# Root $n$ consistent and optimal density estimators for moving average processes


Anton Schick
Binghamton University

Wolfgang Wefelmeyer
Universität Siegen



**Abstract**

The marginal density of a first order moving average process can be written as convolution of two innovation densities. Saavedra & Cao (2000) propose to estimate the marginal density by plugging in kernel density estimators for the innovation densities, based on estimated innovations. They obtain that for an appropriate choice of bandwidth the variance of their estimator decreases at the rate $1/n$.

Their estimator can be interpreted as a specific U-statistic. We suggest a slightly simplified U-statistic as estimator of the marginal density, prove that it is asymptotically normal at the same rate, and describe the asymptotic variance explicitly. We show that the estimator is asymptotically efficient if no structural assumptions are made on the innovation density. For innovation densities known to have mean zero or to be symmetric, we describe improvements of our estimator which are again asymptotically efficient.



*Key words:* Efficient estimator, least dispersed estimator, plug-in estimator, semiparametric model, time series.

*Running head:* Root $n$ consistent density estimators.


## 1 Introduction

Suppose $X_1, \dots, X_n$ are observations from an MA(1) process $X_t = \varepsilon_t - \vartheta\varepsilon_{t-1}$, where $\varepsilon_t$ are i.i.d. innovations with density $f$ and finite second moment, and $|\vartheta| < 1$. We want to estimate the density of $X_t$, say $g$, at a fixed point $x$. A widely used estimator for $g(x)$ is the kernel estimator

$$\overline{g}(x) = \frac{1}{n}\sum_{j=1}^{n} K_b(x - X_j),$$

where $K_b(u) = K(u/b)/b$ with kernel function $K$ and bandwidth $b > 0$. Note that

$$g(x) = \int f(x + \vartheta y)f(y)\,dy.$$

Motivated by this representation, Saavedra & Cao (1999a, 2000) propose the plug-in estimator

$$\hat{g}_{SC}(x) = \int \hat{f}(x + \hat{\vartheta}y)\hat{f}(y)\,dy,$$

where $\hat{\vartheta}$ is $n^{1/2}$-consistent and $\hat{f}$ is a kernel estimator based on estimated innovations $\hat{\varepsilon}_j$. They obtain that their estimator is $n^{1/2}$-consistent for the bandwidth $b = n^{-2/5}$. For this they assume that the innovations have mean zero. They also use rather strong regularity conditions. Saavedra & Cao (1999b) compare the mean integrated squared error of their estimator and the usual kernel estimator. We note that for continuous-time processes, corresponding consistency rates for kernel-type estimators are more common. Castellana & Leadbetter (1986) give conditions for such results to hold in general stationary processes; see also Bosq (1993, 1995), Blanke & Bosq (1997), Bosq et al. (1999) and Veretennikov (1999). Kutoyants (1997, 1998, 1999) obtains efficiency of estimators for the density of diffusion processes.

The parametric convergence rate $n^{-1/2}$ for a density estimator may seem striking but becomes plausible once one notes that $g(x)$ is represented as an integral functional of the innovation density $f$. For models with i.i.d. observations it is well known that certain integral functionals of densities and their derivatives are estimated at the rate $n^{-1/2}$ if appropriate kernel estimators for the density are plugged into the functional. Similar results hold for integral functionals of regression functions and quantile regression functions. Some recent references are Birgé & Massart (1995), Tsybakov & van der Meulen (1996), Chaudhuri et al. (1997) and Efromovich & Samarov (2000). Almost sure representations are obtained by Eggermont & LaRiccia (1999,2001) and Mason (2003). Note that to estimate $g(x)$, we must also plug in an estimator for $\vartheta$.

The estimator $\hat{g}_{SC}(x)$ can be written

$$\hat{g}_{SC}(x) = \frac{1}{n^2 b} \sum_{i,j=1}^{n} L_{\hat{\vartheta}}\Big(\frac{x - \hat{\varepsilon}_i + \hat{\vartheta}\hat{\varepsilon}_j}{b}\Big)$$

with $L_\vartheta(u) = \int K(u + \vartheta v)K(v)\,dv$. A closely related estimator is the U-statistic

$$\hat{g}(x) = \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} K_b(x - \hat{\varepsilon}_i + \hat{\vartheta}\hat{\varepsilon}_j).$$

This is essentially $\hat{g}_{SC}(x)$ with the random kernel $L_{\hat{\vartheta}}$ replaced by $K$. In this paper we prove that $n^{1/2}(\hat{g}(x) - g(x))$ is asymptotically normal, and calculate the asymptotic variance. Our result is valid for bandwidths $b = n^{-a}$ with $a$ strictly between 1/6 and 1/8. We prove it under much weaker regularity conditions. We also do not assume that the innovations have mean zero.

