Estimators for alternating nonlinear autoregression

Ursula U. Müller
Texas A&M University

Anton Schick *
Binghamton University

Wolfgang Wefelmeyer
Universität zu Köln

Abstract

Suppose we observe a time series that alternates between different nonlinear autoregressive processes. We give conditions under which the model is locally asymptotically normal, derive a characterization of efficient estimators for differentiable functionals of the model, and use it to construct efficient estimators for the autoregression parameters and the innovation distributions. Surprisingly, the estimators for the autoregression parameters can be improved if we know that the innovation densities are equal.

Keywords. Convolution theorem, regular estimator, asymptotically linear estimator, Newton–Raphson procedure, weighted least squares estimator, linear autoregression.

1 Introduction

The behavior of a time series may be influenced by periodic (daily, weekly, yearly) changes. If we have observations on a smaller (hourly, daily, monthly) scale, then such changes can be modeled by an *alternating nonlinear autoregressive process of order p and period m*. By this we mean a time series X_i, $i \in \mathbb{Z}$, that alternates periodically between m possibly different nonlinear AR(p) processes,

$$X_{jm+k} = r_{k\vartheta}(\mathbf{X}_{jm+k-1}) + \varepsilon_{jm+k}, \quad j \in \mathbb{Z}, \quad k = 1, \ldots, m, \tag{1.1}$$

where $\mathbf{X}_i = (X_{i-p+1}, \ldots, X_i)$, the autoregression function is known up to a parameter ϑ that varies in an open set $\Theta \subset \mathbb{R}^d$, the innovations ε_i, $i \in \mathbb{Z}$, are independent with mean zero, and ε_{jm+k} has positive density f_k and finite variance σ_k^2. We assume that we have initial observations $X_{-p+1}, \ldots, X_0$ and then observe n periods $X_1, \ldots, X_{nm}$.

Our model includes alternating linear autoregression as a special case, with $r_{k\vartheta}(\mathbf{X}_{k-1}) = \varrho_{k\vartheta}^\top \mathbf{X}_{k-1}$ for a vector $\varrho_{k\vartheta} = (\varrho_{k\vartheta 1}, \ldots, \varrho_{k\vartheta p})^\top$. The case of first-order alternating linear autoregression is studied in Müller et al. (2007). It is shown in Müller et al. (2008) that this case appears in particular when a (non-alternating) first-order linear autoregressive process is observed at certain periodically repeated time points only.

*Supported in part by NSF Grant DMS 0405791.

In Section 2 we give conditions under which an alternating nonlinear autoregressive process is locally asymptotically normal. We describe a characterization of efficient estimators for differentiable functionals of $(\vartheta, f_1, \dots, f_m)$. In Section 3 we study estimation of ϑ. Let $\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})$ denote the gradient of $r_{k\vartheta}(\mathbf{X}_{k-1})$ as a function of ϑ, and write $\mu_k = E[\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})]$ and $R_k = E[\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})\dot{r}_{k\vartheta}^\top(\mathbf{X}_{k-1})]$. The least squares estimator is a solution in ϑ of the martingale estimating equation

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

By Taylor expansion, its asymptotic covariance matrix is seen to be the covariance matrix of $(\sum_{k=1}^m R_k)^{-1}\sum_{k=1}^m \dot{r}_{k\vartheta}(\mathbf{X}_{k-1})\varepsilon_k$,

$$M_{LS} = \Big(\sum_{k=1}^{m} R_k\Big)^{-1}\Big(\sum_{k=1}^{m}\sigma_k^2 R_k\Big)\Big(\sum_{k=1}^{m} R_k\Big)^{-1}.$$

We show that an optimally weighted least squares estimator is a solution in ϑ of the estimating equation

$$\sum_{j=1}^{n}\sum_{k=1}^{m}\tilde{\sigma}_k^{-2}\dot{r}_{k\vartheta}(\mathbf{X}_{jm+k-1})(X_{jm+k} - r_{k\vartheta}(\mathbf{X}_{jm+k-1})) = 0$$

with $\tilde{\sigma}_k^2 = (1/n)\sum_{j=1}^n \tilde{\varepsilon}_{jm+k}^2$ and $\tilde{\varepsilon}_{jm+k} = X_{jm+k} - r_{k\tilde{\vartheta}}(\mathbf{X}_{jm+k-1})$ for some consistent estimator $\tilde{\vartheta}$ of ϑ. By Taylor expansion, its asymptotic covariance matrix is seen to be the covariance matrix of $(\sum_{k=1}^m \sigma_k^{-2} R_k)^{-1}\sum_{k=1}^m \sigma_k^{-2}\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})\varepsilon_k$,

$$M_{LS*} = \Big(\sum_{k=1}^{m}\sigma_k^{-2} R_k\Big)^{-1}.$$

It is straightforward to check that

$$\gamma = \Big(\sum_{k=1}^{m} R_k\Big)^{-1}\sum_{k=1}^{m}\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})\varepsilon_k - \Big(\sum_{k=1}^{m}\sigma_k^{-2} R_k\Big)^{-1}\sum_{k=1}^{m}\sigma_k^{-2}\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})\varepsilon_k$$

is uncorrelated with $\sum_{k=1}^m \sigma_k^{-2}\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})\varepsilon_k$. Hence we can write

$$M_{LS} = M_{LS*} + \Gamma,$$

where Γ is the covariance matrix of γ. In particular, the optimally weighted least squares estimator is strictly better than the ordinary least squares estimator unless $\gamma = 0$, which holds only if $\sigma_1 = \cdots = \sigma_m$. In Section 3 we also construct an efficient estimator for ϑ as one-step improvement of some initial estimator, for example the least squares estimator. The asymptotic covariance matrix of any efficient estimator is shown to equal M with

$$M^{-1} = \sum_{k=1}^{m}\big(J_k(R_k - \mu_k\mu_k^\top) + \sigma_k^{-2}\mu_k\mu_k^\top\big),$$

where $J_k = E[\ell_k^2(\varepsilon_k)]$ with $\ell_k = -f_k'/f_k$. We note that J_k is the Fisher information of the location family generated by f_k, and $R_k - \mu_k\mu_k^\top$ is the covariance matrix of $\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})$. We obtain

$$M^{-1} = M_{LS*}^{-1} + \sum_{k=1}^{m}(J_k - \sigma_k^{-2})(R_k - \mu_k\mu_k^\top).$$

It is known that $J_k \geq \sigma_k^{-2}$. Hence, in general, the optimally weighted least squares estimator is not efficient, except when $J_k = \sigma_k^{-2}$ for all k, which holds if all the f_k are normal densities. In Section 4 we fix $\nu \in \{1, \ldots, m\}$ and use the efficient estimator for ϑ to construct efficient estimators for expectations of functions of ε_ν. They lead to efficient estimators for the innovation distribution functions and quantile functions. We know that $E[\varepsilon_\nu] = 0$ and obtain an efficient estimator by correcting the residual-based empirical estimator, and by basing the residuals on an efficient estimator for ϑ. The correction can be obtained by adding an "estimator of zero", or by introducing random weights, following the empirical likelihood approach of Owen (1988).

In Section 5 we consider the submodel in which the innovation densities are equal. It turns out that this contains information about ϑ. We construct an efficient estimator for ϑ in this submodel. Its asymptotic covariance matrix is M_* with

$$M_*^{-1} = \sum_{k=1}^{m}\big(J(R_k - \mu_k\mu_k^\top) + 2\sigma^{-2}\mu_k\mu_k^\top - \sigma^{-2}\mu_*\mu_*^\top\big),$$

where J and σ^2 are the Fisher information and variance of the common innovation density f and $\mu_* = (1/m)\sum_{k=1}^m \mu_k$. We have

$$M_*^{-1} = M^{-1} + \sigma^{-2}\sum_{k=1}^{m}(\mu_k - \mu_*)(\mu_k - \mu_*)^\top.$$

Hence the efficient estimator for ϑ in the submodel is, in general, strictly better than the efficient estimator in the full model considered in Section 3. The two estimators are asymptotically equivalent, i.e. $M_* = M$, only if $\mu_1 = \cdots = \mu_m$. We also construct efficient estimators for expectations of functions of ε in this submodel.

