We consider Markov chains of order $d$ that satisfy a conditional constraint of the form

$$E(a_\vartheta(X_{i-1}, X_i) \mid X_{i-1}) = 0,$$

where $X_{i-1} = (X_{i-1}, \ldots, X_{i-d})$. These comprise quasi-likelihood models and nonlinear and conditionally heteroscedastic autoregressive models with martingale innovations. Estimators for $\vartheta$ can be obtained from estimating equations

$$\sum_{i=1}^n W_\vartheta(X_{i-1}) a_\vartheta(X_{i-1}, X_i) = 0.$$  

We review different criteria for choosing good weights $W_\vartheta(X_{i-1})$. 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 $\vartheta$, and discuss efficiency of the estimators in the sense of Hájek and Le Cam. 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 $X_1, \ldots, X_n$ be observations from a homogeneous and geometrically ergodic $d$-order Markov chain on some arbitrary state space. Write $X_{i-1} = (X_{i-1}, \ldots, X_{i-d})$, and assume that the chain meets the conditional constraint

$$E(a_\vartheta(X_{i-1}, X_i) \mid X_{i-1}) = 0,$$

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

In Section 2 we derive an asymptotic lower bound for estimators of $\vartheta$ in the sense of Hájek and Le Cam, and give a characterization of efficient estimators.
In Section 3 we consider estimating equations for $\theta$ of the form

$$\sum_{i=1}^{n} W_\theta(X_{i-1})^T a_\theta(X_{i-1}, X_i) = 0,$$

with $W_\theta(x)$ 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 $\theta$. 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,

$$E(X_i | X_{i-1}) = r_\theta(X_{i-1}),$$

$$E((X_i - r_\theta(X_{i-1}))^2 | X_{i-1}) = v_\theta(X_{i-1}).$$

Then $a_\theta(x, y) = (y - r_\theta(x), (y - r_\theta(x))^2 - v_\theta(x))^T$.

Quasi-likelihood models can be written as

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

with innovations $\varepsilon_i$ that are martingale increments, $E(\varepsilon_i | X_{i-1}) = 0$, and that satisfy $E(\varepsilon_i^2 | X_{i-1}) = 1$ for identifiability. The submodel with independent innovations $\varepsilon_i$ 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 $\theta$ in the $d$-order Markov chain model constrained by (1).

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

The nonparametric model is locally asymptotically normal in the following sense. Introduce (Hellinger differentiable) perturbations \( Q_{nh}(x, dy) = Q(x, dy)(1 + n^{-1/2}h(x, y)) \), with \( h \) in the tangent space \( H = \{ h \in L_2(\pi \otimes Q) : Q(x, h) = 0 \text{ for all } x \} \).

Since \( h \) may take large negative values, we cannot simply define \( Q_{nh} \) replacing \( = \) 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 \( Q(x, h) = 0 \) is required for \( Q_{nh} \) to be a transition distribution. Write \( P_{nh} \) and \( P_n \) for the joint law of \( X_1, \ldots, X_n \) under \( Q_{nh} \) and \( Q \), respectively. The log-likelihood ratio has the stochastic expansion

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

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

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

\[
0 = Q_{nh}(x, a_{n\vartheta}) \approx Q(x, a_\vartheta) + n^{-1/2}(Q(x, a_\vartheta h) + Q(x, a_\vartheta)u) = Q(x, a_\vartheta) + n^{-1/2}(Q(x, a_\vartheta h) + Q(x, a_\vartheta)u).
\]

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

\[
H_* = \{ h \in H : Q(x, a_\vartheta h) = -Q(x, a_\vartheta)u \text{ for all } x \}.
\]

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; for Markov chains see also
Wefelmeyer. A $p$-dimensional functional $t(Q)$ is called differentiable at $Q$ with gradient $g$ if $g \in H^p$ and