Of course, these results are valid only if $\vartheta \neq 0$ because for $\vartheta = 0$ the model reduces to i.i.d. observations $X_j = \varepsilon_j$ for which it is well-known that no $n^{1/2}$-consistent estimators exist.

A slightly more difficult argument would show that $\hat{g}_{SC}(x)$, with random kernel $L_{\hat{\vartheta}}$, is asymptotically equivalent to our estimator (under weaker regularity conditions, for our choices of bandwidth, and without the assumption of mean zero innovations).

Since kernel estimators are not $n^{1/2}$-consistent, there is no attainable variance bound for them. Here we have $n^{1/2}$-consistency, and hence a theoretically attainable variance bound. We prove that our estimator attains the variance bound if we choose an efficient estimator for $\vartheta$. More precisely, we then have semiparametric efficiency as described in Bickel et al. (1998) for i.i.d. observations.

If we have additional restrictions on the innovation distributions, such as mean zero or symmetry, we can improve our estimator. For symmetric innovations, we symmetrize the estimator as

$$\frac{1}{4n(n-1)} \sum_{i \neq j} \big( K_b(x - \hat{\varepsilon}_i + \hat{\vartheta}\hat{\varepsilon}_j) + K_b(x + \hat{\varepsilon}_i + \hat{\vartheta}\hat{\varepsilon}_j) + K_b(x - \hat{\varepsilon}_i - \hat{\vartheta}\hat{\varepsilon}_j) + K_b(x + \hat{\varepsilon}_i - \hat{\vartheta}\hat{\varepsilon}_j) \big).$$

For mean zero innovations as in Saavedra & Cao (1999a, 2000), improvements are possible by subtracting from the estimator for $g(x)$ a term of the form $\hat{a}\frac{1}{n}\sum_{j=1}^{n} \hat{\varepsilon}_j$ with properly chosen random coefficient $\hat{a}$.

Estimators of other functionals of MA processes have been considered before. Efficient estimators of $\vartheta$ under various assumptions are in Kreiss (1987), Jeganathan (1995), Drost et al. (1997) and Schick & Wefelmeyer (2000b). Kreiss (1990) also estimates the coefficients of infinite-order MA processes. Efficient estimators of linear functionals of the stationary law are in Schick & Wefelmeyer (2002c). The innovation distribution function of infinite-order MA processes is estimated in Kreiss (1991). Schick & Wefelmeyer (2002d) prove a *functional* central limit theorem and efficiency of a version of $\hat{g}_{SC}$ for higher order moving average processes.

In Section 2 we state our results. In Section 3 we present the results of a small simulation study. Appendix 1 contains auxiliary results on U-statistics and related statistics, Appendices 2 and 3 contain proofs of Theorems 1 and 2.

## 2 Results

We consider an MA(1) process $X_t = \varepsilon_t - \vartheta\varepsilon_{t-1}$, where $0 < |\vartheta| < 1$ and $\varepsilon_t$ are i.i.d. innovations with distribution function $F$, density $f$, mean $\mu$ and finite variance $\sigma^2$. The innovations $\varepsilon_t$ have representation

$$\varepsilon_t = \sum_{s=0}^{\infty} \vartheta^s X_{t-s}; \tag{1}$$

see e.g. Brockwell & Davis (2002, Section 3.1). For later use, we also introduce the derivative of $\varepsilon_t$ with respect to $\vartheta$,

$$\dot{\varepsilon}_t = \sum_{s=1}^{\infty} s\vartheta^{s-1} X_{t-s} = \sum_{s=1}^{\infty} \vartheta^{s-1} \varepsilon_{t-s}. \tag{2}$$

Its expectation is

$$E[\dot\varepsilon] = \sum_{s=1}^{\infty} s\vartheta^{s-1}(1-\vartheta)\mu = \frac{\mu}{1-\vartheta},$$

and its variance is

$$\operatorname{Var}[\dot\varepsilon] = \frac{\sigma^2}{(1-\vartheta)(1+\vartheta)}.$$

