Autoregression, estimating functions, and optimality criteria

Ursula U. Müller[1] and Wolfgang Wefelmeyer[2]

Abstract

We consider d-order Markov chains satisfying a conditional constraint $E(a_\vartheta(\mathbf{X}_{i-1}, X_i) \mid \mathbf{X}_{i-1}) = 0$, where $\mathbf{X}_{i-1} = (X_{i-1}, \ldots, X_{i-d})$. These comprise quasi-likelihood models and nonlinear and conditionally heteroscedastic autoregressive models with martingale innovations. Estimators for ϑ can be obtained from estimating equations $\sum_{i=1}^n W_\vartheta(\mathbf{X}_{i-1})^\top a_\vartheta(\mathbf{X}_{i-1}, X_i) = 0$. We review different criteria for choosing good weights $W_\vartheta(\mathbf{X}_{i-1})$. They usually lead to weights that depend on unknown features of the transition distribution and must be estimated. We compare the approach via estimating functions with other ways of constructing estimators for ϑ, and discuss efficiency of the estimators in the sense of Hájek and LeCam. Analogous comparisons may be made for regression models.

Keywords: generalized quasi-likelihood, extended quasi-likelihood, ARCH model, generalized method of moments, conditional least squares, influence function, gradient, variance bound.

1 Introduction

Let $X_{1-p}, \ldots, X_n$ be observations from a homogeneous and geometrically ergodic d-order Markov chain on some arbitrary state space. Write $\mathbf{X}_{i-1} = (X_{i-1}, \ldots, X_{i-d})$, and assume that the chain meets the conditional constraint

$$E(a_\vartheta(\mathbf{X}_{i-1}, X_i) \mid \mathbf{X}_{i-1}) = 0, \tag{1}$$

where $a_\vartheta(\mathbf{x}, y)$ with $\mathbf{x} = (x_1, \ldots, x_d)$ is a known k-dimensional vector of functions involving an unknown p-dimensional parameter ϑ. We are interested in optimal estimators of ϑ.

In Section 2 we derive an asymptotic lower bound for estimators of ϑ in the sense of Hájek and Le Cam, and give a characterization of efficient estimators.

In Section 3 we consider estimating equations for ϑ of the form

$$\sum_{i=1}^n W_\vartheta(\mathbf{X}_{i-1})^\top a_\vartheta(\mathbf{X}_{i-1}, X_i) = 0,$$

[1]Fachbereich 3 Mathematik, Universität Bremen, Postfach 330 440, 28334 Bremen, Germany. email: uschi@math.uni-bremen.de

[2]Fachbereich 6 Mathematik, Universität Siegen, Walter-Flex-Str. 3, 57068 Siegen, Germany. email: wefelmeyer@mathematik.uni-siegen.de

with $W_\vartheta(\mathbf{x})$ a $k \times p$ matrix of weights. The weights minimizing the asymptotic covariance matrix depend, through conditional expectations of certain functions, on the unknown transition distribution of the chain. Hence the optimal estimating function cannot be used as it stands for estimating ϑ. We indicate that replacing the optimal weights by appropriate estimators does not change the asymptotic covariance matrix, and show that the resulting estimating function with estimated optimal weights is efficient. We also introduce generalized quasi-likelihood estimating functions, replacing the optimal weights by parametric models for the conditional expectations. These estimating functions are easier to calculate, but inefficient both for correctly specified and for misspecified conditional expectations.

We discuss these findings in more specific situations. A particular class of examples of constraints (1) are *quasi-likelihood models*, with real state space and parametric models for the conditional means and variances,

$$E(X_i \mid \mathbf{X}_{i-1}) = r_\vartheta(\mathbf{X}_{i-1}), \tag{2}$$

$$E((X_i - r_\vartheta(X_i))^2 \mid \mathbf{X}_{i-1}) = v_\vartheta(\mathbf{X}_{i-1}). \tag{3}$$

Then $a_\vartheta(\mathbf{x}, y) = (y - r_\vartheta(\mathbf{x}), (y - r_\vartheta(\mathbf{x}))^2 - v_\vartheta(\mathbf{x}))^\top$.

Quasi-likelihood models can be written as

$$X_i = r_\vartheta(\mathbf{X}_{i-1}) + v_\vartheta(\mathbf{X}_{i-1})^{1/2}\varepsilon_i, \tag{4}$$

with innovations ε_i that are martingale increments, $E(\varepsilon_i \mid \mathbf{X}_{i-1}) = 0$, and that satisfy $E(\varepsilon_i^2 \mid \mathbf{X}_{i-1}) = 1$ for identifiability. The submodel with *independent* innovations ε_i is called *nonlinear and heteroscedastic p-order autoregressive model*. We indicate that the estimating function with (estimated) optimal weights is not efficient in this submodel because it does not use the information that the innovations are independent.

2 Efficiency

In this section we derive a characterization of efficient estimators of ϑ in the d-order Markov chain model constrained by (1).

Consider first the *nonparametric* d-order Markov chain model, without constraint (1). Write $Q(\mathbf{x}, dy)$ for the transition distribution of X_i given $\mathbf{X}_{i-1} = \mathbf{x}$, and assume that the chain is geometrically ergodic under Q. Let $\pi(d\mathbf{x})$ be the stationary law of $\mathbf{X}_{i-1}$. Write $(\pi \otimes Q)(d\mathbf{x}, dy) = \pi(d\mathbf{x})Q(\mathbf{x}, dy)$ for the joint law of $(\mathbf{X}_{i-1}, X_i)$, and $Q(\mathbf{x}, f) = \int Q(\mathbf{x}, dy)f(\mathbf{x}, y)$ for the conditional expectation of $f(\mathbf{X}_{i-1}, X_i)$ given $\mathbf{X}_{i-1} = \mathbf{x}$. Whenever the argument $\mathbf{x}$ is omitted, we find it convenient to use the shorter notation Qf for $Q(\cdot, f)$.