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

$$n^{1/2}(\hat{t} - t(Q)) = n^{-1/2} \sum_{i=1}^n f(X_{i-1}, X_i) + o_P(1). \hspace{1cm} (7)$$

Such an estimator is asymptotically normal with covariance matrix $(\pi \otimes Q)(ff\top)$. 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 $\vartheta$. Consider the parameter $\vartheta$ as a functional of the transition distribution by setting $t(Q) = \vartheta$ if $Q(x, a_\vartheta) = 0$. We have

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

Hence, by (6), the canonical gradient is characterized as the vector $g_* \in H^p_u$ 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
\[
\ell(x, y) = -Q(x, \dot{a})Q(x, a\dot{a})^{-1}a(x, y),
\]
\[
J = (\pi \otimes Q)(\ell^\top) = \pi(Q\dot{a}Q(a\dot{a})^{-1}Q\dot{a}).
\]
We have
\[
Q(x, a\ell^\top) = -Q(x, \dot{a}).
\]
Hence the \( j \)-th component \( \ell_j \) of \( \ell \) is in \( H_{e_j} \), where \( e_j \) denotes the \( j \)-th \( p \)-dimensional unit vector. It follows that \( \ell \) and hence \( J^{-1} \ell \) is in \( H_p^* \). Furthermore, for \( h \in H_u \),
\[
(\pi \otimes Q)(J^{-1} \ell \cdot h) = -\pi(Q\dot{a}Q(a\dot{a})^{-1}Q\dot{a})^{-1} \pi(Q\dot{a}Q(a\dot{a})^{-1}Q(a\dot{h})) = u.
\]
This completes the proof that \( J^{-1} \ell \) is the canonical gradient of \( \vartheta \). Using the above characterization of efficient estimators, we arrive at the following result.

**Characterization.** The canonical gradient of \( \vartheta \) is \( g^* = J^{-1} \ell \). Hence an estimator \( \hat{\vartheta} \) for \( \vartheta \) is regular and efficient if and only if
\[
n^{1/2}(\hat{\vartheta} - \vartheta) = J^{-1}n^{-1/2} \sum_{i=1}^n \ell(X_{i-1}, X_i) + o_{P_n}(1).
\]
Its asymptotic covariance matrix is \( J^{-1} \).

We see that \( \ell \) and \( J \) play the roles of score function and Fisher information for \( \vartheta \).

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\(^{34,35} \). 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 \( \pi \) in our model (1) are constructed in Schick and Wefelmeyer\(^{38} \).

### 3 Estimating functions

The characterization (8) of efficient estimators for \( \vartheta \) suggests to construct an efficient estimator as a one-step Newton–Raphson improvement of an initial,
inefficient, estimator \( \hat{\vartheta} \),

\[
\hat{\vartheta} = \vartheta + J^{-1} \frac{1}{n} \sum_{i=1}^{n} \ell(X_{i-1}, X_i),
\]

with appropriate estimators \( J \) and \( \ell \) for \( J \) and \( \ell \). This construction does however not take advantage of the special feature of our model and is not recommended.

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

\[
\sum_{i=1}^{n} W_\vartheta(X_{i-1})^T a_\vartheta(X_{i-1}, X_i) = 0,
\]

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

\[
0 \approx \frac{1}{n} \sum_{i=1}^{n} W_\vartheta(X_{i-1})^T a_\vartheta(X_{i-1}, X_i) + \sum_{i=1}^{n} W_\vartheta(X_{i-1})^T \dot{a}_\vartheta(X_{i-1}, X_i)(\hat{\vartheta} - \vartheta),
\]

with \( \dot{a}_\vartheta(x, y) \) the \( k \times p \)-matrix of partial derivatives of \( a_\vartheta(x, y) \) with respect to \( \vartheta \). Here we have used that \( W_\vartheta(X_{i-1})^T \dot{a}_\vartheta(X_{i-1}, X_i) \) is a martingale increment and therefore negligible. If \( (\pi \otimes Q)(W_\vartheta^T \dot{a}_\vartheta) \) is invertible, we obtain the stochastic approximation

\[
n^{1/2}(\hat{\vartheta} - \vartheta) = -\left( \frac{1}{n} \sum_{i=1}^{n} W_\vartheta(X_{i-1})^T \dot{a}_\vartheta(X_{i-1}, X_i) \right)^{-1} \sum_{i=1}^{n} W_\vartheta(X_{i-1})^T a_\vartheta(X_{i-1}, X_i) + o_p(1). \tag{10}
\]

