Institute of Mathematical Statistics

LECTURE NOTES — MONOGRAPH SERIES

PLUG-IN ESTIMATORS IN SEMIPARAMETRIC STOCHASTIC PROCESS MODELS

Ursula U. Müller
Universität Bremen

Anton Schick
Binghamton University

Wolfgang Wefelmeyer
Universität Siegen

ABSTRACT

Consider a locally asymptotically normal semiparametric model with a real parameter ϑ and a possibly infinite-dimensional parameter F. We are interested in estimating a real-valued functional $a(F)$. If $\hat{a}_\vartheta$ estimates $a(F)$ for known ϑ, and $\hat{\vartheta}$ estimates ϑ, then the plug-in estimator $\hat{a}_{\hat{\vartheta}}$ estimates $a(F)$ if ϑ is unknown. We show that $\hat{a}_{\hat{\vartheta}}$ is asymptotically linear and regular if $\hat{a}_\vartheta$ and $\hat{\vartheta}$ are, and calculate the influence function and the asymptotic variance of $\hat{a}_{\hat{\vartheta}}$. If $a(F)$ can be estimated adaptively with respect to ϑ, then $\hat{a}_{\hat{\vartheta}}$ is efficient if $\hat{a}_\vartheta$ is efficient. If $a(F)$ cannot be estimated adaptively, then for $\hat{a}_{\hat{\vartheta}}$ to be efficient, $\hat{\vartheta}$ must also be efficient. We illustrate the results with stochastic process models, in particular with time series models, and discuss extensions of the results.

Key Words: Empirical estimator, asymptotically linear estimator, influence function, regular estimator, Markov chain model, nonlinear regression, residual distribution, nonlinear autoregression, innovation distribution, stochastic equicontinuity, stochastic differentiability.

1 Introduction

Let $\overline{\mathcal{P}}_n = \{P_{n\vartheta F} : \vartheta \in \Theta, F \in \mathcal{F}\}$ denote a sequence of semiparametric models, with Θ one-dimensional and $\mathcal{F}$ a possibly infinite-dimensional set. We are interested in estimating a real-valued functional $a(F)$. For each ϑ let $\hat{a}_\vartheta$ be an estimator for $a(F)$ when ϑ is known, and let $\hat{\vartheta}$ be an estimator

for ϑ. The *plug-in estimator* for $a(F)$ is $\hat{a}_{\hat{\vartheta}}$, obtained by substituting $\hat{\vartheta}$ for ϑ in $\hat{a}_\vartheta$.

In Section 2 we introduce a semiparametric version of local asymptotic normality, and a concept of asymptotically linear estimators for functionals on such models. This concept requires embedding the semiparametric model into a "nonparametric" supermodel which is to some extent arbitrary, although for specific types of stochastic models there usually is a natural choice. We recall two results which are essentially known, at least for i.i.d. models and Markov chain models. The first characterizes regular estimators among asymptotically linear ones; the second characterizes efficient estimators.

In Section 3 we show that $\hat{a}_{\hat{\vartheta}}$ is asymptotically linear if $\hat{a}_\vartheta$ and $\hat{\vartheta}$ are, and calculate the influence function of $\hat{a}_{\hat{\vartheta}}$.

In Section 4 we assume that both $\hat{a}_\vartheta$ and $\hat{\vartheta}$ are asymptotically linear and regular, and show that then $\hat{a}_{\hat{\vartheta}}$ is also regular. The characterization of regular estimators then allows comparison of the asymptotic variance of the plug-in estimator with the minimal asymptotic variance. In particular, if $a(F)$ can in principle be estimated adaptively with respect to ϑ, i.e., if knowledge of ϑ does not contain information about $a(F)$, then $\hat{a}_{\hat{\vartheta}}$ is efficient whenever $\hat{a}_\vartheta$ is efficient for known ϑ. If $a(F)$ cannot be estimated adaptively with respect to ϑ, then both $\hat{a}_\vartheta$ and $\hat{\vartheta}$ must be efficient for $\hat{a}_{\hat{\vartheta}}$ to be efficient. For i.i.d. observations and estimators of the form $\hat{a}_\vartheta = a(\hat{F}_\vartheta)$, these results are due to Klaassen and Putter (1999).

Sections 5 and 6 contain applications to models with i.i.d. observations and to Markov chain models, respectively. The discussion is heuristic in these two sections. Some extensions of the results are outlined in Section 7.

2 Characterizing regular and efficient estimators

In this section we recall briefly characterizations of regular and of efficient estimators of real-valued functionals in semiparametric models. For the i.i.d. case we refer to Bickel, Klaassen, Ritov and Wellner (1998), Chapter 3 and in particular Section 3.4. For Markov chains see Wefelmeyer (1999); for general models and counting process models see Andersen, Borgan, Gill and Keiding (1993, Chapter VIII); for general parametric models see Le Cam and Yang (1990).

We will embed the semiparametric model in a "nonparametric" supermodel. The reason is that we want a sufficiently rich class of "asymptotically linear" estimators, and asymptotic linearity will be defined in terms of statistics approximating local likelihood ratios; see below.

For $n \in \mathbf{N}$ let $\mathcal{P}_n$ denote a collection of probability measures on some measurable space $(\Omega_n, \mathcal{F}_n)$. Fix a sequence P_n in $\mathcal{P}_n$ and assume that the

model is *locally asymptotically normal* at P_n in the following sense. There is a linear space H with inner product (h, h') and corresponding norm $\|h\|$, there is a sequence of random linear functionals S_n on H, and for each $h \in H$ there are perturbations P_{nh} of P_n within $\mathcal{P}_n$ such that

$$\log \frac{dP_{nh}}{dP_n} = S_n(h) - \frac{1}{2}\|h\|^2 + o_{P_n}(1), \tag{2.1}$$

$$S_n(h) \Rightarrow \|h\|N \quad \text{under } P_n, \tag{2.2}$$

with N denoting a standard normal random variable. The linear space H may be interpreted as (approximate) *tangent space* of $\mathcal{P}_n$ at P_n.

Now consider a sequence of semiparametric submodels $\overline{\mathcal{P}}_n = \{P_{n\vartheta F} : \vartheta \in \Theta, F \in \mathcal{F}\}$ of $\mathcal{P}_n$, with Θ one-dimensional and $\mathcal{F}$ an arbitrary, possibly infinite-dimensional set. We fix ϑ and F and consider perturbations $\vartheta_{nu} = \vartheta + n^{-1/2}u$ of ϑ and F_{nv} of F, with v in some linear space V. In general, the appropriate rate may be different from $n^{-1/2}$, but it will be $n^{-1/2}$ in our applications, which is why we have taken it to be $n^{-1/2}$ here. We assume that (2.1) and (2.2) hold at $P_n = P_{n\vartheta F}$, and that the semiparametric model is locally asymptotically normal in the following sense. There are an element m in H and a linear operator $D : V \to H$ such that for $P_{nuv} = P_{n\vartheta_{nu}F_{nv}}$,

$$\log \frac{dP_{nuv}}{dP_{n\vartheta F}} = S_n(um + Dv) - \frac{1}{2}\|um + Dv\|^2 + o_{P_{n\vartheta F}}(1).$$

From (2.2),

$$S_n(um + Dv) \Rightarrow \|um + Dv\|N \quad \text{under } P_{n\vartheta F}.$$

The tangent space of the semiparametric model is

$$\overline{H} = [m] + DV,$$

with $[m]$ denoting the linear span of m. The one-dimensional space $[m]$ is the tangent space for known F, obtained by perturbing ϑ. The space DV is the tangent space for known ϑ, obtained by perturbing F. We assume that DV is closed, and that m does not belong to DV. We may think of m as the *score function* for ϑ.