The nonparametric model is *locally asymptotically normal* in the following sense. Introduce (Hellinger differentiable) perturbations

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

with h in the *tangent space*

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

Since h may take large negative values, we cannot simply define Q_{nh} replacing $\doteq$ by an equality sign. There are three ways to take care of this problem: truncation of h, transformation of the density, or, simplest, restriction to bounded h (which are dense in H). The condition $Q(\mathbf{x}, h) = 0$ is required for Q_{nh} to be a transition distribution. Write P_{nh} and P_n for the joint law of $X_{1-p}, \ldots, X_n$ under Q_{nh} and Q, respectively. The log-likelihood ratio has the stochastic expansion

$$\log \frac{dP_{nh}}{dP_n} = n^{-1/2} \sum_{i=1}^{n} h(\mathbf{X}_{i-1}, X_i) - \frac{1}{2}(\pi \otimes Q)(h^2) + o_{P_n}(1).$$

For bounded h see Penev [37]. For general Hellinger differentiable perturbations, the stochastic expansion may be obtained by modifying Höpfner [21]. See also Höpfner, Jacod and Ladelli [23] and Höpfner [22]. By a martingale central limit theorem, $n^{-1/2} \sum_{i=1}^{n} h(\mathbf{X}_{i-1}, X_i)$ is asymptotically normal with variance $(\pi \otimes Q)(h^2)$.

Now suppose that the model is constrained by (1). Relation (1) may be written $Q(\mathbf{x}, a_\vartheta) = 0$. The perturbed transition distribution Q_{nh} must also fulfill the constraint, possibly with perturbed parameter, say $\vartheta_{nu} \doteq \vartheta + n^{-1/2}u$:

$$0 = Q_{nh}(\mathbf{x}, a_{\vartheta_{nu}}) \doteq Q(\mathbf{x}, a_\vartheta) + n^{-1/2}(Q(\mathbf{x}, a_\vartheta h) + Q(\mathbf{x}, \dot{a}_\vartheta)u). \tag{5}$$

Hence the tangent space of the constrained model is the union, call it H_*, of the affine spaces

$$H_u = \{h \in H : Q(\mathbf{x}, a_\vartheta h) = -Q(\mathbf{x}, \dot{a}_\vartheta)u \text{ for all } \mathbf{x}\}.$$

We recall the following definitions and results from Le Cam's and Hájek's theory of efficient estimation. The standard reference for the i.i.d. case is Bickel, Klaassen, Ritov and Wellner [1]; for Markov chains see also Wefelmeyer [42]. A p-dimensional functional $t(Q)$ is called *differentiable* at Q with *gradient* g if $g \in H^p$ and

$$n^{1/2}(t(Q_{nh}) - t(Q)) \to (\pi \otimes Q)(gh) \quad \text{for } h \in H_*. \tag{6}$$

The *canonical* gradient g_* is the componentwise projection of g onto the tangent space H_*. An estimator $\hat{t}$ for $t(Q)$ is called *regular* at Q with *limit* L if

$$n^{1/2}(\hat{t} - t(Q_{nh})) \Rightarrow L \quad \text{under } P_{nh} \quad \text{for } h \in H_*.$$

The Convolution Theorem says that if $\hat{t}$ is regular for $t(Q)$ with limit L, then

$$L = (\pi \otimes Q)(g_* g_*^\top)^{1/2} N + M \quad \text{in distribution,}$$

where N a p-dimensional standard normal random vector, and M a random vector independent of N. This justifies calling a regular estimator *efficient* for $t(Q)$ if its limit is

$$L = (\pi \otimes Q)(g_* g_*^\top)^{1/2} N \quad \text{in distribution.}$$

An estimator $\hat{t}$ for $t(Q)$ is called *asymptotically linear* at P with *influence function* f if $f \in H^p$ and

$$n^{1/2}(\hat{t} - t(Q)) = n^{-1/2} \sum_{i=1}^{n} f(\mathbf{X}_{i-1}, X_i) + o_{P_n}(1). \tag{7}$$

Such an estimator is asymptotically normal with covariance matrix $(\pi \otimes Q)(f f^\top)$. We have the following two characterizations.

1. An asymptotically linear estimator for $t(Q)$ is regular if and only if its influence function is a gradient for $t(Q)$.

2. An estimator for $t(Q)$ is (regular and) efficient if and only if it is asymptotically linear with influence function equal to the canonical gradient of $t(Q)$.

Now we apply these results to estimation of ϑ. Consider the parameter ϑ as a functional of the transition distribution by setting $t(Q) = \vartheta$ if $Q(\mathbf{x}, a_\vartheta) = 0$. We have

$$n^{1/2}(t(Q_{nh}) - t(Q)) \doteq n^{1/2}(\vartheta_{nu} - \vartheta) = u \quad \text{for } h \in H_u.$$

Hence, by (6), the canonical gradient is characterized as the vector $g_* \in H_*^p$ such that

$$(\pi \otimes Q)(g_* h) = u \quad \text{for } h \in H_u.$$

We show that the canonical gradient is $g_* = J^{-1}\ell$ with

$$\begin{aligned} \ell(\mathbf{x}, y) &= -Q(\mathbf{x}, \dot{a}_\vartheta^\top) Q(\mathbf{x}, a_\vartheta a_\vartheta^\top)^{-1} a_\vartheta(\mathbf{x}, y), \\ J &= (\pi \otimes Q)(\ell \ell^\top) = \pi(Q\dot{a}_\vartheta^\top Q(a_\vartheta a_\vartheta^\top)^{-1} Q\dot{a}_\vartheta). \end{aligned}$$

