# Variance bounds for estimators in autoregressive models with constraints

Ursula U. Müller*
Texas A&M University

Anton Schick†
Binghamton University

Wolfgang Wefelmeyer
Universität zu Köln

**Abstract**

We consider nonlinear and heteroscedastic autoregressive models whose residuals are martingale increments with conditional distributions that fulfill certain constraints. We treat two classes of constraints: residuals depending on the past through some function of the past observations only, and residuals that are invariant under some finite group of transformations. We determine the efficient influence function for estimators of the autoregressive parameter in such models, calculate variance bounds, discuss information gains, and suggest how to construct efficient estimators. Without constraints, efficient estimators can be given by weighted least squares estimators. With the constraints considered here, efficient estimators are obtained differently, as one-step improvements of some initial estimator, similarly as in autoregressive models with independent increments.

**Keywords.** Constrained autoregression; martingale estimating equation; M-estimator; Cramér–Rao bound; convolution theorem; efficient score function; information matrix; Newton–Raphson improvement.

## 1 Introduction

Let $X_{1-p}, \dots, X_n$ be observations of a Markov chain of order $p$ with a parametric model for the conditional mean,

$$E(X_i|\mathbf{X}_{i-1}) = r_\beta(\mathbf{X}_{i-1}), \tag{1.1}$$

where $\mathbf{X}_{i-1} = (X_{i-p}, \dots, X_{i-1})$ and $\beta$ is an unknown $d$-dimensional parameter. An efficient estimator for $\beta$ in this model is a randomly weighted least squares estimator that solves the estimating equation

$$\sum_{i=1}^{n} \tilde{\sigma}^{-2}(\mathbf{X}_{i-1}) \dot{r}_\beta(\mathbf{X}_{i-1}) \left(X_i - r_\beta(\mathbf{X}_{i-1})\right) = 0, \tag{1.2}$$

*Supported by NSF Grant DMS 0907014.
†Supported by NSF Grant DMS 0906551.

where $\dot{r}_\beta$ is the vector of partial derivatives of $r_\beta$ with respect to $\beta$, and $\tilde{\sigma}^2(\mathbf{X}_{i-1})$ estimates the conditional variance $\sigma^2(\mathbf{X}_{i-1}) = E((X_i - r_\beta(\mathbf{X}_{i-1}))^2|\mathbf{X}_{i-1})$. We refer to Wefelmeyer (1996, 1997) and, more generally, to Müller and Wefelmeyer (2002).

The Markov chain model (1.1) can be described as having a transition distribution from $\mathbf{X}_{i-1} = \mathbf{x}$ to $X_i = y$ of the form

$$A(\mathbf{x}, dy) = T(\mathbf{x}, dy - r_\beta(\mathbf{x})) \tag{1.3}$$

with $\int T(\mathbf{x}, dy)y = 0$ for (almost all) $\mathbf{x} = (x_1, \ldots, x_p)$. It can also be written as a nonlinear autoregressive model

$$X_i = r_\beta(\mathbf{X}_{i-1}) + \varepsilon_i, \tag{1.4}$$

where the residual $\varepsilon_i$ depends on the past through $\mathbf{X}_{i-1}$ only and has conditional distribution $T(\mathbf{X}_{i-1}, dy)$ with $\int T(\mathbf{x}, dy)y = 0$.

Suppose now that $T$ is known to fulfill certain additional restrictions. Then it is not useful to describe the model through the conditional constraint (1.1). We will instead use description (1.3), which depends explicitly on the conditional distribution of the residuals. This is also the approach in the well-studied degenerate case in which $T(\mathbf{x}, dy) = f(y)\,dy$ does not depend on the past at all, so that the autoregressive process (1.4) is driven by *independent innovations* $\varepsilon_i$ with density $f$. Efficient estimators of $\beta$ in the latter case are constructed as one-step improvements of some initial estimator; see Kreiss (1987a, 1987b), Koul and Schick (1997) and Schick (2001). For regression models $Y_i = r_\beta(\mathbf{X}_i) + \varepsilon_i$, corresponding results are in Schick (1993). We follow their approach in our model (1.3) with restrictions on $T$. We consider in particular the following two types of constraints.

(1) The conditional distribution of the residuals $\varepsilon_i$ is *partially independent* of the past, i.e., $T(\mathbf{x}, dy) = T_0(B\mathbf{x}, dy)$ for a known function $B : \mathbb{R}^p \to \mathbb{R}^q$ with $0 \le q \le p$. For example, the dependence is *lagged* by $p - q$ time points, $B\mathbf{x} = (x_1, \ldots, x_q)$; or the residuals have *shorter memory*, of length $q$ only, $B\mathbf{x} = (x_{p-q}, \ldots, x_p)$. Also contained here is the nonparametric case, with no restriction on $T$, described by $q = p$ and $B\mathbf{x} = \mathbf{x}$, and the case with independent innovations, described by $q = 0$ and $B\mathbf{x} = 0$.

(2) The conditional distribution of the residuals $\varepsilon_i$ is *invariant* under some finite group of transformations $B_j : \mathbb{R}^{p+1} \to \mathbb{R}^{p+1}$, $j = 1, \ldots, m$, i.e. $T$ has density $t$ with $t(\mathbf{z}) = t(B_j\mathbf{z})$ for $\mathbf{z} = (\mathbf{x}, y)$ and $j = 1, \ldots, m$. For example, the residuals are *symmetric about zero* in the sense that $t(\mathbf{x}, y) = t(-\mathbf{x}, -y)$, or the conditional distribution of the residuals $\varepsilon_i$ is *symmetric about the previous observation* $X_{i-1}$, i.e., $t(\mathbf{x}, y) = t(\mathbf{x}, 2x_p - y)$.

Our approach also covers models having both types of restrictions. Examples are the linear and nonlinear autoregressive models with independent and symmetric innovations, treated in Kreiss (1987b) and Koul and Schick (1997).

In Section 2 we consider model (1.3) with one or both of the above two constraints on $T$. We determine efficient influence functions for estimators of $\beta$, calculate variance bounds,

and specify the information about $\beta$ that is contained in the constraints. We also indicate how to construct efficient estimators of $\beta$ in these models.

In Section 3 we treat model (1.3) when, besides a parametric model for the conditional mean of the Markov chain, one also has a parametric model for the conditional variance,

$$E(X_i|\mathbf{X}_{i-1}) = r_\beta(\mathbf{X}_{i-1}),$$
$$E((X_i - r_\beta(\mathbf{X}_{i-1}))^2|\mathbf{X}_{i-1}) = s_\beta^2(\mathbf{X}_{i-1}).$$

This is a Markov chain with transition distribution of the form

$$A(\mathbf{x}, dy) = \frac{1}{s_\beta(\mathbf{x})} T\Big(\mathbf{x}, \frac{dy - r_\beta(\mathbf{x})}{s_\beta(\mathbf{x})}\Big), \tag{1.5}$$

with $T(\mathbf{X}_{i-1}, dy)$ now having conditional mean zero *and* variance one. It can also be written as a nonlinear autoregressive model

$$X_i = r_\beta(\mathbf{X}_{i-1}) + s_\beta(\mathbf{X}_{i-1})\varepsilon_i \tag{1.6}$$

with residual $\varepsilon_i$ having conditional distribution $T(\mathbf{X}_{i-1}, dy)$ with $\int T(\mathbf{x}, dy)y = 0$ and $\int T(\mathbf{x}, dy)y^2 = 1$. When there are no constraints on $T$, an efficient estimator for $\beta$ is obtained as a least squares estimator that solves an estimating equation of the form

$$\sum_{i=1}^n \Big(\hat{a}(\mathbf{X}_{i-1})\big((X_i - r_\beta(X_{i-1})\big) + \hat{b}(\mathbf{X}_{i-1})\big(\big(X_i - r_\beta(X_{i-1})\big)^2 - s_\beta^2(X_{i-1})\big)\Big) = 0$$

with appropriate $d$-dimensional vectors $\hat{a}(\mathbf{X}_{i-1})$ and $\hat{b}(\mathbf{X}_{i-1})$ of random weights as described in Wefelmeyer (1996) and Müller and Wefelmeyer (2002); see also (3.2) below. When the model has *independent innovations*, $T(\mathbf{x}, dy) = f(y)\,dy$ for a density $f$ having mean zero and variance one, then the description via (1.5) works, and efficient estimators of $\beta$ can be constructed as one-step improvements of some initial estimator; see Drost, Klaassen and Werker (1997). Ngatchou-Wandji (2008) studies weighted least squares estimators. For regression models $Y_i = r_\beta(\mathbf{X}_i) + s_\beta(\mathbf{X}_i)\varepsilon_i$, corresponding efficient estimators are in Schick (1993). We determine efficient influence functions and variance bounds for estimators of $\beta$ in model (1.5) under the two types of constraints on $T$ described above, again including the two known cases of no constraints on $T$, and of independent innovations. A special case is the *homoscedastic* autoregressive model with $E(X_i|\mathbf{X}_{i-1}) = r_\beta(\mathbf{X}_{i-1})$ and with conditional variance $E((X_i - r_\beta(\mathbf{X}_{i-1}))^2|\mathbf{X}_{i-1}) = \sigma_\beta^2$ independent of the past $\mathbf{X}_{i-1}$. Since Section 3 is parallel to Section 2, we will be brief.

## 2 Parametric conditional mean

For parametric models and independent observations, a lower variance bound for unbiased estimators is given by the Cramér–Rao inequality; Fréchet (1943), Darmois, (1945), Rao

(1945, 1962, 1963) and Cramér (1946). If the model is locally asymptotically normal, and we consider the larger class of regular estimators, then the Cramér–Rao bound is asymptotically attainable, at least locally. Furthermore, an asymptotically optimal estimator is asymptotically more concentrated in every symmetric interval about the true parameter. This follows from the convolution theorem of Kaufman (1966), Inagaki (1970), Hájek (1970) and Le Cam (1972). We also refer to the monograph of Le Cam (1986) and the historical article of Le Cam (2000). Following Stein (1956), the convolution theorem was extended to regular estimators of differentiable real-valued functionals on semiparametric models; see e.g. Pfanzagl and Wefelmeyer (1982). Then the asymptotic variance bound is that of the least favorable one-dimensional submodel. Similar extensions of the Cramér–Rao inequality, for estimators that are only asymptotically unbiased, are in Pfanzagl (2001) and Janssen (2003).

