

Efficient estimators for alternating quasi-likelihood models

Ursula U. Müller
Texas A&M University

Anton Schick
Binghamton University

Wolfgang Wefelmeyer
Universität zu Köln

Abstract

We consider time series described by Markov chains that alternate periodically between different transition distributions, with conditional constraints involving unknown parameters. We obtain variance bounds and characterize efficient estimators for these parameters. Efficient estimators can be obtained as solutions of randomly weighted martingale estimating equations. Our model includes alternating heteroskedastic nonlinear autoregressive models whose innovations are martingale increments, in other words, alternating quasi-likelihood models. We consider in particular submodels of these in which the transition distributions do not alternate except for the conditional means and variances, and show that this information leads to better estimators for the parameters.

Keywords. Variance bound, efficient influence function, weighted least squares estimator, Newton–Raphson improvement.

1 Introduction

Consider a time series that evolves in a periodically changing environment, with observations on a smaller time scale. This can be modeled by a periodic change in the transition distribution. Müller, Schick and Wefelmeyer (2007, 2009) consider homoskedastic and heteroskedastic alternating linear and nonlinear autoregression models with *independent* innovations. The assumption of independent innovations is not always realistic, and typically not made in the econometric literature. This is why we now consider, instead, alternating Markov chains with conditional constraints, including alternating quasi-likelihood models and alternating (homoskedastic) nonlinear and linear regression models with innovations depending on the past.

Specifically, let X_i , $i \in \mathbb{Z}$, be a Markov chain of order p on an arbitrary state space. Assume that the chain alternates periodically between m possibly different transition distributions with possibly different conditional constraints. Here p is allowed to be larger than the length m of the period. At time $jm + k$ with $j \in \mathbb{Z}$ and $k = 1, \dots, m$, the transition distribution from $\mathbf{X}_{jm+k-1} = \mathbf{x} = (x_1, \dots, x_p)$ to $X_{jm+k} = y$ is $Q_k(\mathbf{x}, dy)$, with conditional

constraints

$$(1.1) \quad Q_k(\mathbf{x}, a_{k\vartheta}) = E(a_{k\vartheta}(\mathbf{X}_{jm+k-1}, X_{jm+k}) | \mathbf{X}_{jm+k-1} = \mathbf{x}) = 0,$$

where $\mathbf{X}_i = (X_{i-p+1}, \dots, X_i)$ and $a_{k\vartheta}(\mathbf{x}, y)$ is a known q -dimensional vector of functions involving an unknown d -dimensional parameter ϑ . We assume that we have initial observations X_{-p+1}, \dots, X_0 and then observe n periods X_1, \dots, X_{nm} .

In Section 2 we characterize efficient estimators of general differentiable vector-valued functionals of (Q_1, \dots, Q_k) . In Section 3 we describe an efficient estimator for ϑ as the solution of a randomly weighted martingale estimating equation. The weights involve estimators of conditional moments of the time series. There is a large literature on such estimators, and we will be brief here.

A special case of model (1.1) are alternating heteroskedastic nonlinear autoregressive models whose innovations are martingale increments, which we may also call alternating quasi-likelihood models. Here the state space is the real line, and we have parametric models for the conditional means and variances,

$$(1.2) \quad E(X_{jm+k} | \mathbf{X}_{jm+k-1}) = r_{k\vartheta}(\mathbf{X}_{jm+k-1}),$$

$$(1.3) \quad E((X_{jm+k} - r_{k\vartheta}(\mathbf{X}_{jm+k-1}))^2 | \mathbf{X}_{jm+k-1}) = s_{k\vartheta}^2(\mathbf{X}_{jm+k-1}).$$

Then (1.1) holds with

$$(1.4) \quad a_{k\vartheta}(\mathbf{x}, y) = \begin{pmatrix} y - r_{k\vartheta}(\mathbf{x}) \\ (y - r_{k\vartheta}(\mathbf{x}))^2 - s_{k\vartheta}^2(\mathbf{x}) \end{pmatrix}.$$

This alternating quasi-likelihood model can also be written as the alternating nonlinear and heteroskedastic autoregressive model

$$X_{jm+k} = r_{k\vartheta}(\mathbf{X}_{jm+k-1}) + s_{k\vartheta}(\mathbf{X}_{jm+k-1})\varepsilon_{jm+k}, \quad j \in \mathbb{Z}, \quad k = 1, \dots, m,$$

with innovations ε_{jm+k} that are martingale increments with conditional distribution of ε_{jm+k} given $\mathbf{X}_{jm+k-1} = \mathbf{x}$ of the form $T_k(\mathbf{x}, dy)$ and fulfilling the conditional constraints $\int T_k(\mathbf{x}, dy)y = 0$ and $\int T_k(\mathbf{x}, dy)y^2 = 1$. Then the transition distribution from $\mathbf{X}_{jm+k-1} = \mathbf{x}$ to X_{jm+k} is given by

$$(1.5) \quad Q_k(\mathbf{x}, dy) = \frac{1}{s_{k\vartheta}(\mathbf{x})} T_k\left(\mathbf{x}, \frac{dy - r_{k\vartheta}(\mathbf{x})}{s_{k\vartheta}(\mathbf{x})}\right).$$

Description (1.1) of the model is convenient if we make no structural assumptions on T_k . However, if we have constraints on T_k for $k = 1, \dots, m$, the description (1.5) is more appropriate because the T_k appear explicitly as ‘‘parameters’’ of the model. For non-alternating quasi-likelihood models, Müller, Schick and Wefelmeyer (2011) consider constraints on the transition distribution $T = T_1 = \dots = T_k$. Specifically they assume that the conditional

distribution $T(\mathbf{x}, dy)$ is symmetric about x_p , or that it does not depend on the last q arguments x_{p-q+1}, \dots, x_p . A degenerate case is $q = p$, in which case $T(\mathbf{x}, dy) = T(dy)$, so the innovations are independent. For efficient estimation of ϑ in the latter model see Drost, Klaassen and Werker (1997). Koul and Schick (1997) treat the case with a constant scale function.

In Section 4 we use description (1.5) of the model to give another characterization of efficient estimators for ϑ that covers constraints on the T_k . For our alternating quasi-likelihood model, we are in particular interested in the constraint that the conditional distribution of the innovations does not alternate, $T_k = T$. We show that this contains information about ϑ . An efficient estimator can now be constructed by the one-step (or Newton–Raphson) improvement of a consistent initial estimator, for example a least squares estimator. This method is useful for general semiparametric models.