We have

$$Q(\mathbf{x}, a_\vartheta \ell^\top) = -Q(\mathbf{x}, \dot{a}_\vartheta).$$

Hence the j-th component ℓ_j of ℓ is in H_{e_j}, where e_j denotes the j-th p-dimensional unit vector. It follows that ℓ and hence $J^{-1}\ell$ is in H_*^p. Furthermore, for $h \in H_u$,

$$(\pi \otimes Q)(J^{-1}\ell \cdot h) = -\pi(Q\dot{a}_\vartheta^\top Q(a_\vartheta a_\vartheta^\top)^{-1} Q\dot{a}_\vartheta)^{-1}\ \pi(Q\dot{a}_\vartheta^\top Q(a_\vartheta a_\vartheta^\top)^{-1} Q(a_\vartheta h)) = u.$$

This completes the proof that $J^{-1}\ell$ is the canonical gradient of ϑ. Using the above characterization of efficient estimators, we arrive at the following result.

Characterization. *The canonical gradient of ϑ is $g_* = J^{-1}\ell$. Hence an estimator $\hat\vartheta$ for ϑ is regular and efficient if and only if*

$$n^{1/2}(\hat\vartheta - \vartheta) = J^{-1}n^{-1/2}\sum_{i=1}^{n}\ell(\mathbf{X}_{i-1}, X_i) + o_{P_n}(1). \tag{8}$$

Its asymptotic covariance matrix is J^{-1}.

We see that ℓ and J play the roles of *score function* and *Fisher information* for ϑ.

The characterization sketched in this section is analogous to that obtained in Müller and Wefelmeyer [33] for the corresponding *regression* model, with i.i.d. observations (X_i, Y_i) meeting the conditional constraint $E(a_\vartheta(X_i, Y_i) \mid X_i) = 0$. A (different) derivation of the asymptotic variance bound J^{-1} is already sketched in Chamberlain [3], with generalizations in [4]. Reviews are Newey [34] and [35]. Similar arguments as above are used in Müller and Wefelmeyer [32] for models with i.i.d. observations X_i satisfying an unconditional constraint $Ea_\vartheta(X_i) = 0$. Estimators of the *stationary law* π in our model (1) are constructed in Schick and Wefelmeyer [38].

3 Estimating functions

The characterization (8) of efficient estimators for ϑ suggests a construction as one-step Newton–Raphson improvement of an initial, inefficient, estimator $\bar\vartheta$,

$$\hat\vartheta = \bar\vartheta + \bar J^{-1}\frac{1}{n}\sum_{i=1}^{n}\bar\ell(\mathbf{X}_{i-1}, X_i),$$

with appropriate estimators $\bar J$ and $\bar\ell$ for J and ℓ. This construction does not take advantage of the special feature of our model and is not recommended.

The conditional constraint (1) says that $a_\vartheta(\mathbf{X}_{i-1}, X_i)$ is a martingale increment. This suggests estimating ϑ by solutions $\hat\vartheta$ of martingale estimating equations

$$\sum_{i=1}^{n} W_\vartheta(\mathbf{X}_{i-1})^\top a_\vartheta(\mathbf{X}_{i-1}, X_i) = 0, \tag{9}$$

with $W_\vartheta(\mathbf{x})$ a $k \times p$-matrix of weights. The asymptotic distribution of $\hat\vartheta$ is obtained from a Taylor expansion

$$0 = \sum_{i=1}^{n} W_\vartheta(\mathbf{X}_{i-1})^\top a_\vartheta(\mathbf{X}_{i-1}, X_i) + \sum_{i=1}^{n} W_\vartheta(\mathbf{X}_{i-1})^\top \dot a_\vartheta(\mathbf{X}_{i-1}, X_i)(\hat\vartheta - \vartheta) + \dots,$$

with $\dot{a}_\vartheta(\mathbf{x}, y)$ the $k \times p$-matrix of partial derivatives of $a_\vartheta(\mathbf{x}, y)$ with respect to ϑ. If $(\pi \otimes Q)(W_\vartheta^\top \dot{a}_\vartheta)$ is invertible, we obtain the stochastic approximation

$$\begin{aligned} n^{1/2}(\hat{\vartheta} - \vartheta) &= -\Big(\frac{1}{n}\sum_{i=1}^n W_\vartheta(\mathbf{X}_{i-1})^\top \dot{a}_\vartheta(\mathbf{X}_{i-1}, X_i)\Big)^{-1} \\ &\qquad n^{-1/2}\sum_{i=1}^n W_\vartheta(\mathbf{X}_{i-1})^\top a_\vartheta(\mathbf{X}_{i-1}, X_i) + o_{P_n}(1). \end{aligned} \tag{10}$$

By ergodicity, we may replace the average in (10) by its mean $(\pi \otimes Q)(W_\vartheta^\top \dot{a}_\vartheta)$. Then $\hat{\vartheta}$ is seen to be asymptotically linear (7) with influence function

$$f(\mathbf{x}, y) = -(\pi \otimes Q)(W_\vartheta^\top \dot{a}_\vartheta)^{-1} W_\vartheta(\mathbf{x})^\top a_\vartheta(\mathbf{x}, y).$$