In this section we consider model (1.3) with constraints on $T$. This means that the observations $X_{1-p}, \dots, X_n$ come from a Markov chain of order $p$ with transition distribution $A(\mathbf{x}, dy) = T(\mathbf{x}, y - r_\beta(\mathbf{x}))$ such that $\beta$ is a $d$-dimensional parameter and $\int T(\mathbf{x}, dy)y = 0$ for all $\mathbf{x} = (x_1, \dots, x_p)$, with an additional constraint on $T$. We calculate efficient influence functions and variance bounds for estimators of $\beta$ in this model. This requires the time series to be locally asymptotically normal. We fix $\beta$ and $T$ and assume that the time series is strictly stationary and positive Harris recurrent. Let $G$ denote the stationary law of $\mathbf{X}_{i-1}$. Then $G \otimes T$ is the stationary law of $(\mathbf{X}_{i-1}, \varepsilon_i)$. We will briefly write $(\mathbf{X}, \varepsilon)$ for a random vector with this law. We write $e(y) = y$ for the identity function on $\mathbb{R}$. In the following we write conditional expectations as $T(\mathbf{x}, v) = \int T(\mathbf{x}, dy)v(\mathbf{x}, y) = E(v(\mathbf{X}, \varepsilon)|\mathbf{X} = \mathbf{x})$. We also write $T(\mathbf{x}, ve) = \int T(\mathbf{x}, dy)v(\mathbf{x}, y)y$ etc. We make the following assumptions on $r_\beta$ and $T$.

**Assumption 1.** *There is a $G$-square-integrable function $\dot{r} = \dot{r}_\beta$ such that for each $C > 0$,*

$$\sup_{\|\Delta\| \le Cn^{-1/2}} \sum_{i=1}^{n} \Big( r_{\beta+\Delta}(X_{i-1}) - r_\beta(X_{i-1}) - \Delta^\top \dot{r}(X_{i-1}) \Big)^2 = o_{P_n}(1).$$

**Assumption 2.** *For each $\mathbf{x}$, the conditional distribution $T(\mathbf{x}, dy)$ has a positive and absolutely continuous density $t(\mathbf{x}, y)$, and $E[\varepsilon^2]$ and $E[\ell_1^2(\mathbf{X}, \varepsilon)]$ are finite, where $\ell_1(\mathbf{x}, y) = -t'(\mathbf{x}, y)/t(\mathbf{x}, y)$ with derivative taken with respect to $y$.*

Since $T$ has a density $t$, the distribution $G$ of $\mathbf{X}$ also has a density, say $g$. Introduce perturbations $\beta_{nu} = \beta + n^{-1/2}u$ with $u \in \mathbb{R}^d$ and $t_{nv}(\mathbf{x}, y) = t(\mathbf{x}, y)(1 + n^{-1/2}v(\mathbf{x}, y))$ with $v$ a bounded and measurable function on $\mathbb{R}^{p+1}$. We must have $T_{nv}(\mathbf{x}, 1) = 1$ and $T_{nv}(\mathbf{x}, e) = 0$ and hence $T(\mathbf{x}, v) = 0$ and $T(\mathbf{x}, ve) = 0$. Let $V$ denote the set of functions $v$ with $T(\mathbf{x}, v) = 0$, and $V_1$ the subset of functions $v$ for which also $T(\mathbf{x}, ve) = 0$. Write $g_{nuv}$ for the density of $\mathbf{X}$ under $(\beta_{nu}, t_{nv})$. Write $P_n$ and $P_{nuv}$ for the joint law of the observations $X_{d-1}, \dots, X_n$ under $(\beta, t)$ and $(\beta_{nu}, t_{nv})$, respectively. Their log-likelihood ratio is

$$\log \frac{dP_{nuv}}{dP_n} = \log \frac{g_{nuv}(\mathbf{X}_0)}{g(\mathbf{X}_0)} + \sum_{i=1}^{n} \log \frac{t_{nv}(\mathbf{X}_{i-1}, \varepsilon_i(\beta_{nu}))}{t(\mathbf{X}_{i-1}, \varepsilon_i(\beta))}$$

with $\varepsilon_i(\beta) = X_i - r_\beta(\mathbf{X}_{i-1})$. We can prove local asymptotic normality similarly as in Koul and Schick (1997) who treat independent innovations.

**Theorem 1.** *Let $(u,v) \in \mathbb{R}^d \times V_1$. Suppose Assumptions 1 and 2 hold and $g$ depends smoothly on the parameters in the sense that $\int |g_{nuv}(\mathbf{x}) - g(\mathbf{x})|\, d\mathbf{x} \to 0$. Then*

$$\log \frac{dP_{nuv}}{dP_n} = n^{-1/2} \sum_{i=1}^n s_{uv}(\mathbf{X}_{i-1}, \varepsilon_i) - \tfrac{1}{2}\|(u,v)\|^2 + o_{P_n}(1), \tag{2.1}$$

$$n^{-1/2} \sum_{i=1}^n s_{uv}(\mathbf{X}_{i-1}, \varepsilon_i) \Rightarrow \|(u,v)\| N \quad \textit{under } P_n, \tag{2.2}$$

*where $N$ is a standard normal random variable and*

$$s_{uv}(\mathbf{x}, y) = u^\top \dot{r}(\mathbf{x}) \ell_1(\mathbf{x}, y) + v(\mathbf{x}, y),$$
$$\|(u,v)\|^2 = E[s_{uv}^2(\mathbf{X}, \varepsilon)].$$

The norm $\|(u,v)\|$ determines how difficult it is, asymptotically, to distinguish between $(\beta, t)$ and $(\beta_{nu}, t_{nv})$ on the basis of the observations. It induces an inner product

$$((u', v'), (u, v)) = E[s_{u'v'}(\mathbf{X}, \varepsilon) s_{uv}(\mathbf{X}, \varepsilon)].$$

Consider now a *model* for $T$, i.e. a family $\mathcal{T}$ of conditional distributions $T$ with $T(\mathbf{x}, e) = 0$. Assume that the fixed $T$ belongs to $\mathcal{T}$. Let $W$ denote the set of all $v$ in $V_1$ such that $T_{nv}$ lies in $\mathcal{T}$. Assume that $W$ is a linear space, the *local parameter space* of $\mathcal{T}$ at $T$. Let $\bar{W}$ and $\bar{V}_1$ denote the closures of $W$ and $V_1$ in $L_2(G \otimes T)$. We can then characterize efficient estimators of real-valued functions of $(\beta, t)$ as follows, using results of Hájek and LeCam, for which we refer to Section 3.3 of Bickel, Klaassen, Ritov and Wellner (1998).

**Definition 1.** A real-valued functional $\varphi$ of $(\beta, t)$ is called *differentiable* at $(\beta, t)$ with *gradient* $s_{u_\varphi v_\varphi}$ if $(u_\varphi, v_\varphi) \in \mathbb{R}^d \times \bar{V}_1$ and

$$n^{1/2}(\varphi(\beta_{nu}, t_{nv}) - \varphi(\beta, t)) \to ((u_\varphi, v_\varphi), (u, v)), \quad (u,v) \in \mathbb{R}^d \times W.$$

If $v_\varphi = w_\varphi$ is in $\bar{W}$, then $s_{u_\varphi w_\varphi}$ is called the *canonical gradient* of $\varphi$.

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

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

The convolution theorem says that for such an estimator, $L = \|(u_\varphi, w_\varphi)\| N + M$ in distribution, with $M$ independent of $N$. This justifies the following definition.

**Definition 3.** An estimator $\hat{\varphi}$ of $\varphi$ is called *efficient* at $(\beta, t)$ if

$$n^{1/2}(\hat{\varphi} - \varphi(\beta, t)) \Rightarrow \|(u_\varphi, w_\varphi)\| N.$$

**Definition 4.** An estimator $\hat{\varphi}$ of $\varphi$ is called *asymptotically linear* at $(\beta, t)$ with *influence function* $s_{u_\varphi v_\varphi}$ if $(u_\varphi, v_\varphi) \in \mathbb{R}^d \times \bar{V}_1$ and

$$n^{1/2}(\hat{\varphi} - \varphi(\beta, t)) = n^{-1/2} \sum_{i=1}^{n} s_{u_\varphi v_\varphi}(\mathbf{X}_{i-1}, \varepsilon_i) + o_{P_n}(1).$$

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 $\beta$, i.e., to the $d$-dimensional functional $\varphi(\beta, t) = \beta$. Differentiability of multivariate functionals $\varphi$ 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 then also meant componentwise.

We have two parameters $(\beta, t)$ and know that the canonical gradient of $\beta$ must be orthogonal to the tangent space for *fixed* $\beta$. Our approach will be to identify first the direction of the efficient influence function for $\beta$ through an orthogonal decomposition of the full tangent space.

**Residual distribution partially independent of the past.** Let $q \in \{0, \ldots, p\}$. Suppose $\mathcal{T}$ consists of the conditional distributions $T$ with $\int T(\mathbf{x}, dy)y = 0$ and $T(\mathbf{x}, dy) = T_0(B\mathbf{x}, dy)$ for some known function $B : \mathbb{R}^p \to \mathbb{R}^q$. Fix $T_0$. Let $t_0$ denote the density of $T_0$. Then we can write $\ell_1(\mathbf{x}, y) = \ell_{01}(B\mathbf{x}, y)$ with $\ell_{01} = -t_0'/t_0$, with derivative $t_0'(B\mathbf{x}, y) = \partial_y t_0(B\mathbf{x}, y)$. The local parameter space $W$ of $\mathcal{T}$ consists of the functions $v \in V_1$ of the form $v(\mathbf{x}, y) = v_0(B\mathbf{x}, y)$. For $u \in \mathbb{R}^d$ and $v \in W$ of the form $v(\mathbf{x}, y) = v_0(B\mathbf{x}, y)$ we have

$$s_{uv}(\mathbf{X}, \varepsilon) = u^\top \dot{r}(\mathbf{X}) \ell_{01}(B\mathbf{X}, \varepsilon) + v_0(B\mathbf{X}, \varepsilon).$$

The projection of $\ell_{01}(B\mathbf{x}, y)$ onto $\bar{W}$ is

$$\ell_{01}^*(B\mathbf{x}, y) = \ell_{01}(B\mathbf{x}, y) - \sigma_0^{-2}(B\mathbf{x})y$$

with $\sigma_0^2(B\mathbf{X}) = E(\varepsilon^2 | B\mathbf{X})$. We set $\varrho(B\mathbf{X}) = E(\dot{r}(\mathbf{X}) | B\mathbf{X})$ and write

$$\dot{r}(\mathbf{X})\ell_{01}(B\mathbf{X}, \varepsilon) = (\dot{r}(\mathbf{X}) - \varrho(B\mathbf{X}))\ell_{01}(B\mathbf{X}, \varepsilon) + \varrho(B\mathbf{X})\ell_{01}(B\mathbf{X}, \varepsilon).$$