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

\[
f(x, y) = -(\pi \otimes Q)(W_\vartheta^T \dot{a}_\vartheta)^{-1}W_\vartheta(x)^T a_\vartheta(x, y).
\]

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

\[
(\pi \otimes Q)(W_\vartheta^T \dot{a}_\vartheta)^{-1}(\pi \otimes Q)(W_\vartheta^T a_\vartheta a_\vartheta^T W_\vartheta)(\pi \otimes Q)(\dot{a}_\vartheta W_\vartheta^T)^{-1} = (\pi W_\vartheta^T Q \dot{a}_\vartheta)^{-1}(\pi(W_\vartheta^T Q a_\vartheta a_\vartheta^T W_\vartheta)(\pi(Q \dot{a}_\vartheta W_\vartheta)^{-1}.
\]

\[
(\pi \otimes Q)(W_\vartheta^T \dot{a}_\vartheta)^{-1}(\pi \otimes Q)(W_\vartheta^T a_\vartheta a_\vartheta^T W_\vartheta)(\pi \otimes Q)(\dot{a}_\vartheta W_\vartheta^T)^{-1} = (\pi W_\vartheta^T Q \dot{a}_\vartheta)^{-1}(\pi(W_\vartheta^T Q a_\vartheta a_\vartheta^T W_\vartheta)(\pi(Q \dot{a}_\vartheta W_\vartheta)^{-1}.
\]

\[
(\pi \otimes Q)(W_\vartheta^T \dot{a}_\vartheta)^{-1}(\pi \otimes Q)(W_\vartheta^T a_\vartheta a_\vartheta^T W_\vartheta)(\pi \otimes Q)(\dot{a}_\vartheta W_\vartheta^T)^{-1} = (\pi W_\vartheta^T Q \dot{a}_\vartheta)^{-1}(\pi(W_\vartheta^T Q a_\vartheta a_\vartheta^T W_\vartheta)(\pi(Q \dot{a}_\vartheta W_\vartheta)^{-1}.
\]
By the Cauchy–Schwarz inequality, the optimal weights are
\[ W_\vartheta(x) = W_\vartheta^*(x) = Q(x, a_\vartheta a_\vartheta^\top)^{-1}Q(x, \hat{a}_\vartheta). \] (12)

For these weights, the covariance matrix (11) is
\[ \pi(Q\hat{a}_\vartheta \hat{a}_\vartheta^\top)^{-1}Q(x, \hat{a}_\vartheta)^{-1}. \]

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 \( \frac{1}{n} \sum_{i=1}^{n} W_\vartheta(x_{i-1})^\top a_\vartheta(x_{i-1}, x_i) \) in (10) may also be replaced by \( \frac{1}{n} \sum_{i=1}^{n} W_\vartheta(x_{i-1})^\top Q(x_{i-1}, \hat{a}_\vartheta) \). The asymptotic optimality criterion of Godambe and Heyde\(^{15}\) suggests minimizing the matrix
\[
\left( \sum_{i=1}^{n} W_\vartheta(x_{i-1})^\top Q(x_{i-1}, \hat{a}_\vartheta) \right)^{-1} \sum_{i=1}^{n} W_\vartheta(x_{i-1})^\top Q(x_{i-1}, a_\vartheta a_\vartheta^\top) W_\vartheta(x_{i-1}) \left( \sum_{i=1}^{n} Q(x_{i-1}, \hat{a}_\vartheta)^\top W_\vartheta(x_{i-1}) \right)^{-1}.
\] (13)

This leads to the same optimal weights. We refer to Heyde\(^{20}\) for uses of this criterion.