Hence $\hat{\vartheta}$ is asymptotically normal with covariance matrix

$$\begin{aligned} &(\pi \otimes Q)(W_\vartheta^\top \dot{a}_\vartheta)^{-1}(\pi \otimes Q)(W_\vartheta^\top a_\vartheta a_\vartheta^\top W_\vartheta)(\pi \otimes Q)(\dot{a}_\vartheta W_\vartheta^\top)^{-1} \\ &= \pi(W_\vartheta^\top Q \dot{a}_\vartheta)^{-1}\pi(W_\vartheta^\top Q(a_\vartheta a_\vartheta^\top)W_\vartheta)\pi(Q\dot{a}_\vartheta^\top W_\vartheta)^{-1}. \end{aligned} \tag{11}$$

By the Cauchy–Schwarz inequality, the optimal weights are

$$W_\vartheta(\mathbf{x}) = W_\vartheta^*(\mathbf{x}) = Q(\mathbf{x}, a_\vartheta a_\vartheta^\top)^{-1} Q(\mathbf{x}, \dot{a}_\vartheta). \tag{12}$$

For these weights, the covariance matrix (11) is

$$\pi(Q\dot{a}_\vartheta^\top Q(a_\vartheta a_\vartheta^\top)^{-1} Q\dot{a}_\vartheta)^{-1}.$$

This is the asymptotic variance bound J^{-1} obtained in Section 2.

Minimizing the matrix (11) is also suggested by the *non-asymptotic optimality criterion* of Godambe [13] and Godambe and Heyde [15].

The average in (10) may also be replaced by $\frac{1}{n}\sum_{i=1}^n W_\vartheta(\mathbf{X}_{i-1})^\top Q(\mathbf{X}_{i-1}, \dot{a}_\vartheta)$. The *asymptotic optimality criterion* of Godambe and Heyde [15] suggests minimizing the matrix

$$\begin{aligned} &\Big(\sum_{i=1}^n W_\vartheta(\mathbf{X}_{i-1})^\top Q(\mathbf{X}_{i-1}, \dot{a}_\vartheta)\Big)^{-1} \\ &\quad \sum_{i=1}^n W_\vartheta(\mathbf{X}_{i-1})^\top Q(\mathbf{X}_{i-1}, a_\vartheta a_\vartheta^\top) W_\vartheta(\mathbf{X}_{i-1}) \\ &\qquad \Big(\sum_{i=1}^n Q(\mathbf{X}_{i-1}, \dot{a}_\vartheta)^\top W_\vartheta(\mathbf{X}_{i-1})\Big)^{-1}. \end{aligned} \tag{13}$$

This leads to the same optimal weights. We refer to Heyde [20] for uses of this criterion.

The optimal weights depend, through $Q(\mathbf{X}_{i-1}, a_\vartheta a_\vartheta^\top)$ and $Q(\mathbf{X}_{i-1}, \dot{a}_\vartheta)$, on the unknown transition distribution of the Markov chain. Hence the corresponding

optimal estimating function cannot be used as it stands for estimating ϑ. We will call such an estimating function *undetermined*.

Generalized method of moments. The martingale estimating equation (9) results in an estimator that is asymptotically equivalent to the *GMM estimator* obtained from the *generalized method of moments*, the minimizer $\hat{\vartheta}$ of

$$\sum_{i=1}^{n} a_\vartheta(\mathbf{X}_{i-1}, X_i)^\top W_\vartheta(\mathbf{X}_{i-1}) M_n \sum_{i=1}^{n} W_\vartheta(\mathbf{X}_{i-1})^\top a_\vartheta(\mathbf{X}_{i-1}, X_i), \tag{14}$$

where M_n is a random $p \times p$ matrix converging to a deterministic matrix M, say. To prove the asymptotic equivalence, we write the GMM estimator as solution of an estimating equation. Taking partial derivatives with respect to ϑ, we see that $\hat{\vartheta}$ solves

$$\sum_{i=1}^{n} \dot{a}_{\hat{\vartheta}}(\mathbf{X}_{i-1}, X_i)^\top W_\vartheta(\mathbf{X}_{i-1}) M_n \sum_{i=1}^{n} W_\vartheta(\mathbf{X}_{i-1})^\top a_{\hat{\vartheta}}(\mathbf{X}_{i-1}, X_i) = 0.$$

A Taylor expansion gives

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \dot{a}_\vartheta(\mathbf{X}_{i-1}, X_i)^\top W_\vartheta(\mathbf{X}_{i-1}) M_n \sum_{i=1}^{n} W_\vartheta(\mathbf{X}_{i-1})^\top a_\vartheta(\mathbf{X}_{i-1}, X_i) \\ &\quad + \sum_{i=1}^{n} \dot{a}_\vartheta(\mathbf{X}_{i-1}, X_i)^\top W_\vartheta(\mathbf{X}_{i-1}) M_n \sum_{i=1}^{n} W_\vartheta(\mathbf{X}_{i-1})^\top \dot{a}_\vartheta(\mathbf{X}_{i-1}, X_i)(\hat{\vartheta} - \vartheta) + \dots . \end{aligned}$$

If M and $(\pi \otimes Q)(W_\vartheta^\top \dot{a}_\vartheta)$ are invertible, we obtain

$$\begin{aligned} n^{1/2}(\hat{\vartheta} - \vartheta) &= -\Big((\pi \otimes Q)(\dot{a}_\vartheta^\top W_\vartheta) \cdot M \cdot (\pi \otimes Q)(W_\vartheta^\top \dot{a}_\vartheta)\Big)^{-1} \\ &\qquad (\pi \otimes Q)(\dot{a}_\vartheta^\top W_\vartheta) \cdot M \cdot n^{-1/2} \sum_{i=1}^{n} W_\vartheta(\mathbf{X}_{i-1})^\top a_\vartheta(\mathbf{X}_{i-1}, X_i) + o_{P_n}(1) \\ &= -(\pi \otimes Q)(W_\vartheta^\top \dot{a}_\vartheta)^{-1} n^{-1/2} \sum_{i=1}^{n} W_\vartheta(\mathbf{X}_{i-1})^\top a_\vartheta(\mathbf{X}_{i-1}, X_i) + o_{P_n}(1). \end{aligned}$$

Hence the GMM estimator has the same influence function as the estimator obtained from estimating equation (9). The optimal weights are therefore again given by (12). The generalized method of moments was developed by Hansen [17] and [18]. The optimal weights for this method were first obtained by Newey [35]. For reviews see Newey and McFadden [36] and Wooldridge [43]. Note that the influence function of the GMM estimator does not involve the matrix M. Hence the random matrix M_n in (14) plays no role.