We obtain the orthogonal decomposition

$$s_{uv}(\mathbf{X}, \varepsilon) = u^\top \tau(\mathbf{X}, \varepsilon) + u^\top \varrho(B\mathbf{X}) \ell_{01}^*(B\mathbf{X}, \varepsilon) + v_0(B\mathbf{X}, \varepsilon) \tag{2.3}$$

with

$$\tau(\mathbf{X}, \varepsilon) = (\dot{r}(\mathbf{X}) - \varrho(B\mathbf{X}))\ell_{01}(B\mathbf{X}, \varepsilon) + \varrho(B\mathbf{X})\sigma_0^{-2}(B\mathbf{X})\varepsilon.$$

The random variable $\tau(\mathbf{X}, \varepsilon)$ is called the *score function* for $\beta$ at $T_0$. By construction, $\varrho(B\mathbf{x})\ell_{01}^*(B\mathbf{x}, y)$ is in $\bar{W}$, and the functions $(\dot{r}(\mathbf{x}) - \varrho(B\mathbf{x}))\ell_{01}(B\mathbf{x}, y)$ and $\varrho(B\mathbf{x})\sigma_0^{-2}(B\mathbf{x})y$

are orthogonal to $W$. It follows that the efficient influence function for $\beta$ is $\gamma(\mathbf{x}, y) = \Lambda^{-1}\tau(\mathbf{x}, y)$, where

$$\begin{aligned}\Lambda &= E[\tau(\mathbf{X}, \varepsilon)\tau^\top(\mathbf{X}, \varepsilon)] \\ &= E[(R(B\mathbf{X}) - \varrho(B\mathbf{X})\varrho^\top(B\mathbf{X}))J_{01}(B\mathbf{X})] + E[\varrho(B\mathbf{X})\varrho^\top(B\mathbf{X})\sigma_0^{-2}(B\mathbf{X})]\end{aligned}$$

with $R(B\mathbf{X}) = E(\dot{r}(\mathbf{X})\dot{r}^\top(\mathbf{X})|B\mathbf{X})$ and with $J_{01}(B\mathbf{X}) = E(\ell_{01}^2(B\mathbf{X}, \varepsilon)|B\mathbf{X}))$ the conditional Fisher information for location of $\varepsilon$. By the characterization above, an estimator $\hat{\beta}$ of $\beta$ is regular and efficient at $(\beta, t)$ if

$$n^{1/2}(\hat{\beta} - \beta) = \Lambda^{-1}n^{-1/2}\sum_{i=1}^{n}\tau(\mathbf{X}_{i-1}, \varepsilon_i) + o_{P_n}(1).$$

Its asymptotic covariance matrix is $\Lambda^{-1}$. The matrix $\Lambda$ is called the *information matrix* for $\beta$ at $T_0$. To construct an efficient estimator $\hat{\beta}$ of $\beta$, choose a $n^{1/2}$-consistent *initial estimator* $\tilde{\beta}$ of $\beta$, for example the least squares estimator or a weighted version. Estimate the residual $\varepsilon_i$ by $\tilde{\varepsilon}_i = X_i - r_{\tilde{\beta}}(\mathbf{X}_{i-1})$. Under appropriate regularity conditions, an efficient estimator $\hat{\beta}$ is obtained as a *one-step improvement* (or Newton–Raphson estimator)

$$\hat{\beta} = \tilde{\beta} + \tilde{\Lambda}^{-1}\frac{1}{n}\sum_{i=1}^{n}\tilde{\tau}(\mathbf{X}_{i-1}, \tilde{\varepsilon}_i) \tag{2.4}$$

with some estimator $\tilde{\tau}$ of the score function $\tau$, and with

$$\tilde{\Lambda} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\tau}(\mathbf{X}_{i-1}, \tilde{\varepsilon}_i)\tilde{\tau}^\top(\mathbf{X}_{i-1}, \tilde{\varepsilon}_i). \tag{2.5}$$

An estimator $\tilde{\tau}$ of $\tau$ requires estimators for $\dot{r}$, $\varrho$, $\ell_{01}$ and $\sigma_0^2$ as follows. The gradient $\dot{r}_\beta$ is estimated by $\dot{r}_{\tilde{\beta}}$. We can estimate $\varrho(\mathbf{b})$ by a generalization of the Nadaraya–Watson estimator,

$$\tilde{\varrho}(\mathbf{b}) = \frac{\int_{B\mathbf{x}=\mathbf{b}}\dot{r}_{\tilde{\beta}}(\mathbf{x})\tilde{g}(\mathbf{x})\,d\mathbf{x}}{\int_{B\mathbf{x}=\mathbf{b}}\tilde{g}(\mathbf{x})\,d\mathbf{x}},$$

where $\tilde{g}$ is a density estimator of $g$ based on $\mathbf{X}_i$, $i = 1, \ldots, n$. To estimate $\ell_{01}$, we express this function in terms of the stationary density of $(B\mathbf{X}, \varepsilon)$. Let $g_0$ denote the stationary density of $B\mathbf{X}$. The stationary density of $(B\mathbf{X}, \varepsilon)$ at $(B\mathbf{x}, y)$ is $h_0(B\mathbf{x}, y) = g_0(B\mathbf{x})t_0(B\mathbf{x}, y)$. Set $h_0'(B\mathbf{x}, y) = \partial_y h_0(B\mathbf{x}, y)$. Then $\ell_{01} = -t_0'/t_0 = -h_0'/h_0$. Estimate $h_0$ by a density estimator $\tilde{h}_0$ based on $(B\mathbf{X}_{i-1}, \tilde{\varepsilon}_i)$, $i = 1, \ldots, n$. An estimator of $\ell_{01}$ is $\tilde{\ell}_{01} = -\tilde{h}_0'/\tilde{h}_0$. An estimator of $\sigma_0^2(\mathbf{b})$ is the Nadaraya–Watson estimator

$$\tilde{\sigma}_0^2(\mathbf{b}) = \frac{\int y^2\tilde{h}_0(\mathbf{b}, y)\,dy}{\int \tilde{h}_0(\mathbf{b}, y)\,dy}.$$

Our estimator for the score function $\tau(\mathbf{X}_{i-1}, \varepsilon_i)$ is then $\tilde{\tau}(\mathbf{X}_{i-1}, \tilde{\varepsilon}_i)$ with

$$\tilde{\tau}(\mathbf{x}, y) = (\dot{r}_{\tilde{\beta}}(\mathbf{x}) - \tilde{\varrho}(B\mathbf{x}))\tilde{\ell}_{01}(B\mathbf{x}, y) + \tilde{\varrho}(B\mathbf{x})\tilde{\sigma}_0^{-2}(B\mathbf{x})y.$$

*Special case: Nonparametric model.* Suppose we have no structural information on the conditional distribution $T$. Then $\mathcal{T}$ consists of *all* conditional distributions $T$ with $T(\mathbf{x}, e) = 0$. Setting $q = p$ and $B\mathbf{x} = \mathbf{x}$, this can be expressed as $T(\mathbf{x}, dy) = T(B\mathbf{x}, y)$. The autoregressive model is then determined by the conditional constraint (1.1) alone and was treated before, in particular in Wefelmeyer (1997). We give an alternative characterization and construction of an efficient estimator of $\beta$ via the Markov chain description (1.3) of the model. Fix $T$ with density $t$. We have $t_0 = t$, $\ell_{01} = \ell_1 = -t'/t$ and $W = V_1$. The projection of $\ell_1(\mathbf{x}, y)$ onto $\bar{V}_1$ is $\ell_1^*(\mathbf{x}, y) = \ell_1(\mathbf{x}, y) - \sigma^{-2}(\mathbf{x})y$ with $\sigma^2(\mathbf{X}) = E(\varepsilon^2|\mathbf{X})$. Now $\varrho(B\mathbf{x}) = \dot{r}(\mathbf{x})$, and the orthogonal decomposition (2.3) simplifies to

$$s_{uv}(\mathbf{X}, \varepsilon) = u^\top \tau(\mathbf{X}, \varepsilon) + u^\top \dot{r}(\mathbf{X})\ell_1^*(\mathbf{X}, \varepsilon) + v(\mathbf{X}, \varepsilon)$$

with

$$\tau(\mathbf{X}, \varepsilon) = \dot{r}(\mathbf{X})\sigma^{-2}(\mathbf{X})\varepsilon.$$

It follows that the efficient influence function for $\beta$ is $\gamma(\mathbf{x}, y) = \Lambda^{-1}\tau(\mathbf{x}, y)$ with

$$\Lambda = E[\tau(\mathbf{X}, \varepsilon)\tau^\top(\mathbf{X}, \varepsilon)] = E[\dot{r}(\mathbf{X})\dot{r}^\top(\mathbf{X})\sigma^{-2}(\mathbf{X})].$$

Wefelmeyer (1997) and, more generally, Müller and Wefelmeyer (2002) show that an efficient estimator of $\beta$ is obtained as a solution of the estimating equation (1.2). An alternative efficient estimator $\hat{\beta}$ of $\beta$ can be obtained as one-step improvement of a $n^{1/2}$-consistent initial estimator $\tilde{\beta}$,

$$\hat{\beta} = \tilde{\beta} + \tilde{\Lambda}^{-1}\frac{1}{n}\sum_{i=1}^{n} \dot{r}_{\tilde{\beta}}(\mathbf{X}_{i-1})\tilde{\sigma}^{-2}(\mathbf{X}_{i-1})(X_i - r_{\tilde{\beta}}(X_{i-1}))$$

with

$$\tilde{\Lambda} = \frac{1}{n}\sum_{i=1}^{n} \dot{r}_{\tilde{\beta}}(\mathbf{X}_{i-1})\dot{r}_{\tilde{\beta}}^\top(\mathbf{X}_{i-1})\tilde{\sigma}^{-2}(\mathbf{X}_{i-1}).$$

**Remark 1.** The information gain of partial independence over the nonparametric model is calculated as follows. Fix a conditional distribution $T$ in the smaller model. It is of the form $T(\mathbf{x}, dy) = T_0(B\mathbf{x}, dy)$. The nonparametric score function for $\beta$ is then

$$\tau(\mathbf{X}, \varepsilon) = \dot{r}(\mathbf{X})\sigma_0^{-2}(B\mathbf{X})\varepsilon.$$

The corresponding information matrix is

$$E[\tau(\mathbf{X}, \varepsilon)\tau^\top(\mathbf{X}, \varepsilon)] = E[R(B\mathbf{X})\sigma_0^{-2}(B\mathbf{X})].$$

The information gain is obtained by subtracting from this the information matrix in the smaller model, resulting in

$$E[(R(B\mathbf{X}) - \varrho(B\mathbf{X})\varrho^\top(B\mathbf{X}))(J_{01}(B\mathbf{X}) - \sigma_0^{-2}(B\mathbf{X})].$$

Here

$$R(B\mathbf{X}) - \varrho(B\mathbf{X})\varrho^\top(B\mathbf{X}) = E\big((\dot r(\mathbf{X}) - \varrho(B\mathbf{X}))(\dot r(\mathbf{X}) - \varrho(B\mathbf{X}))^\top \big| B\mathbf{X}\big)$$

is a conditional covariance matrix and positive semidefinite. By the Cauchy–Schwarz inequality,

$$1 = E(\ell_{01}(B\mathbf{X},\varepsilon)\varepsilon|B\mathbf{X}) \le E(\ell_{01}^2(B\mathbf{X},\varepsilon)|B\mathbf{X})^{1/2} E(\varepsilon^2|B\mathbf{X})^{1/2} = J_{01}^{1/2}(B\mathbf{X})\sigma_0(B\mathbf{X}).$$

Hence $J_{01}(B\mathbf{X}) - \sigma_0^{-2}(B\mathbf{X}) \ge 0$. There is no information gain if $B\mathbf{X} = \mathbf{X}$ or if the conditional distribution of $\varepsilon$ given $B\mathbf{X}$ is normal, in which case $J_{01} = \sigma_0^{-2}$.