Let $a(\vartheta, F)$ be a real-valued functional. It is called *differentiable* at (ϑ, F) with *gradient* g if $g \in H$ and

$$n^{1/2}(a(\vartheta_{nu}, F_{nv}) - a(\vartheta, F)) \to (g, um + Dv) \quad \text{for } u \in \mathbf{R}, v \in V. \tag{2.3}$$

The *canonical gradient* $\overline{g}$ is the projection of g onto $\overline{H}$; it is therefore of the form $\overline{u}m + D\overline{v}$.

An estimator $\hat{a}$ is called *asymptotically linear* for $a(\vartheta, F)$ with *influence function* b if $b \in H$ and

$$n^{1/2}(\hat{a} - a(\vartheta, F)) = S_n(b) + o_{P_{n\vartheta F}}(1). \tag{2.4}$$

An estimator $\hat{a}$ is called *regular* for $a(\vartheta, F)$ with *limit* L if L is a random variable such that

$$n^{1/2}(\hat{a} - a(\vartheta_{nu}, F_{nv})) \Rightarrow L \quad \text{under } P_{nuv} \text{ for all } u \in \mathbf{R}, v \in V. \tag{2.5}$$

The convolution theorem of Hájek (1970) says that

$$L = \|\overline{g}\|N + M \quad \text{in distribution,}$$

with M independent of N. This justifies calling a regular estimator $\hat{a}$ *efficient* for $a(\vartheta, F)$ if it is asymptotically normal with variance $\|\overline{g}\|^2$,

$$n^{1/2}(\hat{a} - a(\vartheta, F)) \Rightarrow \|\overline{g}\|N \quad \text{under } P_{n\vartheta F}.$$

We have the following two characterizations:

(1) *An asymptotically linear estimator is regular for $a(\vartheta, F)$ if and only if its influence function is a gradient of $a(\vartheta, F)$.*

(2) *A regular estimator is efficient for $a(\vartheta, F)$ if and only if it is asymptotically linear with influence function equal to the canonical gradient of $a(\vartheta, F)$.*

Now it becomes clear why we have introduced the "nonparametric" model. If we restrict attention to the semiparametric model, then there is only one gradient, the canonical one, and all regular and asymptotically linear estimators are asymptotically equivalent. In the examples of Sections 5 and 6 we will need to consider estimators whose influence functions are non-canonical gradients. The concept of asymptotically linear estimators is arbitrary in that it depends on the choice of "nonparametric" model; see Wefelmeyer (1991).

3 Asymptotic linearity of plug-in estimators

Consider the problem of estimating a real-valued functional $a(F)$ in the semiparametric model $\overline{\mathcal{P}}_n = \{P_{n\vartheta F} : \vartheta \in \Theta, F \in \mathcal{F}\}$. Fix ϑ and F. Suppose that for each τ near ϑ we have an asymptotically linear estimator $\hat{a}_\tau$ of $a(F)$, with influence function b_τ. We assume that asymptotic linearity holds *locally uniformly* in shrinking neighborhoods of ϑ,

$$\sup_{|u| \le \Delta} |n^{1/2}(\hat{a}_{\vartheta_{nu}} - a(F)) - S_n(b_{\vartheta_{nu}})| = o_{P_{n\vartheta F}}(1), \tag{3.1}$$

and that $S_n(b_{\vartheta_{nu}})$ is *stochastically differentiable* at $u = 0$,

$$\sup_{|u|\le\Delta} |S_n(b_{\vartheta_{nu}}) - S_n(b_\vartheta) + u(m, b_\vartheta)| = o_{P_{n\vartheta F}}(1). \tag{3.2}$$

Now let $\hat\vartheta$ be asymptotically linear for ϑ, with influence function c,

$$n^{1/2}(\hat\vartheta - \vartheta) = S_n(c) + o_{P_{n\vartheta F}}(1).$$

By conditions (3.1) and (3.2), the plug-in estimator $\hat a_{\hat\vartheta}$ fulfills

$$\begin{aligned} n^{1/2}(\hat a_{\hat\vartheta} - a(F)) &= S_n(b_{\hat\vartheta}) + o_{P_{n\vartheta F}}(1) \\ &= S_n(b_\vartheta) - (m, b_\vartheta) n^{1/2}(\hat\vartheta - \vartheta) + o_{P_{n\vartheta F}}(1) \\ &= S_n(b_\vartheta - (m, b_\vartheta)c) + o_{P_{n\vartheta F}}(1). \end{aligned}$$

This means that $\hat a_{\hat\vartheta}$ is asymptotically linear for $a(F)$ with influence function

$$b = b_\vartheta - (m, b_\vartheta)c. \tag{3.3}$$

If $(m, b_\vartheta) = 0$, then we can relax the assumption that $\hat\vartheta$ is asymptotically linear to $n^{1/2}$-consistency and still get that the plug-in estimator is asymptotically linear, now with influence function $b = b_\vartheta$.

Asymptotic linearity (3.3) of the plug-in estimator also follows if we replace conditions (3.1) and (3.2) by a non-uniform version of (3.1),

$$n^{1/2}(\hat a_\vartheta - a(F)) = S_n(b_\vartheta) + o_{P_{n\vartheta F}}(1), \tag{3.4}$$

and an expansion of $\hat a_{\hat\vartheta}$,

$$n^{1/2}(\hat a_{\hat\vartheta} - \hat a_\vartheta) = -(m, b_\vartheta) n^{1/2}(\hat\vartheta - \vartheta) + o_{P_{n\vartheta F}}(1). \tag{3.5}$$

An application is Example 3 in Section 6.

Remark 1. (*Plug-in and sample splitting.*) Our requirements (3.1) and (3.2) are stronger than the following conditions. For every bounded sequence u_n,

$$\begin{aligned} n^{1/2}(\hat a_{\vartheta_{nu_n}} - a(F)) &= S_n(b_{\vartheta_{nu_n}}) + o_{P_{n\vartheta F}}(1), & (3.6)\\ S_n(b_{\vartheta_{nu_n}}) - S_n(b_\vartheta) &= -u_n(m, b_\vartheta) + o_{P_{n\vartheta F}}(1). & (3.7) \end{aligned}$$

Property (3.7) appears quite frequently in the literature. It has been been verified by Drost, Klaassen and Werker (1997), Jeganathan (1995), Koul and Schick (1997) and Kreiss (1987) for some time series models. A simple sufficient condition for (3.7) is given in Schick (1999a) in the context of i.i.d. observations.

It was shown by Klaassen and Putter (1999) in the context of i.i.d. observations that under these weaker conditions one can use sample splitting techniques to construct a modified version of the plug-in estimator with influence function b as in (3.3). Their construction can be generalized to stationary and ergodic Markov chains using the sample splitting techniques developed in Schick (1998). But this will not be pursued here.