Let $g$ denote the density of $X_t$. We want to estimate $g$ at a fixed point $x$, based on observations $X_1,\dots,X_n$. Let $\hat\vartheta$ be a $n^{1/2}$-consistent estimator of $\vartheta$. Representation (1) motivates estimating $\varepsilon_j$ by

$$\hat\varepsilon_j = \sum_{s=0}^{r} \hat\vartheta^s X_{j-s}, \quad j = r+1,\dots,n,$$

for some integer $r$. We modify the estimator given in the Introduction correspondingly and estimate $g(x)$ by the U-statistic

$$\hat g(x) = \frac{1}{(n-r)(n-r-1)} \sum_{\substack{i,j=r+1\\ i\neq j}}^{n} K_b(x - \hat\varepsilon_i + \hat\vartheta\hat\varepsilon_j),$$

where $K_b(u) = K(u/b)/b$ for some kernel function $K$ and some bandwidth $b > 0$.

**Condition 1**

The kernel $K$ has compact support, is twice continuously differentiable, and satisfies

$$\int K(u)\,du = 1 \quad \text{and} \quad \int u^i K(u)\,du = 0 \quad \text{for } i = 1,2,3.$$

**Condition 2**

The innovation density $f$ has an absolutely continuous derivative, and the almost everywhere derivative $f''$ of $f'$ is integrable and bounded.

The density of $\vartheta\varepsilon$ is $f_\vartheta(u) = f(u/\vartheta)/|\vartheta|$. We have the representations

$$g(x) = \int f(x+\vartheta y) f(y)\,dy = \int f_\vartheta(y-x) f(y)\,dy.$$

The derivatives of $g$ with respect to $\vartheta$ and $x$ are

$$\begin{aligned} \dot g(x) &= \int y f'(x+\vartheta y) f(y)\,dy, \\ g'(x) &= \int f'(x+\vartheta y) f(y)\,dy. \end{aligned}$$

We set

$$\psi(y) = f(x+\vartheta y) + f_\vartheta(y-x) - 2g(x), \quad y \in \mathbb{R}.$$

**Theorem 1**
*Assume Conditions 1 and 2. Let $\hat{\vartheta}$ be a $n^{1/2}$-consistent estimator of $\vartheta$. Let the bandwidth $b$ fulfill $nb^6 \to \infty$ and $nb^8 \to 0$. Let $r/\log n \to \infty$ and $r/n \to 0$. Then*

$$n^{1/2}(\hat{g}(x) - g(x)) = n^{-1/2} \sum_{j=1}^{n} \psi(\varepsilon_j) + \big(\dot{g}(x) - \mu g'(x)\big) n^{1/2}(\hat{\vartheta} - \vartheta) + o_p(1).$$

*If $f$ also has a finite third moment and $K$ has a bounded third derivative, then we can relax $nb^6 \to \infty$ to $nb^4 \to \infty$.*

Typical estimators $\hat{\vartheta}$ of $\vartheta$ are such that $n^{1/2}(\hat{\vartheta} - \vartheta)$ is asymptotically normal with mean zero and variance $\zeta^2$, say, and asymptotically independent of $n^{-1/2} \sum_{j=1}^{n} \psi(\varepsilon_j)$. In this case, it follows from Theorem 1 that $n^{1/2}(\hat{g}(x) - g(x))$ is asymptotically normal with mean zero and variance

$$\tau^2 = \int \psi^2 dF + \zeta^2 \big(\dot{g}(x) - \mu g'(x)\big)^2.$$

We assume two derivatives for $f$. The corresponding kernel estimator has optimal bandwidth proportional to $n^{-1/5}$. For our plug-in estimator, a larger bandwidth, with rate between $n^{-1/6}$ and $n^{-1/8}$, say $n^{-1/7}$, is needed, unless we require the additional assumptions at the end of Theorem 1. Because we have two derivatives for $f$, we have four derivatives for $g$. The optimal bandwidth for a kernel estimator of $g$ would therefore be proportional to $n^{-1/9}$.