*Special case: Lagged residuals.* Suppose the conditional distribution $T$ is lagged by $p - q$ time points, i.e. $T(\mathbf{x}, dy) = T_0(\mathbf{x}_0, dy)$ for $\mathbf{x}_0 = (x_1, \dots, x_q)$. Setting $B\mathbf{x} = \mathbf{x}_0$, this can be expressed as $T(\mathbf{x}, dy) = T_0(B\mathbf{x}, dy)$. The efficient influence function for $\beta$ is calculated as in the general model with residual distribution partially independent of the past. Setting $\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1)$, we now have

$$\varrho(\mathbf{x}_0) = E(\dot r(\mathbf{X})|\mathbf{X}_0 = \mathbf{x}_0) = \frac{\int \dot r(\mathbf{x}_0, \mathbf{x}_1) g(\mathbf{x}_0, \mathbf{x}_1)\, d\mathbf{x}_1}{\int g(\mathbf{x}_0, \mathbf{x}_1)\, d\mathbf{x}_1}.$$

This can be estimated by a Nadaraya–Watson estimator. Let $g_0$ denote the stationary density of $\mathbf{X}_0$, and $t_0$ the density of $T_0$. The stationary density of $(\mathbf{X}_0, \varepsilon)$ is $h_0(\mathbf{x}_0, y) = g_0(\mathbf{x}_0) t_0(\mathbf{x}_0, y)$, and we can write

$$\sigma_0^2(\mathbf{x}_0) = E(\varepsilon^2|\mathbf{X}_0 = \mathbf{x}_0) = \frac{\int y^2 h_0(\mathbf{x}_0, y)\, dy}{\int h_0(\mathbf{x}_0, y)\, dy}.$$

This can also be estimated by a Nadaraya–Watson estimator.

*Special case: Residuals with shorter memory.* Suppose the conditional distribution $T$ is of order $q$ only, i.e. $T(\mathbf{x}, dy) = T_0(\mathbf{x}_1, dy)$ for $\mathbf{x}_1 = (x_{q+1}, \dots, x_p)$. Setting $B\mathbf{x} = \mathbf{x}_1$, this can be expressed as $T(\mathbf{x}, dy) = T_0(B\mathbf{x}, dy)$. The efficient influence function for $\beta$ is calculated as in the previous case, now with

$$\varrho(\mathbf{x}_1) = E(\dot r(\mathbf{X})|\mathbf{X}_1 = \mathbf{x}_1) = \frac{\int \dot r(\mathbf{x}_0, \mathbf{x}_1) g(\mathbf{x}_0, \mathbf{x}_1)\, d\mathbf{x}_0}{\int g(\mathbf{x}_0, \mathbf{x}_1)\, d\mathbf{x}_0}.$$

This can be estimated by a Nadaraya–Watson estimator. Let $g_0$ denote the stationary density of $\mathbf{X}_1$, and $t_0$ the density of $T_0$. The stationary density of $(\mathbf{X}_1, \varepsilon)$ is $h_0(\mathbf{x}_1, y) = g_0(\mathbf{x}_1) t_0(\mathbf{x}_1, y)$, and we can write

$$\sigma_0^2(\mathbf{x}_1) = E(\varepsilon^2|\mathbf{X}_1 = \mathbf{x}_1) = \frac{\int y^2 h_0(\mathbf{x}_1, y)\, dy}{\int h_0(\mathbf{x}_1, y)\, dy}.$$

This can also be estimated by a Nadaraya–Watson estimator.

*Special case: Independent innovations.* Suppose the autoregressive model (1.3) has independent innovations. Then $\mathcal{T}$ consists of all conditional distributions $T$ of the form $T(\mathbf{x}, dy) = f(y)\,dy$ with $\int y f(y)\,dy = 0$. Setting $q = 0$ and $B\mathbf{x} = 0$, this can be expressed as $T(\mathbf{x}, dy) = T(B\mathbf{x}, dy) = T(0, dy)$. Fix $f$. Then $t(\mathbf{x}, y) = f(y)$ and $\ell_1(y) = -f'(y)/f(y)$. The local parameter space $W$ of $\mathcal{T}$ consists of the bounded measurable functions $v$ on $\mathbb{R}$ with $E[v(\varepsilon)] = 0$ and $E[v(\varepsilon)\varepsilon] = 0$.

Assumption 2 then says that $\varepsilon$ has a positive and absolutely continuous density $f(y)$, and $E[\varepsilon^2]$ and $J_1 = E[\ell_1^2(\varepsilon)]$ are finite, where $\ell_1(y) = -f'(y)/f(y)$. Efficient estimation in this model is treated extensively, but it is instructive to compare this model with the nonparametric model above. ¿From Assumption 2 we obtain $E[\ell_1(\varepsilon)] = 0$ and $E[\ell_1(\varepsilon)\varepsilon] = 1$. The projection of $\ell_1(y)$ onto $\bar{W}$ is $\ell_1^*(y) = \ell_1(y) - \sigma^{-2}y$ with $\sigma^2 = E[\varepsilon^2]$. Now $\varrho(B\mathbf{x}) = E[\dot{r}(X)] = \varrho$, say, and the orthogonal decomposition (2.3) becomes

$$s_{uv}(\mathbf{X}, \varepsilon) = u^\top \tau(\mathbf{X}, \varepsilon) + u^\top \varrho \ell_1^*(\varepsilon) + v(\varepsilon)$$

with

$$\tau(\mathbf{X}, \varepsilon) = (\dot{r}(\mathbf{X}) - \varrho)\ell_1(\varepsilon) + \varrho\sigma^{-2}\varepsilon.$$

It follows that the efficient influence function for $\beta$ is $\gamma(\mathbf{x}, y) = \Lambda^{-1}\tau(\mathbf{x}, y)$, where

$$\Lambda = J_1(R - \varrho\varrho^\top) + \sigma^{-2}\varrho\varrho^\top$$

with $R = E[\dot{r}(\mathbf{X})\dot{r}^\top(\mathbf{X})]$. Koul and Schick (1997) construct an efficient estimator $\hat{\beta}$ as a one-step improvement of some $n^{1/2}$-consistent estimator $\tilde{\beta}$ of $\beta$.

**Residual distribution invariant under transformations.** Let $B_1, \ldots, B_m$ be a group of transformations on $\mathbb{R}^{p+1}$. Suppose $\mathcal{T}$ consists of all conditional distributions $T$ with $\int T(\mathbf{x}, dy)y = 0$ and density $t$ that is invariant under these transformations, i.e., $t(\mathbf{z}) = t(B_j\mathbf{z})$ for $\mathbf{z} = (\mathbf{x}, y)$ and $j = 1, \ldots, m$. Fix $t$. The local parameter space $W$ of $\mathcal{T}$ consists of the functions $v$ in $V_1$ with $v(\mathbf{z}) = v(B_j\mathbf{z})$ for $j = 1, \ldots, m$. Let $t'(\mathbf{x}, y) = \partial_y t(\mathbf{x}, y)$. Note that $t'$ and hence $\ell_1 = -t'/t$ are in general not invariant under the transformations. The projection of $\ell_1(\mathbf{x}, y)$ onto $\bar{V}_1$ is $\ell_1(\mathbf{x}, y) - \sigma^{-2}(\mathbf{x})y$ with $\sigma^2(\mathbf{x}) = E(\varepsilon^2|\mathbf{X} = \mathbf{x})$. In order to decompose $s_{uv}(\mathbf{x}, y) = u^\top\dot{r}(\mathbf{x})\ell_1(\mathbf{x}, y) + v(\mathbf{x}, y)$, we set

$$\lambda(\mathbf{x}, y) = \dot{r}(\mathbf{x})\ell_1(\mathbf{x}, y), \qquad \mu(\mathbf{x}, y) = \dot{r}(\mathbf{x})\sigma^{-2}(\mathbf{x})y.$$

The components of $\lambda(\mathbf{x}, y) - \mu(\mathbf{x}, y) = \dot{r}(\mathbf{x})(\ell_1(\mathbf{x}, y) - \sigma^{-2}(\mathbf{x})y)$ are in $\bar{V}_1$. The componentwise projection onto $\bar{W}$ is $\lambda_0(\mathbf{x}, y) - \mu_0(\mathbf{x}, y)$ with

$$\lambda_0(\mathbf{z}) = \frac{1}{m}\sum_{j=1}^m \lambda(B_j\mathbf{z}), \qquad \mu_0(\mathbf{z}) = \frac{1}{m}\sum_{j=1}^m \mu(B_j\mathbf{z}).$$

We arrive at the decomposition $s_{uv} = u^\top \tau + u^\top(\lambda_0 - \mu_0) + v$ with $\tau = \lambda - \lambda_0 + \mu_0$. It follows that the efficient influence function for $\beta$ is $\gamma(\mathbf{x}, y) = \Lambda^{-1}\tau(\mathbf{x}, y)$ with $\Lambda = E[\tau(\mathbf{X}, \varepsilon)\tau^\top(\mathbf{X}, \varepsilon)]$. Since $\lambda - \lambda_0$ and $\mu - \mu_0$ are componentwise orthogonal to $W$, we have

$$\begin{aligned}\Lambda &= E[(\lambda(\mathbf{X}, \varepsilon) - \lambda_0(\mathbf{X}, \varepsilon))(\lambda(\mathbf{X}, \varepsilon) - \lambda_0(\mathbf{X}, \varepsilon))^\top] + E[\mu_0(\mathbf{X}, \varepsilon)\mu_0^\top(\mathbf{X}, \varepsilon)] \\ &= E[\lambda(\mathbf{X}, \varepsilon)(\lambda(\mathbf{X}, \varepsilon) - \lambda_0(\mathbf{X}, \varepsilon))^\top] + E[\mu(\mathbf{X}, \varepsilon)\mu_0^\top(\mathbf{X}, \varepsilon)].\end{aligned}$$