Remark 2. (*Stochastic differentiability.*) We will check stochastic differentiability (3.2) for specific types of processes in Sections 5 and 6, using stochastic equicontinuity of empirical processes. Here we indicate that (3.2) also follows from a locally uniform version of local asymptotic normality. Compare the proof of Theorem 6.2 in Bickel (1982). Since F is fixed in (3.2), we will omit it from the notation. Fix ϑ and set $\tau = \vartheta_{nu} = \vartheta + n^{-1/2}u$. Assume that the parametric family $\{P_{n\vartheta} : \vartheta \in \Theta\}$ is locally asymptotically normal at ϑ,

$$\log \frac{dP_{n\tau}}{dP_{n\vartheta}} = uS_{n\vartheta}(m) - \frac{1}{2}u^2\|m\|_\vartheta^2 + o_{P_{n\vartheta}}(1).$$

Assume also that for τ near ϑ there is a tangent space H_τ such that for each $b_\tau \in H_\tau$ we have local asymptotic normality at τ,

$$\log \frac{dP_{n\tau b_\tau}}{dP_{n\tau}} = S_{n\tau}(b_\tau) - \frac{1}{2}\|b_\tau\|_\tau^2 + o_{P_{n\tau}}(1).$$

Then

$$\log \frac{dP_{n\tau b_\tau}}{dP_{n\vartheta}} = S_{n\tau}(b_\tau) + uS_{n\vartheta}(m) - \frac{1}{2}\|b_\tau\|_\tau^2 - \frac{1}{2}u^2\|m\|_\vartheta^2 + o_{P_{n\vartheta}}(1). \quad (3.8)$$

If b_τ is continuous at $\tau = \vartheta$ in an appropriate sense, $P_{n\tau b_\tau}$ will be approximately equal to $P_{n\vartheta,um+b_\vartheta}$. (For a more explicit argument it would be convenient if the sequence of "nonparametric" supermodels $\mathcal{P}_n$ were indexed by an infinite-dimensional parameter; see LeCam (1986, Chapter 11) and Greenwood and Wefelmeyer (1991, Section 4).) Hence

$$\begin{aligned} \log \frac{dP_{n\tau b_\tau}}{dP_{n\vartheta}} &\doteq \log \frac{dP_{n\vartheta,um+b_\vartheta}}{dP_{n\vartheta}} \\ &= S_{n\vartheta}(um + b_\vartheta) - \frac{1}{2}\|um + b_\vartheta\|_\vartheta^2 + o_{P_{n\vartheta}}(1). \end{aligned} \quad (3.9)$$

If both $S_{n\tau}$ and $\|\ \|_\tau$ are continuous at $\tau = \vartheta$ in an appropriate sense, we obtain from (3.8) and (3.9):

$$S_{n\vartheta}(b_\tau) - S_{n\vartheta}(b_\vartheta) + u(m, b_\vartheta)_\vartheta = o_{P_{n\vartheta}}(1).$$

Stochastic differentiability (3.2) requires that the *supremum* over $|u| \le \Delta$ is stochastically negligible. For this, we need a corresponding *strong* version of local asymptotic normality, as introduced in Fabian and Hannan (1985, Section 9.1).

4 Efficient and adaptive plug-in estimators

We continue the discussion of plug-in estimators under additional assumptions. As in Section 3, let $a(F)$ be a real-valued functional of F. Fix ϑ and F. For τ near ϑ let $\hat{a}_\tau$ be a locally uniformly asymptotically linear estimator of $a(F)$ in the sense of (3.1), and let $\hat{\vartheta}$ be an asymptotically linear estimator of ϑ. Now assume, in addition, that $a(F)$ is differentiable (2.3) at (ϑ, F), that $\hat{a}_\vartheta$ is regular at (ϑ, F) for known ϑ, and that $\hat{\vartheta}$ is regular at (ϑ, F). Then we can decompose the canonical gradient of $a(F)$ and the influence function of the plug-in estimator $\hat{a}_{\hat{\vartheta}}$ as follows.

Let $\overline{c}_F$ denote the canonical gradient of ϑ when F is *known*. It is of the form $\overline{c}_F = tm$ with t determined by

$$n^{1/2}(\vartheta_{nu} - \vartheta) = u \stackrel{!}{=} (tm, um),$$

i.e., $t = \|m\|^{-2}$ and

$$\overline{c}_F = \|m\|^{-2} m.$$

The squared length $\|m\|^2$ of m may be called the *Fisher information* for ϑ when F is known.

Let Dv_m be the projection of m onto DV. When ϑ is unknown, the canonical gradient of ϑ is characterized by three properties. It is in $\overline{H} = [m] + DV$, orthogonal to DV, and its projection onto $[m]$ is $\overline{c}_F$. Hence it is of the form $\overline{c} = t(m - Dv_m)$, with t determined by $(\overline{c} - \overline{c}_F, m) = 0$, i.e., $t = (m - Dv_m, m)^{-1} = \|m - Dv_m\|^{-2}$. Hence the canonical gradient of ϑ is

$$\overline{c} = \|m - Dv_m\|^{-2}(m - Dv_m).$$

The squared length $\|m - Dv_m\|^2$ of $m - Dv_m$ is the *Fisher information* for ϑ.

Let $\overline{g}_\vartheta$ denote the canonical gradient of $a(F)$ when ϑ is known. The canonical gradient $\overline{g}$ of $a(F)$ for unknown ϑ is characterized by three properties. It is in $\overline{H} = [m] + DV$, orthogonal to m, and its projection onto DV is $\overline{g}_\vartheta$. Hence $\overline{g} - \overline{g}_\vartheta$ is of the form $t\overline{c}$, with t determined by $(m, \overline{g}) = 0$, i.e., $t = -(m, \overline{g}_\vartheta)$, i.e.,

$$\overline{g} = \overline{g}_\vartheta - (m, \overline{g}_\vartheta)\overline{c}. \tag{4.1}$$

Since $\hat{\vartheta}$ is regular for ϑ, we obtain from characterization (1) of Section 2 that the influence function of $\hat{\vartheta}$ is a gradient of ϑ, say c. Hence

$$c - \overline{c} \perp \overline{H} = [m] + DV. \tag{4.2}$$

Since $\hat{a}_\vartheta$ is regular for $a(F)$ when ϑ is known, we obtain from characterization (1) of Section 2 that the influence function of $\hat{a}_\vartheta$ is a gradient of $a(F)$ for known ϑ, say g_ϑ. Hence

$$g_\vartheta - \overline{g}_\vartheta \perp DV. \tag{4.3}$$

From (3.3) we obtain that the plug-in estimator $\hat{a}_{\hat{\vartheta}}$ is asymptotically linear for $a(F)$, with influence function

$$g = g_\vartheta - (m, g_\vartheta)c. \tag{4.4}$$

From (4.1) and (4.4) we obtain

$$g - \overline{g} = g_\vartheta - \overline{g}_\vartheta - (m, g_\vartheta - \overline{g}_\vartheta)\overline{c} - (m, g_\vartheta)(c - \overline{c}). \tag{4.5}$$

It follows from $(m, \overline{c}) = 1$ and relations (4.2) and (4.3) that $g - \overline{g}$ is orthogonal to $\overline{H}$. Hence g is a gradient. Characterization (1) of Section 2 now implies that the plug-in estimator $\hat{a}_{\hat{\vartheta}}$ is regular.

