsilon_\vartheta)$. The efficient estimator is not obtained by modifying the empirical estimator. Instead we suggest a one-step improvement which requires a better initial estimator than the empirical estimator, namely $\pi_{\hat\vartheta}(f)$.

As mentioned in the introduction, efficient estimators for ϑ rather than $\pi(f)$ in nonlinear autoregression models have been constructed by Wefelmeyer [31], [32] for martingale increment innovations and, most recently, by Drost et al. [5] and Koul and Schick [19] for independent innovations. We note that again the efficient estimators are quite different in the two models. With martingale increment innovations, a simple weighted least squares estimator with random weights is efficient. With independent innovations, an efficient estimator is obtained by a one-step improvement of an initial estimator.

References

[1] H. Z. An and F. C. Huang, *The geometrical ergodicity of nonlinear autoregressive models*, Statist. Sinica, 6 (1996), pp. 943–956.

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

[3] R. Bhattacharya and C. Lee, *On geometric ergodicity of nonlinear autoregressive models*, Statist. Probab. Lett., 22 (1995), pp. 311–315.

[4] F. C. Drost, C. A. J. Klaassen and B. J. M. Werker, *Adaptiveness in time series models*, In: Asymptotic Statistics, P. Mandl and M. Hušková, eds., Physica-Verlag, Heidelberg, 1994, pp. 467–474.

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

[6] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds., *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London, 1996.

[7] M. I. Gordin, *The central limit theorem for stationary processes*, Soviet Math. Dokl., 10 (1969), pp. 1174–1176.

[8] M. I. Gordin and B. A. Lifšic, *The central limit theorem for stationary Markov processes*, Soviet Math. Dokl., 19 (1987), pp. 392–394.

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

[10] P. E. Greenwood and W. Wefelmeyer, *Reversible Markov chains and optimality of symmetrized empirical estimators*, Bernoulli, 5 (1999), pp. 109-123.

[11] D. Guegan and J. Diebolt, *Probabilistic properties of the β-ARCH model*, Statist. Sinica, 4 (1994), pp. 71–87.

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

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

[14] R. Höpfner, *Asymptotic inference for Markov step processes: observation up to a random time*, Stochastic Process. Appl., 48 (1993), pp. 295–310.

[15] S. Y. Hwang and I. V. Basawa, *Asymptotic optimal inference for a class of nonlinear time series models*, Stochastic Process. Appl., 46 (1993), pp. 91–113.

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

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

[18] M. Kessler, A. Schick and W. Wefelmeyer, *The information in the marginal law of a Markov chain*, Submitted.

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

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

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

[22] L. LeCam, *Locally asymptotically normal families of distributions*, Univ. California Publ. Statist., 3 (1960), pp. 37–98.

[23] O. Linton, *Adaptive estimation in ARCH models*, Econometric Theory, 9 (1993), pp. 539–569.

[24] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer, London, 1994.

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

[26] J. Pfanzagl and W. Wefelmeyer, *Contributions to a General Asymptotic Statistical Theory*, Springer, New York, 1982.

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

[28] A. Schick, *On efficient estimation in regression models with unknown scale functions*, Math. Methods Statist., 3 (1994), pp. 171-212.

[29] A. Schick, *Sample splitting with Markov chains*, Submitted.

[30] W. Wefelmeyer, *An efficient estimator for the expectation of a bounded function under the residual distribution of an autoregressive process*, Ann. Inst. Statist. Math., 46 (1994), pp. 309–315.

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

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

Anton Schick
SUNY Binghamtom
Department of Mathematical Sciences
Binghamton, NY 13902-6000
USA
anton@@math.binghamton.edu
http://math.binghamton.edu/anton/index.html

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