2 Characterizing efficiency in the general model

Consider first the *nonparametric* alternating Markov chain model with period m and transition distributions $Q_k(\mathbf{x}, dy)$ of order p for $k = 1, \dots, m$ about which we do not make any structural assumptions. It can be viewed as a non-alternating m -dimensional Markov chain

$$\mathbf{Y}_j = (X_{(j-1)m+1}, \dots, X_{jm})^\top, \quad j \in \mathbb{Z}.$$

This is a homogeneous Markov chain of order $c = \lceil p/m \rceil$. Its transition distribution from $\mathbf{Y}_{j-c}, \dots, \mathbf{Y}_{j-1}$ to $\mathbf{Y}_j = (x_1, \dots, x_m)^\top$ depends only on the values of the last p components of the vector $(\mathbf{Y}_{j-c}^\top, \dots, \mathbf{Y}_{j-1}^\top)^\top$, which form the vector

$$\mathbf{X}_{j-1} = (X_{(j-1)m-p+1}, \dots, X_{(j-1)m})^\top.$$

Setting $\mathbf{x}_{k-1} = (x_{k-p}, \dots, x_{k-1})$, it is given by

$$Q(\mathbf{x}_0, dx_1, \dots, dx_m) = Q_1 \otimes \dots \otimes Q_m(\mathbf{x}_0, dx_1, \dots, dx_m) = \prod_{k=1}^m Q_k(\mathbf{x}_{k-1}, dx_k).$$

We observe $\mathbf{X}_0, \mathbf{Y}_1, \dots, \mathbf{Y}_n$. Assume that \mathbf{Y}_j , $j \in \mathbb{Z}$, is strictly stationary and positive Harris recurrent. Write $Q_k(\mathbf{x}, v) = \int Q_k(\mathbf{x}, dy)v(\mathbf{x}, y)$ for the conditional expectation of a random variable $v(\mathbf{X}_{k-1}, X_k)$ given $\mathbf{X}_{k-1} = \mathbf{x}$. When the argument \mathbf{x} is omitted, we use the abbreviation $Q_k v = Q_k(\cdot, v)$. Let G_{k-1} denote the joint law of \mathbf{X}_{k-1} . Then $G_{k-1} \otimes Q_k$ is the stationary law of (\mathbf{X}_{k-1}, X_k) .

The nonparametric model is locally asymptotically normal in the following sense. Fix ϑ and $Q = Q_1 \otimes \dots \otimes Q_m$. For $k = 1, \dots, m$ introduce

$$H_k = \{v_k \in L_2(G_{k-1} \otimes Q_k) : Q_k v_k = 0\}.$$

For $v_k \in H_k$ choose Hellinger differentiable perturbations $Q_{k nv_k}$ of Q_k

$$\iint \left(\left(\frac{dQ_{k nv_k}(\mathbf{x}, \cdot)}{Q_k(\mathbf{x}, \cdot)}(y) \right)^{1/2} - 1 - \frac{1}{2} n^{-1/2} v_k(\mathbf{x}, y) \right)^2 G_{k-1}(d\mathbf{x}) Q_k(\mathbf{x}, dy) = o(n^{-1}).$$

Since $Q_{k nv_k}$ must be a conditional distribution, we must have $Q_k v_k = 0$. It follows that the *local parameter space* of the nonparametric model is $H = H_1 \times \cdots \times H_m$. Write P_n and P_{nv} for the joint law of $(\mathbf{X}_0, \mathbf{Y}_1, \dots, \mathbf{Y}_n)$ under Q and $Q_{nv} = Q_{1 nv_1} \otimes \cdots \otimes Q_{m nv_m}$, respectively. As in Penev (1991), Höpfner et al. (1990) and Höpfner (1993) we obtain *local asymptotic normality* for $v = (v_1, \dots, v_m) \in H$,

$$\begin{aligned} \log \frac{dP_{nv}}{dP_n} &= n^{-1/2} \sum_{j=1}^n \sum_{k=1}^m v_k(\mathbf{X}_{jm+k-1}, X_{jm+k}) - \frac{1}{2} \|v\|^2 + o_{P_n}(1), \\ n^{-1/2} \sum_{j=1}^n \sum_{k=1}^m v_k(\mathbf{X}_{jm+k-1}, X_{jm+k}) &\Rightarrow \|v\| N \quad \text{under } P_n, \end{aligned}$$

where N is a standard normal random variable and

$$\|v\|^2 = \sum_{k=1}^m G_{k-1} \otimes Q_k v_k^2.$$

Here we have used that $v_1(\mathbf{X}_0, X_1), \dots, v_m(\mathbf{X}_{m-1}, X_m)$ are martingale increments and therefore orthogonal. The norm $\|v\|$ induces on H an inner product

$$(v, v') = \sum_{k=1}^m G_{k-1} \otimes Q_k (v_k v'_k).$$

This inner product determines how difficult it is, asymptotically, to distinguish between Q and Q_{nv} on the basis of observations $\mathbf{X}_0, \mathbf{Y}_1, \dots, \mathbf{Y}_n$.

Consider now the submodel with conditional constraints (1.1). They can be written

$$Q_k(\mathbf{x}, a_{k\vartheta}) = \int Q_k(\mathbf{x}, dy) a_{k\vartheta}(\mathbf{x}, y) = 0, \quad k = 1, \dots, m.$$

They must also hold for the perturbed transition distribution $Q_{nv} = Q_{1 nv_1} \otimes \cdots \otimes Q_{m nv_m}$, possibly with perturbed parameter, say $\vartheta_{nu} = \vartheta + n^{-1/2} u + o(n^{-1/2})$ for some $u \in \mathbb{R}^d$. Under appropriate differentiability conditions on $a_{k\vartheta}$ with respect to ϑ , we obtain the pointwise expansion

$$0 = Q_{k nv_k} a_{k\vartheta_{nu}} = Q_k a_{k\vartheta} + n^{-1/2} (Q_k(a_{k\vartheta} v_k) + (Q_k \dot{a}_{k\vartheta}) u) + o(n^{-1/2}),$$

where $\dot{a}_{k\vartheta}$ is the $q \times d$ matrix of partial derivatives of $a_{k\vartheta}$ with respect to ϑ . Hence the perturbation v_k must be in the affine space

$$V_{ku} = \{v_k \in H_k : Q_k(a_{k\vartheta} v_k) + (Q_k \dot{a}_{k\vartheta}) u = 0\}.$$

Set $V_u = V_{1u} \times \cdots \times V_{mu}$. The local parameter space of the constrained alternating Markov chain model is the union V of the sets V_u , $u \in \mathbb{R}^p$. In model (1.1), we can now characterize efficient estimators of finite-dimensional functionals of Q as follows, using results originally due to Hájek and Le Cam, for which we refer to Theorem 2 in Section 3.3 of the monograph by Bickel et al. (1998). Note that this theorem holds for general locally asymptotically normal models even though it is stated only for the i.i.d. case.

A d -dimensional functional $\varphi(Q)$ is called *differentiable* at Q with *gradient* g if $g = (g_1, \dots, g_m) \in H^d = (H_1 \times \cdots \times H_m)^d$ and

$$n^{1/2}(\varphi(Q_{nv}) - \varphi(Q)) \rightarrow (g, v) = \sum_{k=1}^m G_{k-1} \otimes Q_k(g_k v_k), \quad v = (v_1, \dots, v_m) \in V.$$

We always regard an element in $H^d = (H_1 \times \cdots \times H_m)^d$ as a $d \times m$ matrix. The *canonical gradient* g^* is the componentwise projection of g onto V^d . An estimator $\hat{\varphi}$ is called *regular* for φ at Q with *limit* L if L is a d -dimensional random vector such that

$$n^{1/2}(\hat{\varphi} - \varphi(Q_{nv})) \Rightarrow L \quad \text{under } P_{nv}, \quad v \in V.$$

The convolution theorem says that if $\hat{\varphi}$ is regular for φ at Q with limit L , then

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

where N is a d -dimensional standard normal random vector and M a random vector independent of N . This justifies calling a regular estimator *efficient* for φ if its limit is

$$L = (g^*, g^{*\top})^{1/2} N \quad \text{in distribution.}$$

Its asymptotic covariance matrix is $(g^*, g^{*\top})$. An estimator $\hat{\varphi}$ is called *asymptotically linear* for φ at Q with *influence function* f if $f = (f_1, \dots, f_m) \in H^d$ and

$$n^{1/2}(\hat{\varphi} - \varphi(Q)) = n^{-1/2} \sum_{j=1}^n \sum_{k=1}^m f_k(\mathbf{X}_{j_{m+k-1}}, X_{j_{m+k}}) + o_{P_n}(1).$$

Such an estimator is asymptotically normal with covariance matrix $\sum_{k=1}^m G_{k-1} \otimes Q_k(f_k f_k^\top)$. By Theorem 2 of Bickel et al. (1998) mentioned above, we have the following two characterizations.