The optimal weights depend, through \( Q(x_{i-1}, a_\vartheta a_\vartheta^\top) \) and \( Q(x_{i-1}, \hat{a}_\vartheta) \), on the unknown transition distribution of the Markov chain. Hence the corresponding optimal estimating function cannot be used as it stands for estimating \( \vartheta \). 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
\[
\sum_{i=1}^{n} a_\vartheta(x_{i-1}, x_i)^\top W_\vartheta(x_{i-1}) M_n \sum_{i=1}^{n} W_\vartheta(x_{i-1})^\top a_\vartheta(x_{i-1}, x_i),
\] (14)

where \( M_n \) is a random symmetric \( p \times p \) 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 \( \hat{\vartheta} \), we see that \( \hat{\vartheta} \) solves
\[
\sum_{i=1}^{n} \dot{a}_{\vartheta}(X_{i-1}, X_{i})^T \hat{W}_{\vartheta}(X_{i-1}) M_{n} \sum_{i=1}^{n} \hat{W}_{\vartheta}(X_{i-1})^T a_{\vartheta}(X_{i-1}, X_{i}) \\
+ \sum_{i=1}^{n} a_{\hat{\vartheta}}(X_{i-1}, X_{i})^T \hat{W}_{\hat{\vartheta}}(X_{i-1}) M_{n} \sum_{i=1}^{n} \hat{W}_{\hat{\vartheta}}(X_{i-1})^T a_{\hat{\vartheta}}(X_{i-1}, X_{i}) = 0.
\]
Again, the term involving \( \dot{\hat{W}}_{\hat{\vartheta}} \) is negligible because \( \hat{W}_{\vartheta}(X_{i-1})^T a_{\vartheta}(X_{i-1}, X_{i}) \)
is a martingale increment. Using this argument repeatedly, we obtain by a Taylor expansion,
\[
0 = n \sum_{i=1}^{n} \dot{a}_{\vartheta}(X_{i-1}, X_{i})^T \hat{W}_{\vartheta}(X_{i-1}) M_{n} \sum_{i=1}^{n} \hat{W}_{\vartheta}(X_{i-1})^T a_{\vartheta}(X_{i-1}, X_{i}) \\
+ n \sum_{i=1}^{n} a_{\hat{\vartheta}}(X_{i-1}, X_{i})^T \hat{W}_{\hat{\vartheta}}(X_{i-1}) M_{n} \sum_{i=1}^{n} \hat{W}_{\hat{\vartheta}}(X_{i-1})^T a_{\hat{\vartheta}}(X_{i-1}, X_{i}).
\]
If \( M \) and \((\pi \otimes Q)(W_{\vartheta}^T \dot{a}_{\vartheta})\) are invertible, we obtain
\[
n^{1/2}(\hat{\vartheta} - \vartheta) = - \left( (\pi \otimes Q)(\dot{a}_{\vartheta}^T W_{\vartheta}) \cdot M \cdot (\pi \otimes Q)(W_{\vartheta}^T \dot{a}_{\vartheta}) \right)^{-1} \\
(\pi \otimes Q)(\dot{a}_{\vartheta}^T W_{\vartheta}) \cdot M \cdot n^{-1/2} \sum_{i=1}^{n} W_{\vartheta}(X_{i-1})^T a_{\vartheta}(X_{i-1}, X_{i}) \\
+ o_{P_{n}}(1) \\
= - (\pi \otimes Q)(W_{\vartheta}^T \dot{a}_{\vartheta})^{-1} n^{-1/2} \sum_{i=1}^{n} W_{\vartheta}(X_{i-1})^T a_{\vartheta}(X_{i-1}, X_{i}) \\
+ o_{P_{n}}(1).
\]
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\(^{17,18}\). 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:
\[
\Sigma_{\vartheta}(x) = Q(x, a_{\vartheta} a_{\vartheta}^T) \quad \text{and} \quad A_{\vartheta}(x) = Q(x, \dot{a}_{\vartheta}).
\]
This leads to the estimating equation
\[ n \sum_{i=1}^{n} A_\varphi(X_{i-1})^\top \Sigma_\varphi(X_{i-1})^{-1} a_\varphi(X_{i-1}, X_i) = 0. \] (15)