Since $g - \overline{g}$ and $c - \overline{c}$ are orthogonal to $\overline{H}$, the asymptotic variance of the plug-in estimator $\hat{a}_{\hat{\vartheta}}$ is

$$\begin{aligned} \|g\|^2 &= \|\overline{g}\|^2 + \|g_\vartheta - \overline{g}_\vartheta\|^2 + \|c - \overline{c}\|^2 (m, g_\vartheta)^2 \\ &\quad - \|m - Dv_m\|^{-2}(m, g_\vartheta - \overline{g}_\vartheta)^2 - 2(c - \overline{c}, g_\vartheta - \overline{g}_\vartheta)(m, g_\vartheta). \end{aligned} \tag{4.6}$$

If $\hat{a}_\vartheta$ is efficient for $a(F)$ when ϑ is known, and $\hat{\vartheta}$ is efficient for ϑ, then $g_\vartheta = \overline{g}_\vartheta$ and $c = \overline{c}$. Hence $g = \overline{g}$ by (4.5), and the plug-in estimator $\hat{a}_{\hat{\vartheta}}$ is efficient for $a(F)$. By (4.1), its asymptotic variance is

$$\|\overline{g}\|^2 = \|\overline{g}_\vartheta\|^2 + \|m - Dv_m\|^{-2}(m, \overline{g}_\vartheta)^2. \tag{4.7}$$

If $\hat{a}_\vartheta$ is efficient for $a(F)$ when ϑ is known, then by (4.5) the influence function of the plug-in estimator $\hat{a}_{\hat{\vartheta}}$ is

$$g = \overline{g} - (m, g_\vartheta)(c - \overline{c}), \tag{4.8}$$

and by (4.6) the asymptotic variance of $\hat{a}_{\hat{\vartheta}}$ is

$$\|\overline{g}\|^2 + \|c - \overline{c}\|^2 (m, g_\vartheta)^2. \tag{4.9}$$

If $\hat{\vartheta}$ is efficient for ϑ, then by (4.5) the influence function of the plug-in estimator $\hat{a}_{\hat{\vartheta}}$ is

$$g = \overline{g} + g_\vartheta - \overline{g}_\vartheta - (m, g_\vartheta - \overline{g}_\vartheta)\overline{c}, \tag{4.10}$$

and by (4.6) the asymptotic variance of $\hat{a}_{\hat{\vartheta}}$ is

$$\|\overline{g}\|^2 + \|g_\vartheta - \overline{g}_\vartheta\|^2 - \|m - Dv_m\|^{-2}(m, g_\vartheta - \overline{g}_\vartheta)^2. \tag{4.11}$$

We say that $a(F)$ can be estimated *adaptively* with respect to ϑ if the asymptotic variance bound for $a(F)$ is not decreased by knowing ϑ. This is the case if and only if $\overline{g} = \overline{g}_\vartheta$. By (4.1), this is equivalent to $(m, \overline{g}_\vartheta) = 0$. Then $\overline{g}$ does not depend on $\overline{c}$. Hence the plug-in estimator $\hat{a}_{\hat{\vartheta}}$ is efficient whenever $\hat{a}_\vartheta$ is efficient for known ϑ, and the asymptotic variance is $\|\overline{g}_\vartheta\|^2$.

We say that F can be estimated *adaptively* with respect to ϑ if for every differentiable functional $a(F)$ the asymptotic variance bound for $a(F)$ is not decreased by knowing ϑ. This is equivalent to orthogonality of m and DV. Then ϑ can also be estimated adaptively with respect to F, and $\overline{c} = \overline{c}_F = \|m\|^{-2} m$.

5 The i.i.d. case

If we have i.i.d. observations $X_1, \ldots, X_n$, then a natural candidate for the "nonparametric" model of Section 2 is the usual nonparametric model, with completely unknown distribution $Q(dx)$ of X_i. (Larger nonparametric models are obtained by allowing the observations to be dependent.) Fix Q. Set $Qh = \int h(x) Q(dx)$ and

$$H = L_{2,0}(Q) = \{h \in L_2(Q) : Qh = 0\}.$$

For $h \in H$ set

$$Q_{nh}(dx) \doteq Q(dx)(1 + n^{-1/2} h(x)).$$

The approximation is meant in the sense of Hellinger differentiability. The joint law of $X_1, \ldots, X_n$ is $P_n = Q^n$. We have local asymptotic normality:

$$\begin{aligned} \log \frac{dP_{nh}}{dP_n} &= n^{-1/2} \sum_{i=1}^n h(X_i) - \frac{1}{2} Qh^2 + o_{P_n}(1), \\ n^{-1/2} \sum_{i=1}^n h(X_i) &\Rightarrow (Qh^2)^{1/2} N, \end{aligned}$$

with N standard normal. This is (2.1) and (2.2) with

$$S_n(h) = n^{-1/2} \sum_{i=1}^n h(X_i), \quad (h, h') = Q(hh').$$

Remark 3. (*Stochastic differentiability.*) In the stochastic differentiability condition (3.2), the parameter F is fixed, and we may omit it. Let $\{Q_\vartheta : \vartheta \in \Theta\}$ be a parametric family of distributions of X_i. For τ near ϑ let

$b_\tau \in L_{2,0}(Q_\tau)$. Stochastic differentiability (3.2) follows if b_τ is differentiable at $\tau = \vartheta$ in an appropriate sense. By Taylor expansion,

$$\sup_{|u|\le\Delta} \Big| n^{-1/2} \sum_{i=1}^n b_{\vartheta_{nu}}(X_i) - n^{-1/2} \sum_{i=1}^n b_\vartheta(X_i) - u Q_\vartheta b'_\vartheta \Big| = o_{P_{n\vartheta}}(1). \tag{5.1}$$

From $Q_\tau b_\tau = 0$ we obtain by taking the derivative under the integral,

$$Q_\vartheta b'_\vartheta = -Q_\vartheta(m_\vartheta b_\vartheta), \tag{5.2}$$

with $m_\vartheta = \partial_{\tau=\vartheta} dQ_\tau / dQ_\vartheta$ the score function for ϑ. Relations (5.1) and (5.2) imply stochastic differentiability (3.2). A proof for fixed bounded sequences $u = u_n$ is in Schick (1999a). For (5.2) it is essential that $Q_\tau b_\tau = 0$, in other words, that b_τ is in the nonparametric tangent space, i.e., that $n^{-1/2} \sum_{i=1}^n b_\tau(X_i)$ is a statistic which approximates a local likelihood.