1. *An asymptotically linear estimator is regular for φ at Q if and only if its influence function is a gradient for φ at Q .*
2. *An estimator is regular and efficient for φ at Q if and only if it is asymptotically linear with influence function equal to the canonical gradient of φ at Q .*

Now we apply these results to the parameter ϑ , considered as a functional of the transition distribution by setting $\varphi(Q) = \vartheta$ if $Q_k a_{k\vartheta} = 0$ for $k = 1, \dots, m$. We have

$$n^{1/2}(\varphi(Q_{nv}) - \varphi(Q)) = n^{1/2}(\vartheta_{nu} - \vartheta) + o(1) \rightarrow u, \quad v \in V_u.$$

Hence the canonical gradient of ϑ is characterized as the vector $g^* = (g_1^*, \dots, g_m^*) \in V^d$ for which

$$(g^*, v) = \sum_{k=1}^m G_{k-1} \otimes Q_k(g_k^* v_k) = u, \quad v \in V_u.$$

We now prove that $g^* = J^{-1}\lambda$ with $\lambda = (\lambda_1, \dots, \lambda_m)$ and $J = \sum_{k=1}^m J_k$, where

$$\begin{aligned} \lambda_k(\mathbf{x}_{k-1}, x_k) &= -Q_k(\mathbf{x}_k, \dot{a}_{k\vartheta}^\top) Q_k(\mathbf{x}_k, a_{k\vartheta} a_{k\vartheta}^\top)^{-1} a_{k\vartheta}(\mathbf{x}_{k-1}, x_k), \\ J_k &= G_{k-1} \otimes Q_k(\lambda_k \lambda_k^\top) = G_{k-1} (Q_k \dot{a}_{k\vartheta}^\top Q_k (a_{k\vartheta} a_{k\vartheta}^\top)^{-1} Q_k \dot{a}_{k\vartheta}). \end{aligned}$$

For the proof note first that $Q_k(a_{k\vartheta} \lambda_k^\top) = -Q_k \dot{a}_{k\vartheta}$. This means that the i -th row of λ is in V_{e_i} , where e_i denotes the i -th d -dimensional unit vector. We obtain $Q_k(a_{k\vartheta} \lambda_k^\top) J^{-1} = -Q_k \dot{a}_{k\vartheta} J^{-1}$. Hence the i -th row of $g^* = J^{-1}\lambda$ is in V_{u_i} , where u_i denotes the i -th column of J^{-1} . Hence g^* is in V^d . Finally, for $v \in V_u$ we have

$$\begin{aligned} (g^*, v) &= \sum_{k=1}^m G_{k-1} \otimes Q_k(g_k^* v_k) = J^{-1} \sum_{k=1}^m G_{k-1} \otimes Q_k(\lambda_k v_k) \\ &= -J^{-1} \sum_{k=1}^m Q_k \dot{a}_{k\vartheta}^\top Q_k (a_{k\vartheta} a_{k\vartheta}^\top)^{-1} Q_k (a_{k\vartheta} v_k) = J^{-1} \sum_{k=1}^m J_k u = u. \end{aligned}$$

Hence $g^* = J^{-1}\lambda$ is the canonical gradient of ϑ .

It follows that an efficient estimator $\hat{\vartheta}$ of ϑ has the asymptotic expansion

$$\begin{aligned} n^{1/2}(\hat{\vartheta} - \vartheta) &= n^{-1/2} \sum_{j=1}^n \sum_{k=1}^m g_k^*(\mathbf{X}_{jm+k-1}, X_{jm+k}) + o_{P_n}(1) \\ &= J^{-1} n^{-1/2} \sum_{j=1}^n \sum_{k=1}^m \lambda_k(\mathbf{X}_{jm+k-1}, X_{jm+k}) + o_{P_n}(1) \end{aligned}$$

and asymptotic covariance matrix

$$(g^*, g^{*\top}) = J^{-1} \sum_{k=1}^m G_{k-1} \otimes Q_k(\lambda_k \lambda_k^\top) J^{-1} = J^{-1}.$$

3 Estimators in the general model

For $q = d$, a simple estimator of ϑ is the *least squares estimator* $\hat{\vartheta}_{LS}$, which we define as a solution of the d -dimensional martingale estimating equation

$$\sum_{j=1}^n \sum_{k=1}^m a_{k\vartheta}(\mathbf{X}_{jm+k-1}, X_{jm+k}) = 0.$$

For arbitrary q , we define *weighted least squares estimators* $\hat{\vartheta}_{WLS}$ as solutions of the martingale estimating equations of the form

$$\sum_{j=1}^n \sum_{k=1}^m W_{k\vartheta}^\top(\mathbf{X}_{jm+k-1}) a_{k\vartheta}(\mathbf{X}_{jm+k-1}, X_{jm+k}) = 0,$$

where $W_{k\vartheta}$ is a $q \times d$ matrix of weight functions. Under appropriate differentiability assumptions on $a_{k\vartheta}$ with respect to ϑ , the asymptotic distribution of $\hat{\vartheta}_{WLS}$ is obtained from a Taylor expansion

$$\begin{aligned} 0 &= \frac{1}{n} \sum_{j=1}^n \sum_{k=1}^m W_{k\vartheta}^\top(\mathbf{X}_{jm+k-1}) a_{k\vartheta}(\mathbf{X}_{jm+k-1}, X_{jm+k}) \\ &\quad + \frac{1}{n} \sum_{j=1}^n \sum_{k=1}^m W_{k\vartheta}^\top(\mathbf{X}_{jm+k-1}) \dot{a}_{k\vartheta}(\mathbf{X}_{jm+k-1}, X_{jm+k}) (\hat{\vartheta}_{WLS} - \vartheta) + o_{P_n}(n^{-1/2}). \end{aligned}$$

If $\sum_{k=1}^m G_{k-1}(W_{k\vartheta}^\top Q_k \dot{a}_{k\vartheta})$ is invertible, we can rewrite the stochastic expansion as

$$\begin{aligned} &n^{1/2}(\hat{\vartheta}_{WLS} - \vartheta) \\ &= - \left(\sum_{k=1}^m G_{k-1}(W_{k\vartheta}^\top Q_k \dot{a}_{k\vartheta}) \right)^{-1} n^{-1/2} \sum_{j=1}^n \sum_{k=1}^m W_{k\vartheta}^\top(\mathbf{X}_{jm+k-1}) a_{k\vartheta}(\mathbf{X}_{jm+k-1}, X_{jm+k}) \\ &\quad + o_{P_n}(1). \end{aligned}$$