We show now that $\hat{g}(x)$ is efficient for $g(x)$ if $\hat{\vartheta}$ is efficient for $\vartheta$. This is an instance of a parametric "plug-in principle": If $\hat{\kappa}(\vartheta)$ is an efficient estimator for some functional $\kappa$ of $(\vartheta, f)$ when $\vartheta$ is known, and $\hat{\vartheta}$ is efficient for $\vartheta$, then the plug-in estimator $\hat{\kappa}(\hat{\vartheta})$ will be efficient when $\vartheta$ is not known. We refer to Klaassen & Putter (2002) for sufficient conditions in the i.i.d. case. We make the following assumption.

**Condition 3**
The innovation density $f$ has *finite Fisher information for location.* This means that $f$ is absolutely continuous and

$$J = \int \ell^2(y) f(y)\, dy < \infty, \quad \text{where } \ell = f'/f.$$

Then we can write

$$g'(x) = \int \ell(y) f_\vartheta(y - x) f(y)\, dy = -\frac{1}{\vartheta} \int \ell(y) f(x + \vartheta y) f(y)\, dy. \tag{3}$$

The first identity follows from a substitution, while the second follows from integration by parts. It is well-known that the MA(1) model is locally asymptotically normal under Condition 3. For various versions see Kreiss (1987), Jeganathan (1995), Drost et al. (1997), Koul & Schick

(1997) and Schick & Wefelmeyer (2002b). Here we work with the following version. Introduce a local model by perturbing $\vartheta$ as $\vartheta_{nc} = \vartheta + n^{-1/2}c$, and $f$ as $f_{nh}$ such that

$$\int \Big(n^{1/2}\big(f_{nh}^{1/2}(y) - f^{1/2}(y)\big) - \frac{1}{2}h(y)f^{1/2}(y)\Big)^2 dy \to 0.$$

For convenience, we choose $f_{nh}$ such that, in addition, $\|f_{nh} - f\|_\infty \to 0$. For an explicit construction of $f_{nh}$ see Schick & Wefelmeyer (2002b). Here the local parameter $c$ is real, and $h$ belongs to

$$L_{2,0}(F) = \{h \in L_2(F) : \int h dF = 0\}.$$

Write $P_n$ and $P_{nch}$ for the joint distribution of $(X_1, \dots, X_n)$ under $(\vartheta, f)$ and $(\vartheta_{nc}, f_{nh})$, respectively. Then the local log-likelihood has stochastic approximation

$$\begin{aligned} \log \frac{dP_{nch}}{dP_n} &= n^{-1/2}\sum_{j=1}^n \big(c\dot\varepsilon_j \ell(\varepsilon_j) + h(\varepsilon_j)\big) \\ &\quad -\frac{1}{2}\Big(c^2 E[\dot\varepsilon^2]J + 2cE[\dot\varepsilon]\int \ell h dF + \int h^2 dF\Big) + o_p(1). \end{aligned} \tag{4}$$

The inner product induced by (4) is

$$\begin{aligned} ((c,h),(c_1,h_1)) &= E[(c\dot\varepsilon\ell(\varepsilon) + h(\varepsilon))(c_1\dot\varepsilon\ell(\varepsilon) + h_1(\varepsilon))] \\ &= cc_1 E[\dot\varepsilon^2]J + cE[\dot\varepsilon]\int \ell h_1 dF + c_1 E[\dot\varepsilon]\int \ell h dF + \int h h_1 dF. \end{aligned}$$

A real-valued functional $\kappa$ is *differentiable* at $(\vartheta, f)$ if there exist $c_* \in \mathbb{R}$ and $h_* \in L_{2,0}(F)$ such that

$$n^{1/2}\big(\kappa(\vartheta_{nc}, f_{nh}) - \kappa(\vartheta, f)\big) \to cc_* + \int hh_* dF \quad \text{for all } (c,h) \in \mathbb{R} \times L_{2,0}(F).$$

To characterize efficient estimators of $\kappa$, we need to express the right-hand side in terms of the above inner product, i.e., we must find $(c_\kappa, h_\kappa) \in \mathbb{R} \times L_{2,0}(F)$ such that

$$cc_* + \int hh_* dF = ((c,h),(c_\kappa, h_\kappa)) \quad \text{for all } (c,h) \in \mathbb{R} \times L_{2,0}(F).$$