2 Local asymptotic normality

The alternating nonlinear autoregressive process (1.1) is parametrized by ϑ and the vector of innovation densities $f = (f_1, \ldots, f_m)$. In order to apply results on non-alternating processes, we view it as an m-dimensional process $\mathbf{Y}_j = (X_{(j-1)m+1}, \ldots, X_{jm})^\top$, $j \in \mathbb{Z}$. This is a homogeneous Markov chain of order $q = \lceil p/m \rceil$. Its transition density from $\mathbf{Y}_{j-q}, \ldots, \mathbf{Y}_{j-1}$ to $\mathbf{Y}_j = (x_1, \ldots, x_m)^\top$ depends only on the values of the last p components of $\mathbf{Y}_{j-q}, \ldots, \mathbf{Y}_{j-1}$, say $\mathbf{x}_0 = (x_{1-p}, \ldots, x_0)$, and is given by

$$q(\mathbf{x}_0; x_1, \ldots, x_m) = \prod_{k=1}^{m} f_k(x_k - r_{k\vartheta}(\mathbf{x}_{k-1})).$$

Note that this is not a multivariate nonlinear autoregressive process of order q, which would require a representation $\mathbf{Y}_j = R_\vartheta(\mathbf{Y}_{j-1}, \ldots, \mathbf{Y}_{j-q}) + \boldsymbol{\varepsilon}_j$ for an m-dimensional function R_ϑ and i.i.d. m-dimensional innovation vectors $\boldsymbol{\varepsilon}_j$.

To prove local asymptotic normality, fix ϑ and f such that $\mathbf{Y}_j$, $j \in \mathbb{Z}$, is strictly stationary and positive Harris recurrent under (ϑ, f). Write g for the stationary density of $\mathbf{X}_j$ under (ϑ, f). Introduce perturbations $\vartheta_{nu} = \vartheta + n^{-1/2}u$ with $u \in \mathbb{R}^d$ and $f_{knv_k}(x) = f_k(x)(1 + n^{-1/2}v_k(x))$ with v_k in the space V_k of bounded measurable functions such that $E[v_k(\varepsilon_k)] = 0$ and $E[\varepsilon_k v_k(\varepsilon_k)] = 0$. These two conditions imply that f_{knv_k} is a positive mean zero probability density for n sufficiently large. Note that if v is a bounded measurable function, then v_k defined by

$$v_k = v - E[v(\varepsilon_k)] - \frac{E[\varepsilon_k v(\varepsilon_k)]}{E[\varepsilon_k w(\varepsilon_k)]}(w - E[w(\varepsilon_k)])$$

belongs to V_k for every bounded measurable function w for which $E[\varepsilon_k w(\varepsilon_k)]$ is not zero. A possible choice of w is given by $w(x) = x\mathbf{1}[|x| \le a]$ with a sufficiently large. Write $v = (v_1, \ldots, v_m)$, $V = V_1 \times \cdots \times V_m$ and $f_{nv} = (f_{1nv_1}, \ldots, f_{mnv_m})$. Suppose that we have observations $\mathbf{X}_0, \mathbf{Y}_1, \ldots, \mathbf{Y}_n$, and write P_n and P_{nuv} for their joint law under (ϑ, f) and (ϑ_{nu}, f_{nv}), respectively. Let g_{nuv} denote the stationary density of $\mathbf{X}_j$ under (ϑ_{nu}, f_{nv}). We make the following assumptions.

Assumption 1. For $k = 1, \ldots, m$, the innovation density f_k is absolutely continuous with a.e. derivative f_k' and finite Fisher information $J_k = E[\ell_k^2(\varepsilon_k)]$, where $\ell_k = -f_k'/f_k$.

Assumption 2. For $k = 1, \ldots, m$, there is $\dot{r}_{k\vartheta} \in L_2^d(g)$ such that, for each constant C,

$$\sup_{|\tau - \vartheta| \le Cn^{-1/2}} \sum_{j=1}^{n} \Big(r_{k\tau}(\mathbf{X}_{jm+k-1}) - r_{k\vartheta}(\mathbf{X}_{jm+k-1}) - (\tau - \vartheta)^\top \dot{r}_{k\vartheta}(\mathbf{X}_{jm+k-1}) \Big)^2 = o_{P_n}(1).$$

Then we have local asymptotic normality as follows. The proof is obtained as in Koul and Schick (1997), who treat the non-alternating case.

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

$$\log \frac{dP_{nuv}}{dP_n} = n^{-1/2} \sum_{j=1}^{n} \sum_{k=1}^{m} s_{kuv_k}(\mathbf{X}_{jm+k-1}, \varepsilon_{jm+k}) - \frac{1}{2}\|(u, v)\|^2 + o_{P_n}(1), \tag{2.1}$$

$$n^{-1/2} \sum_{j=1}^{n} \sum_{k=1}^{m} s_{kuv_k}(\mathbf{X}_{jm+k-1}, \varepsilon_{jm+k}) \Rightarrow \|(u, v)\| N \quad \text{under } P_n, \tag{2.2}$$

where N is a standard normal random variable and

$$s_{kuv_k}(\mathbf{X}_{k-1}, \varepsilon_k) = u^\top \dot{r}_{k\vartheta}(\mathbf{X}_{k-1})\ell_k(\varepsilon_k) + v_k(\varepsilon_k),$$

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

Here we have used that $s_{1uv_1}(\mathbf{X}_0, \varepsilon_1), \ldots, s_{muv_m}(\mathbf{X}_{m-1}, \varepsilon_m)$ are uncorrelated.

A sufficient condition for positive Harris recurrence and geometric ergodicity in L_1 of the m-dimensional Markov chain $\mathbf{Y}_j$, $j \in \mathbb{Z}$, is

$$|r_{k\vartheta}(\mathbf{x})| \leq c_k + \alpha_k |\mathbf{x}|, \quad \mathbf{x} \in \mathbb{R}^p, \quad k = 1, \ldots, m,$$

with $\prod_{k=1}^m \alpha_k < 1$. This follows as in the non-alternating case; see e.g. Bhattacharya and Lee (1995) or An and Huang (1996). Geometric ergodicity implies that at (ϑ, f) the stationary density of $\mathbf{Y}_j$, $j \in \mathbb{Z}$, depends continuously in L_1 on the transition density. This implies the continuity condition on g in Theorem 1 above and in Theorem 4 below.

Let $\bar{V}_k$ denote the closure of V_k in $L_2(f_k)$ and set $\bar{V} = \bar{V}_1 \times \cdots \times \bar{V}_m$. The *tangent space* of our model is

$$S = \Big\{ \sum_{k=1}^m s_{kuv_k}(\mathbf{X}_{k-1}, \varepsilon_k) : (u, v) \in \mathbb{R}^d \times \bar{V} \Big\}.$$

Let T denote the space of random variables $t(\mathbf{Y}_{1-q}, \ldots, \mathbf{Y}_1)$ such that

$$E[t^2(\mathbf{Y}_{1-q}, \ldots, \mathbf{Y}_1)] < \infty \quad \text{and} \quad E(t(\mathbf{Y}_{1-q}, \ldots, \mathbf{Y}_1) | \mathbf{Y}_{1-q}, \ldots, \mathbf{Y}_0) = 0.$$

Then T is a Hilbert space, and S is a closed linear subspace of T.

We can think of T as the tangent space of the larger, nonparametric, model of all homogeneous Markov chains of order q on $\mathbb{R}^m$. In this model, a perturbation of a transition distribution Q is of the form

$$Q_{nt}(\mathbf{y}_{1-q}, \ldots, \mathbf{y}_0, d\mathbf{y}_1) = Q(\mathbf{y}_{1-q}, \ldots, \mathbf{y}_0, d\mathbf{y}_1)(1 + n^{-1/2} t(\mathbf{y}_{1-q}, \ldots, \mathbf{y}_1))$$

with bounded $t(\mathbf{Y}_{1-q}, \ldots, \mathbf{Y}_1) \in T$, and we have local asymptotic normality

$$\log \frac{dP_{nt}}{dP_n} = n^{-1/2} \sum_{j=1}^n t(\mathbf{Y}_{j-q}, \ldots, \mathbf{Y}_j) - \frac{1}{2} E[t^2(\mathbf{Y}_{1-q}, \ldots, \mathbf{Y}_1)] + o_{P_n}(1),$$

so the tangent space is T. We note that $t(\mathbf{Y}_{j-q}, \ldots, \mathbf{Y}_j)$ are martingale increments on the natural filtration. For local asymptotic normality of general Markov chain models and Markov step processes (of order one) see Roussas (1965), Höpfner, Jacod and Ladelli (1990), Penev (1991) and Höpfner (1993).

The norm $\|(u, v)\|$ is the norm induced on $\mathbb{R}^d \times \bar{V}$ by the L_2-norm on $S \subset T$. It determines how difficult it is, asymptotically, to distinguish between (ϑ, f) and (ϑ_{nu}, f_{nv}) on the basis of the observations. It induces an inner product on $\mathbb{R}^d \times \bar{V}$,

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

We can now characterize efficient estimators of real-valued functionals of (ϑ, f) as follows, using results originally due to Hájek and LeCam, for which we refer to Section 3.3 of Bickel et al. (1998).