Hence $\hat{\vartheta}_{WLS}$ is asymptotically linear with influence function $f = (f_1, \dots, f_m)$ given by

$$f_k(\mathbf{x}_{k-1}, x_k) = - \left(\sum_{k=1}^m G_{k-1}(W_{k\vartheta}^\top Q_k \dot{a}_{k\vartheta}) \right)^{-1} W_{k\vartheta}^\top(\mathbf{x}_{k-1}) a_{k\vartheta}(\mathbf{x}_{k-1}, x_k).$$

The asymptotic covariance matrix is

$$\left(\sum_{k=1}^m G_{k-1}(W_{k\vartheta}^\top Q_k \dot{a}_{k\vartheta}) \right)^{-1} \sum_{k=1}^m G_{k-1}(W_{k\vartheta}^\top Q_k (a_{k\vartheta} a_{k\vartheta}^\top) W_{k\vartheta}) \left(\sum_{k=1}^m G_{k-1}(W_{k\vartheta}^\top Q_k \dot{a}_{k\vartheta}) \right)^{-1}.$$

By the Cauchy–Schwarz inequality, the optimal weights are

$$W_{k\vartheta} = W_{k\vartheta}^* = Q_k (a_{k\vartheta} a_{k\vartheta}^\top)^{-1} Q_k \dot{a}_{k\vartheta}.$$

For these weights, the asymptotic covariance matrix becomes J^{-1} with J defined in Section 2. The weights $W_{k\vartheta}^*$ depend on the unknown transition distributions Q_k . In order to make the corresponding estimating equation meaningful, the weights must be replaced by estimators, say $\hat{W}_{k\vartheta}^*$. If the state space is \mathbb{R} , such estimators are obtained by replacing the conditional expectations $Q_k(\mathbf{x}, a_{k\vartheta} a_{k\vartheta}^\top)$ and $Q_k(\mathbf{x}, \dot{a}_{k\vartheta})$ appearing in $W_{k\vartheta}^*(\mathbf{x})$ by modified versions of Nadaraya–Watson estimators

$$\begin{aligned} \hat{Q}_k(\mathbf{x}, a_{k\vartheta} a_{k\vartheta}^\top) &= \frac{\sum_{i=1}^n K_b(\mathbf{x} - \mathbf{X}_{im+k-1}) a_{k\vartheta}(\mathbf{x}, X_{im+k}) a_{k\vartheta}^\top(\mathbf{x}, X_{im+k})}{\sum_{i=1}^n K_b(\mathbf{x} - \mathbf{X}_{im+k-1})}, \\ \hat{Q}_k(\mathbf{x}, \dot{a}_{k\vartheta}) &= \frac{\sum_{i=1}^n K_b(\mathbf{x} - \mathbf{X}_{im+k-1}) \dot{a}_{k\vartheta}(\mathbf{x}, \mathbf{X}_{im+k})}{\sum_{i=1}^n K_b(\mathbf{x} - \mathbf{X}_{im+k-1})}, \end{aligned}$$

where $K_b(\mathbf{x}) = K(x_1/b, \dots, x_p/b)$ with p -dimensional kernel K and bandwidth $b = b_n \rightarrow 0$. Such modifications may be necessitated by technical considerations. Nadaraya–Watson

estimators typically perform poorly at points \mathbf{x} at which the density of \mathbf{X}_{k-1} is close to zero. Thus one may be forced to exclude such points. This can be achieved by replacing the conditional expectations $\hat{Q}_k(\mathbf{x}, h)$ by the density weighted versions

$$\hat{Q}_k(\mathbf{x}, h) \mathbf{1} \left[\sum_{i=1}^n K_b(\mathbf{x} - \mathbf{X}_{im+k-1}) > \eta n b^p \right]$$

with $\eta = \eta_n$ slowly tending to zero.

An *optimally weighted least squares estimator* $\hat{\vartheta}^*$ is then obtained as a solution of

$$(3.1) \quad \sum_{j=1}^n \sum_{k=1}^m \hat{W}_{k\vartheta}^{*\top}(\mathbf{X}_{jm+k-1}) a_{k\vartheta}(\mathbf{X}_{jm+k-1}, X_{jm+k}) = 0.$$

Under appropriate conditions, the stochastic expansion of $\hat{\vartheta}^*$ is not changed when the weights $W_{k\vartheta}^*$ are replaced by the estimators $\hat{W}_{k\vartheta}^*$. It follows that the influence function of $\hat{\vartheta}^*$ equals the canonical gradient $g^* = J^{-1}\lambda$ of ϑ . Hence $\hat{\vartheta}^*$ is efficient.

Remark. In particular, for $q = d$, the least squares estimator $\hat{\vartheta}_{LS}$, with weights W_k given by the $d \times d$ unit matrix, is also asymptotically linear. Its influence function is $f^{LS} = (f_1^{LS}, \dots, f_m^{LS})$ with

$$f_k^{LS}(\mathbf{x}_{k-1}, x_k) = - \left(\sum_{k=1}^m G_{k-1} \otimes Q_k \dot{a}_{k\vartheta} \right)^{-1} a_{k\vartheta}(\mathbf{x}_{k-1}, x_k),$$

and its asymptotic covariance matrix is

$$\left(\sum_{k=1}^m G_{k-1} \otimes Q_k \dot{a}_{k\vartheta} \right)^{-1} \sum_{k=1}^m G_{k-1} \otimes Q_k (a_{k\vartheta} a_{k\vartheta}^\top) \left(\sum_{k=1}^m G_{k-1} \otimes Q_k \dot{a}_{k\vartheta} \right)^{-1}.$$

It is straightforward to check that $f^{LS} - g^*$ and g^* are uncorrelated,

$$(f^{LS}, g^{*\top}) = \sum_{k=1}^m G_{k-1} \otimes Q_k (f_k^{LS} g_k^{*\top}) = J^{-1} = (g^*, g^{*\top}).$$

Hence the difference between the asymptotic covariance matrices of $\hat{\vartheta}_{LS}$ and $\hat{\vartheta}_*$ is nonnegative and equals

$$(f^{LS} - g^*, (f^{LS} - g^*)^\top) = (f^{LS} - g^*, f^{LS\top}).$$

It follows that the optimally weighted least squares estimator $\hat{\vartheta}^*$ is strictly better than $\hat{\vartheta}_{LS}$ unless $f^{LS} = g^*$.

4 Characterizing efficiency in quasi-likelihood models

Now we consider the alternating nonlinear and heteroskedastic autoregressive model

$$(4.1) \quad X_{jm+k} = r_{k\vartheta}(\mathbf{X}_{jm+k-1}) + s_{k\vartheta}(\mathbf{X}_{jm+k-1})\varepsilon_{jm+k}, \quad j \in \mathbb{Z}, \quad k = 1, \dots, m.$$

We assume that the conditional distribution $T_k(\mathbf{x}, dy)$ of ε_{jm+k} given $\mathbf{X}_{jm+k-1} = \mathbf{x}$ fulfills $T_k(\mathbf{x}, e) = 0$ and $T_k(\mathbf{x}, e^2) = 1$, where $e(y) = y$ is the identity on \mathbb{R} . In Section 2 we have shown local asymptotic normality for alternating Markov chain models with conditional constraints (1.1). For the special case of an alternating quasi-likelihood model (4.1) we now give a different proof, using the parametrization by ϑ and T_k . As noted, this will allow us to treat constraints on the T_k , in particular the special constraint $T_1 = \dots = T_m = T$.