Generalized quasi-likelihood. One way of dealing with the problem of undetermined estimating functions is to specify parametric models for the conditional expectations involved in the optimal weights:

$$\Sigma_\vartheta(\mathbf{x}) = Q(\mathbf{x}, a_\vartheta a_\vartheta^\top) \quad \text{and} \quad A_\vartheta(\mathbf{x}) = Q(\mathbf{x}, \dot{a}_\vartheta).$$

This leads to the estimating equation

$$\sum_{i=1}^{n} A_\vartheta(\mathbf{X}_{i-1})^\top \Sigma_\vartheta(\mathbf{X}_{i-1})^{-1} a_\vartheta(\mathbf{X}_{i-1}, X_i) = 0. \tag{15}$$

We call the estimating function on the left (score function of the) *generalized quasi-likelihood*.

If Σ_ϑ and A_ϑ are correctly specified, we can find new estimating functions besides (9) by using, besides $a_\vartheta(\mathbf{X}_{i-1}, X_i)$, further martingale increments

$$a_\vartheta(\mathbf{X}_{i-1}, X_i) a_\vartheta(\mathbf{X}_{i-1}, X_i)^\top - \Sigma_\vartheta(\mathbf{X}_{i-1}) \quad \text{and} \quad \dot{a}_\vartheta(\mathbf{X}_{i-1}, X_i) - A_\vartheta(\mathbf{X}_{i-1}).$$

Hence the generalized quasi-likelihood is inefficient, in general. If Σ_ϑ and A_ϑ are misspecified, then the generalized quasi-likelihood still gives a consistent estimator, but is again inefficient, in general, now in model (1), since the weights will be different from the optimal ones.

We note that since $Q(\mathbf{x}, a_\vartheta a_\vartheta^\top)$ is $k \times k$ and symmetric, and $Q(\mathbf{x}, \dot{a}_\vartheta)$ is $k \times p$, the generalized quasi-likelihood requires modeling up to $\frac{1}{2}k(k+1) + kp$ functions in addition to the k components of a_ϑ.

We can summarize the above discussion in the following statement.

Dichotomy. *The estimating equation (9) with optimal weights (12) is undetermined; the generalized quasi-likelihood (15) is inefficient.*

Another, more satisfactory way of dealing with the problem of undetermined optimal weights is to replace them with estimators. It is not difficult to see that the stochastic approximation (10) remains valid if we replace the weights $W_\vartheta(\mathbf{X}_{i-1})$ by appropriate estimators. The reason is that the weights are predictable. This argument is well known. For heteroscedastic linear models $Y_{ij} = \vartheta^\top x_i + H(x_i)\varepsilon_{ij}$ and $Y_{ij} = \vartheta^\top x_i + H(\vartheta^\top x_i)\varepsilon_{ij}$ with unknown function H see Carroll [2]. For quasi-likelihood models (2) and (3) see Wefelmeyer [40] and [41]. For nonparametric regression models $Y_i = g(\vartheta^\top x_i) + v(g(\vartheta^\top x_i))^{1/2}\varepsilon_i$ with unknown function v and unknown or known function g see Chiou and Müller [6] and [7]. We arrive at the following result.

Estimated weights. *If $\hat{W}^*_\vartheta(\mathbf{x})$ is an appropriate estimator for*

$$W^*_\vartheta(\mathbf{x}) = Q(\mathbf{x}, a_\vartheta a_\vartheta^\top)^{-1} Q(\mathbf{x}, \dot{a}_\vartheta),$$

then an efficient estimator for ϑ is obtained from the estimating equation with estimated optimal weights,

$$\sum_{i=1}^{n} \hat{W}^*_\vartheta(\mathbf{x})^\top a_\vartheta(\mathbf{X}_{i-1}, X_i) = 0.$$

Müller and Wefelmeyer [33] obtain an analogous result for the corresponding regression model, with i.i.d. observations (X_i, Y_i) satisfying $E(a_\vartheta(X_i, Y_i) \mid X_i) = 0$. Let us briefly sketch two specific methods of estimating the optimal weights $W^*_\vartheta(\mathbf{x})$.