For stochastic differentiability (3.2) to hold, the function b_ϑ need not be differentiable. For τ near ϑ consider the empirical process

$$\nu_{n\tau} = n^{-1/2} \sum_{i=1}^n \Big(b_\tau(X_i) - Q_\vartheta b_\tau \Big).$$

If the collection of functions b_τ, τ near ϑ, fulfills an appropriate bracketing condition, then $\nu_{n\tau}$ is *stochastically equicontinuous* at $\tau = \vartheta$: For each $\varepsilon, \eta > 0$ there is $\delta > 0$ such that

$$\limsup_n P_{n\vartheta}\Big(\sup_{|\tau-\vartheta|\le\delta} |\nu_{n\tau} - \nu_{n\vartheta}| > \eta \Big) \le \varepsilon.$$

Such a result was first proved by Daniels (1961) and Huber (1967) to obtain asymptotic normality of the maximum likelihood estimator under weak conditions on the score function; for a general version see Pollard (1985). For $\tau = \vartheta_{nu}$ we obtain

$$\begin{aligned} \sup_{|u|\le\Delta} \Big| & n^{-1/2} \sum_{i=1}^n b_{\vartheta_{nu}}(X_i) - n^{-1/2} \sum_{i=1}^n b_\vartheta(X_i) \\ & - n^{1/2}(Q_\vartheta b_{\vartheta_{nu}} - Q_\vartheta b_\vartheta) \Big| = o_{P_{n\vartheta}}(1). \end{aligned} \tag{5.3}$$

From $Q_\tau b_\tau = 0$ we obtain

$$\begin{aligned} Q_\vartheta b_\tau - Q_\vartheta b_\vartheta &= Q_\vartheta b_\tau - Q_\tau b_\tau \doteq -(\tau-\vartheta) Q_\tau(m_\tau b_\tau) \\ &\doteq -(\tau-\vartheta) Q_\vartheta(m_\vartheta b_\vartheta). \end{aligned} \tag{5.4}$$

Relations (5.3) and (5.4) imply stochastic differentiability (3.2).

Example 1. (*Nonlinear regression.*) Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be pairs of real observations with

$$Y_i = r(\vartheta, X_i) + \varepsilon_i.$$

The ε_i are i.i.d. with density $f(y)$ which has mean zero and finite variance but is unknown otherwise. The X_i are independent and independent of the ε_i and have unknown distribution function $G(y)$. The pair (X_i, Y_i) plays the role of X_i in the general setting. The model is semiparametric. The distribution of (X_i, Y_i) is

$$Q_{\vartheta f G}(dx, dy) = dG(x) f(y - r(\vartheta, x)) dy.$$

Fix ϑ, f and G. We introduce perturbations

$$\begin{aligned} \vartheta_{nu} &= \vartheta + n^{-1/2} u, \\ f_{nz}(y) &\doteq f(y)(1 + n^{-1/2} z(y)), \\ dG_{nw}(x) &\doteq dG(x)(1 + n^{-1/2} w(x)). \end{aligned}$$

The density f_{nz} must integrate to one and have mean zero, so z runs through

$$Z = \{z \in L_2(f) : \int z(y) f(y) dy = 0, \int y z(y) f(y) dy = 0\}.$$

The function G_{nw} must be a distribution function, so w runs through

$$L_{2,0}(G) = \{w \in L_2(G) : \int w(x) dG(x) = 0\}.$$

The perturbed distribution of (X_i, Y_i) is

$$\begin{aligned} &Q_{nuzw}(dx, dy) = Q_{\vartheta_{nu} f_{nz} G_{nw}}(dx, dy) = dG_{nw}(x) f_{nz}(y - r(\vartheta_{nu}, x)) dy \\ &\doteq Q_{\vartheta f G}(dx, dy) \Big(1 + n^{-1/2} \Big(u \dot{r}(\vartheta, x) \ell(y - r(\vartheta, x)) \\ &\qquad\qquad + z(y - r(\vartheta, x)) + w(x) \Big) \Big), \end{aligned}$$

where $\dot{r}(\vartheta, x)$ is the derivative of $r(\vartheta, x)$ with respect to ϑ, and $\ell(y) = -f'(y)/f(y)$ is the score function for location of the error distribution. Hence the tangent space $\overline{H}$ of the nonlinear regression model consists of functions

$$h(x, y) = u \dot{r}(\vartheta, x) \ell(y - r(\vartheta, x)) + z(y - r(\vartheta, x)) + w(x).$$

It is therefore of the form $\overline{H} = [m] + DV$ of Section 2, with $V = Z \times L_{2,0}(G)$,

$$m(x, y) = \dot{r}(\vartheta, x) \ell(y - r(\vartheta, x)), \quad Dv(x, y) = z(y - r(\vartheta, x)) + w(x).$$

Note that by taking the derivative under the integral,

$$E(\varepsilon\ell(\varepsilon)) = \int x\ell(x)f(x)dx = -\int xf'(x)dx = 1. \tag{5.5}$$

Note also that $w(x)$ is orthogonal to both $m(x,y)$ and $z(y-r(\vartheta,x))$, so that both ϑ and f are adaptive with respect to G.

We want to estimate the expectation

$$a(f) = Ek(\varepsilon) = \int k(y)f(y)dy$$

of an f-square-integrable function k under the error distribution. The usual estimator is the empirical estimator based on the *estimated* errors,

$$\hat{a} = \frac{1}{n}\sum_{i=1}^{n} k(\hat{\varepsilon}_i),$$

with $\hat{\varepsilon}_i = Y_i - r(\hat{\vartheta}, X_i)$. A natural estimator of ϑ is the least squares estimator $\hat{\vartheta}$, which solves

$$\sum_{i=1}^{n} \dot{r}(\vartheta, X_i)(Y_i - r(\vartheta, X_i)) = 0.$$

The least squares estimator is asymptotically linear with influence function

$$c(x,y) = (E\dot{r}(\vartheta,X)^2)^{-1}\dot{r}(\vartheta,x)(y - r(\vartheta,x)).$$

We have $(m,c) = E\dot{r}(\vartheta,X)^2$ by (5.5), and $c \perp DV$ since $\int yz(y)f(y)dy = 0$ for $z \in Z$. Hence c is a gradient of ϑ. The empirical estimator of $a(f) = Ek(\varepsilon)$ based on *true* innovations is

$$\hat{a}_\vartheta = \frac{1}{n}\sum_{i=1}^{n} k(Y_i - r(\vartheta, X_i)).$$

Its influence function is

$$g_\vartheta(x,y) = k(y - r(\vartheta,x)) - Ek(\varepsilon).$$

For $v = (z,w) \in Z \times L_{2,0}(G)$ we have

$$n^{1/2}(a(f_{nz}) - a(f)) \to E(k(\varepsilon)z(\varepsilon)) = (g_\vartheta, Dv).$$

Hence g_ϑ is a gradient of $a(f)$ when ϑ is known. It fulfills

$$(m, g_\vartheta) = E\dot{r}(\vartheta,X)E(\ell(\varepsilon)k(\varepsilon)).$$

Hence by Remark 3, an appropriate bracketing condition on the collection of functions $b_\tau(x,y) = k(y-\tau x) - Ek(\varepsilon)$ implies stochastic differentiability (3.2) of the form

$$\sup_{|u|\le\Delta} \Big| n^{-1/2} \sum_{i=1}^n g_{\vartheta_{nu}}(X_i,Y_i) - n^{-1/2} \sum_{i=1}^n g_\vartheta(X_i,Y_i) + \; uE\dot r(\vartheta,X)E(\ell(\varepsilon)k(\varepsilon)) \Big| = o_{P_{n\vartheta fG}}(1).$$

It follows from (3.3) that the plug-in estimator $\hat a_{\hat\vartheta}$ is asymptotically linear for $a(f)$ with influence function

$$\begin{aligned} g &= g_\vartheta - (m,g_\vartheta)c \\ &= k(\varepsilon) - Ek(\varepsilon) - E\dot r(\vartheta,X)E(\ell(\varepsilon)k(\varepsilon))(E\dot r(\vartheta,X)^2)^{-1}\dot r(\vartheta,x)\varepsilon. \end{aligned}$$

Efficient estimators for ϑ are constructed in Schick (1993). The canonical gradient $\overline{g}$ and an efficient estimator for $Ek(\varepsilon)$ are in Müller and Wefelmeyer (2000a).