Fix ϑ and $\mathbf{T} = (T_1, \dots, T_m)$. Assume that $\mathbf{Y}_j = (X_{(j-1)m+1}, \dots, X_{jm})^\top$, $j \in \mathbb{Z}$, is strictly stationary and positive Harris recurrent. Following Müller et al. (2011), we make the following assumptions on $r_{k\vartheta}$, $s_{k\vartheta}$ and \mathbf{T} .

Assumption 1. For $k = 1, \dots, m$ there are G_{k-1} -square-integrable functions $\dot{r}_k = \dot{r}_{k\vartheta}$ and $\dot{s}_k = \dot{s}_{k\vartheta}$ such that, for each constant C ,

$$\begin{aligned} \sup_{\|\Delta\| \leq Cn^{-1/2}} \sum_{j=1}^n \left(r_{k,\vartheta+\Delta}(\mathbf{X}_{jm+k-1}) - r_{k\vartheta}(\mathbf{X}_{jm+k-1}) - \Delta^\top \dot{r}_k(\mathbf{X}_{jm+k-1}) \right)^2 &= o_{P_n}(1), \\ \sup_{\|\Delta\| \leq Cn^{-1/2}} \sum_{j=1}^n \left(s_{k,\vartheta+\Delta}(\mathbf{X}_{jm+k-1}) - s_{k\vartheta}(\mathbf{X}_{jm+k-1}) - \Delta^\top \dot{s}_k(\mathbf{X}_{jm+k-1}) \right)^2 &= o_{P_n}(1), \end{aligned}$$

and the function $s_{k\vartheta}$ is bounded away from zero locally uniformly in ϑ .

Assumption 2. For $k = 1, \dots, m$ and each \mathbf{x} , the conditional distribution $T_k(\mathbf{x}, dy)$ has a positive and absolutely continuous density $t_k(\mathbf{x}, y)$, and $E[\varepsilon_k^4]$ and $E[\ell_{k1}^2(\mathbf{X}_{k-1}, \varepsilon_k)(1 + \varepsilon_k^2)]$ are finite, where $\ell_{k1}(\mathbf{x}, y) = -t'_k(\mathbf{x}, y)/t_k(\mathbf{x}, y)$, with derivative taken with respect to y .

Since the T_k have densities, the G_{k-1} also have densities, say g_{k-1} . Introduce perturbations $\vartheta_{nu} = \vartheta + n^{-1/2}u$ with $u \in \mathbb{R}^d$ and $t_{k\nu v_k}(\mathbf{x}, y) = t_k(\mathbf{x}, y)(1 + n^{-1/2}v_k(\mathbf{x}, y))$ with v_k a bounded measurable function on \mathbb{R}^{p+1} . The constraints $T_k(\mathbf{x}, 1) = 1$, $T_k(\mathbf{x}, e) = 0$, $T_k(\mathbf{x}, e^2) = 1$ must also hold for $T_k = T_{k\nu v_k}$. We obtain

$$T_k(\mathbf{x}, v_k) = 0, \quad T_k(\mathbf{x}, v_k e) = 0, \quad T_k(\mathbf{x}, v_k e^2) = 0.$$

Let V_k denote the set of these v_k . Write $\mathbf{v} = (v_1, \dots, v_m)$, $\mathbf{V} = V_1 \times \dots \times V_m$ and $\mathbf{T}_{nv} = (T_{1nv_1}, \dots, T_{mnv_m})$. Suppose we observe $\mathbf{X}_0, \mathbf{Y}_1, \dots, \mathbf{Y}_n$. Write P_n and $P_{nu\mathbf{v}}$ for the joint law of the observations under (ϑ, \mathbf{T}) and $(\vartheta_{nu}, \mathbf{T}_{nv})$, respectively. Write $g_{0nu\mathbf{v}}$ for the density of \mathbf{X}_0 under $(\vartheta_{nu}, \mathbf{T}_{nv})$. Then we have local asymptotic normality as follows. The proof is similar to the proof in Drost, Klaassen and Werker (1997), who treat the non-alternating

case and independent observations. For $k = 1, \dots, m$ set $\ell_{k2}(\mathbf{x}, y) = \ell_{k1}(\mathbf{x}, y)y - 1$ and $\lambda_k = (\ell_{k1}, \ell_{k2})^\top$, and define the $d \times 2$ matrix

$$M_k(\mathbf{x}) = M_{k\vartheta}(\mathbf{x}) = \frac{1}{s_{k\vartheta}(\mathbf{x})}(\dot{r}_k(\mathbf{x}), \dot{s}_k(\mathbf{x})).$$

Theorem 1. *Let $(u, \mathbf{v}) \in \mathbb{R}^d \times \mathbf{V}$. Suppose Assumptions 1 and 2 hold and the stationary density g_0 depends smoothly on the parameters in the sense that $\int |g_{0nu\mathbf{v}}(\mathbf{x}) - g_0(\mathbf{x})| d\mathbf{x} \rightarrow 0$. Then*

$$(4.2) \quad \log \frac{dP_{nu\mathbf{v}}}{dP_n} = n^{-1/2} \sum_{j=1}^n \sum_{k=1}^m s_{kuv_k}(\mathbf{X}_{jm+k-1}, \varepsilon_{jm+k}) - \frac{1}{2} \|(u, \mathbf{v})\|^2 + o_{P_n}(1),$$

$$(4.3) \quad n^{-1/2} \sum_{j=1}^n \sum_{k=1}^m s_{kuv_k}(\mathbf{X}_{jm+k-1}, \varepsilon_{jm+k}) \Rightarrow \|(u, \mathbf{v})\|N \quad \text{under } P_n,$$

where N is a standard normal random variable and

$$s_{kuv_k}(\mathbf{X}_{k-1}, \varepsilon_k) = u^\top M_k(\mathbf{X}_{k-1}) \lambda_k(\mathbf{X}_{k-1}, \varepsilon_k) + v_k(\mathbf{X}_{k-1}, \varepsilon_k),$$

$$\|(u, \mathbf{v})\|^2 = \sum_{k=1}^m E[s_{kuv_k}^2(\mathbf{X}_{k-1}, \varepsilon_k)].$$

Here we have used that $s_{1uv_1}(\mathbf{X}_0, \varepsilon_1), \dots, s_{muv_m}(\mathbf{X}_{m-1}, \varepsilon_m)$ are uncorrelated. The norm $\|(u, \mathbf{v})\|$ determines how difficult it is, asymptotically, to distinguish between (ϑ, \mathbf{t}) and $(\vartheta_{nu}, \mathbf{t}_{n\mathbf{v}})$ on the basis of the observations. It induces an inner product

$$((u', \mathbf{v}'), (u, \mathbf{v})) = \sum_{k=1}^m E[s_{ku'v'_k}(\mathbf{X}_{k-1}, \varepsilon_k) s_{kuv_k}(\mathbf{X}_{k-1}, \varepsilon_k)].$$

Consider now a *model* for \mathbf{T} . It is given by a family of vectors \mathcal{T} of conditional distributions $\mathbf{T} = (T_1, \dots, T_m)$. Assume that the fixed \mathbf{T} belongs to \mathcal{T} . Let \mathbf{W} be the set of all $\mathbf{v} \in \mathbf{V}$ such that $\mathbf{T}_{n\mathbf{v}}$ lies in \mathcal{T} . Assume that \mathbf{W} is a linear space, the *local parameter space* of \mathcal{T} at \mathbf{T} . Let \bar{V}_k denote the closure of V_k in $L_2(G_{k-1} \otimes T_k)$, and set $\bar{\mathbf{V}} = \bar{V}_1 \times \dots \times \bar{V}_m$. Let $\bar{\mathbf{W}}$ denote the closure of \mathbf{W} in $\bar{\mathbf{V}}$.