Definition 1. A real-valued functional φ of (ϑ, f) is called *differentiable* at (ϑ, f) with *gradient* t_φ if $t_\varphi(\mathbf{Y}_{1-q}, \ldots, \mathbf{Y}_1) \in T$ and

$$n^{1/2}(\varphi(\vartheta_{nu}, f_{nv}) - \varphi(\vartheta, f)) \to \sum_{k=1}^{m} E[t_\varphi(\mathbf{Y}_{1-q}, \ldots, \mathbf{Y}_1) s_{kuv_k}(\mathbf{X}_{k-1}, \varepsilon_k)], \quad (u, v) \in \mathbb{R}^d \times V.$$

The *canonical gradient* is the projection of any gradient $t_\varphi(\mathbf{Y}_{1-q}, \ldots, \mathbf{Y}_1)$ onto S.

The canonical gradient is of the form $\sum_{k=1}^{m} s_{ku_\varphi v_{\varphi k}}(\mathbf{X}_{k-1}, \varepsilon_k)$. Since the random variables $s_{1uv_1}(\mathbf{X}_0, \varepsilon_1), \ldots, s_{muv_m}(\mathbf{X}_{m-1}, \varepsilon_m)$ are uncorrelated, it is determined by

$$n^{1/2}(\varphi(\vartheta_{nu}, f_{nv}) - \varphi(\vartheta, f)) \to \sum_{k=1}^{m} E[s_{ku_\varphi v_{\varphi k}}(\mathbf{X}_{k-1}, \varepsilon_k) s_{kuv_k}(\mathbf{X}_{k-1}, \varepsilon_k)] \tag{2.3}$$

for $(u, v) \in \mathbb{R}^d \times V$.

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

$$n^{1/2}(\hat{\varphi} - \varphi(\vartheta_{nu}, f_{nv})) \Rightarrow L \quad \text{under } P_{nuv}, \quad (u, v) \in \mathbb{R}^d \times V.$$

Definition 3. An estimator $\hat{\varphi}$ of φ is called *asymptotically linear* at (ϑ, f) with *influence function* χ if $\chi(\mathbf{Y}_{1-q}, \ldots, \mathbf{Y}_1) \in T$ and

$$n^{1/2}(\hat{\varphi} - \varphi(\vartheta, f)) = n^{-1/2} \sum_{j=1}^{n} \chi(\mathbf{Y}_{j-q}, \ldots, \mathbf{Y}_j) + o_{P_n}(1).$$

Theorem 2. *Suppose we have local asymptotic normality (2.1) and (2.2) at (ϑ, f). Let φ be differentiable at (ϑ, f) with canonical gradient $\sum_{k=1}^{m} s_{ku_\varphi v_{\varphi k}}(\mathbf{X}_{k-1}, \varepsilon_k)$. Let $\hat{\varphi}$ be regular at (ϑ, f) with limit L. Then there is a random variable M independent of N such that $L = \|(u_\varphi, v_\varphi)\| N + M$ in distribution. We have $M = 0$ if and only if $\hat{\varphi}$ is asymptotically linear at (ϑ, f) with influence function equal to the canonical gradient.*

An estimator $\hat{\varphi}$ with limit $L = \|(u_\varphi, v_\varphi)\| N$ at (ϑ, f) is least dispersed in intervals symmetric about zero among all regular estimators of φ. We call such an estimator *efficient* at (ϑ, f).

Theorem 3. *Suppose we have local asymptotic normality (2.1) and (2.2) at (ϑ, f). Let φ be differentiable at (ϑ, f), and let $\hat{\varphi}$ be asymptotically linear for φ at (ϑ, f). Then $\hat{\varphi}$ is regular at (ϑ, f) if and only if its influence function is a gradient of φ at (ϑ, f).*

It follows from Theorems 2 and 3 that an estimator $\hat{\varphi}$ is regular and efficient if and only if it is asymptotically linear with influence function equal to the canonical gradient,

$$n^{1/2}(\hat{\varphi} - \varphi(\vartheta, f)) = n^{-1/2} \sum_{j=1}^{n} \sum_{k=1}^{m} s_{ku_\varphi v_{\varphi k}}(\mathbf{X}_{jm+k-1}, \varepsilon_{jm+k}) + o_{P_n}(1). \tag{2.4}$$

To calculate gradients, it is convenient to decompose $s_{kuv_k}(\mathbf{X}_{k-1}, \varepsilon_k)$ into orthogonal components. Under Assumption 1 we have $E[\varepsilon_k \ell_k(\varepsilon_k)] = 1$. Hence the projection of $\ell_k(\varepsilon_k)$ onto $\bar{V}_k$ is

$$\ell_k^*(\varepsilon_k) = \ell_k(\varepsilon_k) - \frac{E[\varepsilon_k \ell_k(\varepsilon_k)]}{E[\varepsilon_k^2]}\varepsilon_k = \ell_k(\varepsilon_k) - \sigma_k^{-2}\varepsilon_k.$$

Set $\mu_k = E[\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})]$ and

$$s_k(\mathbf{X}_{k-1}, \varepsilon_k) = (\dot{r}_{k\vartheta}(\mathbf{X}_{k-1}) - \mu_k)\ell_k(\varepsilon_k) + \sigma_k^{-2}\mu_k\varepsilon_k.$$

We can write

$$s_{kuv_k}(\mathbf{X}_{k-1}, \varepsilon_k) = u^\top s_k(\mathbf{X}_{k-1}, \varepsilon_k) + u^\top \mu_k \ell_k^*(\varepsilon_k) + v_k(\varepsilon_k). \tag{2.5}$$

By construction, $\ell_k^* \in \bar{V}_k$, and $s_k(\mathbf{X}_{k-1}, \varepsilon_k)$ is orthogonal to $\bar{V}_k$ in the sense that

$$E[s_{kuv_k}(\mathbf{X}_{k-1}, \varepsilon_k)v_k(\varepsilon_k)] = 0, \quad v_k \in \bar{V}_k. \tag{2.6}$$

We arrive at an orthogonal decomposition $S = S_0 + S_V$ of the tangent space, with

$$S_0 = \Big\{ \sum_{k=1}^m u^\top s_k(\mathbf{X}_{k-1}, \varepsilon_k) : u \in \mathbb{R}^p \Big\}, \qquad S_V = \Big\{ \sum_{k=1}^m v_k(\varepsilon_k) : v \in \bar{V} \Big\}.$$

Set

$$\Lambda_k = E[s_k(\mathbf{X}_{k-1}, \varepsilon_k)s_k^\top(\mathbf{X}_{k-1}, \varepsilon_k)] = J_k(R_k - \mu_k\mu_k^\top) + \sigma_k^{-2}\mu_k\mu_k^\top$$

and $\Lambda = \sum_{k=1}^m \Lambda_k$. Relation (2.3) can be rewritten as follows.

Proposition 1. *Let φ be differentiable at (ϑ, f). Then its canonical gradient is of the form*

$$\sum_{k=1}^m u_\varphi^\top s_k(\mathbf{X}_{k-1}, \varepsilon_k) + \sum_{k=1}^m v_{\varphi k}(\varepsilon_k)$$

with $(u_\varphi, v_\varphi) \in \mathbb{R}^d \times \bar{V}$ determined by

$$n^{1/2}(\varphi(\vartheta_{nu}, f_{nv}) - \varphi(\vartheta, f)) \to u_\varphi^\top \Lambda u + \sum_{k=1}^m E[v_{\varphi k}(\varepsilon_k)\ell_k^*(\varepsilon_k)]\mu_k^\top u + \sum_{k=1}^m E[v_{\varphi k}(\varepsilon_k)v_k(\varepsilon_k)]$$

for $(u, v) \in \mathbb{R}^d \times V$.