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

If \( \Sigma_\varphi \) and \( A_\varphi \) are correctly specified, this amounts to an additional restriction on the model. In this case, we can find new estimating functions besides (9) by using, in addition to \( a_\varphi(X_{i-1}, X_i) \), further martingale increments
\[ a_\varphi(X_{i-1}, X_i) a_\varphi(X_{i-1}, X_i)^\top - \Sigma_\varphi(X_{i-1}) \quad \text{and} \quad \dot{a}_\varphi(X_{i-1}, X_i) - A_\varphi(X_{i-1}). \]

Hence the generalized quasi-likelihood is inefficient, in general.

If \( \Sigma_\varphi \) and \( A_\varphi \) 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 \( Q(x, a_\varphi a_\varphi^\top) \) is \( k \times k \) and symmetric, and \( Q(x, \dot{a}_\varphi) \) 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 \( a_\varphi \).

We can summarize the above discussion in the following statement.

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

Another, more satisfactory way of dealing with the problem of underdetermined 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 \( W_\varphi(X_{i-1}) \) by appropriate estimators \( \hat{W}_\varphi(X_{i-1}) \). The reason is that the weights are predictable. This argument is well known: \( n^{-1/2} \sum_{i=1}^{n} (\hat{W}_\varphi(X_{i-1}) - W_\varphi(X_{i-1})) a_\varphi(X_{i-1}, X_i) \) is (approximately) conditionally centered with negligible terms, and therefore negligible. For heteroscedastic linear models \( Y_{ij} = \varphi^\top x_i + H(x_i) \xi_{ij} \) and \( Y_{ij} = \varphi^\top x_i + H(\varphi^\top x_i) \xi_{ij} \) with unknown function \( H \) see Carroll\textsuperscript{2}. For quasi-likelihood models (2) and (3) see Wefelmeyer\textsuperscript{40,41}. For nonparametric regression models \( Y_i = g(\varphi^\top x_i) + v(g(\varphi^\top x_i))^{1/2} \xi_i \) with unknown function \( v \) and unknown or known function \( g \) see Chiou and Müller\textsuperscript{6,7}. We arrive at the following result.

Estimated weights. If \( \hat{W}_\varphi^*(x) \) is an appropriate consistent estimator (possibly depending on \( \varphi \)) for
\[ \hat{W}_\varphi^*(x) = Q(x, a_\varphi a_\varphi^\top)^{-1} Q(x, \dot{a}_\varphi), \]
then an efficient estimator for $\vartheta$ is obtained from the estimating equation with estimated optimal weights,

$$
\sum_{i=1}^{n} W^*_\vartheta(X_{i-1})^T a_\vartheta(X_{i-1}, X_i) = 0.
$$

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

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

$$
\left( \frac{1}{n} \sum_{i=1}^{n} W_\vartheta(X_{i-1})^T \hat{a}_\vartheta(X_{i-1}, X_i) \right)^{-1}
\left( \frac{1}{n} \sum_{i=1}^{n} W_\vartheta(X_{i-1})^T a_\vartheta(X_{i-1}, X_i)^T \hat{W}_\vartheta(X_{i-1}) + \lambda \mathbf{I} \right)^{-1}
\left( \frac{1}{n} \sum_{i=1}^{n} \hat{a}_\vartheta(X_{i-1}, X_i)^T W_\vartheta(X_{i-1}) \right)^{-1}.
$$