Definition 1. A real-valued functional φ of (ϑ, \mathbf{t}) is called *differentiable* at (ϑ, \mathbf{t}) with *gradient* $\sum_{k=1}^m s_{ku\varphi v_{\varphi k}}(\mathbf{X}_{k-1}, \varepsilon_k)$ if $(u_\varphi, \mathbf{v}_\varphi) \in \mathbb{R}^d \times \bar{\mathbf{V}}$ and

$$n^{1/2}(\varphi(\vartheta_{nu}, \mathbf{t}_{n\mathbf{v}}) - \varphi(\vartheta, \mathbf{t})) \rightarrow ((u_\varphi, \mathbf{v}_\varphi), (u, \mathbf{v})), \quad (u, \mathbf{v}) \in \mathbb{R}^d \times \mathbf{W}.$$

If $\mathbf{v}_\varphi = \mathbf{w}_\varphi$ is in $\bar{\mathbf{W}}$, then $s_{u_\varphi \mathbf{w}_\varphi}$ is called the *canonical gradient* of φ .

Definition 2. An estimator $\hat{\varphi}$ of φ is called *regular* at (ϑ, \mathbf{t}) with *limit* L if L is a random variable such that

$$n^{1/2}(\hat{\varphi} - \varphi(\vartheta_{nu}, \mathbf{t}_{n\mathbf{v}})) \Rightarrow L \quad \text{under } P_{nu\mathbf{v}}, \quad (u, \mathbf{v}) \in \mathbb{R}^d \times \mathbf{W}.$$

As in Section 2 we obtain from the convolution theorem that for such an estimator, $L = \|(u_\varphi, \mathbf{w}_\varphi)\|N + M$ in distribution, with M independent of N .

Definition 3. An estimator $\hat{\varphi}$ of φ is called *asymptotically linear* at (ϑ, \mathbf{t}) with *influence function* $\sum_{k=1}^m s_{ku_\varphi v_{\varphi k}}(\mathbf{X}_{k-1}, \varepsilon_k)$ if $(u_\varphi, \mathbf{v}_\varphi) \in \mathbb{R}^d \times \bar{\mathbf{V}}$ and

$$n^{1/2}(\hat{\varphi} - \varphi(\vartheta, \mathbf{t})) = \sum_{j=1}^n \sum_{k=1}^m s_{ku_\varphi v_{\varphi k}}(\mathbf{X}_{jm+k-1}, \varepsilon_{jm+k}) + o_{P_n}(1).$$

As in Section 2, we have the following characterization: An estimator is regular and efficient if and only if it is asymptotically linear with influence function equal to the canonical gradient. We apply the theory to several models \mathcal{T} and to estimating ϑ , i.e., to the d -dimensional functional $\varphi(\vartheta, \mathbf{t}) = \vartheta$. Then differentiability of multivariate functionals φ and asymptotic linearity of multivariate estimators $\hat{\varphi}$ are understood componentwise. Regularity and the convolution theorem have obvious multivariate versions. The characterization of efficient estimators is also meant componentwise.

First we consider the nonparametric model. This is a special case of the model which was treated in Sections 2 and 3 by a different approach.

4.1 Nonparametric model

Suppose we have no structural information on T_1, \dots, T_m . Then $\mathbf{W} = \mathbf{V}$. In order to calculate the canonical gradient of ϑ , we use the orthogonal decomposition of s_{kuv_k} described in Müller, Schick and Wefelmeyer (2011). From $T_k(\mathbf{x}, 1) = 1$ and Assumption 2 we obtain $T_k(\mathbf{x}, \ell_{k1}e) = 1$, $T_k(\mathbf{x}, \ell_{k1}e^2) = 0$, $T_k(\mathbf{x}, \ell_{k1}e^3) = 3$. Setting $\psi(y) = (y, y^2 - 1)^\top$, we have

$$T_k(\mathbf{x}, \psi \lambda_k^\top) = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

We obtain

$$(4.4) \quad s_{kuv_k}(\mathbf{X}_{k-1}, \varepsilon_k) = u^\top M_k(\mathbf{X}_{k-1}) A_k^\top(\mathbf{X}_{k-1}) \psi(\varepsilon_k) + u^\top M_k(\mathbf{X}_{k-1}) \lambda_k^*(\mathbf{X}_{k-1}, \varepsilon_k) + v_k(\mathbf{X}_{k-1}, \varepsilon_k),$$

where the components of $\lambda_k^*(\mathbf{x}, y) = \lambda_k(\mathbf{x}, y) - A_k^\top(\mathbf{x})\psi(y)$ are in V_k , and

$$A_k(\mathbf{x}) = T_k(\mathbf{x}, \psi \psi^\top)^{-1} T_k(\mathbf{x}, \psi \lambda_k^\top) = c_k(\mathbf{x}) \begin{pmatrix} T_k(\mathbf{x}, e^4) - 1 & -2T_k(\mathbf{x}, e^3) \\ -T_k(\mathbf{x}, e^3) & 2 \end{pmatrix}$$

with $1/c_k(\mathbf{x}) = \det T_k(\mathbf{x}, \psi \psi^\top) = T_k(\mathbf{x}, e^4) - 1 - T_k(\mathbf{x}, e^3)^2$. The canonical gradient of ϑ is then

$$\Lambda^{-1} \sum_{k=1}^m M_k(\mathbf{X}_{k-1}) A_k^\top(\mathbf{X}_{k-1}) \psi(\varepsilon_k),$$

where $\Lambda = \sum_{k=1}^m \Lambda_k$ with $\Lambda_k = E[c_k(\mathbf{X})M_k(\mathbf{X})B_k(\mathbf{X})M_k^\top(\mathbf{X})]$ and

$$B_k(\mathbf{x}) = \begin{pmatrix} T_k(\mathbf{x}, e^4) - 1 & -2T_k(\mathbf{x}, e^3) \\ -2T_k(\mathbf{x}, e^3) & 4 \end{pmatrix}.$$

An efficient estimator $\hat{\vartheta}$ of ϑ is obtained as a solution of the estimating equations

$$\sum_{j=1}^n \sum_{k=1}^m M_{k\vartheta}(\mathbf{X}_{jm+k-1}) \tilde{A}_k^\top(\mathbf{X}_{jm+k-1}) \psi\left(\frac{X_{jm+k} - r_{k\vartheta}(\mathbf{X}_{jm+k-1})}{s_{k\vartheta}(\mathbf{X}_{jm+k-1})}\right) = 0.$$