Remark 1. Alternating *linear* autoregression is a degenerate case. Let $r_{k\vartheta}(\mathbf{X}_{k-1}) = \varrho_{k\vartheta}^\top \mathbf{X}_{k-1}$ for a vector $\varrho_{k\vartheta} = (\varrho_{k\vartheta 1}, \ldots, \varrho_{k\vartheta p})^\top$ of functions of ϑ. Let $\dot{\varrho}_{k\vartheta}$ denote the $d \times p$ matrix whose columns are the gradients of $\varrho_{k\vartheta 1}, \ldots, \varrho_{k\vartheta p}$. Then $\dot{r}_{k\vartheta}(\mathbf{X}_{k-1}) = \dot{\varrho}_{k\vartheta}\mathbf{X}_{k-1}$. Since $E[\varepsilon_k] = 0$, we have $\mu_k = \dot{\varrho}_{k\vartheta}E[\mathbf{X}_{k-1}] = 0$ and hence

$$s_k(\mathbf{X}_{k-1}, \varepsilon_k) = \dot{\varrho}_{k\vartheta}\mathbf{X}_{k-1}\ell_k(\varepsilon_k).$$

We obtain

$$s_{kuv_k}(\mathbf{X}_{k-1}, \varepsilon_k) = u^\top \dot{\varrho}_{k\vartheta} \mathbf{X}_{k-1} \ell_k(\varepsilon_k) + v_k(\varepsilon_k).$$

The canonical gradient of φ is therefore of the form

$$\sum_{k=1}^{m} u_\varphi^\top \dot{\varrho}_{k\vartheta} \mathbf{X}_{k-1} \ell_k(\varepsilon_k) + \sum_{k=1}^{m} v_{\varphi k}(\varepsilon_k)$$

with $(u_\varphi, v_\varphi) \in \mathbb{R}^d \times \bar{V}$ determined by

$$n^{1/2}(\varphi(\vartheta_{nu}, f_{nv}) - \varphi(\vartheta, f)) \to \sum_{k=1}^{m} u_\varphi^\top \dot{\varrho}_{k\vartheta} \Sigma_k \dot{\varrho}_{k\vartheta}^\top u + \sum_{k=1}^{m} E[v_{\varphi k}(\varepsilon_k) v_k(\varepsilon_k)]$$

for $(u, v) \in \mathbb{R}^d \times V$, with $\Sigma_k = E[\mathbf{X}_{k-1}\mathbf{X}_{k-1}^\top]$. This implies that for a functional φ depending on ϑ only we obtain $v_{\varphi k} = 0$. Hence the canonical gradient of such a functional is the same for each submodel in which some or all of the f_k are known. Then we cannot estimate φ better, asymptotically, in these submodels. In this sense, functionals depending on ϑ only are *adaptive* with respect to the parameter f. Similarly, functionals of one or some of the f_k are adaptive with respect to the other parameters.

In the following sections we apply characterization (2.4) to various functionals. A version of the characterization also holds for multivariate functionals as follows. The proof reduces to the case of one-dimensional functionals. Let $\varphi = (\varphi_1, \ldots, \varphi_q)^\top$ be a functional of (ϑ, f). Differentiability of φ is then understood componentwise. The canonical gradient is obtained by componentwise projection of gradients of $\varphi_1, \ldots, \varphi_q$. Regularity of an estimator $\hat{\varphi}$ of φ is defined as before, now with L a q-dimensional random vector. Asymptotic linearity of $\hat{\varphi}$ is understood componentwise. Theorem 2 now says that

$$L = ((u_\varphi, v_\varphi), (u_\varphi, v_\varphi)^\top)^{1/2} N_q + M \quad \text{in distribution,}$$

where $(u_\varphi, v_\varphi) = ((u_{\varphi_1}, v_{\varphi_1}), \ldots, (u_{\varphi_q}, v_{\varphi_q}))^\top$, and where N_q is a q-dimensional standard normal random vector and M is independent of N_q. Theorem 3 remains unchanged, and characterization (2.4) is again understood componentwise.

3 Autoregression parameters

Before we construct an efficient estimator for ϑ, we begin with some results on least squares estimators. An estimator for ϑ is the *least squares estimator* $\tilde{\vartheta}$, the minimum in ϑ of

$$\sum_{j=1}^{n} \sum_{k=1}^{m} (X_{jm+k} - r_{k\vartheta}(\mathbf{X}_{jm+k-1}))^2. \tag{3.1}$$

It is a solution of a q-dimensional martingale estimating equation

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

Under appropriate conditions, the least squares estimator is asymptotically linear with influence function

$$\xi(\mathbf{Y}_{1-q},\ldots,\mathbf{Y}_1)=R^{-1}\sum_{k=1}^{m}\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})\varepsilon_k,$$

where $R=\sum_{k=1}^{m}R_k$ with $R_k=E[\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})\dot{r}_{k\vartheta}^{\top}(\mathbf{X}_{k-1})]$. Hence $\tilde{\vartheta}$ is asymptotically normal with covariance matrix $M_{LS}=R^{-1}\big(\sum_{k=1}^{m}\sigma_k^2R_k\big)R^{-1}$. Here $\mathbf{Y}_j=(X_{(j-1)m+1},\ldots,X_{jm})^{\top}$ is defined as in Section 2.

The least squares estimator can be improved by weighting the martingale increments. Let $W_{k\vartheta}(\mathbf{x})$ be a $d\times d$ matrix of weights and $\tilde{\vartheta}_W$ a solution of the martingale estimating equation

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

Under appropriate conditions, a Taylor expansion shows that $\tilde{\vartheta}_W$ has influence function

$$\xi_W(\mathbf{Y}_{1-q},\ldots,\mathbf{Y}_1)=R_W^{-1}\sum_{k=1}^{m}W_{k\vartheta}(\mathbf{X}_{k-1})\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})\varepsilon_k,$$

where $R_W=\sum_{k=1}^{m}R_{kW}$ with $R_{kW}=E[W_{k\vartheta}(\mathbf{X}_{k-1})\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})\dot{r}_{k\vartheta}^{\top}(\mathbf{X}_{k-1})]$. Hence $\tilde{\vartheta}_W$ is asymptotically normal with covariance matrix $R_W^{-1}Q_WR_W^{-1}$, where

$$Q_W=\sum_{k=1}^{m}\sigma_k^2E[W_{k\vartheta}(\mathbf{X}_{k-1})\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})\dot{r}_{k\vartheta}^{\top}(\mathbf{X}_{k-1})W_{k\vartheta}^{\top}(\mathbf{X}_{k-1})].$$

The covariance matrix is minimized for $W_{k\vartheta}^{*}(\mathbf{x})=\sigma_k^{-2}I_d$ with I_d the $d\times d$ unit matrix. This follows from the fact that $\xi_W-\xi_{W_*}$ is orthogonal to ξ_{W_*}, which in turn is seen by straightforward calculation. We have

$$\xi_{W_*}(\mathbf{Y}_{1-q},\ldots,\mathbf{Y}_1)=R_{W_*}^{-1}\sum_{k=1}^{m}\sigma_k^{-2}\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})\varepsilon_k,$$

where $R_{W_*}=\sum_{k=1}^{m}\sigma_k^{-2}R_k$. An optimal weighted least squares estimator $\tilde{\vartheta}^{*}$ is obtained as solution of the estimating equation

$$\sum_{j=1}^{n}\sum_{k=1}^{m}\tilde{\sigma}_k^{-2}\dot{r}_{k\vartheta}(\mathbf{X}_{jm+k-1})(X_{jm+k}-r_{k\vartheta}(\mathbf{X}_{jm+k-1}))=0,$$

where $\tilde{\sigma}_k^2$ is a consistent estimator of σ_k^2, for example the estimator

$$\tilde{\sigma}_k^2 = \frac{1}{n}\sum_{j=1}^{n} \tilde{\varepsilon}_{jm+k}^2$$

based on residuals $\tilde{\varepsilon}_{jm+k} = X_{jm+k} - r_{k\tilde{\vartheta}}(\mathbf{X}_{jm+k-1})$, with $\tilde{\vartheta}$ the least squares estimator minimizing (3.1). The asymptotic covariance matrix of $\tilde{\vartheta}^*$ is $M_{LS*} = R_{W_*}^{-1}$. The estimator $\tilde{\vartheta}^*$ weights the squared martingale increments in (3.1) by the inverses of their variances, which is a plausible result.

In ordinary (non-alternating) nonlinear autoregression, the least squares estimator is not efficient, except when the innovations are normally distributed. We expect that our optimally weighted least squares estimator is also inefficient. To see this, and to construct an efficient estimator of ϑ, we first determine the canonical gradient of the d-dimensional functional $\varphi(\vartheta, f) = \vartheta$, for which

$$n^{1/2}(\varphi(\vartheta_{nu}, f_{nv}) - \varphi(\vartheta, f)) = u.$$

Assume that $\Lambda = \sum_{k=1}^{m} \Lambda_k$ is positive definite. From the d-dimensional version of Proposition 1, the canonical gradient of ϑ is obtained as

$$\Lambda^{-1}\sum_{k=1}^{m} s_k(\mathbf{X}_{k-1}, \varepsilon_k).$$

This is different from the influence function ξ_{W_*} of the optimally weighted least squares estimator $\tilde{\vartheta}^*$, except in the following case. Suppose that for $k = 1, \ldots, m$ the innovation densities f_k are normal with mean zero and variance σ_k^2. Then $\ell_k(\varepsilon_k) = \sigma_k^{-2}\varepsilon_k$ and $s_k(\mathbf{X}_{k-1}, \varepsilon_k) = \sigma_k^{-2}\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})\varepsilon_k$. Hence $\Lambda_k = \sigma_k^{-2}R_k$ and $\Lambda = \sum_{k=1}^{m}\Lambda_k = R_{W_*}$, and the canonical gradient equals ξ_{W_*}.