With $\tilde{\beta}$ a $n^{1/2}$-consistent initial estimator of $\beta$, an efficient estimator $\hat{\beta}$ of $\beta$ can be obtained as a one-step improvement of the form (2.4) and (2.5), where $\tilde{\tau}$ can be constructed as follows. Estimate $\dot{r}_\beta$ by $\dot{r}_{\tilde{\beta}}$. Estimate the residual $\varepsilon_i$ by $\tilde{\varepsilon}_i = X_i - r_{\tilde{\beta}}(\mathbf{X}_{i-1})$. Estimate the stationary density $h(\mathbf{x}, y) = g(\mathbf{x})t(\mathbf{x}, y)$ of $(\mathbf{X}, \varepsilon)$ by a density estimator based on $(\mathbf{X}_{i-1}, \tilde{\varepsilon}_i)$, $i = 1, \ldots, n$. Then $\ell_1 = -t'/t = -h'/h$ is estimated by $\tilde{\ell}_1 = -\tilde{h}'/\tilde{h}$. An estimator of $\sigma^2$ is given by

$$\tilde{\sigma}^2(\mathbf{x}) = \frac{\int y^2 \tilde{h}(\mathbf{x}, y)\, dy}{\int \tilde{h}(\mathbf{x}, y)\, dy}.$$

Set $\tilde{\lambda}(\mathbf{x}, y) = r_{\tilde{\beta}}(\mathbf{x})\tilde{\ell}_1(\mathbf{x}, y)$ and $\tilde{\mu} = r_{\tilde{\beta}}(\mathbf{x})\tilde{\sigma}^{-2}(\mathbf{x})y$. Their symmetrizations are

$$\tilde{\lambda}_0(\mathbf{z}) = \frac{1}{m}\sum_{j=1}^{m} \tilde{\lambda}(B_j\mathbf{z}), \qquad \tilde{\mu}_0(\mathbf{z}) = \frac{1}{m}\sum_{j=1}^{m} \tilde{\mu}(B_j\mathbf{z}).$$

Our estimator for the score function $\tau(\mathbf{X}_{i-1}, \varepsilon_i)$ is then $\tilde{\tau}(\mathbf{X}_{i-1}, \tilde{\varepsilon}_i)$ with $\tilde{\tau} = \tilde{\lambda} - \tilde{\lambda}_0 + \tilde{\mu}_0$.

*Special case: Nonparametric model.* Suppose we have no structural information on the conditional distribution $T$. This case is known and was also considered above as a degenerate case of partial information. It is also a degenerate case of invariance of $T$ under transformations if the group of transformations consists only of the identity. Then

$$\begin{aligned}\lambda_0(\mathbf{x}, y) &= \lambda(\mathbf{x}, y) = \dot{r}(\mathbf{x})\ell_1(\mathbf{x}, y), \\ \mu_0(\mathbf{x}, y) &= \mu(\mathbf{x}, y) = \dot{r}(\mathbf{x})\sigma^{-2}(\mathbf{x})y.\end{aligned}$$

Hence $\lambda - \lambda_0 = 0$ and $\tau(\mathbf{x}, y) = \mu(\mathbf{x}, y) = \dot{r}(\mathbf{x})\sigma^{-2}(\mathbf{x})y$, as already seen.

**Remark 2.** The information gain of group invariance over the nonparametric model is calculated as follows. Fix a conditional density $t$ in the smaller model. It fulfills $t(B_j\mathbf{z}) = t(B_j\mathbf{z})$ for $j = 1, \ldots, m$. The nonparametric score function for $\beta$ is

$$\tau(\mathbf{X}, \varepsilon) = \dot{r}(\mathbf{X})\sigma^{-2}(\mathbf{X})\varepsilon = \mu(\mathbf{X}, \varepsilon).$$

The corresponding information matrix is

$$E[\tau(\mathbf{X}, \varepsilon)\tau^\top(\mathbf{X}, \varepsilon)] = E[\mu(\mathbf{X}, \varepsilon)\mu^\top(\mathbf{X}, \varepsilon)] = E[\dot{r}(\mathbf{X})\dot{r}^\top(\mathbf{X})\sigma^{-2}(\mathbf{X})].$$

The information gain is obtained by subtracting from this the information matrix in the smaller model, resulting in

$$E[(\lambda(\mathbf{X},\varepsilon)-\lambda_0(\mathbf{X},\varepsilon))(\lambda(\mathbf{X},\varepsilon)-\lambda_0(\mathbf{X},\varepsilon))^\top]-E[(\mu(\mathbf{X},\varepsilon)-\mu_0(\mathbf{X},\varepsilon))(\mu(\mathbf{X},\varepsilon)-\mu_0(\mathbf{X},\varepsilon))^\top].$$

There is no information gain if $\lambda$ is group invariant or if the conditional density $t(\mathbf{x},\cdot)$ is normal, in which case $\ell_1(\mathbf{x},y)=\sigma^{-2}(\mathbf{x})y$.

*Special case: Symmetric residuals.* Suppose $\mathcal{T}$ consists of all conditional distributions $T$ with $T(\mathbf{x},e)=0$ that are symmetric about zero, i.e., with density $t$ fulfilling $t(\mathbf{x},y)=t(-\mathbf{x},-y)$. This can be expressed as $t(\mathbf{z})=t(B\mathbf{z})$ for $B\mathbf{z}=-\mathbf{z}$. The group of transformations consists of $B$ and the identity. Fix $t$. The local parameter space $W$ consists of the functions $v\in V_1$ with $v(\mathbf{z})=v(-\mathbf{z})$. We have $t'(\mathbf{x},y)=-t'(-\mathbf{x},-y)$ and hence $\ell_1(\mathbf{x},y)=-\ell_1(-\mathbf{x},-y)$. We obtain

$$\begin{aligned}\lambda_0(\mathbf{x},y)&=\frac{1}{2}\big(\dot r(\mathbf{x})\ell_1(\mathbf{x},y)+\dot r(-\mathbf{x})\ell_1(-\mathbf{x},-y)\big)\\&=\frac{1}{2}(\dot r(\mathbf{x})-\dot r(-\mathbf{x}))\ell_1(\mathbf{x},y).\end{aligned}$$

We have $\sigma^2(\mathbf{x})=\int y^2t(\mathbf{x},y)\,dy=\sigma^2(-\mathbf{x})$ and obtain

$$\mu_0(\mathbf{x},y)=\frac{1}{2}(\dot r(\mathbf{x})-\dot r(-\mathbf{x}))\sigma^{-2}(\mathbf{x})y.$$

In particular,

$$\begin{aligned}\lambda(\mathbf{x},y)-\lambda_0(\mathbf{x},y)&=\frac{1}{2}(\dot r(\mathbf{x})+\dot r(-\mathbf{x}))\ell_1(\mathbf{x},y),\\\mu(\mathbf{x},y)-\mu_0(\mathbf{x},y)&=\frac{1}{2}(\dot r(\mathbf{x})+\dot r(-\mathbf{x}))\sigma^{-2}(\mathbf{x})y.\end{aligned}$$

and

$$\Lambda=\frac{1}{4}E[(\dot r(\mathbf{X})+\dot r(-\mathbf{X}))(\dot r(\mathbf{X})+\dot r(-\mathbf{X}))^\top(J_1(\mathbf{X})+\sigma^{-2}(\mathbf{X}))]$$

with $J_1(\mathbf{X})=E(\ell_1^2(\mathbf{X},\varepsilon)|\mathbf{X})$. The information gain over the nonparametric model is

$$\frac{1}{4}E[(\dot r(\mathbf{X})+\dot r(-\mathbf{X}))(\dot r(\mathbf{X})+\dot r(-\mathbf{X}))^\top(J_1(\mathbf{X})-\sigma^{-2}(\mathbf{X}))].$$

This is zero if the autoregression function is antisymmetric about zero, $r(\mathbf{x})=-r(-\mathbf{x})$, or if the conditional density $t(\mathbf{x},\cdot)$ is normal.

*Special case: Conditionally symmetric residuals.* Suppose $\mathcal{T}$ consists of all conditional distributions $T$ with density $t$ fulfilling $t(\mathbf{x},2x_p-y)=t(\mathbf{x},y)$. Setting $B(\mathbf{x},y)=(\mathbf{x},2x_p-y)$, this can be expressed as $t(\mathbf{z})=t(B\mathbf{z})$. The group of transformations consists of $B$ and the identity. The local parameter space $W$ of $\mathcal{T}$ consists of all $v\in V_1$ with $v(\mathbf{x},2x_p-y)=v(\mathbf{x},y)$. Since $t$ is symmetric, $t'$ is antisymmetric, $t'(\mathbf{x},2x_p-y)=-t'(\mathbf{x},y)$. Hence $\ell_1=-t'/t$ is also

antisymmetric and therefore orthogonal to $W$. Hence $s_{uv}(\mathbf{X},\varepsilon) = u^\top \dot{r}(\mathbf{X})\ell_1(\mathbf{X},\varepsilon) + v(\mathbf{X},\varepsilon)$ is an orthogonal decomposition. The covariance matrix of $\dot{r}(\mathbf{X})\ell_1(\mathbf{X},\varepsilon)$ is

$$\Lambda = E[\dot{r}(\mathbf{X})\dot{r}^\top(\mathbf{X})J_1(\mathbf{X})]$$

with $J_1(\mathbf{X}) = E(\ell_1^2(\mathbf{X},\varepsilon)|\mathbf{X})$. Hence the efficient influence function for $\beta$ is given by $\gamma(\mathbf{x},y) = \Lambda^{-1}\dot{r}(\mathbf{x})\ell_1(\mathbf{x},y)$.

**Remark 3.** Our approach applies also to models with both types of restriction on $T$. An example is the nonlinear autoregressive model $X_i = r_\beta(\mathbf{X}_{i-1}) + \varepsilon_i$ with independent and symmetric innovations $\varepsilon_i$. If the innovations have density $f$, then $\mathcal{T}$ consists of all (conditional) distributions $T$ with density $t(\mathbf{x},y) = f(y)$ and $f(y) = f(-y)$. The local parameter space $W$ consists of the even bounded measurable functions $v$ on $\mathbb{R}$. Now $\ell_1 = -f'/f$ is odd and therefore orthogonal to $\bar{W}$. It follows that the score function for $\beta$ is $\tau(\mathbf{X},\varepsilon) = (\dot{r}(\mathbf{X}) - \varrho)\ell_1(\varepsilon)$ with $\varrho = E[\dot{r}(\mathbf{X})]$. Hence the efficient influence function for $\beta$ is $\Lambda^{-1}\tau(\mathbf{x},y)$ with $\Lambda = J_1(R - \varrho\varrho^\top)$, where $J_1 = E[\ell_1^2(\varepsilon)]$ and $R = E[\dot{r}(\mathbf{X})\dot{r}^\top(\mathbf{X})]$. Efficient estimators for $\beta$ are constructed in Kreiss (1987b) and Koul and Schick (1997) for linear and nonlinear autoregression, respectively.