Here $\tilde{A}_k(\mathbf{x})$ is an estimator of $A_k(\mathbf{x})$ obtained by replacing the conditional third and fourth moments $T_k(\mathbf{x}, e^i)$, $i = 3, 4$, by potentially density weighted versions of Nadaraya–Watson estimators

$$\tilde{T}_k(\mathbf{x}, e^i) = \frac{\sum_{j=1}^n K_b(\mathbf{x} - \mathbf{X}_{jm+k-1}) \tilde{\varepsilon}_{jm+k}^i}{\sum_{j=1}^n K_b(\mathbf{x} - \mathbf{X}_{jm+k-1})}$$

with $K_b(\mathbf{x}) = K(x_1/b, \dots, x_p/b)$ for a p -dimensional kernel K and a bandwidth b , based on residuals $\tilde{\varepsilon}_{jm+k} = (X_{jm+k} - r_{k\vartheta}(\mathbf{X}_{jm+k-1}))/s_{k\vartheta}(\mathbf{X}_{jm+k-1})$. The estimator $\tilde{\vartheta}$ can be any $n^{1/2}$ -consistent estimator of ϑ , for example the least squares estimator, which solves the estimating equation

$$\sum_{j=1}^n \sum_{k=1}^m M_{k\vartheta}(\mathbf{X}_{jm+k-1}) \psi\left(\frac{X_{jm+k} - r_{k\vartheta}(\mathbf{X}_{jm+k-1})}{s_{k\vartheta}(\mathbf{X}_{jm+k-1})}\right) = 0.$$

The asymptotic covariance matrix of $\hat{\vartheta}$ is Λ^{-1} .

The efficient estimator $\hat{\vartheta}$ is asymptotically equivalent to the efficient estimator $\hat{\vartheta}^*$ solving equation (3.1) of Section 3 when $a_{k\vartheta}$ is given by (1.4). For this $a_{k\vartheta}$ we have

$$\dot{a}_k(\mathbf{x}, y) = - \begin{pmatrix} \dot{r}_k(\mathbf{x}) \\ 2\dot{r}_k(\mathbf{x})(y - r_{k\vartheta}(\mathbf{x})) + 2\dot{s}_k(\mathbf{x})s_{k\vartheta}^2(\mathbf{x}) \end{pmatrix}.$$

Hence the optimal weights $\hat{W}_{k\vartheta}^*$ are estimators of $W_{k\vartheta}^* = Q_k(a_{k\vartheta}a_{k\vartheta}^\top)^{-1}Q_k\dot{a}_k$ which now involve the two matrices of conditional expectations

$$Q_k(\mathbf{x}, a_{k\vartheta}a_{k\vartheta}^\top) = \begin{pmatrix} s_{k\vartheta}^2(\mathbf{x}) & Q_k(\mathbf{x}, (e - r_{k\vartheta}(\mathbf{x}))^3) \\ Q_k(\mathbf{x}, (e - r_{k\vartheta}(\mathbf{x}))^3) & Q_k(\mathbf{x}, (e - r_{k\vartheta}(\mathbf{x}))^4) - s_{k\vartheta}^4(\mathbf{x}) \end{pmatrix},$$

$$Q_k(\mathbf{x}, \dot{a}_k) = - \begin{pmatrix} \dot{r}_k(\mathbf{x}) \\ 2\dot{s}_k(\mathbf{x})s_{k\vartheta}^2(\mathbf{x}) \end{pmatrix}.$$

The centered conditional moments $Q_k(\mathbf{x}, (e - r_{k\vartheta}(\mathbf{x}))^i)$, $i = 3, 4$, can again be estimated by potentially density weighted versions of Nadaraya–Watson estimators.

4.2 Equal conditional innovation distributions

Suppose the conditional distributions of the innovations ε_k given \mathbf{X}_{k-1} are known to be equal, $T_1 = \dots = T_m = T$. Then $\mathbf{T} = (T, \dots, T)$. A perturbation of the density t of T is of the form $t_{nv}(\mathbf{x}, y) = t(\mathbf{x}, y)(1 + n^{-1/2}v(\mathbf{x}, y))$, where v belongs to the set V of bounded measurable functions on \mathbb{R}^{p+1} such that $T(\mathbf{x}, v) = 0$, $T(\mathbf{x}, ve) = 0$, $T(\mathbf{x}, ve^2) = 0$. The local parameter space \mathbf{W} of \mathcal{T} now consists of the vectors $\mathbf{w} = (v, \dots, v)$ with $v \in V$. Set $\ell_1(\mathbf{x}, y) = -t'(\mathbf{x}, y)/t(\mathbf{x}, y)$, $\ell_2(\mathbf{x}, y) = \ell_1(\mathbf{x}, y)y - 1$ and $\lambda = (\ell_1, \ell_2)^\top$. Local asymptotic normality (4.2), (4.3) now holds with $v_k = v$ and $\mathbf{v} = \mathbf{w} = (v, \dots, v)$, and with $\lambda_k = \lambda$. The decomposition (4.4) reduces to

$$(4.5) \quad s_{kuv_k}(\mathbf{X}_{k-1}, \varepsilon_k) = u^\top M_k(\mathbf{X}_{k-1})A^\top(\mathbf{X}_{k-1})\psi(\varepsilon_k) \\ + u^\top M_k(\mathbf{X}_{k-1})\lambda_*(\mathbf{X}_{k-1}, \varepsilon_k) + v(\mathbf{X}_{k-1}, \varepsilon_k),$$

where the two components of $\lambda_*(\mathbf{x}, y) = \lambda(\mathbf{x}, y) - A^\top(\mathbf{x})\psi(y)$ are in V , and

$$A(\mathbf{x}) = T(\mathbf{x}, \psi\psi^\top)^{-1}T(\mathbf{x}, \psi\lambda^\top) = c(\mathbf{x}) \begin{pmatrix} T(\mathbf{x}, e^4) - 1 & -2T(\mathbf{x}, e^3) \\ -T(\mathbf{x}, e^3) & 2 \end{pmatrix}$$

with $1/c(\mathbf{x}) = \det T(\mathbf{x}, \psi\psi^\top) = T(\mathbf{x}, e^4) - 1 - T(\mathbf{x}, e^3)^2$. However, the d -dimensional functions $M_k(\mathbf{x})\lambda_*(\mathbf{x}, y)$ still depend on k , so $(M_1(\mathbf{x})\lambda_*(\mathbf{x}, y), \dots, M_m(\mathbf{x})\lambda_*(\mathbf{x}, y))$ is not in $\bar{\mathbf{W}}^d$. The row-wise projection onto $\bar{\mathbf{W}}^d$ is $(M_*(\mathbf{x})\lambda_*(\mathbf{x}, y), \dots, M_*(\mathbf{x})\lambda_*(\mathbf{x}, y))$ with

$$M_*(\mathbf{x}) = \frac{1}{m} \sum_{k=1}^m M_k(\mathbf{x}).$$

We arrive at the orthogonal decomposition

$$s_{kuv_k}(\mathbf{X}_{k-1}, \varepsilon_k) = u^\top s_k^*(\mathbf{X}_{k-1}, \varepsilon_k) + u^\top M_*(\mathbf{X}_{k-1})\lambda_*(\mathbf{X}_{k-1}, \varepsilon_k) + v(\mathbf{X}_{k-1}, \varepsilon_k)$$

with

$$s_k^*(\mathbf{x}, y) = M_k(\mathbf{x})A^\top(\mathbf{x})\psi(y) + (M_k(\mathbf{x}) - M_*(\mathbf{x}))\lambda_*(\mathbf{x}, y) \\ = M_k(\mathbf{x})\lambda(\mathbf{x}, y) - M_*(\mathbf{x})\lambda_*(\mathbf{x}, y).$$