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,

$$
E(X_i | X_{i-1}) = r_\vartheta(X_{i-1}).
$$

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

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

$$
\sum_{i=1}^{n} (X_i - r_\vartheta(X_{i-1}))^2.
$$
See Klimko and Nelson and Tjøstheim. Taking partial derivatives with respect to $\vartheta$, we see that $\hat{\vartheta}$ solves
\[
\sum_{i=1}^{n} \dot{r}_\vartheta(X_{i-1})^\top (X_i - r_\vartheta(X_{i-1})) = 0.
\]
Here $\dot{r}_\vartheta(x)$ is the row vector of partial derivatives with respect to $\vartheta$.

The martingale estimating equations (9) corresponding to model (16) are
\[
\sum_{i=1}^{n} W_\vartheta(X_{i-1})^\top (X_i - r_\vartheta(X_{i-1})) = 0,
\]
with $W_\vartheta$ a $p \times 1$ vector of weights. Here $Q(x, \dot{a}_\vartheta) = -\dot{r}_\vartheta(x)$ does not involve the (unknown) transition distribution $Q$. The optimal weights (12) are
\[
W^*_\vartheta(x) = -\left( \int Q(x, dy)(y - r_\vartheta(x))^2 \right)^{-1} \dot{r}_\vartheta(x).
\]
An efficient estimator for $\vartheta$ is obtained from the estimating function
\[
\sum_{i=1}^{n} \hat{r}_\vartheta(X_{i-1})^\top \hat{v}_\vartheta(X_{i-1})^{-1} (X_i - r_\vartheta(X_{i-1})) = 0,
\]
with $\hat{v}_\vartheta(x)$ an appropriate estimator of the conditional variance $\int Q(x, dy)(y - r_\vartheta(x))^2$; see Wefelmeyer. The quasi-likelihood estimator replaces $\hat{v}_\vartheta(x)$ by a parametric model
\[
v_\vartheta(x) = \int Q(x, dy)(y - r_\vartheta(x))^2.
\]

The discussion of estimating equation (15) has shown that the quasi-likelihood estimator does not use the information about $\vartheta$ 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 $a_\vartheta(x, y) = (y - r_\vartheta(x), (y - r_\vartheta(x))^2 - v_\vartheta(x))^\top$. Hence
\[
Q(x, \dot{a}_\vartheta) = -\begin{pmatrix} \dot{r}_\vartheta(x) \\ \hat{v}_\vartheta(x) \end{pmatrix},
\]
\[
Q(x, a_\vartheta a_\vartheta^\top) = \begin{pmatrix} v_\vartheta(x) \\ \mu_3(x) \\ \mu_3(x) \mu_4(x) - v_\vartheta(x)^2 \end{pmatrix},
\]
where $\mu_j(x) = \int Q(x, dy)(y - r_\vartheta(x))^j$, $j = 3, 4$, are the third and fourth centered conditional moments of the chain. An efficient estimator for $\vartheta$ is
obtained from the corresponding estimating equation with estimated optimal weights; see Wefelmeyer\textsuperscript{40}. It requires estimators for $\mu_3(x)$ and $\mu_4(x)$. The extended quasi-likelihood estimator replaces these moments by parametric models; again it does not use the information about $\vartheta$ in the additional specifications. For the extended quasi-likelihood estimator in the case when $\mu_3(x) = 0$, see Crowder\textsuperscript{8,9}, Godambe\textsuperscript{14}, and Godambe and Thompson\textsuperscript{16}; for the general case see Heyde\textsuperscript{19,20}.

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

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

where the innovations are i.i.d. with density $f$ having mean 0 and variance $\sigma^2$, say. Then $Q(x, dy) = f(y - r_\vartheta(x))dy$. The conditional variance $\int Q(x, dy)(y - r_\vartheta(x))^2$ reduces to $\sigma^2$, and the optimal estimating equation (17) simplifies to the equation defining the conditional least squares estimator,

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

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 $\vartheta$ are constructed in Hwang and Basawa\textsuperscript{24}, Jeganathan\textsuperscript{25}, Drost, Klaassen and Werker\textsuperscript{11}, and Koul and Schick\textsuperscript{27}.