# 3 Parametric conditional mean and variance

In this section we consider model (1.5) with constraints $T$. This means that $X_{1-p},\ldots,X_n$ are observations of a Markov chain of order $p$ with transition distribution $A(\mathbf{x},dy) = T(\mathbf{x},(y - r_\beta(\mathbf{x}))/s_\beta(\mathbf{x}))$ such that $\beta$ is a $d$-dimensional parameter and $T(\mathbf{x},e) = 0$ and $T(\mathbf{x},e^2) = 1$ for almost all $\mathbf{x} = (x_1,\ldots,x_p)$, with an additional constraint on $T$ of the type described in the Introduction. In order to characterize efficient estimators in this model, we prove it to be locally asymptotically normal. This is parallel to Section 2. We fix $\beta$ and $T$ and assume that the time series is strictly stationary and positive Harris recurrent. Let $G$ again denote the stationary law of $\mathbf{X}_{i-1}$. We make the following assumptions on $r_\beta$ and $s_\beta$ and on $T$.

**Assumption 3.** *The functions $r_\beta$ and $s_\beta$ are differentiable in the sense of Assumption 1, with gradients $\dot{r} = \dot{r}_\beta$ and $\dot{s} = \dot{s}_\beta$, respectively. Furthermore, the function $s_\beta$ is bounded away from zero locally uniformly in $\beta$.*

**Assumption 4.** *For each $\mathbf{x}$, the conditional distribution $T(\mathbf{x},dy)$ has a positive and absolutely continuous density $t(\mathbf{x},y)$, and $E[\varepsilon^4]$ and $E[\ell_1^2(\mathbf{X},\varepsilon)(1+\varepsilon^2)]$ are finite.*

Assumption 4 implies that $E[\ell_1^2(\mathbf{X},\varepsilon)]$ and $E[\ell_2^2(\mathbf{X},\varepsilon)]$ are finite, where $\ell_2(\mathbf{x},y) = \ell_1(\mathbf{x},y)y - 1$. As in Section 2, we introduce perturbations $\beta_{nu} = \beta + n^{-1/2}u$ with $u \in \mathbb{R}^d$ and $t_{nv}(\mathbf{x},y) = t(\mathbf{x},y)(1 + n^{-1/2}v(\mathbf{x},y))$ with $v$ a bounded and measurable function on $\mathbb{R}^{p+1}$. Now we must have $T_{nv}(\mathbf{x},1) = 1$, $T_{nv}(\mathbf{x},e) = 0$, $T_{nv}(\mathbf{x},e^2) = 1$ and hence $T(\mathbf{x},v) = 0$,

$T(\mathbf{x}, ve) = 0$, $T(\mathbf{x}, ve^2) = 0$. Let $V_2$ denote the set of functions $v$ with these three properties. For $u \in \mathbb{R}^p$ and $v \in V_2$ the log-likelihood ratio is

$$\log \frac{dP_{nuv}}{dP_n} = \log \frac{g_{nuv}(\mathbf{X}_0)}{g(\mathbf{X}_0)} + \sum_{i=1}^{n} \log \frac{t_{nv}(\mathbf{X}_{i-1}, \varepsilon_i(\beta_{nu}))}{t(\mathbf{X}_{i-1}, \varepsilon_i(\beta))}$$

with $\varepsilon_i(\beta) = (X_i - r_\beta(\mathbf{X}_{i-1}))/s_\beta(\mathbf{X}_{i-1})$. Similarly as in Theorem 1 we obtain local asymptotic normality. We introduce the two-dimensional vector $\ell = (\ell_1, \ell_2)^\top$ and the $d \times 2$ matrix

$$M(\mathbf{x}) = M_\beta(\mathbf{x}) = \frac{1}{s_\beta(\mathbf{x})}(\dot{r}(\mathbf{x}), \dot{s}(\mathbf{x})).$$

**Theorem 2.** *Let $(u, v) \in \mathbb{R}^d \times V_2$. Suppose Assumptions 3 and 4 hold and $g$ depends smoothly on the parameters in the sense that $\int |g_{nuv}(\mathbf{x}) - g(\mathbf{x})|\, d\mathbf{x} \to 0$. Then local asymptotic normality (2.1), (2.2) holds with*

$$s_{uv}(\mathbf{x}, y) = u^\top M(\mathbf{x})\ell(\mathbf{x}, y) + v(\mathbf{x}, y).$$

Consider now a *model* for $T$, i.e. a family $\mathcal{T}$ of conditional distributions $T$ with $T(\mathbf{x}, e) = 0$ and $T(\mathbf{x}, e^2) = 1$. Assume that the fixed $T$ is in $\mathcal{T}$. Let $W$ denote the set of all $v \in V_2$ such that $T_{nv}$ lies in $\mathcal{T}$. Assume that $W$ is a linear space. Let $\bar{W}$ and $\bar{V}_2$ denote the closures of $W$ and $V_2$ in $L_2(G \otimes T)$. The definitions of regular and efficient estimators are the same as in Section 2. In the definitions of asymptotically linear estimator and differentiable functional, replace $\bar{V}_1$ by $\bar{V}_2$. Again, *an estimator is regular and efficient if and only if it is asymptotically linear with influence function equal to the canonical gradient.*

**Residual distribution partially independent of the past.** Let $q \in \{0, \ldots, p\}$. Suppose $\mathcal{T}$ consists of the conditional distributions $T$ with $T(\mathbf{x}, dy) = T_0(B\mathbf{x}, dy)$ and $T_0(B\mathbf{x}, e) = 0$, $T_0(B\mathbf{x}, e^2) = 1$ for some known function $B : \mathbb{R}^p \to \mathbb{R}^q$. Fix $T_0$. Let $t_0$ denote the density of $T_0$. Set $\ell_{01} = -t_0'/t_0$ and $\ell_{02}(B\mathbf{x}, y) = y\ell_{01}(B\mathbf{x}, y) - 1$. Write $\ell_0 = (\ell_{01}, \ell_{02})^\top$. The local parameter space $W$ of $\mathcal{T}$ consists of the functions $v \in V_2$ of the form $v(\mathbf{x}, y) = v_0(B\mathbf{x}, y)$. For $u \in \mathbb{R}^d$ and $v \in W$ of the form $v(\mathbf{x}, y) = v_0(B\mathbf{x}, y)$ we have

$$s_{uv}(\mathbf{X}, \varepsilon) = u^\top M(\mathbf{X})\ell_0(B\mathbf{X}, \varepsilon) + v_0(B\mathbf{X}, \varepsilon).$$

Set $\psi(y) = (y, y^2 - 1)^\top$. The conditions $T_0(B\mathbf{x}, e) = 0$ and $T_0(B\mathbf{x}, e^2) = 1$ can be written $E(\psi(\varepsilon)|B\mathbf{X}) = 0$. Hence the componentwise projection of $\ell_0(B\mathbf{x}, y)$ onto $\bar{W}$ is

$$\ell_0^*(B\mathbf{x}, y) = \ell_0(B\mathbf{x}, y) - U_0^\top(B\mathbf{x})\psi(y),$$

where $U_0(B\mathbf{X}) = \Psi_0^{-1}(B\mathbf{X})L_0(B\mathbf{X})$ with $\Psi_0(B\mathbf{X}) = E(\psi(\varepsilon)\psi^\top(\varepsilon)|B\mathbf{X})$ and $L_0(B\mathbf{X}) = E(\psi(\varepsilon)\ell_0^\top(B\mathbf{X}, \varepsilon)|B\mathbf{X})$. We have

$$\Psi_0(B\mathbf{X}) = \begin{pmatrix} 1 & E(\varepsilon^3|B\mathbf{X}) \\ E(\varepsilon^3|B\mathbf{X}) & E(\varepsilon^4|B\mathbf{X}) - 1 \end{pmatrix}$$

and hence

$$\Psi_0^{-1}(B\mathbf{X}) = c(B\mathbf{X}) \begin{pmatrix} E(\varepsilon^4|B\mathbf{X}) - 1 & -E(\varepsilon^3|B\mathbf{X}) \\ -E(\varepsilon^3|B\mathbf{X}) & 1 \end{pmatrix}$$

with

$$1/c(B\mathbf{X}) = \det \Psi_0(B\mathbf{X}) = E(\varepsilon^4|B\mathbf{X}) - 1 - E(\varepsilon^3|B\mathbf{X})^2.$$

¿From $T_0(B\mathbf{x}, 1) = 1$ and Assumption 4 we obtain

$$E(\varepsilon \ell_{01}(B\mathbf{X}, \varepsilon)|B\mathbf{X}) = 1, \quad E(\varepsilon^2 \ell_{01}(B\mathbf{X}, \varepsilon)|B\mathbf{X}) = 0, \quad E(\varepsilon^3 \ell_{01}(B\mathbf{X}, \varepsilon)|B\mathbf{X}) = 3,$$

and hence

$$L_0(B\mathbf{X}) = L = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

We arrive at

$$U_0(B\mathbf{X}) = c(B\mathbf{X}) \begin{pmatrix} E(\varepsilon^4|B\mathbf{X}) - 1 & -2E(\varepsilon^3|B\mathbf{X}) \\ -2E(\varepsilon^3|B\mathbf{X}) & 2 \end{pmatrix}.$$

Set

$$M_0(B\mathbf{X}) = E(M(\mathbf{X})|B\mathbf{X}) = (E\dot{r}(\mathbf{X})/s_\beta(\mathbf{X})|B\mathbf{X}), E\dot{s}(\mathbf{X})/s_\beta(\mathbf{X})|B\mathbf{X})).$$

We obtain the orthogonal decomposition

$$s_{uv}(\mathbf{X}, \varepsilon) = u^\top \tau(\mathbf{X}, \varepsilon) + u^\top M_0(B\mathbf{X}) \ell_0^*(B\mathbf{X}, \varepsilon) + v(B\mathbf{X}, \varepsilon), \tag{3.1}$$

where

$$\tau(\mathbf{X}, \varepsilon) = (M(\mathbf{X}) - M_0(B\mathbf{X}))\ell_0(B\mathbf{X}, \varepsilon) + M_0(B\mathbf{X}) U_0^\top(B\mathbf{X}) \psi(\varepsilon)$$

is the score function for $\beta$ at $T_0$. It follows that the efficient influence function for $\beta$ is $\Lambda^{-1}\tau(\mathbf{x}, y)$ with information matrix

$$\begin{aligned} \Lambda &= E[\tau(\mathbf{X}, \varepsilon)\tau^\top(\mathbf{X}, \varepsilon)] \\ &= E[(M(\mathbf{X}) - M_0(B\mathbf{X}))J_0(B\mathbf{X})(M(\mathbf{X}) - M_0(B\mathbf{X}))^\top] \\ &\quad + E[M_0(B\mathbf{X})L^\top \Psi_0^{-1}(B\mathbf{X}) L M_0^\top(B\mathbf{X})], \end{aligned}$$

where $J_0(B\mathbf{X}) = E(\ell_0(B\mathbf{X}, \varepsilon)\ell_0^\top(B\mathbf{X}, \varepsilon)|B\mathbf{X})$ is the conditional Fisher information matrix for location and scale of $\varepsilon$. An efficient estimator of $\beta$ is now obtained as a one-step improvement of a $n^{1/2}$-consistent initial estimator, similarly as in Section 2.