Hence the canonical gradient of ϑ is $\Lambda_*^{-1} \sum_{k=1}^m s_k^*(\mathbf{X}_{k-1}, \varepsilon_k)$ with $\Lambda_* = \sum_{k=1}^m \Lambda_k^*$ and

$$\Lambda_k^* = E[s_k^*(\mathbf{X}_{k-1}, \varepsilon_k)s_k^{*\top}(\mathbf{X}_{k-1}, \varepsilon_k)] \\ = \Lambda_k + E[(M_k(\mathbf{X}_{k-1}) - M_*(\mathbf{X}_{k-1}))J_*(\mathbf{X}_{k-1})(M_k(\mathbf{X}_{k-1}) - M_*(\mathbf{X}_{k-1}))^\top].$$

Here we have used $T(\mathbf{x}, \lambda_*\psi^\top) = 0$, and we have set

$$J_*(\mathbf{x}) = T(\mathbf{x}, \lambda_*\lambda_*^\top) \\ = T(\mathbf{x}, \lambda\lambda^\top) - A^\top(\mathbf{x})T(\mathbf{x}, \psi\lambda^\top) - T(\mathbf{x}, \lambda\psi^\top)A(\mathbf{x}) + A^\top(\mathbf{x})T(\mathbf{x}, \psi\psi^\top)A(\mathbf{x}) \\ = T(\mathbf{x}, \lambda\lambda^\top) - c(\mathbf{x}) \begin{pmatrix} T(\mathbf{x}, e^4) - 1 & -2T(\mathbf{x}, e^3) \\ -2T(\mathbf{x}, e^3) & 4 \end{pmatrix}.$$

The canonical gradient of ϑ now involves the score functions $\ell_1(\mathbf{x}, y) = -t'(\mathbf{x}, y)/t(\mathbf{x}, y)$ and $\ell_2(\mathbf{x}, y) = \ell_1(\mathbf{x}, y)y - 1$ for (conditional) location and scale of $T(\mathbf{x}, dy)$. This makes it difficult to estimate ϑ by an estimating equation. We can follow Müller et al. (2007) and construct an efficient estimator as one-step improvement of a consistent initial estimator $\tilde{\vartheta}$,

$$\hat{\vartheta}^* = \tilde{\vartheta} + \tilde{\Lambda}_*^{-1} \frac{1}{n} \sum_{j=1}^n \sum_{k=1}^m \tilde{s}_k^*(\mathbf{X}_{jm+k-1}, \tilde{\varepsilon}_{jm+k}),$$

The right-hand side requires estimators for ℓ_1 and ℓ_2 and for the conditional moments $T(\mathbf{x}, e^i)$, $i = 3, 4$. An estimator for ℓ_1 is \tilde{t}'/\tilde{t} , where \tilde{t} is a kernel estimator based on $(\mathbf{X}_i, \tilde{\varepsilon}_i)$, $i = 1, \dots, nm$, with residuals $\tilde{\varepsilon}_{jm+k} = (X_{jm+k} - r_{k\vartheta}(\mathbf{X}_{jm+k-1}))/s_{k\vartheta}(\mathbf{X}_{jm+k-1})$, $j = 1, \dots, n$, $k = 1, \dots, m$. The conditional moments can be estimated by plugging \tilde{t} into $T(\mathbf{x}, e^i) = \int t(\mathbf{x}, y)y^i dy$, or by potentially density weighted Nadaraya–Watson estimators based on the residuals.

When we know that $T_1 = \dots = T_m$, then the asymptotic covariance matrix of an efficient estimator of ϑ is Λ_*^{-1} . This is strictly smaller than the smallest asymptotic covariance matrix Λ^{-1} of estimators that do not use this information, unless $\Lambda_k^* = \Lambda_k$ for $k = 1, \dots, m$, or equivalently, unless

$$E[(M_k(\mathbf{X}_{k-1}) - M_*(\mathbf{X}_{k-1}))(M_k(\mathbf{X}_{k-1}) - M_*(\mathbf{X}_{k-1}))^\top] = 0, \quad k = 1, \dots, m.$$

This may happen in degenerate cases only.

One is that $J_* = 0$. This is the case if $T(\mathbf{x}, \cdot)$ is standard normal, which means that the innovations ε_i are independent and standard normal. Then $\ell_1(\mathbf{x}, y) = y$, so $\lambda = \psi$ and hence $T(\mathbf{x}, \psi\psi^\top) = T(\mathbf{x}, \psi\lambda^\top)$, which implies that A is the 2×2 identity matrix. Then $\lambda_* = 0$ and therefore $J_* = 0$.

Another degenerate case is that $M_k - M_* = 0$ for $k = 1, \dots, m$, i.e. $M_{1\vartheta} = \dots = M_{m\vartheta}$, so $\dot{r}_{k\vartheta}/s_{k\vartheta}$ and $\dot{s}_{k\vartheta}/s_{k\vartheta}$ do not depend on k . This happens of course when the time series is *not* alternating. It can also happen for alternating time series, for example when the conditional variances $s_{k\vartheta}^2$ of the innovations do not depend on ϑ , so $\dot{s}_{k\vartheta} = 0$; and the conditional means do not alternate, $r_{1\vartheta} = \dots = r_{m\vartheta}$.

Acknowledgments

Ursula U. Müller was supported by NSF Grant DMS 0907014. Anton Schick was supported by NSF Grant DMS 0906551.

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] Drost, F. C., Klaassen, C. A. J. and Werker, B. J. M. (1997). Adaptive estimation in time-series models. *Ann. Statist.* **25**, 786–817.
- [3] 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.
- [4] 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.
- [5] Koul, H. L. and Schick, A. (1997). Efficient estimation in nonlinear autoregressive time-series models. *Bernoulli* **3**, 247–277.
- [6] Müller, U. U., Schick, A. and Wefelmeyer, W. (2007). Inference for alternating time series. In: *Recent Advances in Stochastic Modeling and Data Analysis* (C. H. Skiadas, ed.), 589–596, World Scientific, Singapore.
- [7] Müller, U. U., Schick, A. and Wefelmeyer, W. (2009). Estimators for alternating non-linear autoregression. *J. Multivariate Anal.* **100**, 266–277.
- [8] Müller, U. U., Schick, A. and Wefelmeyer, W. (2011). Variance bounds for estimators in autoregressive models with constraints. To appear in: *Statistics*.
- [9] Penev, S. (1991) Efficient estimation of the stationary distribution for exponentially ergodic Markov chains. *J. Statist. Plann. Inference* **27**, 105-123.