As in Koul and Schick (1997), Section 6, we obtain an efficient estimator for ϑ under additional conditions on $\dot{r}_{k\vartheta}$ as follows. Let $\tilde{\vartheta}$ be root-n consistent and *discretized*, i.e. with values on a rectangular grid with side lengths of order $n^{-1/2}$. For $c = c_n \to \infty$ introduce the truncation

$$\bar{x} = x\mathbf{1}[|x| \le c] + c\frac{x}{|x|}\mathbf{1}[|x| > c], \quad x \in \mathbb{R}^d.$$

Estimate $\mu_k = E[\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})]$ and $R_k = E[\dot{r}_{k\vartheta}(\mathbf{X}_{k-1})\dot{r}_{k\vartheta}^{\top}(\mathbf{X}_{k-1})]$ by truncated empirical estimators

$$\tilde{\mu}_k = \frac{1}{n}\sum_{j=1}^{n} \bar{\dot{r}}_{k\tilde{\vartheta}}(\mathbf{X}_{jm+k-1}), \qquad \tilde{R}_k = \frac{1}{n}\sum_{j=1}^{n} \bar{\dot{r}}_{k\tilde{\vartheta}}(\mathbf{X}_{jm+k-1})\bar{\dot{r}}_{k\tilde{\vartheta}}^{\top}(\mathbf{X}_{jm+k-1}).$$

Estimate σ_k^2 by $\tilde{\sigma}_k^2 = (1/n)\sum_{j=1}^{n}\tilde{\varepsilon}_{jm+k}^2$. Let K be a kernel fulfilling Condition K of Schick (1993), for example the logistic density. For a bandwidth $b = b_n \to 0$, set $K_b(x) = K(x/b)/b$.

Then $K_b'(x) = K'(x/b)/b^2$. Estimate f_k and f_k' by

$$\tilde{f}_k(x) = \frac{1}{n}\sum_{j=1}^{n} K_b(x - \tilde{\varepsilon}_{jm+k}), \qquad \tilde{f}_k'(x) = \frac{1}{n}\sum_{j=1}^{n} K_b'(x - \tilde{\varepsilon}_{jm+k}).$$

Let $a = a_n \downarrow 0$ and estimate ℓ_k and J_k by

$$\tilde{\ell}_k = -\frac{\tilde{f}_k'}{\tilde{f}_k + a}, \qquad \tilde{J}_k = \frac{1}{n}\sum_{j=1}^{n} \tilde{\ell}_k^2(\tilde{\varepsilon}_{jm+k}).$$

Our estimator for ϑ is now obtained by the *Newton–Raphson procedure*, a one-step improvement of $\tilde{\vartheta}$, as

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

with $\tilde{\Lambda} = \sum_{k=1}^{m} \tilde{\Lambda}_k$ and

$$\tilde{\Lambda}_k = \tilde{J}_k(\tilde{R}_k - \tilde{\mu}_k\tilde{\mu}_k^\top) + \tilde{\sigma}_k^{-2}\tilde{\mu}_k\tilde{\mu}_k^\top,$$
$$\tilde{s}_k(\mathbf{X}_{k-1}, \varepsilon_k) = (\bar{\dot{r}}_{k\tilde{\vartheta}}(\mathbf{X}_{k-1}) - \tilde{\mu}_k)\tilde{\ell}_k(\varepsilon_k) + \tilde{\sigma}_k^{-2}\tilde{\mu}_k\varepsilon_k.$$

For appropriate choices of a, b and c, the influence function of the estimator $\hat{\vartheta}$ equals the canonical gradient; hence $\hat{\vartheta}$ is efficient for ϑ. This follows as in the non-alternating case, Koul and Schick (1997), which in turn uses results of Schick (1987). The asymptotic covariance matrix of $\hat{\vartheta}$ is $M = \Lambda^{-1}$.

Remark 2. The case of *equal* autoregression functions does not lead to noticeable simplifications. The expectations $\mu_k = E[\dot{r}_\vartheta(\mathbf{X}_{k-1})]$ and the covariance matrices

$$R_k = E[\dot{r}_\vartheta(\mathbf{X}_{k-1})\dot{r}_\vartheta^\top(\mathbf{X}_{k-1})]$$

still depend on k. The optimally weighted least squares estimator now solves

$$\sum_{j=1}^{n}\sum_{k=1}^{m} \tilde{\sigma}_k^{-2}\dot{r}_\vartheta(\mathbf{X}_{jm+k-1})(X_{jm+k} - r_\vartheta(\mathbf{X}_{jm+k-1})) = 0,$$

and its asymptotic covariance matrix is $(\sum_{k=1}^{m}\sigma_k^{-2}R_k)^{-1}$. The efficient estimator for ϑ remains unchanged except that now $\dot{r}_{k\vartheta} = \dot{r}_\vartheta$ for $k = 1, \ldots, m$.

Remark 3. The case of *linear* autoregression functions $r_{k\vartheta}(\mathbf{X}_{k-1}) = \varrho_{k\vartheta}^\top\mathbf{X}_{k-1}$ leads to considerable simplifications. We have

$$s_k(\mathbf{X}_{k-1}, \varepsilon_k) = \dot{\varrho}_{k\vartheta}\mathbf{X}_{k-1}, \qquad \Lambda_k = J_k\dot{\varrho}_{k\vartheta}\Sigma_k\dot{\varrho}_{k\vartheta}^\top,$$

and we can take residuals $\tilde{\varepsilon}_{jm+k} = X_{jm+k} - \varrho_{k\tilde{\vartheta}}^\top\mathbf{X}_{jm+k-1}$. An efficient estimator of ϑ is now obtained as

$$\hat{\vartheta} = \tilde{\vartheta} + \Big(\sum_{k=1}^{m}\tilde{J}_k\dot{\varrho}_{k\tilde{\vartheta}}\tilde{\Sigma}_k\dot{\varrho}_{k\tilde{\vartheta}}^\top\Big)^{-1}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{m}\dot{\varrho}_{k\tilde{\vartheta}}\mathbf{X}_{jm+k-1}\tilde{\ell}_k(\tilde{\varepsilon}_{jm+k}),$$

where

$$\tilde{\Sigma}_k = \frac{1}{n}\sum_{j=1}^{n} \mathbf{X}_{jm+k-1}\mathbf{X}_{jm+k-1}^\top.$$

Compare this with Remark 1. For alternating autoregression of order $p = 1$ see Müller, Schick and Wefelmeyer (2007).

4 Innovation distributions

In this section we fix one of the indices $\nu \in \{1, \dots, m\}$ within a period and consider estimators of linear functionals $\varphi(f_\nu) = E[h(\varepsilon_\nu)] = \int h(x) f_\nu(x)\, dx$.

If ϑ were known, a simple estimator would be $\mathbb{E}h = (1/n)\sum_{j=1}^n h(\varepsilon_{jm+\nu})$, the empirical estimator based on the innovations. Since $E[\varepsilon_\nu] = 0$, we obtain new unbiased estimators $\mathbb{E}h_a$ with $h_a(x) = h(x) - ax$. Their asymptotic variance is minimized for $a = a_* = \sigma_\nu^{-2} E[\varepsilon_\nu h(\varepsilon_{jm+\nu})]$. Since a_* depends on the unknown distribution of ε_ν, we must replace it by an estimator, for example a ratio of empirical estimators, and arrive at the estimator

$$\frac{1}{n}\sum_{j=1}^{n} h(\varepsilon_{jm+\nu}) - \frac{\sum_{j=1}^n \varepsilon_{jm+\nu} h(\varepsilon_{jm+\nu})}{\sum_{j=1}^n \varepsilon_{jm+\nu}^2} \frac{1}{n}\sum_{j=1}^{n} \varepsilon_{jm+\nu} \tag{4.1}$$

for $E[h(\varepsilon_\nu)]$, which is known to be efficient.

Here we have improved the empirical estimator by an additive correction. Following Owen (1988, 2001), we can also choose random weights $w_{jm+\nu}$ such that the weighted empirical distribution has mean zero, $\sum_{j=1}^n w_{jm+\nu}\varepsilon_{jm+\nu} = 0$, and estimate $E[h(\varepsilon_\nu)]$ by the *weighted empirical estimator*

$$\frac{1}{n}\sum_{j=1}^{n} w_{jm+\nu} h(\varepsilon_{jm+\nu}).$$

By the method of Lagrange multipliers, the weights are seen to be of the form $w_{jm+\nu} = 1/(1 + \lambda_\nu \varepsilon_{jm+\nu})$. This implies $\lambda_\nu = \sigma_\nu^{-2}(1/n)\sum_{j=1}^n \varepsilon_{jm+\nu} + o_{P_n}(n^{-1/2})$ and therefore

$$\frac{1}{n}\sum_{j=1}^{n} w_{jm+\nu} h(\varepsilon_{jm+\nu}) = \frac{1}{n}\sum_{j=1}^{n} h(\varepsilon_{jm+\nu}) - \sigma_\nu^{-2} E[\varepsilon_\nu h(\varepsilon_\nu)] \frac{1}{n}\sum_{j=1}^{n} \varepsilon_{jm+\nu} + o_{P_n}(n^{-1/2}).$$

Hence the weighted empirical estimator is asymptotically equivalent to the additively corrected empirical estimator (4.1).