Kernel estimators and penalized empirical variance. The optimal weights $W^*_\vartheta(\mathbf{x})$ involve conditional expectations. One way of estimating them is to use kernel estimators $\hat\Sigma_\vartheta(\mathbf{x})$ and $\hat A_\vartheta(\mathbf{x})$ for $Q(\mathbf{x}, a_\vartheta a_\vartheta^\top)$ and $Q(\mathbf{x}, \dot a_\vartheta)$. Such estimators require fairly large sample sizes. A different approach is developed by Li [28] and [29], exploiting ideas of Lindsay [30]. Li considers i.i.d. observations (X_i, Y_i) with $E(Y_i \mid X_i) = \mu(\vartheta^\top X_i)$ and $E((Y_i - \mu(\vartheta^\top X_i))^2 \mid X_i) = \nu(\vartheta^\top X_i)$. For our constrained model (1), the approach consists in determining, for fixed ϑ, weights $\hat W^*_\vartheta(\mathbf{x})$ that minimize the appropriately penalized empirical version of the covariance matrix (11),

$$\Big(\frac{1}{n}\sum_{i=1}^n W_\vartheta(\mathbf{X}_{i-1})^\top \dot a_\vartheta(\mathbf{X}_{i-1}, X_i)\Big)^{-1} \Big(\frac{1}{n}\sum_{i=1}^n W_\vartheta(\mathbf{X}_{i-1})^\top a_\vartheta(\mathbf{X}_{i-1}, X_i) a_\vartheta(\mathbf{X}_{i-1}, X_i)^\top W_\vartheta(\mathbf{X}_{i-1}) + \lambda I\Big) \Big(\frac{1}{n}\sum_{i=1}^n \dot a_\vartheta(\mathbf{X}_{i-1}, X_i)^\top W_\vartheta(\mathbf{X}_{i-1})\Big)^{-1}.$$

In the following we illustrate the above remarks on optimal estimating functions with five examples.

Quasi-likelihood. Suppose the state space is real, and we have a parametric model for the conditional mean of the Markov chain,

$$E(X_i \mid \mathbf{X}_{i-1}) = r_\vartheta(\mathbf{X}_{i-1}). \tag{16}$$

This is a conditional constraint with $a_\vartheta(\mathbf{x}, y) = y - r_\vartheta(\mathbf{x})$.

A simple estimator for ϑ is the *conditional least squares estimator*, the minimizer $\hat\vartheta$ of

$$\sum_{i=1}^n (X_i - r_\vartheta(\mathbf{X}_{i-1}))^2.$$

See Klimko and Nelson [26] and Tjøstheim [39]. Taking partial derivatives with respect to ϑ, we see that $\hat\vartheta$ solves

$$\sum_{i=1}^n \dot r_\vartheta(\mathbf{X}_{i-1})^\top (X_i - r_\vartheta(\mathbf{X}_{i-1})) = 0.$$

Here $\dot r_\vartheta(\mathbf{x})$ is the *row* vector of partial derivatives with respect to ϑ.

The martingale estimating equations (9) corresponding to model (16) are

$$\sum_{i=1}^n W_\vartheta(\mathbf{X}_{i-1})^\top (X_i - r_\vartheta(\mathbf{X}_{i-1})) = 0,$$

with W_ϑ a $p \times 1$ vector of weights. Here $Q(\mathbf{x}, \dot{a}_\vartheta) = -\dot{r}_\vartheta(\mathbf{x})$ does not involve the (unknown) transition distribution Q. The optimal weights (12) are

$$W^*_\vartheta(\mathbf{x}) = -\Big(\int Q(\mathbf{x}, dy)(y - r_\vartheta(\mathbf{x}))^2 \Big)^{-1} \dot{r}_\vartheta(\mathbf{x}).$$

An efficient estimator for ϑ is obtained from the estimating function

$$\sum_{i=1}^{n} \dot{r}_\vartheta(\mathbf{X}_{i-1})^\top \hat{v}_\vartheta(\mathbf{X}_{i-1})^{-1}(X_i - r_\vartheta(\mathbf{X}_{i-1})) = 0, \tag{17}$$

with $\hat{v}_\vartheta(\mathbf{x})$ an appropriate estimator of the conditional variance $\int Q(\mathbf{x}, dy)(y - r_\vartheta(\mathbf{x}))^2$; see Wefelmeyer [41]. The *quasi-likelihood estimator* replaces $\hat{v}_\vartheta(\mathbf{x})$ by a parametric model

$$v_\vartheta(\mathbf{x}) = \int Q(\mathbf{x}, dy)(y - r_\vartheta(\mathbf{x}))^2. \tag{18}$$

We have seen that it does not use the information about ϑ in the additional specification (18).

Extended quasi-likelihood. Suppose the state space is real, and we have parametric models (16) and (18) for the conditional mean and variance of the Markov chain. Then $a_\vartheta(\mathbf{x}, y) = (y - r_\vartheta(\mathbf{x}), (y - r_\vartheta(\mathbf{x}))^2 - v_\vartheta(\mathbf{x}))^\top$. Hence

$$\begin{aligned} Q(\mathbf{x}, \dot{a}_\vartheta) &= -\begin{pmatrix} \dot{r}_\vartheta(\mathbf{x}) \\ \dot{v}_\vartheta(\mathbf{x}) \end{pmatrix}, \\ Q(\mathbf{x}, a_\vartheta a_\vartheta^\top) &= \begin{pmatrix} v_\vartheta(\mathbf{x}) & \mu_3(\mathbf{x}) \\ \mu_3(\mathbf{x}) & \mu_4(\mathbf{x}) - v_\vartheta(\mathbf{x})^2 \end{pmatrix}, \end{aligned}$$

where $\mu_j(\mathbf{x}) = \int Q(\mathbf{x}, dy)(y - r_\vartheta(\mathbf{x}))^j$, $j = 3, 4$, are the third and fourth centered conditional moments of the chain. An efficient estimator for ϑ is obtained from the corresponding estimating equation with estimated optimal weights; see Wefelmeyer [40]. It requires estimators for $\mu_3(\mathbf{x})$ and $\mu_4(\mathbf{x})$. The *extended quasi-likelihood estimator* replaces these moments by parametric models; again it does not use the information about ϑ in the additional specifications. For the extended quasi-likelihood estimator in the case when $\mu_3(\mathbf{x}) = 0$, see Crowder [8] and [9], Godambe [14], and Godambe and Thompson [16]; for the general case see Heyde [19] and [20].