It is easy to check that

$$\begin{aligned} c_\kappa &= \frac{c_* - E[\dot\varepsilon]\int \ell h_* dF}{\mathrm{Var}[\dot\varepsilon]J} = \frac{(1-\vartheta^2)c_* - (1+\vartheta)\mu\int \ell h_* dF}{\sigma^2 J}, \\ h_\kappa &= h_* - c_\kappa E[\dot\varepsilon]\ell = h_* - c_\kappa \frac{\mu}{1-\vartheta}\ell. \end{aligned}$$

An estimator $\hat\kappa$ of $\kappa$ is called *regular* at $(\vartheta, f)$ with *limit* $L$ if $L$ is a random variable such that

$$n^{1/2}(\hat\kappa - \kappa(\vartheta_{nc}, f_{nh})) \Rightarrow L \quad \text{under } P_{nch} \quad \text{for all } (c,h) \in \mathbb{R} \times L_{2,0}(F).$$

The convolution theorem says that $L$ is the convolution of a normal random variable with mean zero and variance $\|(c_\kappa, h_\kappa)\|^2 = ((c_\kappa, h_\kappa), (c_\kappa, h_\kappa))$, and some other random variable. This justifies calling $\hat\kappa$ *efficient* for $\kappa$ at $(\vartheta, f)$ if $\hat\kappa$ is regular and asymptotically normal with mean zero and variance $\|(c_\kappa, h_\kappa)\|^2$. It also follows from the convolution theorem that $\hat\kappa$ is efficient if and only if

$$n^{1/2}(\hat\kappa - \kappa(\vartheta, f)) = n^{-1/2} \sum_{j=1}^n \big(c_\kappa \dot\varepsilon_j \ell(\varepsilon_j) + h_\kappa(\varepsilon_j)\big) + o_p(1).$$

Here we have used a time series version of the convolution theorem from Bickel et al. (1998, Section 3.3).

For efficient estimation of $g(x)$, we interpret $g(x)$ as functional of $(\vartheta, f)$ via

$$g(x) = \int f(x + \vartheta y) f(y)\, dy =: \kappa_g(\vartheta, f).$$

**Theorem 2**

*Assume Condition 3. Then the functional $\kappa_g$ is differentiable with $c_* = \dot g(x)$ and $h_* = \psi$:*

$$n^{1/2}\Big(\int f_{nh}(x + \vartheta_{nc} y) f_{nh}(y)\, dy - \int f(x + \vartheta y) f(y)\, dy\Big) \to c\dot g(x) + \int h\psi dF.$$

Using (3) we find that $\int \ell\psi\, dF = (1 - \vartheta) g'(x)$. Thus, by Theorem 2, the pair $(c_\kappa, h_\kappa)$ for the functional $\kappa = \kappa_g$ is given by

$$c_\kappa = (\dot g(x) - \mu g'(x)) \frac{1 - \vartheta^2}{\sigma^2 J} \quad \text{and} \quad h_\kappa = \psi - c_\kappa \frac{\mu}{1 - \vartheta} \ell.$$

Consequently, an estimator $\hat g(x)$ of $g(x)$ is efficient if and only if

$$n^{1/2}(\hat g(x) - g(x)) = n^{-1/2} \sum_{j=1}^n \Big(\psi(\varepsilon_j) + (\dot g(x) - \mu g'(x)) \frac{1 - \vartheta^2}{\sigma^2 J} \Big(\dot\varepsilon_j - \frac{\mu}{1 - \vartheta}\Big)\ell(\varepsilon_j)\Big) + o_p(1).$$

Such an estimator has asymptotic variance

$$\tau_*^2 = \int \psi^2 dF + (\dot g(x) - \mu g'(x))^2 \frac{1 - \vartheta^2}{\sigma^2 J}. \tag{5}$$

From this we see that our kernel estimator $\hat g(x)$ is efficient if

$$n^{1/2}(\hat\vartheta - \vartheta) = n^{-1/2} \sum_{j=1}^n \frac{1 - \vartheta^2}{\sigma^2 J} \Big(\dot\varepsilon_j - \frac{\mu}{1 - \vartheta}\Big)\ell(\varepsilon_j) + o_p(1). \tag{6}$$