6 Markov chain models

Let $X_0,\dots,X_n$ be observations from a homogeneous and uniformly ergodic Markov chain with transition distribution $Q(x,dy)$ and invariant law $\pi(dx)$. Assume for simplicity that the chain is stationary. The natural "nonparametric" model of Section 2 is described by the collection of *all* such transition distributions. Fix Q. Set $Q_x h = \int Q(x,dy)h(x,y)$ and

$$H = \{h \in L_2(\pi\otimes Q) : Q_x h = 0\}.$$

For $h \in H$ set

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

The approximation is meant in the sense of Hellinger differentiability for Markov chains. The joint law of $X_0,\dots,X_n$ is

$$P_n(dx_0,\dots,dx_n) = \pi(dx_0)Q(x_0,dx_1)\cdots Q(x_{n-1},dx_n).$$

We have local asymptotic normality:

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

$$n^{-1/2}\sum_{i=1}^n h(X_{i-1},X_i) \;\Rightarrow\; (\pi\otimes Qh^2)^{1/2}N,$$

with N standard normal. This is (2.1) and (2.2) with

$$S_n(h) = n^{-1/2} \sum_{i=1}^{n} h(X_{i-1}, X_i), \quad (h, h') = \pi \otimes Q(hh').$$

Remark 4. (*Stochastic differentiability.*) The arguments of Remark 3 translate to stochastic process models. Stochastic equicontinuity for Markov chains was obtained by Ogata (1980) in connection with asymptotic normality of maximum likelihood estimators. Results for general discrete-time stochastic processes are in Andrews (1994) and Andrews and Pollard (1994). Let $\{Q_\vartheta : \vartheta \in \Theta\}$ be a parametric family of transition distributions of X_i. For τ near ϑ let b_τ be $\pi_\tau \otimes Q_\tau$-square-integrable with $\int Q_\tau(x, dy) b_\tau(y) = 0$ for all x. The score function for ϑ is

$$m_\vartheta(x, y) = \partial_{\tau=\vartheta} \frac{dQ_\tau(x, \cdot)}{dQ_\vartheta(x, \cdot)}(y).$$

If the functions b_τ fulfill an appropriate bracketing condition for τ near ϑ, we have stochastic differentiability (3.2) of the form

$$\sup_{|u| \le \Delta} \Big| n^{-1/2} \sum_{i=1}^{n} b_{\vartheta_{nu}}(X_{i-1}, X_i) - n^{-1/2} \sum_{i=1}^{n} b_\vartheta(X_{i-1}, X_i) + u(m_\vartheta, b_\vartheta) \Big| = o_{P_{n\vartheta}}(1). \tag{6.1}$$

Example 2. (*Nonlinear autoregression.*) The observations $X_0, \ldots, X_n$ are real with

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

The ε_i are i.i.d. with density $f(x)$ which has mean zero and finite variance but is unknown otherwise. Conditions for uniform ergodicity are in Bhattacharya and Lee (1995) and An and Huang (1996). The model is semiparametric, with transition distribution

$$Q_{\vartheta f}(x, dy) = f(y - r(\vartheta, x)) dy.$$

Fix ϑ and f. We introduce perturbations

$$\begin{aligned} \vartheta_{nu} &= \vartheta + n^{-1/2} u, \\ f_{nv}(x) &\doteq f(x)(1 + n^{-1/2} v(x)). \end{aligned}$$

As in the regression example, Example 1, the function v runs through

$$V = \{v \in L_2(f) : \int v(x)f(x)dx = 0, \int xv(x)f(x)dx = 0\}.$$

The perturbed transition distribution is

$$\begin{aligned} &Q_{nuv}(x, dy) = Q_{\vartheta_{nu} f_{nv}}(x, dy) = f_{nv}(y - r(\vartheta_{nu}, x))dy \\ &\doteq Q_{\vartheta f}(x, dy)\Big(1 + n^{-1/2}\Big(u\dot{r}(\vartheta, x)\ell(y - r(\vartheta, x)) + v(y - r(\vartheta, x))\Big)\Big), \end{aligned}$$

where $\dot{r}(\vartheta, x)$ is the derivative of $r(\vartheta, x)$ with respect to ϑ, and $\ell(x) = -f'(x)/f(x)$ is the score function for location of the innovation distribution. Hence the tangent space $\overline{H}$ of the nonlinear autoregressive model consists of functions

$$h(x, y) = u\dot{r}(\vartheta, x)\ell(y - r(\vartheta, x)) + v(y - r(\vartheta, x)).$$

It is therefore of the form $\overline{H} = [m] + DV$ of Section 2, with

$$m(x, y) = \dot{r}(\vartheta, x)\ell(y - r(\vartheta, x)), \quad Dv(x, y) = v(y - r(\vartheta, x)).$$

We want to estimate the expectation

$$a(f) = Ek(\varepsilon) = \int k(x)f(x)dx$$

of an f-square-integrable function k under the innovation distribution. The usual estimator is the empirical estimator based on the *estimated* innovations,

$$\hat{a} = \frac{1}{n}\sum_{i=1}^{n} k(\hat{\varepsilon}_i),$$

with $\hat{\varepsilon}_i = X_i - r(\hat{\vartheta}, X_{i-1})$. A natural estimator of ϑ is the least squares estimator $\hat{\vartheta}$, which solves

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

The least squares estimator is asymptotically linear with influence function

$$c(x, y) = (E\dot{r}(\vartheta, X)^2)^{-1}\dot{r}(\vartheta, x)(y - r(\vartheta, x)).$$

We have $(m, c) = E\dot{r}(\vartheta, X)^2$ by (5.5), and $c \perp DV$ since $\int xv(x)f(x)dx = 0$ for $v \in V$. Hence c is a gradient of ϑ. The empirical estimator of $a(f) = Ek(\varepsilon)$ based on *true* innovations is

$$\hat{a}_\vartheta = \frac{1}{n}\sum_{i=1}^{n} k(X_i - r(\vartheta, X_{i-1})).$$

Its influence function is

$$g_\vartheta(x,y) = k(y-\vartheta x) - Ek(\varepsilon).$$

We have

$$n^{1/2}(a(f_{nv}) - a(f)) \to E(k(\varepsilon)v(\varepsilon)) = (g_\vartheta, Dv) \quad \text{for } v \in V.$$

Hence g_ϑ is a gradient of $a(f)$ when ϑ is known. It fulfills

$$(m, g_\vartheta) = E\dot{r}(\vartheta, X)E(\ell(\varepsilon)k(\varepsilon)).$$

By Remark 4, an appropriate bracketing condition on the functions

$$b_\tau(x,y) = k(y-\tau x) - Ek(\varepsilon)$$

implies stochastic differentiability (6.1). It follows from (3.3) that the plug-in estimator $\hat{a}_{\hat{\vartheta}}$ is asymptotically linear for $a(f)$ with influence function

$$\begin{aligned} g &= g_\vartheta - (m, g_\vartheta)c \\ &= k(\varepsilon) - Ek(\varepsilon) - E\dot{r}(\vartheta, X)E(\ell(\varepsilon)k(\varepsilon))(E\dot{r}(\vartheta, X)^2)^{-1}\dot{r}(\vartheta, x)\varepsilon. \end{aligned}$$

Efficient estimators for ϑ are constructed in Drost, Klaassen and Werker (1997) and Koul and Schick (1997). The canonical gradient $\overline{g}$ and an efficient estimator for $Ek(\varepsilon)$ is in Schick and Wefelmeyer (2000a).