*Special case: Nonparametric model.* Suppose $\mathcal{T}$ consists of *all* conditional distributions $T$ with $T(\mathbf{x}, e) = 0$ and $T(\mathbf{x}, e^2) = 1$. Then $B\mathbf{x} = \mathbf{x}$, so that $t_0 = t$, $\ell_{01} = \ell_1 = -t'/t$, $\ell_{02}(\mathbf{x}, y) = \ell_2(\mathbf{x}, y) = y\ell_1(\mathbf{x}, y) - 1$, and $\ell_0 = \ell = (\ell_1, \ell_2)^\top$. The local parameter space of $\mathcal{T}$ is $W = V_2$. Now $M_0 = M$, $U_0 = U = \Psi^{-1}L$ with $\Psi(\mathbf{X}) = E(\psi(\varepsilon)\psi^\top(\varepsilon)|\mathbf{X})$. The componentwise projection of $\ell$ onto $\bar{V}_2$ is

$$\ell^*(\mathbf{x}, y) = \ell(\mathbf{x}, y) - U^\top(\mathbf{x})\psi(y).$$

The orthogonal decomposition (3.1) now simplifies to

$$s_{uv}(\mathbf{X},\varepsilon) = u^\top \tau(\mathbf{X},\varepsilon) + u^\top M(\mathbf{X})\ell^*(\mathbf{X},\varepsilon) + v(\mathbf{X},\varepsilon)$$

with

$$\tau(\mathbf{X},\varepsilon) = M(\mathbf{X})U^\top(\mathbf{X})\psi(\varepsilon).$$

It follows that the efficient influence function for $\beta$ is $\Lambda^{-1}\tau(\mathbf{x},y)$ with

$$\Lambda = E[\tau(\mathbf{X},\varepsilon)\tau^\top(\mathbf{X},\varepsilon)] = E[M(\mathbf{X})L^\top\Psi^{-1}(\mathbf{X})LM^\top(\mathbf{X})].$$

For a different derivation see Wefelmeyer (1996) and, more generally, Müller and Wefelmeyer (2002). They also show that an efficient estimator $\hat\beta$ of $\beta$ is obtained as a solution of the martingale estimating equations

$$\sum_{i=1}^n M_\beta(\mathbf{X}_{i-1})\hat U^\top(\mathbf{X}_{i-1})\psi\Big(\frac{X_i - r_\beta(\mathbf{X}_{i-1})}{s_\beta(\mathbf{X}_{i-1})}\Big) = 0, \tag{3.2}$$

where the estimator $\hat U(\mathbf{x})$ of $U(\mathbf{x})$ is obtained by replacing the conditional third and fourth moments $T(\mathbf{x},e^3)$ and $T(\mathbf{x},e^4)$ by Nadaraya–Watson estimators based on estimated residuals $\tilde\varepsilon_i = (X_i - r_{\tilde\beta}(X_{i-1}))/s_{\tilde\beta}(X_{i-1})$. Here $\tilde\beta$ is some $n^{1/2}$-consistent estimator of $\beta$.

**Remark 4.** The information gain of partial independence over the nonparametric model is calculated as follows. Fix a conditional distribution $T$ in the smaller model. It is of the form $T(\mathbf{x},dy) = T_0(B\mathbf{x},dy)$. The nonparametric score function for $\beta$ is then

$$\tau(\mathbf{X},\varepsilon) = M(\mathbf{X})U_0^\top(B\mathbf{X})\psi(\varepsilon).$$

Let $\tau_0$ denote the efficient score function in the smaller model. Then

$$\tau(\mathbf{X},\varepsilon) - \tau_0(\mathbf{X},\varepsilon) = (M(\mathbf{X}) - M_0(B\mathbf{X}))(\ell_0(B\mathbf{X},\varepsilon) - U_0^\top(B\mathbf{X})\psi(\varepsilon).$$

The information gain is the variance of $\tau(\mathbf{X},\varepsilon) - \tau_0(\mathbf{X},\varepsilon)$,

$$E[(M(\mathbf{X}) - M_0(B\mathbf{X}))(J_0(B\mathbf{X}) - L_0^\top(B\mathbf{X})\Psi_0(B\mathbf{X})L_0(B\mathbf{X}))(M(\mathbf{X}) - M_0(B\mathbf{X}))^\top].$$

*Special case: Lagged residuals or shorter memory.* Suppose the conditional distribution $T$ is lagged by $p-q$ time points, i.e. $T(\mathbf{x},dy) = T_0(\mathbf{x}_0,dy)$ for $\mathbf{x}_0 = (x_1,\dots,x_q)$. Setting $B\mathbf{x} = \mathbf{x}_0$, this can be expressed as $T(\mathbf{x},dy) = T_0(B\mathbf{x},dy)$. Setting $\mathbf{x} = (\mathbf{x}_0,\mathbf{x}_1)$, we now have

$$E(\dot r(\mathbf{X})/s(\mathbf{x})|\mathbf{X}_0 = \mathbf{x}_0) = \frac{\int(\dot r(\mathbf{x}_0,\mathbf{x}_1)/s(\mathbf{x}_0,\mathbf{x}_1))g(\mathbf{x}_0,\mathbf{x}_1)\,d\mathbf{x}_1}{\int g(\mathbf{x}_0,\mathbf{x}_1)\,d\mathbf{x}_1};$$

similarly for $E(\dot s(\mathbf{X})/s(\mathbf{x})|\mathbf{X}_0 = \mathbf{x}_0)$. These conditional expectations can be estimated by Nadaraya–Watson estimators. The case of shorter memory is treated analogously.

*Special case: Independent innovations.* Suppose the autoregressive model (1.5) has independent innovations. Then $\mathcal{T}$ consists of all conditional distributions $T$ of the form $T(\mathbf{x}, dy) = f(y)\,dy$ with $\int yf(y)\,dy = 0$ and $\int y^2 f(y)\,dy = 1$. Setting $q = 0$ and $B\mathbf{x} = 0$, this can be expressed as $T(\mathbf{x}, dy) = T(B\mathbf{x}, dy) = T(0, dy)$. Fix $f$. Then $t(\mathbf{x}, y) = f(y)$, $\ell_1(y) = -f'(y)/f(y)$, $\ell_2(y) = y\ell_1(y) - 1$. The local parameter space $W$ of $\mathcal{T}$ consists of the bounded measurable functions $v$ on $\mathbb{R}$ with $E[v(\varepsilon)] = 0$, $E[v(\varepsilon)\varepsilon] = 0$ and $E[v(\varepsilon)\varepsilon^2] = 0$. The projection of $\ell_1(y)$ onto $\bar{W}$ is $\ell^*(y) = \ell(y) - U_0^\top y$ with $U_0 = \Psi_0^{-1}L$ and

$$\Psi_0 = E[\psi(\varepsilon)\psi^\top(\varepsilon)] = \begin{pmatrix} 1 & E[\varepsilon^3] \\ E[\varepsilon^3] & E[\varepsilon^4] - 1 \end{pmatrix}.$$

With $1/c = \det \Psi_0 = E[\varepsilon^4] - 1 - E[\varepsilon^3]^2$ we obtain

$$U_0 = \Psi_0^{-1}L = c \begin{pmatrix} E[\varepsilon^4] - 1 & -2E[\varepsilon^3] \\ -2E[\varepsilon^3] & 2 \end{pmatrix}.$$

Set

$$M_0 = E[M(\mathbf{X})] = (E[\dot{r}(\mathbf{X})/s_\beta(\mathbf{X})], E[\dot{s}(\mathbf{X})/s_\beta(\mathbf{X})]).$$

It follows that the efficient influence function for $\beta$ is $\Lambda^{-1}\tau(\mathbf{x}, y)$ with score function

$$\tau(\mathbf{X}, \varepsilon) = (M(\mathbf{X}) - M_0)\ell(\varepsilon) + M_0 U_0^\top \psi(\varepsilon)$$

and information matrix

$$\Lambda = E[\tau(\mathbf{X}, \varepsilon)\tau^\top(\mathbf{X}, \varepsilon)] = E[(M(\mathbf{X}) - M_0)J(M(\mathbf{X}) - M_0)^\top] + M_0 L^\top \Psi_0^{-1} L M_0^\top,$$

where $J = E[\ell(\varepsilon)\ell^\top(\varepsilon)]$ is the Fisher information matrix for location and scale of $\varepsilon$. Drost, Klaassen and Werker (1997) construct an efficient estimator $\hat{\beta}$ of $\beta$ as a one-step improvement of some $n^{1/2}$-consistent estimator.