This is the characterization of an efficient estimator of $\vartheta$. Indeed, the functional $\kappa(\vartheta, f) = \vartheta$ is differentiable with $c_* = 1$ and $h_* = 0$, which yields

$$c_\kappa = \frac{1 - \vartheta^2}{\sigma^2 J} \quad \text{and} \quad h_\kappa = -c_\kappa \frac{\mu}{1 - \vartheta} \ell.$$

Efficient estimators of $\vartheta$ were constructed in Kreiss (1987) under the assumption of symmetry, and in Drost et al. (1997), Koul & Schick (1997) and Schick & Wefelmeyer (2002b) under the assumption that $\mu = 0$. These constructions can be adapted to our slightly more general situation. Just replace $\dot\xi_j(\vartheta)$ by $\dot\xi_j(\vartheta) - \frac{1}{n-r_n(\vartheta)} \sum_{i=r_n(\vartheta)+1}^n \dot\xi_i(\vartheta)$ in Schick & Wefelmeyer (2002b, (5.4) and (5.6)).

*Remark 1.* We have made no structural assumptions on $f$. If such are made, the local parameter $h$ must be restricted correspondingly, to some subset $H$, say, of $L_{2,0}(F)$. Then the above $h_*$ and $h_\kappa$ must be chosen in $H$ rather than in $L_{2,0}(F)$.

For example, if $f$ has expectation zero, then $H = \{h \in L_{2,0}(F) : \int y h(y) f(y)\, dy = 0\}$, and $c_\kappa$ and $h_\kappa$ must be replaced by

$$c_\kappa^0 = \dot g(x) \frac{1-\vartheta^2}{\sigma^2 J} \quad \text{and} \quad h_\kappa^0(y) = \psi(y) - a_* y,$$

with

$$a_* = \frac{\int z\psi(z) f(z)\, dz}{\sigma^2}. \tag{7}$$

If $f$ is known to be symmetric about zero, then $H = \{h \in L_{2,0} : h \text{ is symmetric about zero}\}$, and $c_\kappa$ and $h_\kappa$ must be replaced by

$$c_\kappa^s = \dot g(x) \frac{1-\vartheta^2}{\sigma^2 J} \quad \text{and} \quad h_\kappa^s(y) = \frac{1}{2}(\psi(y) + \psi(-y)).$$

In both submodels we have $\mu = 0$, and the stochastic approximation (6) reduces to

$$n^{1/2}(\hat\vartheta - \vartheta) = n^{-1/2} \sum_{j=1}^n \frac{1-\vartheta^2}{\sigma^2 J} \dot\varepsilon_j \ell(\varepsilon_j) + o_p(1).$$

It can be checked that this characterizes efficient estimators in both submodels. Hence an estimator $\hat\vartheta$ satisfying (6) is already efficient in these submodels. The estimator $\hat\vartheta$ mentioned above is thus efficient in the submodels as well. Under symmetry it would however be better to use a symmetrized kernel estimator in the construction of $\hat\vartheta$; see Kreiss (1987), Jeganathan (1995), Drost et al. (1997) and Koul & Schick (1997).

In the following remarks we construct improved estimators for $g(x)$ in submodels obtained by restricting the innovation density $f$.

*Remark 2.* Suppose that $f$ is known to have mean zero, as in Saavedra & Cao (1999a, 2000). In this case, the estimators $\hat g(x)$ and $\hat g_{SC}(x)$ for $g(x)$ are not efficient any more. To see this, consider the class of estimators

$$\hat g_a(x) = \hat g(x) - a \frac{1}{n-r} \sum_{j=r+1}^n \hat\varepsilon_j,$$

with $\hat\vartheta$ taken to be an efficient estimator for $\vartheta$. As $\mu = 0$, we have $E[\dot\varepsilon] = 0$ and, from (13) below and $r/n \to 0$,

$$\frac{1}{n-r}\sum_{j=r+1}^{n}\hat\varepsilon_j = \frac{1}{n-r}\sum_{j=r+1}^{n}\varepsilon_j + (\hat\vartheta-\vartheta)\frac{1}{n-r}\sum_{j=r+1}^{n}\dot\varepsilon_j + o_p(n^{-1/2}) = \frac{1}{n}\sum_{j=1}^{n}\varepsilon_j + o_p(n^{-1/2}).$$