Nonlinear autoregression. A submodel of the Markov chain model with parametric specification (16) of the conditional mean is the *nonlinear d-order autoregressive model*

$$X_i = r_\vartheta(\mathbf{X}_{i-1}) + \varepsilon_i,$$

where the innovations are i.i.d. with density f having mean 0 and variance σ^2, say. Then $Q(\mathbf{x}, dy) = f(y - r_\vartheta(\mathbf{x}))dy$. The conditional variance $\int Q(\mathbf{x}, dy)(y - r_\vartheta(\mathbf{x}))^2$

reduces to σ^2, and the optimal estimating equation (17) simplifies to the equation defining the conditional least squares estimator,

$$\sum_{i=1}^{n} \dot{r}_\vartheta(\mathbf{X}_{i-1})^\top (X_i - r_\vartheta(\mathbf{X}_{i-1})) = 0.$$

This estimating equation does not require estimators for the weights. It is not efficient because it does not use the information that the innovations are i.i.d. Efficient estimators for ϑ are constructed in Hwang and Basawa [24], Jeganathan [25], Drost, Klaassen and Werker [11], and Koul and Schick [27].

Nonlinear and heteroscedastic autoregression. A submodel of the quasi-likelihood model (16) and (18) is the *nonlinear and heteroscedastic d-order autoregressive model*

$$X_i = r_\vartheta(\mathbf{X}_{i-1}) + v_\vartheta(\mathbf{X}_{i-1})^{1/2}\varepsilon_i,$$

where the innovations are i.i.d. with density f having mean 0 and variance 1. Then

$$\begin{aligned} Q(\mathbf{x}, dy) &= \frac{1}{v_\vartheta(\mathbf{x})^{1/2}} f\Big(\frac{y - r_\vartheta(\mathbf{x})}{v_\vartheta(\mathbf{x})^{1/2}}\Big) dy, \\ Q(\mathbf{x}, a_\vartheta a_\vartheta^\top) &= \begin{pmatrix} v_\vartheta(\mathbf{x}) & v_\vartheta(\mathbf{x})^{3/2}\mu_3 \\ v_\vartheta(\mathbf{x})^{3/2}\mu_3 & v_\vartheta(\mathbf{x})^2(\mu_4 - 1) \end{pmatrix}, \end{aligned}$$

where μ_3 and μ_4 are the third and fourth (centered) moments of the innovation distribution. The optimal weights are therefore easy to estimate: simply replace μ_j by the empirical estimator

$$\hat{\mu}_{j\vartheta} = \frac{1}{n}\sum_{i=1}^{n} (X_i - r_\vartheta(\mathbf{X}_{i-1}))^j, \quad j = 3, 4.$$

Then the estimating equation with estimated optimal weights is

$$\begin{aligned} \sum_{i=1}^{n} (\dot{r}_\vartheta(\mathbf{X}_{i-1})^\top, \dot{v}_\vartheta(\mathbf{X}_{i-1})^\top) \begin{pmatrix} v_\vartheta(\mathbf{X}_{i-1}) & v_\vartheta(\mathbf{X}_{i-1})^{3/2}\hat{\mu}_{3\vartheta} \\ v_\vartheta(\mathbf{X}_{i-1})^{3/2}\hat{\mu}_{3\vartheta} & v_\vartheta(\mathbf{X}_{i-1})^2(\hat{\mu}_{4\vartheta} - 1) \end{pmatrix}^{-1} \\ \begin{pmatrix} X_i - r_\vartheta(\mathbf{X}_{i-1}) \\ (X_i - r_\vartheta(\mathbf{X}_{i-1}))^2 - v_\vartheta(\mathbf{X}_{i-1}) \end{pmatrix} = 0. \end{aligned} \tag{19}$$

Again this estimator is not efficient. See Drost, Klaassen and Werker [11] for efficient estimators of ϑ.

ARCH. A special case of the heteroscedastic d-order autoregressive model is the *ARCH(d) model*

$$X_i = v_\vartheta(\mathbf{X}_{i-1})^{1/2}\varepsilon_i \quad \text{with} \quad v_\vartheta(\mathbf{x}) = \vartheta_0 + \sum_{j=1}^{d} \vartheta_j x_j^2,$$

with $(d+1)$-dimensional parameter $\vartheta = (\vartheta_0, \ldots, \vartheta_d)$. The innovations are again assumed i.i.d. with mean 0 and variance 1. It is convenient to introduce $\mathbf{Y}_{i-1} = (1, X_{i-1}^2, \ldots, X_{i-d}^2)$. Then $v_\vartheta(\mathbf{X}_{i-1}) = \vartheta^\top \mathbf{Y}_{i-1}$. The optimal estimating equation (19) reduces to

$$\sum_{i=1}^{n} (\vartheta^\top \mathbf{Y}_{i-1})^{-2} \mathbf{Y}_{i-1} (X_i^2 - \vartheta^\top \mathbf{Y}_{i-1}) = 0.$$

Since the weights $(\vartheta^\top \mathbf{Y}_{i-1})^{-2}$ depend on ϑ, we cannot solve the equation explicitly. However, as seen above, we may replace the weights by estimators without changing the influence function of the solution of the estimating equation. A simple estimator for ϑ is the conditional least squares estimator

$$\overline{\vartheta} = \Big(\sum_{i=1}^{n} \mathbf{Y}_{i-1} \mathbf{Y}_{i-1}^\top \Big)^{-1} \sum_{i=1}^{n} X_{i-1}^2 \mathbf{Y}_{i-1}.$$

The solution of the estimating equation with estimated optimal weights is

$$\hat{\vartheta} = \Big(\sum_{i=1}^{n} (\overline{\vartheta}^\top \mathbf{Y}_{i-1})^{-2} \mathbf{Y}_{i-1} \mathbf{Y}_{i-1}^\top \Big)^{-1} \sum_{i=1}^{n} (\overline{\vartheta}^\top \mathbf{Y}_{i-1})^{-2} X_{i-1}^2 \mathbf{Y}_{i-1}.$$

For a direct derivation see Chandra and Taniguchi [5]. The estimator is not efficient. For efficient estimators see Engle and Gonzáles-Rivera [12], Linton [31], and Drost and Klaassen [10].