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

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

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

$$Q(x, dy) = \frac{1}{v_\vartheta(x)^{1/2}} f \left( \frac{y - r_\vartheta(x)}{v_\vartheta(x)^{1/2}} \right) dy,$$

$$Q(x, a_\vartheta a_\vartheta^\top) = \begin{pmatrix} v_\vartheta(x) & v_\vartheta(x)^{3/2} \mu_3 \\ v_\vartheta(x)^{3/2} \mu_3 & v_\vartheta(x)^2 (\mu_4 - 1) \end{pmatrix},$$

where $\mu_3$ and $\mu_4$ are the third and fourth (centered) moments of the innovation distribution. The optimal weights are therefore easy to estimate: simply
replace $\mu_j$ by the empirical estimator
\[
\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} (X_i - r_\theta(X_{i-1}))^j , \quad j = 3, 4.
\]
Then the estimating equation with estimated optimal weights is
\[
\sum_{i=1}^{n} (\hat{r}_\theta(X_{i-1})^\top, \hat{v}_\theta(X_{i-1})^\top) \left( \begin{array}{cc}
v_\theta(X_{i-1}) & v_\theta(X_{i-1})^{3/2} \hat{\mu}_{3\theta} \\
v_\theta(X_{i-1})^{3/2} \hat{\mu}_{3\theta} & v_\theta(X_{i-1})^2 (\hat{\mu}_{4\theta} - 1) \end{array} \right)^{-1} 
\begin{array}{c}
X_i - r_\theta(X_{i-1}) \\
(X_i - r_\theta(X_{i-1}))^2 - v_\theta(X_{i-1})
\end{array} = 0.
\]
(19)

Again this estimator is not efficient. See Drost, Klaassen and Werker\textsuperscript{11} for efficient estimators of $\vartheta$.

**ARCH.** A special case of the heteroscedastic d-order autoregressive model is the ARCH(d) model
\[
X_i = v_\theta(X_{i-1})^{1/2} \varepsilon_i \quad \text{with} \quad v_\theta(x) = \vartheta_0 + \sum_{j=1}^{d} \vartheta_j x_j^2,
\]
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 $Y_{i-1} = (1, X_{i-1}^2, \ldots, X_{i-1}^d)$. Then $v_\theta(X_{i-1}) = \vartheta^\top Y_{i-1}$. The optimal estimating equation (19) reduces to
\[
\sum_{i=1}^{n} (\vartheta^\top Y_{i-1})^{-2} Y_{i-1} (X_i^2 - \vartheta^\top Y_{i-1}) = 0.
\]
Since the weights $(\vartheta^\top Y_{i-1})^{-2}$ depend on $\vartheta$, 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 $\vartheta$ is the conditional least squares estimator
\[
\hat{\vartheta} = \left( \sum_{i=1}^{n} Y_{i-1} Y_{i-1}^\top \right)^{-1} \sum_{i=1}^{n} X_{i-1}^2 Y_{i-1}.
\]
The solution of the estimating equation with estimated optimal weights is
\[
\hat{\vartheta} = \left( \sum_{i=1}^{n} (\hat{\vartheta}^\top Y_{i-1})^{-2} Y_{i-1} Y_{i-1}^\top \right)^{-1} \sum_{i=1}^{n} (\hat{\vartheta}^\top Y_{i-1})^{-2} X_{i-1}^2 Y_{i-1}.
\]
For a direct derivation see Chandra and Taniguchi\textsuperscript{5}. The estimator is not efficient. For efficient estimators see Engle and González-Rivera\textsuperscript{12}, Linton\textsuperscript{31}, and Drost and Klaassen\textsuperscript{10}.

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

References