Example 3. (*Heteroscedastic linear autoregression.*) The observations $X_0, \ldots, X_n$ are real with

$$X_i = \vartheta X_{i-1} + s(X_{i-1})\varepsilon_i.$$

The ε_i are independent and, for simplicity, standard normal. Conditions for uniform ergodicity and efficient estimators for ϑ are in Maercker (1997) and Schick (1999b). The model is semiparametric, with transition distribution

$$Q_{\vartheta s}(x, dy) = \frac{1}{s(x)}\varphi\Big(\frac{y-\vartheta x}{s(x)}\Big)dy,$$

where φ is the standard normal density. Fix ϑ and s. Introduce perturbations

$$\vartheta_{nu} = \vartheta + n^{-1/2}u, \qquad s_{nv}(x) = s(x)(1 + n^{-1/2}v(x)).$$

The function v runs through $V = L_2(f)$, where f is the stationary density. The perturbed transition distribution is

$$\begin{aligned} Q_{nuv}(x, dy) &= Q_{\vartheta_{nu}s_{nv}}(x, dy) = \frac{1}{s_{nv}(x)}\varphi\Big(\frac{y-\vartheta_{nu}x}{s_{nv}(x)}\Big)dy \\ &\doteq Q_{\vartheta s}(x, dy)\Big(1 + n^{-1/2}\Big(um(x,y) + Dv(x,y)\Big)\Big), \end{aligned}$$

with

$$m(x,y) = \frac{x}{s(x)}\frac{y-\vartheta x}{s(x)}, \qquad Dv(x,y) = v(x)\Big(\Big(\frac{y-\vartheta x}{s(x)}\Big)^2 - 1\Big).$$

Since the normal distribution is symmetric, m and DV are orthogonal, and s can be estimated adaptively with respect to ϑ.

Suppose we want to estimate the functional

$$a(s) = \int_0^1 s(x)^2 dx.$$

For all $u \in \mathbf{R}$ and $v \in V$ we have

$$n^{1/2}(a(s_{nv}) - a(s)) \to 2\int_0^1 s(x)^2 v(x) dx = (Dv_a, Dv + um)$$

with $v_a = 1_{[0,1]} s^2/f$. Hence $a(s)$ is differentiable at (ϑ, s), with canonical gradient

$$g(x,y) = g_\vartheta(x,y) = Dv_a(x,y) = 1_{[0,1]}(x)\frac{s(x)^2}{f(x)}\Big(\Big(\frac{y-\vartheta x}{s(x)}\Big)^2 - 1\Big).$$

Assume first that ϑ is known. Then we can estimate $a(s)$ by

$$\hat{a}_\vartheta = \int_0^1 \hat{s}_\vartheta(x)^2 dx,$$

where

$$\hat{s}_\vartheta(x)^2 = \frac{\sum_{i=1}^n (X_i - \vartheta X_{i-1})^2 w_n(X_{i-1} - x)}{\sum_{i=1}^n w_n(X_{i-1} - x)}.$$

Here $w_n(x) = c_n^{-1} w(c_n^{-1} x)$, where w is a continuously differentiable symmetric density with compact support $[-1,1]$, and c_n is a bandwidth of order $n^{-1/3}$. We show that $\hat{a}_\vartheta$ is asymptotically linear with influence function Dv_a. We do so under the assumption that s is twice continuously differentiable. Write

$$(X_i - \vartheta X_{i-1})^2 = s(X_{i-1})^2(\varepsilon_i^2 - 1) + s(X_{i-1})^2.$$

Expand $s(X_{i-1})$ around $s(x)$ to obtain

$$\hat{a}_\vartheta - a(s) = \int_0^1 \frac{\hat{A}(x) + 2s(x)s'(x)\hat{f}_1(x)}{\hat{f}_0(x)} dx + O_{P_{n\vartheta s}}(c_n^2),$$

where

$$\begin{aligned}\hat{A}(x) &= \frac{1}{n}\sum_{i=1}^{n} s(X_{i-1})^2(\varepsilon_i^2-1)w_n(X_{i-1}-x),\\ \hat{f}_j(x) &= \frac{1}{n}\sum_{i=1}^{n}(X_{i-1}-x)^j w_n(X_{i-1}-x).\end{aligned}$$

The assumptions imply that f is twice continuously differentiable. Hence we obtain uniformly for $x \in [0,1]$,

$$\begin{aligned}&E\hat{A}(x)^2 = O(n^{-1}c_n^{-1}) = O(n^{-2/3}),\\ &E(\hat{f}_0(x)-f(x))^2 = O(n^{-1}c_n^{-1}+c_n^4) = O(n^{-2/3}),\\ &E(\hat{f}_1(x)-c_n f'(x))^2 = O(n^{-1}c_n^{-1}+c_n^4) = O(n^{-2/3}).\end{aligned}$$

We can also show that $\sup_{0\le x\le 1}|\hat{f}_0(x)-f(x)|$ converges to zero in probability. From this and the fact that f is bounded away from zero on $[0,1]$, we can conclude that

$$\hat{a}_\vartheta - a(s) = \int_0^1 \frac{\hat{A}(x)}{f(x)}dx + O_{P_{n\vartheta s}}(c_n^2).$$

Now write

$$\int_0^1 \frac{\hat{A}(x)}{f(x)}dx = \frac{1}{n}\sum_{i=1}^{n}\frac{s(X_{i-1})^2}{f(X_{i-1})}(\varepsilon_i^2-1)I_n(X_{i-1})$$

with

$$I_n(y) = \int_0^1 \frac{f(y)}{f(x)}w_n(y-x)dx.$$

It is easy to check that I_n converges in $L_2(f)$ to the indicator of $[0,1]$. Combining the above lets us conclude that $\hat{a}_\vartheta$ has influence function Dv_a.

Suppose now that ϑ is unknown. Let $\hat{\vartheta}$ be a $n^{1/2}$-consistent estimator of ϑ. We prove that the plug-in estimator $\hat{a}_{\hat{\vartheta}}$ is efficient. We have already shown above that $\hat{a}_\vartheta$ fulfills (3.4) with $b_\vartheta = Dv_a$. By the argument of Section 3, it remains to show (3.5). Since $(m, Dv_a) = 0$ by adaptivity, (3.5) reduces to asymptotic equivalence of $\hat{a}_{\hat{\vartheta}}$ and $\hat{a}_\vartheta$, i.e., $n^{1/2}(\hat{a}_{\hat{\vartheta}} - \hat{a}_\vartheta) = o_{P_{n\vartheta s}}(1)$. To prove this, we note first that

$$n^{1/2}(\hat{a}_{\hat{\vartheta}} - \hat{a}_\vartheta) = 2n^{1/2}(\hat{\vartheta}-\vartheta)\int_0^1 \frac{\hat{B}(x)}{\hat{f}_0(x)}dx + O_{P_{n\vartheta s}}(n^{-1/2}),$$

where

$$\hat{B}(x) = \frac{1}{n}\sum_{i=1}^{n} s(X_{i-1})X_{i-1}\varepsilon_i w_n(X_{i-1}-x).$$

Since $\int_0^1 \hat{B}(x)/\hat{f}_0(x)dx$ converges to zero in probability, we obtain the desired result.