Hence we obtain

$$n^{1/2}(\hat g_a(x) - g(x)) = n^{-1/2}\sum_{j=1}^{n}\Big(\psi(\varepsilon_j) - a\varepsilon_j + \dot g(x)\,\frac{1-\vartheta^2}{\sigma^2 J}\,\dot\varepsilon_j \ell(\varepsilon_j)\Big) + o_p(1).$$

Thus the asymptotic variance of $n^{1/2}(\hat g_a(x) - g(x))$ is

$$\tau^2(a) = \int (\psi(y) - ay)^2 f(y)\,dy + \dot g(x)^2\,\frac{1-\vartheta^2}{\sigma^2 J}$$

which is uniquely minimized by $a = a_*$ as in (7), with minimal value

$$\tau^2(a_*) = \tau^2(0) - a_*^2\sigma^2 = \tau_*^2 - a_*^2\sigma^2.$$

This shows that $\hat g_{a_*}(x)$ has smaller asymptotic variance than $\hat g(x)$. The constant $a_*$ depends on $\vartheta$ and $f$ and must be estimated. If $\hat a_*$ is a consistent estimator for $a_*$, then $\hat g_{\hat a_*}(x)$ is asymptotically equivalent to $\hat g_{a_*}(x)$ and therefore better than $\hat g(x)$. Furthermore,

$$n^{1/2}(\hat g_{\hat a_*}(x) - g(x)) = n^{-1/2}\sum_{j=1}^{n}\Big(\psi(\varepsilon_j) - a_*\varepsilon_j + \dot g(x)\,\frac{1-\vartheta^2}{\sigma^2 J}\,\dot\varepsilon_j \ell(\varepsilon_j)\Big) + o_p(1).$$

It follows from Remark 1 that $\hat g_{\hat a_*}(x)$ is efficient for $g(x)$ in the restricted model.

Efficiency results for such corrections by "estimators of zero" under linear constraints like $E[\varepsilon] = 0$ have a long history. See Levit (1975) for the i.i.d. case, Wefelmeyer (1994) for AR(1), Schick & Wefelmeyer (2002a,b,c) for more general autoregressive and linear processes, and Müller et al. (2001, 2002) for nonparametric Markov chain models and nonparametric regression.

*Remark 3.* Suppose $f$ is known to be symmetric about zero. Then $-\varepsilon_t$ is distributed as $\varepsilon_t$, and $\mu = 0$. Hence we can use the symmetrized estimator

$$\frac{1}{4(n-r+1)(n-r)}\sum_{\substack{i,j=r+1\\ i\neq j}}^{n}\big(K_b(x - \hat\varepsilon_i + \hat\vartheta\hat\varepsilon_j) + K_b(x + \hat\varepsilon_i + \hat\vartheta\hat\varepsilon_j) + K_b(x - \hat\varepsilon_i - \hat\vartheta\hat\varepsilon_j) + K_b(x + \hat\varepsilon_i - \hat\vartheta\hat\varepsilon_j)\big)$$

with $\hat\vartheta$ chosen to be efficient in this restricted model. Under the assumptions of Theorem 1, this estimator satisfies

$$n^{1/2}(\hat g_s(x) - g(x)) = n^{-1/2}\sum_{j=1}^{n}\Big(\frac{1}{2}(\psi(\varepsilon_j) + \psi(-\varepsilon_j)) + \dot g(x)\,\frac{1-\vartheta^2}{\sigma^2 J}\,\dot\varepsilon_j \ell(\varepsilon_j)\Big) + o_p(1).$$

By Remark 1, the symmetrized estimator $\hat g_s(x)$ is efficient for $g(x)$ under symmetry.

## 3 Simulations

In this section we report the results of a simulation study in which we compare the estimator $\hat{g}(x)$ with the usual kernel estimator $\overline{g}(x)$. We restrict ourselves to standard normal innovations, to two choices of parameter values $\vartheta = .3, .7$, three arguments $x = 0, .5, 1$, and sample sizes $n = 30, 60$. Even for these small sample sizes, $\hat{g}(x)$ turns out to be noticeably better.