Acknowledgment. We are most grateful to the referee for several useful suggestions that have made the paper more readable.

References

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

[2] Carroll, R. J. (1982), Adapting for heteroscedasticity in linear models, Ann. Statist. 10, 1224–1233.

[3] Chamberlain, G. (1987), Asymptotic efficiency in estimation with conditional moment restrictions, J. Econometrics 34, 305–334.

[4] Chamberlain, G. (1992), Efficiency bounds for semiparametric regression, Econometrica 60, 567–596.

[5] Chandra, S. A. and Taniguchi, M. (2001), Estimating functions for nonlinear time series models, Ann. Inst. Statist. Math. 53, 125–141.

[6] Chiou, J.-M. and Müller, H.-G. (1998), Quasi-likelihood regression with unknown link and variance functions, J. Amer. Statist. Assoc. 93, 1376–1387.

[7] Chiou, J.-M. and Müller, H.-G. (1999), Nonparametric quasi-likelihood, Ann. Statist. 27, 36–64.

[8] Crowder, M. (1986), On consistency and inconsistency of estimating equations, Econometric Theory 2, 305–330.

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

[10] Drost, F. C. and Klaassen, C. A. (1997), Efficient estimation in semiparametric GARCH models, J. Econometrics 81, 193–221.

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

[12] Engle, R. F. and González-Rivera, G. (1991), Semiparametric ARCH models, J. Business Economic Statist. 9, 345–359.

[13] Godambe, V. P. (1985), The foundations of finite sample estimation in stochastic processes, Biometrika 72, 419–428.

[14] Godambe, V. P. (1987), The foundations of finite sample estimation in stochastic processes II, in: Proceedings of the 1st World Congress of the Bernoulli Society (Y. Prohorov and V. V. Sazonov, eds.), 49–54, VNU Science Press, Utrecht.

[15] Godambe, V. P. and Heyde, C. C. (1987), Quasi-likelihood and optimal estimation, Internat. Statist. Rev. 55, 231–244.

[16] Godambe, V. P. and Thompson, M. E. (1989), An extension of quasi-likelihood estimation, J. Statist. Plann. Inference 22, 137–152.

[17] Hansen, L. P. (1982), Large sample properties of generalized method of moments estimators, Econometrica 50, 1029–1054.

[18] Hansen, L. P. (1985), A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators, J. Econometrics 30, 203–238.

[19] Heyde, C. C. (1987), On combining quasi-likelihood estimating functions, Stochastic Process. Appl. 25, 281–287.

[20] Heyde, C. C. (1997), Quasi-Likelihood And Its Application, A General Approach to Optimal Parameter Estimation, Springer Series in Statistics, Springer, New York.

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

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

[23] 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

[24] 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.

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

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

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

[28] Li, B. (2000), Nonparametric estimating equations based on a penalized information criterion, Canad. J. Statist. 28, 621–639.

[29] Li, B. (2001), On quasilikelihood equations with nonparametric weights, Scand. J. Statist. 28, 577–602.

[30] Lindsay, B. G. (1985), Using empirical partially Bayes inference for increased efficiency, Ann. Statist. 13, 914–931.

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

[32] Müller, U. U. and Wefelmeyer, W. (2001a), Estimators for models with constraints involving unknown parameters, to appear in: Math. Methods Statist. http://www.math.uni-siegen.de/statistik/wefelmeyer.html.

[33] Müller, U. U. and Wefelmeyer, W. (2001b), Regression type models and optimal estimators, in preparation.

[34] Newey, W. K. (1990), Semiparametric efficiency bounds, J. Appl. Econometrics 5, 99–135.

[35] Newey, W. K. (1993), Efficient estimation of models with conditional moment restrictions, in: Handbook of Statistics 11: Econometrics (G. S. Maddala, C. R. Rao and H. D. Vinod, eds.), 419–454. Elsevier, Amsterdam.

[36] Newey, W. K. and McFadden, D. L. (1994), Large sample estimation and hypothesis testing, in: Handbook of Econometrics 4 (R. F. Engle and D. L. McFadden, eds.), 2111–2245, Elsevier, Amsterdam.

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

[38] Schick, A. and Wefelmeyer, W. (1999), Efficient estimation of invariant distributions of some semiparametric Markov chain models, Math. Methods Statist. 8, 119–134.

[39] Tjøstheim, D. (1986), Estimation in nonlinear time series models, Stochastic Process. Appl. 21, 251–273.

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

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

[42] Wefelmeyer, W. (1999), Efficient estimation in Markov chain models: an introduction, in: Asymptotics, Nonparametrics, and Time Series (S. Ghosh, ed.), 427–459, Statistics: Textbooks and Monographs 158, Dekker, New York.

[43] Wooldridge, J. M. (1994), Estimation and inference for dependent processes, in: Handbook of Econometrics 4 (R. F. Engle and D. L. McFadden, eds.), 2639–2738, Elsevier, Amsterdam.