However, we do not know ϑ and must replace the innovations $\varepsilon_{jm+\nu}$ by residuals $\hat{\varepsilon}_{jm+\nu} = X_{jm+\nu} - r_{\nu\hat{\vartheta}}(\mathbf{X}_{jm+\nu-1})$ for some estimator $\hat{\vartheta}$. By the so-called plug-in principle, see e.g. Müller, Schick and Wefelmeyer (2001) and Klaassen and Putter (2005), we expect to obtain an efficient estimator for $E[h(\varepsilon_\nu)]$ as

$$\hat{\varphi}_a = \frac{1}{n}\sum_{j=1}^{n} h(\hat{\varepsilon}_{jm+\nu}) - \frac{\sum_{j=1}^n \hat{\varepsilon}_{jm+\nu} h(\hat{\varepsilon}_{jm+\nu})}{\sum_{j=1}^n \hat{\varepsilon}_{jm+\nu}^2} \frac{1}{n}\sum_{j=1}^{n} \hat{\varepsilon}_{jm+\nu}$$

if we use an efficient estimator $\hat{\vartheta}$ for the residuals.

Again, instead of $\hat{\varphi}_a$ we can use a *weighted residual-based empirical estimator*

$$\hat{\varphi}_w = \frac{1}{n}\sum_{j=1}^{n} \hat{w}_{jm+\nu} h(\hat{\varepsilon}_{jm+\nu})$$

with random weights $\hat{w}_{jm+\nu}$ determined by $\sum_{j=1}^{n} \hat{w}_{jm+\nu}\hat{\varepsilon}_{jm+\nu} = 0$. It is asymptotically equivalent to $\hat{\varphi}_a$ by similar arguments as above; see Müller, Schick and Wefelmeyer (2005).

Assumption 3. Let $h \in L_2(f_\nu)$ be absolutely continuous with $h' \in L_2(f_\nu)$ and

$$\int \sup_{|a|\le\eta} (h'(x-a) - h'(x))^2 f_\nu(x)\, dx \to 0 \quad \text{as } \eta \to \infty.$$

Set $h_*(x) = h(x) - a_* x$. By Taylor expansion, compare Schick and Wefelmeyer (2002), the influence function of $\hat{\varphi}_a$ and $\hat{\varphi}_w$ is seen to be

$$-E[h_*'(\varepsilon_\nu)]\mu_\nu^\top \Lambda^{-1} \sum_{k=1}^{m} s_k(\mathbf{X}_{k-1}, \varepsilon_k) + h_*(\varepsilon_\nu) - E[h_*(\varepsilon_\nu)].$$

By Theorem 2, an estimator $\hat{\varphi}$ is efficient if its influence function equals the canonical gradient of $\varphi(f_\nu) = E[h(\varepsilon_\nu)]$. To determine the canonical gradient, we note first that for $v \in V_\nu$ we have

$$n^{1/2}(\varphi(f_{\nu nv}) - \varphi(f_\nu)) = E[h(\varepsilon_\nu)v(\varepsilon_\nu)].$$

On the right-hand side, we can replace $h(\varepsilon_\nu)$ by its projection $h_*(\varepsilon_\nu) - E[h_*(\varepsilon_\nu)]$ onto $\bar{V}_\nu$. By Proposition 1, the canonical gradient of $E[h(\varepsilon_\nu)]$ is seen to be of the form

$$\sum_{k=1}^{m} u_\varphi^\top s_k(\mathbf{X}_{k-1}, \varepsilon_k) + h_*(\varepsilon_\nu) - E[h_*(\varepsilon_\nu)]$$

with u_φ so that

$$u_\varphi^\top \Lambda + E[h_*(\varepsilon_\nu)\ell_\nu^*(\varepsilon_\nu)]\mu_\nu^\top = 0.$$

Hence the canonical gradient is

$$-E[h_*(\varepsilon_\nu)\ell_\nu^*(\varepsilon_\nu)]\mu_\nu^\top \Lambda^{-1} \sum_{k=1}^{m} s_k(\mathbf{X}_{k-1}, \varepsilon_k) + h_*(\varepsilon_\nu) - E[h_*(\varepsilon_\nu)].$$

Assumptions 1 and 3 imply in particular that $E[h_*'(\varepsilon_\nu)] = E[h_*(\varepsilon_\nu)\ell_k(\varepsilon_\nu)]$. Hence

$$E[h_*(\varepsilon_\nu)\ell_\nu^*(\varepsilon_\nu)] = E[h_*(\varepsilon_\nu)\ell_\nu(\varepsilon_\nu)] = E[h_*'(\varepsilon_\nu)],$$

and the canonical gradient is seen to be equal to the influence function of $\hat{\varphi}_a$ and $\hat{\varphi}_w$, which are therefore efficient.

Remark 4. We have assumed that h is absolutely continuous. This excludes the interesting case $h(x) = \mathbf{1}[x \le t]$, for which $E[h(\varepsilon_\nu)]$ equals the distribution function $F_\nu(t)$ at t of the innovation density f_ν. If we assume that f_ν is uniformly continuous, then we also obtain uniform stochastic expansions for the additively corrected residual-based empirical distribution function

$$\hat F_a(t) = \frac{1}{n}\sum_{j=1}^n \mathbf{1}[\hat\varepsilon_{jm+\nu} \le t] - \frac{\sum_{j=1}^n \hat\varepsilon_{jm+\nu}\mathbf{1}[\hat\varepsilon_{jm+\nu} \le t]}{\sum_{j=1}^n \hat\varepsilon_{jm+\nu}^2}\frac{1}{n}\sum_{j=1}^n \hat\varepsilon_{jm+\nu}$$

and for the weighted residual-based empirical distribution function

$$\hat F_w(t) = \frac{1}{n}\sum_{j=1}^n \hat w_{jm+\nu}\mathbf{1}[\hat\varepsilon_{jm+\nu} \le t].$$

See Schick and Wefelmeyer (2002). For results on smoothed versions of $\hat F_w$ see also Müller, Schick and Wefelmeyer (2005, Section 4). By Gill (1989), the quantile function is compactly differentiable, and hence we obtain a stochastic expansion of the estimators $\hat F_a$ and $\hat F_w$ for the quantile function F_ν^{-1}.

Remark 5. If the autoregression function $r_{\nu\vartheta}$ is linear, $r_{\nu\vartheta}(\mathbf{X}_{\nu-1}) = \varrho_{\nu\vartheta}^\top \mathbf{X}_{\nu-1}$, then, as already noted in Remark 1,

$$\mu_\nu = E[\dot r_{\nu\vartheta}(\mathbf{X}_{\nu-1})] = \dot\varrho_{\nu\vartheta} E[\mathbf{X}_{\nu-1}] = 0.$$

Let $\tilde\varepsilon_{jm+\nu} = X_{jm+\nu} - \varrho_{\nu\tilde\vartheta}^\top \mathbf{X}_{jm+\nu-1}$. Estimators of $E[h(\varepsilon_\nu)]$ are

$$\tilde\varphi_a = \frac{1}{n}\sum_{j=1}^n h(\tilde\varepsilon_{jm+\nu}) - \frac{\sum_{j=1}^n \tilde\varepsilon_{jm+\nu} h(\tilde\varepsilon_{jm+\nu})}{\sum_{j=1}^n \tilde\varepsilon_{jm+\nu}^2}\frac{1}{n}\sum_{j=1}^n \tilde\varepsilon_{jm+\nu}$$

and

$$\tilde\varphi_w = \frac{1}{n}\sum_{j=1}^n \tilde w_{jm+\nu} h(\tilde\varepsilon_{jm+\nu})$$

with random weights $\tilde w_{jm+\nu}$ determined by $\sum_{j=1}^n \tilde w_{jm+\nu}\tilde\varepsilon_{jm+\nu} = 0$. By Taylor expansion, the influence function of $\tilde\varphi_a$ and $\tilde\varphi_w$ is seen to be $h_*(\varepsilon_\nu) - E[h_*(\varepsilon_\nu)]$ and does not depend on the choice of $\tilde\vartheta$. Similarly, the canonical gradient of $E[h(\varepsilon_\nu)]$ reduces to $h_*(\varepsilon_\nu) - E[h_*(\varepsilon_\nu)]$. Hence $\tilde\varphi_a$ and $\tilde\varphi_w$ are efficient even if an inefficient estimator of ϑ is used. Compare this with Remark 1. For alternating autoregression of order $p = 1$ see Müller, Schick and Wefelmeyer (2007).