**Residual distribution invariant under transformations.** Let $B_1, \ldots, B_m$ be a group of transformations on $\mathbb{R}^{p+1}$. Suppose $\mathcal{T}$ consists of all conditional distributions $T$ with $T(\mathbf{x}, e) = 0$, $T(\mathbf{x}, e^2) = 1$ and density $t$ fulfilling $t(\mathbf{z}) = t(B_j\mathbf{z})$ for $j = 1, \ldots, m$. Fix $t$. The local parameter space consists of the functions $v$ in $V_2$ with $v(\mathbf{z}) = v(B_j\mathbf{z})$ for $j = 1, \ldots, m$. Write $t'(\mathbf{x}, y) = \partial_y t(\mathbf{x}, y)$ and set $\ell_1 = -t'/t$, $\ell_2(\mathbf{x}, y) = y\ell_1(\mathbf{x}, y) - 1$, $\ell = (\ell_1, \ell_2)^\top$. The compoentwise projection of $\ell$ onto $\bar{V}_2$ is $\ell(\mathbf{x}, y) - U^\top(\mathbf{x})\psi(y)$ with $U(\mathbf{x}) = \Psi^{-1}(\mathbf{x})L$ and $\Psi(\mathbf{X}) = E(\psi(\varepsilon)\psi^\top(\varepsilon)|\mathbf{X})$. In order to decompose $s_{uv}(\mathbf{x}, y) = u^\top M(\mathbf{x})\ell(\mathbf{x}, y) + v(\mathbf{x}, y)$, we set

$$\lambda(\mathbf{x}, y) = M(\mathbf{x})\ell(\mathbf{x}, y), \qquad \mu(\mathbf{x}, y) = M(\mathbf{x})U^\top(\mathbf{x})\psi(y).$$

The componentwise projection of $\lambda(\mathbf{x}, y) - \mu(\mathbf{x}, y) = M(\mathbf{x})(\ell(\mathbf{x}, y) - U^\top(\mathbf{x})\psi(y))$ onto $\bar{W}$ is $\lambda_0(\mathbf{x}, y) - \mu_0(\mathbf{x}, y)$ with

$$\lambda_0(\mathbf{x}, y) = \frac{1}{m}\sum_{j=1}^m \lambda(B_j\mathbf{z}), \qquad \mu_0(\mathbf{x}, y) = \frac{1}{m}\sum_{j=1}^m \mu(B_j\mathbf{z}).$$

Hence $\tau = \lambda - \lambda_0 + \mu_0$, and the efficient influence function for $\beta$ is $\Lambda^{-1}\tau(\mathbf{x}, y)$ with

$$\begin{aligned}\Lambda &= E[\tau(\mathbf{X}, \varepsilon)\tau^\top(\mathbf{X}, \varepsilon)] \\ &= E[(\lambda(\mathbf{X}, \varepsilon) - \lambda_0(\mathbf{X}, \varepsilon))(\lambda(\mathbf{X}, \varepsilon) - \lambda_0(\mathbf{X}, \varepsilon))^\top] + E[\mu_0(\mathbf{X}, \varepsilon)\mu_0(\mathbf{X}, \varepsilon)^\top].\end{aligned}$$

The information gain of group invariance over the nonparametric model is analogous to Section 2.

*Special case: Symmetric residuals.* Suppose $\mathcal{T}$ consists of all conditional distributions $T$ with $T(\mathbf{x}, e) = 0$, $T(\mathbf{x}, e^2) = 1$ and density $t$ fulfilling $t(\mathbf{x}, y) = t(\mathbf{x}, -y)$. Fix $t$. The local parameter space $W$ consists of the functions $v$ in $V_2$ with $v(\mathbf{z}) = v(-\mathbf{z})$. We have $t'(\mathbf{x}, y) = -t'(-\mathbf{x}, -y)$ and hence $\ell_1(\mathbf{x}, y) = -\ell_1(-\mathbf{x}, -y)$ and $\ell_2(\mathbf{x}, y) = y\ell_1(\mathbf{x}, y) - 1 = \ell_2(-\mathbf{x}, -y)$, i.e, $\ell_1$ is antisymmetric and $\ell_2$ is symmetric. We obtain

$$\begin{aligned}\lambda_0(\mathbf{x}, y) &= \frac{1}{2}(M(\mathbf{x})\ell(\mathbf{x}, y) + M(-\mathbf{x})\ell(-\mathbf{x}, -y)) \\ &= \frac{1}{2}\Big(\frac{\dot r(\mathbf{x})}{s(\mathbf{x})} - \frac{\dot r(-\mathbf{x})}{s(-\mathbf{x})}\Big)\ell_1(\mathbf{x}, y) + \frac{1}{2}\Big(\frac{\dot s(\mathbf{x})}{s(\mathbf{x})} + \frac{\dot s(-\mathbf{x})}{s(-\mathbf{x})}\Big)\ell_2(\mathbf{x}, y).\end{aligned}$$

We have $\Psi_{jk}(-\mathbf{x}) = (-1)^{j+k}\Psi_{jk}(\mathbf{x})$, i.e., $\Psi(-\mathbf{x})$ is obtained from $\Psi(\mathbf{x})$ by changing the off-diagonal signs. Then $\Psi^{-1}$ and $U$ have the same property. Hence the first component of $U^\top(-\mathbf{x})\psi(-y)$ changes sign, while the second does not. We obtain

$$\begin{aligned}\mu_0(\mathbf{x}, y) = &\frac{1}{2}\Big(\frac{\dot r(\mathbf{x})}{s(\mathbf{x})} - \frac{\dot r(-\mathbf{x})}{s(-\mathbf{x})}\Big)\big((E(\varepsilon^4|\mathbf{X}) - 1)y - 2E(\varepsilon^3|\mathbf{X})(y^2 - 1)\big) \\ &+ \frac{1}{2}\Big(\frac{\dot s(\mathbf{x})}{s(\mathbf{x})} + \frac{\dot s(-\mathbf{x})}{s(-\mathbf{x})}\Big)\big( - 2E(\varepsilon^3|\mathbf{X})y + 2(y^2 - 1)\big).\end{aligned}$$

The score function for $\beta$ is $\tau = \lambda - \lambda_0 + \mu_0$.

*Special case: Conditionally symmetric residuals.* Suppose $\mathcal{T}$ consists of all conditional distributions $T$ with $T(\mathbf{x}, e) = 0$, $T(\mathbf{x}, e^2) = 1$ and density $t$ fulfilling $t(\mathbf{x}, 2x_p - y) = t(\mathbf{x}, y)$. Fix $t$. The local parameter space $W$ consists of the functions $v$ in $V_2$ with $v(\mathbf{x}, 2x_p - y) = v(\mathbf{x}, y)$, and $\ell_1 = -t'/t$ is conditionally antisymmetric and therefore orthogonal to $W$. For $\ell_2(\mathbf{x}, y) = y\ell_1(\mathbf{x}, y) - 1$ we have the orthogonal decomposition

$$\ell_2(\mathbf{x}, y) = x_p\ell_1(\mathbf{x}, y) + (y - x_p)\ell_1(\mathbf{x}, y) - 1.$$

Hence the score function for $\beta$ is $M(\mathbf{x})\delta(\mathbf{x}, y)$ with $\delta(\mathbf{x}, y) = (\ell_1(\mathbf{x}, y), x_p\ell_1(\mathbf{x}, y))^\top$, and the information matrix for $\beta$ is

$$\Lambda = E[M(\mathbf{X})D(\mathbf{X})M^\top(\mathbf{X})]$$

with

$$D(\mathbf{x}) = E(\delta(\mathbf{X}, \varepsilon)\delta^\top(\mathbf{X}, \varepsilon)|\mathbf{X} = \mathbf{x}) = \begin{pmatrix} 1 & x_p \\ x_p & x_p^2 \end{pmatrix} T(\mathbf{x}, \ell_1^2).$$

*Special case: Homoscedastic autoregression.* A *homoscedastic* autoregressive model of order $p$ is a Markov chain with transition distribution (1.5) such that the conditional variance $s_\beta^2(\mathbf{x}) = \sigma_\beta^2$ does not depend on $\mathbf{x}$. Usually the variance $\sigma_\beta^2$ is assumed to vary independently of $\beta$. This can be achieved by replacing the parameter $\beta$ by a pair $\nu = (\beta, \sigma^2)$ such that $r_\nu = r_\beta$ depends only on $\beta$, and $s_\beta^2 = \sigma^2$. We will however treat the more general case $s_\beta^2(\mathbf{x}) = \sigma_\beta^2$ as it fits better into the nonparametric model. Then the efficient influence function has the same form $\gamma(\mathbf{x}, y) = \Lambda^{-1} M(\mathbf{x}) U^\top(\mathbf{x}) \psi(y)$ as in the nonparametric model, but now with

$$M(\mathbf{x}) = M_\beta(\mathbf{x}) = \frac{1}{\sigma_\beta}(\dot{r}(\mathbf{x}), \dot{\sigma}).$$

# 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] Cramér, H. (1946). A contribution to the theory of statistical estimation. *Skand. Aktuarietidskr.* **29**, 85-94.

[3] Darmois, G. (1945). Sur les lois limites de la dispersion de certaines estimations. *Rev. Int. Statist. Inst.* **13**, 9–15.

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

[5] Fréchet, M. (1943). Sur l'extension de certaines evaluations statistiques de petits echantillons. *Rev. Int. Statist. Inst.* **11**, 182–205.

[6] Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* **14**, 323–330.

[7] Inagaki, N. (1970). On the limiting distribution of a sequence of estimators with uniformity property. *Ann. Inst. Statist. Math.* **22**, 1–13.

[8] Janssen, A. (2003). A nonparametric Cramér–Rao inequality for estimators of statistical functionals. *Statist. Probab. Lett.* **64**, 347–358.

[9] Kaufman, S. (1966). Asymptotic efficiency of the maximum likelihood estimator. *Ann. Inst. Statist. Math.* **18**, 155–178.

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

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

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

[13] Le Cam, L. (1972). Limits of experiments. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **1**, 245–261.

[14] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer Series in Statistics, Springer, New York.

[15] Le Cam, L. (2000). La statistique mathématique depuis 1950. In: *Development of Mathematics 1950–2000* (J. P. Pier, ed.), 735–761, Birkhäuser, Basel.

[16] Müller, U. U. and Wefelmeyer, W. (2002). Autoregression, estimating functions, and optimality criteria. In: *Advances in Statistics, Combinatorics and Related Areas* (C. Gulati, Y.-X. Lin, J. Rayner and S. Mishra, eds.), 180-195, World Scientific, Singapore.

[17] Ngatchou-Wandji, J. (2008). Estimation in a class of nonlinear heteroscedastic time series models. *Electron. J. Stat.* **2**, 40–62.

[18] Pfanzagl, J. (2001). A nonparametric asymptotic version of the Cramér–Rao bound. In: *State of the Art in Probability and Statistics* (M. de Gunst, C. A. J. Klaassen and A. W. van der Vaart, eds.), 499–517, IMS Lecture Notes Monogr. Ser. **36**, Institute of Mathematical Statistics, Beachwood.

[19] Pfanzagl, J. and Wefelmeyer, W. (1982). *Contributions to a General Asymptotic Statistical Theory.* Lecture Notes in Statistics **13**, Springer, New York.

[20] Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37**, 81–91.

[21] Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples. *J. Roy. Statist. Soc. Ser. B* **24**, 46–72.

[22] Rao, C. R. (1963). Criteria of estimation in large samples. *Sankhyā Ser. A* **25**, 189–206.

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

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

[25] Schick, A. (2001). Sample splitting with Markov chains. *Bernoulli* **7**, 33–61.

[26] Stein, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1**, 187–195.

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

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