7 Extensions

1. We have assumed ϑ and $a(F)$ to be one-dimensional. Extension to finite-dimensional $a(F)$ is straightforward; infinite-dimensional $a(F)$ require additional technicalities. In nonlinear regression, Example 1, we may, e.g., be interested in estimating the error distribution function F, defined by $F(t) = P(\varepsilon \le t)$. For *linear* regression we refer to Klaassen and Putter (1999). Extension to finite-dimensional ϑ is also straightforward. We note that it may happen that $a(F)$ is adaptive with respect to certain components of ϑ only. For efficiency of $\hat{a}_{\hat{\vartheta}}$, efficient estimators are required only for the non-adaptive components of ϑ. Extensions of nonlinear regression, Example 1, are treated in Müller and Wefelmeyer (2000a). Extensions of nonlinear autoregression, Example 2, are treated in Schick and Wefelmeyer (2000a).

2. We have restricted attention to functionals $a(F)$ of F only. The results may be extended to functionals $a(\vartheta, F)$ which depend also on ϑ. An interesting application is estimation of *invariant* distributions of time series, for example in linear autoregression $X_i = \vartheta X_{i-1} + \varepsilon_i$. Since $\sum_{j=1}^{\infty} \vartheta^j \varepsilon_j$ is distributed as the invariant law, we can write the expectation of a function k under the invariant law as

$$Ek(X) = Ek\Big(\sum_{j=1}^{\infty} \vartheta^j \varepsilon_j\Big) = a(\vartheta, F),$$

where F is the invariant distribution function. Hence $Ek(X)$ can be estimated by a von Mises statistic or a U-statistic based on estimated innovations; see Schick and Wefelmeyer (2000b).

3. The results extend from semiparametric models $\{P_{n\vartheta F} : \vartheta \in \Theta, F \in \mathcal{F}\}$ to parametric families $\{\mathcal{P}_{n\vartheta} : \vartheta \in \Theta\}$ of nonparametric models. This is of interest when we start from a nonparametric model $\mathcal{P}_n$ and impose a restriction which depends on an unknown parameter, say $r_\vartheta(P_n) = 0$, leading to

$$\mathcal{P}_{n\vartheta} = \{P_n : r_\vartheta(P_n) = 0\}.$$

For example, let $X_0, \dots, X_n$ be observations from a Markov chain with transition distribution fulfilling $\int Q(x, dy)y = r(\vartheta, x)$ for some ϑ. This is the nonlinear autoregressive model $X_i = r(\vartheta, X_{i-1}) + \varepsilon_i$, where the ε_i are martingale increments, not i.i.d. as in Example 2. For estimators of ϑ see Wefelmeyer (1994), (1996), (1997a), (1997b); for estimators of the stationary law see Schick and Wefelmeyer (1999). The model may be written as a semiparametric model by introducing transition distributions $F(x, dy)$ with $\int F(x, dy)y = 0$ and writing

$$Q(x, dy) = F(x, dy - r(\vartheta, x)).$$

This is, however, technically inconvenient because we perturb ϑ and would need differentiability of F.

Another example are i.i.d. observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ with joint law fulfilling the constraint $E(a(X, Y, \vartheta)|X) = 0$, where $a(X, Y, \vartheta)$ is a given function. For plug-in estimators in such models see Müller and Wefelmeyer (2000b). A special case is $a(X, Y, \vartheta) = Y - r(\vartheta, X)$, i.e., $Y_i = r(\vartheta, X_i) + \varepsilon_i$, which differs from Example 1 in that we do not assume ε_i and X_i to be independent.

References

Andrews, D. W. K. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica* **62**, 295–314.

Andrews, D. W. K. and Pollard, D. (1994). An introduction to functional central limit theorems for dependent stochastic processes. *Internat. Statist. Rev.* **62**, 119–132.

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

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer Series in Statistics, Springer, Berlin.

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

Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10**, 647–671.

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

Daniels, H. E. (1961). The asymptotic efficiency of a maximum likelihood estimator. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1**, 151–163.

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

Greenwood, P. E. and Wefelmeyer, W. (1991). Efficient estimating equations for nonparametric filtered models. In: *Statistical Inference in Stochastic Processes* (N. U. Prabhu, I. V. Basawa, eds.), 107-141, Marcel Dekker, New York.

Fabian, V. and Hannan, J. (1985). *Introduction to Probability and Mathematical Statistics*. Wiley, New York.

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

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1**, 221–233.

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

Klaassen, C. A. J. and Putter, H. (1999). Efficient estimation of Banach parameters in semiparametric models. Technical Report, Department of Mathematics, University of Amsterdam.

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

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

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.

Le Cam, L. and Yang, G. L. (1990). *Asymptotics in Statistics*. Springer Series in Statistics, Springer, Berlin.

Maercker, G. (1997). *Statistical Inference in Conditional Heteroskedastic Autoregressive Models*. Shaker, Aachen.

Müller, U. U. and Wefelmeyer, W. (2000a). Estimating parameters of the residual distribution in nonlinear regression. In preparation.

Müller, U. U. and Wefelmeyer, W. (2000b). Regression type models and optimal estimators. In preparation.

Ogata, Y. (1980). Maximum likelihood estimates of incorrect Markov models for time series and the derivation of AIC. *J. Appl. Probab.* **17**, 59–72.

Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory* **1**, 295–314.

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

Schick, A. (1998). Sample splitting with Markov chains. To appear in: *Bernoulli*.

http://math.binghamton.edu/anton/preprint.html

Schick, A. (1999a). On asymptotic differentiability of averages. To appear in: *Statist. Probab. Lett.*
http://math.binghamton.edu/anton/preprint.html

Schick, A. (1999b). Efficient estimation in a semiparametric heteroscedastic autoregressive model. *Stat. Inference Stoch. Process.* **2**, 69–98.

Schick, A. and Wefelmeyer, W. (1999). Efficient estimation of invariant distributions of some semiparametric Markov chain models. *Math. Meth. Statist.* **8**, 426–440.

Schick, A. and Wefelmeyer, W. (2000a). Estimating the innovation distribution in nonlinear autoregressive models. Technical Report, Department of Mathematical Sciences, Binghamton University.
http://math.binghamton.edu/anton/preprint.html

Schick, A. and Wefelmeyer, W. (2000b). Estimating invariant laws of linear processes by U-statistics. Technical Report, Department of Mathematics, University of Siegen.
http://www.math.uni-siegen.de/statistik/wefelmeyer.html

Wefelmeyer, W. (1991). A generalization of asymptotically linear estimators. *Statist. Probab. Lett.* **11**, 195–199.

Wefelmeyer, W. (1994). Improving maximum quasi-likelihood estimators. In: *Asymptotic Statistics* (P. Mandl, M. Hušková, eds.), 467-474, Physika-Verlag, Heidelberg.

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

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

Wefelmeyer, W. (1997b). Quasi-likelihood regression models for Markov chains. In: *Selected Proceedings of the Symposium on Estimating Functions* (I. V. Basawa, V. P. Godambe and R. L. Taylor, eds.), 149-173, IMS Lecture Notes–Monograph Series, Institute of Mathematical Statistics, Hayward, California.

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