5 Equal innovation densities

In this section we study the submodel in which all innovation densities are equal, $f_1 = \cdots = f_m = f$, with mean 0 and variance σ^2. To prove local asymptotic normality, we proceed as

in Section 2, now with perturbations $f_{nv}(x) = f(x)(1 + n^{-1/2}v(x))$ with v in the space V_* of bounded measurable functions such that $E[v(\varepsilon)] = 0$ and $E[\varepsilon v(\varepsilon)] = 0$.

Assumption 4. The innovation density f is absolutely continuous with a.e. derivative f' and finite Fisher information $J = E[\ell^2(\varepsilon)]$, where $\ell = -f'/f$.

Theorem 4. *Let $(u, v) \in \mathbb{R}^d \times V_*$. Suppose Assumptions 4 and 2 hold and the stationary density g depends smoothly on the parameters in the sense that $\int |g_{nuv}(\mathbf{x}) - g(\mathbf{x})|\, d\mathbf{x} \to 0$. Then*

$$\log \frac{dP_{nuv}}{dP_n} = n^{-1/2} \sum_{j=1}^{n} \sum_{k=1}^{m} s_{kuv}(\mathbf{X}_{jm+k-1}, \varepsilon_{jm+k}) - \frac{1}{2}\|(u, v)\|^2 + o_{P_n}(1), \tag{5.1}$$

$$n^{-1/2} \sum_{j=1}^{n} \sum_{k=1}^{m} s_{kuv}(\mathbf{X}_{jm+k-1}, \varepsilon_{jm+k}) \Rightarrow \|(u, v)\| N \quad \textit{under } P_n, \tag{5.2}$$

where N is a standard normal random variable and

$$s_{kuv}(\mathbf{X}_{k-1}, \varepsilon_k) = u^\top \dot{r}_{k\vartheta}(\mathbf{X}_{k-1})\ell(\varepsilon_k) + v(\varepsilon_k),$$
$$\|(u, v)\|^2 = \sum_{k=1}^{m} E[s^2_{kuv}(\mathbf{X}_{k-1}, \varepsilon_k)].$$

Let $\bar{V}_*$ denote the closure of V_* in $L_2(f)$. The *tangent space* of the model is

$$S_* = \Big\{ \sum_{k=1}^{m} s_{kuv}(\mathbf{X}_{k-1}, \varepsilon_k) : (u, v) \in \mathbb{R}^d \times \bar{V}_* \Big\}.$$

The tangent space corresponding to *known* ϑ is

$$S_{V_*} = \Big\{ \sum_{k=1}^{m} v(\varepsilon_k) : v \in \bar{V}_* \Big\}.$$

Of course, S_{V_*} is a subspace of $S_V = \{\sum_{k=1}^m v_k(\varepsilon_k) : (v_1, \ldots, v_m)^\top \in \bar{V}\}$, which is the tangent space corresponding to known ϑ but possibly different innovation densities and was introduced in Section 2.

A real-valued functional φ of (ϑ, f) is *differentiable* at (ϑ, f) with *canonical gradient* t_φ if t_φ is of the form $\sum_{k=1}^m s_{ku_\varphi v_\varphi}(\mathbf{X}_{k-1}, \varepsilon_k)$ with $(u_\varphi, v_\varphi) \in \mathbb{R}^d \times \bar{V}_*$ and

$$n^{1/2}(\varphi(\vartheta_{nu}, f_{nv}) - \varphi(\vartheta, f)) \to \sum_{k=1}^{m} E[s_{ku_\varphi v_\varphi}(\mathbf{X}_{k-1}, \varepsilon_k) s_{kuv}(\mathbf{X}_{k-1}, \varepsilon_k)] \tag{5.3}$$

for $(u, v) \in \mathbb{R}^d \times V_*$.

The projection of $\ell(\varepsilon)$ onto $\bar{V}_*$ is $\ell_*(\varepsilon) = \ell(\varepsilon) - \sigma^{-2}\varepsilon$. Set

$$s_k(\mathbf{X}_{k-1}, \varepsilon_k) = (\dot{r}_{k\vartheta}(\mathbf{X}_{k-1}) - \mu_k)\ell(\varepsilon_k) + \sigma^{-2}\mu_k\varepsilon_k.$$

As in Section 2 we have the orthogonal decomposition

$$s_{kuv}(\mathbf{X}_{k-1}, \varepsilon_k) = u^\top s_k(\mathbf{X}_{k-1}, \varepsilon_k) + u^\top \mu_k \ell_*(\varepsilon_k) + v(\varepsilon_k).$$

However, $\sum_{k=1}^m u^\top \mu_k \ell_*(\varepsilon_k)$ is in S_V, but not in S_{V_*}. In order to obtain an orthogonal decomposition of the tangent space S_*, we must project $\sum_{k=1}^m \mu_k \ell_*(\varepsilon_k)$ onto S_{V_*}. In general, the projection of $\sum_{k=1}^m v_k(\varepsilon_k) \in S_V$ onto S_{V_*} is $\sum_{k=1}^m v_*(\varepsilon_k)$ with $v_*(\varepsilon) = (1/m)\sum_{k=1}^m v_k(\varepsilon)$. Hence the projection of $\sum_{k=1}^m \mu_k \ell_*(\varepsilon_k)$ onto S_{V_*} is $\mu_* \sum_{k=1}^m \ell_*(\varepsilon_k)$ with

$$\mu_* = \frac{1}{m}\sum_{k=1}^m \mu_k.$$

We arrive at the orthogonal decomposition

$$\sum_{k=1}^m s_{kuv}(\mathbf{X}_{k-1}, \varepsilon_k) = \sum_{k=1}^m u^\top s_k^*(\mathbf{X}_{k-1}, \varepsilon_k) + \sum_{k=1}^m u^\top \mu_* \ell_*(\varepsilon_k) + \sum_{k=1}^m v(\varepsilon_k)$$

with

$$s_k^*(\mathbf{X}_{k-1}, \varepsilon_k) = s_k(\mathbf{X}_{k-1}, \varepsilon_k) + (\mu_k - \mu_*)\ell_*(\varepsilon_k) = (\dot{r}_{k\vartheta}(\mathbf{X}_{k-1}) - \mu_*)\ell(\varepsilon_k) + \sigma^{-2}\mu_*\varepsilon_k.$$

This implies an orthogonal decomposition $S_* = S_0^* + S_{V_*}$ of the tangent space, with

$$S_0^* = \Big\{ \sum_{k=1}^m u^\top s_k^*(\mathbf{X}_{k-1}, \varepsilon_k) : u \in \mathbb{R}^d \Big\}.$$

Set

$$\begin{aligned}
\Lambda_k^* &= E[s_k^*(\mathbf{X}_{k-1}, \varepsilon_k) s_k^{*\top}(\mathbf{X}_{k-1}, \varepsilon_k)] \\
&= \Lambda_k + \sigma^{-2}(\mu_k - \mu_*)(\mu_k - \mu_*)^\top \\
&= J(R_k - \mu_k \mu_k^\top) + 2\sigma^{-2}\mu_k\mu_k^\top - \sigma^{-2}\mu_*\mu_*^\top
\end{aligned}$$

and $\Lambda_* = \sum_{k=1}^m \Lambda_k^*$. We rewrite (5.3) as follows.

Proposition 2. *Let φ be differentiable at (ϑ, f). Then its canonical gradient is of the form*

$$\sum_{k=1}^m u_\varphi^\top s_k^*(\mathbf{X}_{k-1}, \varepsilon_k) + \sum_{k=1}^m v_\varphi(\varepsilon_k)$$

with $(u_\varphi, v_\varphi) \in \mathbb{R}^d \times \bar{V}_*$ determined by

$$n^{1/2}(\varphi(\vartheta_{nu}, f_{nv}) - \varphi(\vartheta, f)) \to u_\varphi^\top \Lambda_* u + mE[v_\varphi(\varepsilon)\ell_*(\varepsilon)]\mu_*^\top u + mE[v_\varphi(\varepsilon)v(\varepsilon)]$$

for $(u, v) \in \mathbb{R}^d \times V_*$.

Autoregression parameters. The least squares estimator ignores the information of equal innovation densities and remains unchanged. Since now $\sigma_1 = \cdots = \sigma_m = \sigma$, the optimally weighted least squares estimator is asymptotically equivalent to the unweighted one. From the d-dimensional version of Proposition 2, the canonical gradient of ϑ is obtained as $\Lambda_*^{-1} \sum_{k=1}^m s_k^*(\mathbf{X}_{k-1}, \varepsilon_k)$. Introduce residuals $\tilde{\varepsilon}_{jm+k} = X_{jm+k} - r_{k\tilde{\vartheta}}(\mathbf{X}_{jm+k-1})$ for some estimator $\tilde{\vartheta}$. We can now estimate f and f' by

$$\tilde{f}(x) = \frac{1}{nm} \sum_{i=1}^{nm} K_b(x - \tilde{\varepsilon}_i), \qquad \tilde{f}'(x) = \frac{1}{nm} \sum_{i=1}^{nm} K_b'(x - \tilde{\varepsilon}_i),$$

and ℓ, J and σ^2 by

$$\tilde{\ell} = \frac{\tilde{f}'}{\tilde{f} + a}, \quad \tilde{J} = \frac{1}{nm} \sum_{i=1}^{nm} \tilde{\ell}^2(\tilde{\varepsilon}_i), \quad \tilde{\sigma}^2 = \frac{1}{nm} \sum_{i=1}^{nm} \tilde{\varepsilon}_i^2.$$

We estimate μ_* by $\tilde{\mu}_* = \frac{1}{m} \sum_{k=1}^m \tilde{\mu}_k$, and Λ_* and s_k^* by

$$\tilde{\Lambda}_* = \tilde{J} \sum_{k=1}^m (\tilde{R}_k - \tilde{\mu}_k \tilde{\mu}_k^\top) + 2\tilde{\sigma}^{-2} \sum_{k=1}^m \tilde{\mu}_k \tilde{\mu}_k^\top - \tilde{\sigma}^{-2} \sum_{k=1}^m \tilde{\mu}_* \tilde{\mu}_*^\top,$$
$$\tilde{s}_k^*(\mathbf{X}_{k-1}, \varepsilon_k) = (\bar{\dot{r}}_{k\tilde{\vartheta}}(\mathbf{X}_{k-1}) - \tilde{\mu}_*) \tilde{\ell}(\varepsilon_k) + \tilde{\mu}_* \tilde{\sigma}^{-2} \varepsilon_k.$$

Then the one-step improvement of a root-n consistent and discretized initial estimator $\tilde{\vartheta}$ is

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

For appropriate choices of a, b and c, this estimator is efficient for ϑ and has asymptotic covariance matrix $M_* = \Lambda_*^{-1}$. The covariance bound Λ_*^{-1} is strictly smaller than Λ^{-1}, in general, with $\Lambda_* - \Lambda = \sigma^{-2} \sum_{k=1}^m (\mu_k - \mu_*)(\mu_k - \mu_*)^\top$. So equality of the innovation densities carries information about ϑ, except when $\mu_1 = \cdots = \mu_m = \mu_*$. The latter holds of course in alternating *linear* autoregression, for which $\mu_1 = \cdots = \mu_m = 0$.

Innovation distribution. In this subsection we consider estimation of a linear functional $\varphi(f) = E[h(\varepsilon)] = \int h(x) f(x)\, dx$. We can now base it on all residuals. As in Section 4, we expect the estimator

$$\hat{\varphi}_a = \frac{1}{nm} \sum_{i=1}^{nm} h(\hat{\varepsilon}_i) - \frac{\sum_{i=1}^{nm} \hat{\varepsilon}_i h(\hat{\varepsilon}_i)}{\sum_{i=1}^{nm} \hat{\varepsilon}_i^2} \frac{1}{nm} \sum_{i=1}^{nm} \hat{\varepsilon}_i$$

to be efficient for $E[h(\varepsilon)]$ if an efficient estimator $\hat{\vartheta}$ is used for the residuals. Alternatively, we can use the weighted residual-based empirical estimator

$$\hat{\varphi}_w = \frac{1}{nm} \sum_{i=1}^{nm} \hat{w}_i h(\hat{\varepsilon}_i)$$

with random weights $\hat{w}_i$ chosen such that $\sum_{i=1}^{nm} \hat{w}_i \hat{\varepsilon}_i = 0$.

Assumption 5. Let $h \in L_2(f)$ be absolutely continuous with $h' \in L_2(f)$ and

$$\int \sup_{|a| \le \eta} (h'(x-a) - h'(x))^2 f(x)\, dx \to 0 \quad \text{as } \eta \to \infty.$$

Set $h_*(x) = h(x) - a_* x$ with $a_* = \sigma^{-2} E[\varepsilon h(\varepsilon)]$. By Taylor expansion, $\hat{\varphi}_a$ and $\hat{\varphi}_w$ are seen to have influence function

$$-E[h_*'(\varepsilon)]\mu_*^\top \Lambda^{-1} \sum_{k=1}^m s_k^*(\mathbf{X}_{k-1}, \varepsilon_k) + \frac{1}{m}\sum_{k=1}^m h_*(\varepsilon_k) - E[h_*(\varepsilon)].$$

We must show that this is the canonical gradient of $E[h(\varepsilon)]$. To determine the latter, we note first that for $v \in V_*$ we have

$$n^{1/2}(\varphi(f_{nv}) - \varphi(f)) = E[h(\varepsilon)v(\varepsilon)].$$

On the right-hand side, we can replace $h(\varepsilon)$ by its projection $h_*(\varepsilon) - E[h_*(\varepsilon)]$ onto $\bar{V}_*$. By Proposition 2, the canonical gradient of $E[h(\varepsilon)]$ is seen to be of the form

$$\sum_{k=1}^m u_\varphi^\top s_k^*(\mathbf{X}_{k-1}, \varepsilon_k) + \frac{1}{m}\sum_{k=1}^m h_*(\varepsilon_k) - E[h_*(\varepsilon)]$$

with u_φ so that

$$u_\varphi^\top \Lambda_* + E[h_*(\varepsilon)\ell_*(\varepsilon)]\mu_*^\top = 0.$$

Hence the canonical gradient is

$$-E[h_*(\varepsilon)\ell_*(\varepsilon)]\mu_*^\top \Lambda^{-1} \sum_{k=1}^m s_k^*(\mathbf{X}_{k-1}, \varepsilon_k) + \frac{1}{m}\sum_{k=1}^m h_*(\varepsilon_k) - E[h_*(\varepsilon)].$$

Assumptions 1 and 5 imply in particular that $E[h_*'(\varepsilon)] = E[h_*(\varepsilon)\ell(\varepsilon)]$. Hence

$$E[h_*(\varepsilon)\ell_*(\varepsilon)] = E[h_*(\varepsilon)\ell(\varepsilon)] = E[h_*'(\varepsilon)],$$

and the canonical gradient is seen to be equal to the influence function of $\hat{\varphi}_a$ and $\hat{\varphi}_w$, which are therefore efficient.

References

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

Bhattacharya, R. and Lee, C. (1995). On geometric ergodicity of nonlinear autoregressive models. *Statist. Probab. Lett.* **22**, 311–315. Erratum: **41** (1999), 439–440.

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

Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method. I. With a discussion by J. A. Wellner and J. Præstgaard and a reply by the author. *Scand. J. Statist.* **16**, 97–128.

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.

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.

Klaassen, C. A. J. and Putter, H. (2005). Efficient estimation of Banach parameters in semiparametric models. *Ann. Statist.* **33**, 307–346.

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

Müller, U. U., Schick, A. and Wefelmeyer, W. (2001). Plug-in estimators in semiparametric stochastic process models. In: *Selected Proceedings of the Symposium on Inference in Stochastic Processes* (I. V. Basawa, C. C. Heyde and R. L. Taylor, eds), 213-234, IMS Lecture Notes–Monograph Series 37, Institute of Mathematical Statistics, Hayward, California.

Müller, U. U., Schick, A. and Wefelmeyer, W. (2005). Weighted residual-based density estimators for nonlinear autoregressive models. *Statist. Sinica* **15**, 177–195.

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.

Müller, U. U., Schick, A. and Wefelmeyer, W. (2008). Estimators for partially observed Markov chains. In: *Statistical Models and Methods for Biomedical and Technical Systems* (F. Vonta, M. Nikulin, N. Limnios and C. Huber, eds.), 419–434, Birkhäuser, Boston.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.

Owen, A. B. (2001). *Empirical Likelihood.* Monographs on Statistics and Applied Probability **92**, Chapman & Hall / CRC, London.

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

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

Schick, A. (1987). A note on the construction of asymptotically linear estimators. *J. Statist. Plann. Inference* **16**, 89–105. Correction **22** (1989), 269–270.

Schick, A. (1993). On efficient estimation in regression models. *Ann. Statist.* **21**, 1486–1521. Correction **23** (1995), 1862–1863.

Schick, A. and Wefelmeyer, W. (2002). Estimating the innovation distribution in nonlinear autoregressive models. *Ann. Inst. Statist. Math.* **54**, 245–260.