We have used the kernel $K(x) = (3 - x^2)\exp(-x^2/2)/(2\sqrt{2\pi})$. As estimator of $\vartheta$ we have used an approximate minimizer of $\sum_{j=r+1}^n (\varepsilon_{j,r}(\vartheta))^2$, with $\varepsilon_{j,r}(\vartheta) = \sum_{s=0}^r \vartheta^s X_{j-s}$, namely

$$\hat{\vartheta} = \tilde{\vartheta} - \frac{\sum_{j=r+1}^n \dot{\varepsilon}_{j,r}(\tilde{\vartheta})\varepsilon_{j,r}(\tilde{\vartheta})}{\sum_{j=r+1}^n (\dot{\varepsilon}_{j,r}(\tilde{\vartheta}))^2},$$

where $\tilde{\vartheta}$ is the moment estimator defined as the solution of

$$\frac{\tilde{\vartheta}}{1+\tilde{\vartheta}^2} = \Big(-\frac{1}{2}\Big) \vee \frac{\sum_{j=2}^n X_{j-1}X_j}{\sum_{j=2}^n X_{j-1}^2} \wedge \frac{1}{2}.$$

This equation is an empirical version of $E(X_1X_0)/E(X_0^2) = \vartheta/(1+\vartheta^2)$. Here, of course, $\dot{\varepsilon}_{j,r}(\vartheta)$ denotes the derivative of $\varepsilon_{j,r}(\vartheta)$ with respect to $\vartheta$. On the basis of preliminary simulations we have chosen $r = 3$.

TABLE 1 GOES APPROXIMATELY HERE

In Table 1 we report the simulated mean square errors times $10^5$ based on 20,000 repetitions for several choices of bandwidth $b$. For each bandwidth, our estimator $\hat{g}(x)$ has smaller mean square error than the kernel estimator. The improvement is more pronounced for larger values of $\vartheta$. It is around ten percent for $\vartheta = .3$, and around 50 percent for $\vartheta = .7$. The kernel estimator is known to be sensitive to the choice of bandwidth. For the small sample sizes considered here, this appears to be true also for our estimator. Simulations not reported here indicate that our estimator becomes less sensitive to the choice of bandwidth for larger sample sizes. This is consistent with the asymptotic theory, which shows that the asymptotic variance of our estimator is independent of the bandwidth.

## 4 Concluding remarks

In a linear time series driven by independent innovations, the stationary density can be expressed as a smooth function of the innovation density and the parameters. This suggests estimating the stationary density by a plug-in estimator, which replaces the innovation density and the parameters by estimators. By the plug-in principle we expect that the resulting estimator can converge at the parametric rate, even though the estimator of the innovation density has a slower, nonparametric, rate of convergence. In this paper we have shown that this is indeed the case in a moving average process of order one. We expect our approach to work more generally for invertible linear time series models or for nonlinear autoregressive models. A simulation study has shown that the improvement is noticeable already for small sample sizes. Thus our estimator constitutes a practically useful alternative to the usual kernel estimators.

**Acknowledgments**
Anton Schick was partially supported by NSF Grant DMS 0072174.

## References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. & Wellner, J. A. (1998). *Efficient and adaptive estimation for semiparametric models.* Springer, New York.

Birgé, L. & Massart, P. (1995). Estimation of integral functionals of a density. *Ann. Statist.* **23**, 11–29.

Blanke, D. & Bosq, D. (1997). Accurate rates of density estimators for continuous-time processes. *Statist. Probab. Lett.* **33**, 185–191.

Bosq, D. (1993). Vitesses optimales et superoptimales des estimateurs fonctionnels pour les processus à temps continu. *C. R. Acad. Sci. Paris Sér. I. Math.* **317**, 1075–1078.

Bosq, D. (1995). Sur le comportement exotique de l'estimateur à noyau de la densité marginale d'un processus à temps continu. *C. R. Acad. Sci. Paris Sér. I. Math.* **320**, 369–372.

Bosq, D., Merlevède, F. & Peligrad, M. (1999). Asymptotic normality for density kernel estimators in discrete and continuous time. *J. Multivariate Anal.* **68**, 78–95.

Brockwell, P. & Davis, R. A. (2002). *Introduction to time series and forecasting.* Second edition. Springer, New York.

Castellana, J. V. & Leadbetter, M. R. (1986). On smoothed probability density estimation for stationary processes. *Stochastic Process. Appl.* **21**, 179–193.

Chaudhuri, P., Doksum, K. & Samarov, A. (1997). On average derivative quantile regression. *Ann. Statist.* **25**, 715–744.

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

Eggermont, P. P. B. & LaRiccia, V. N. (1999). Best asymptotic normality of the kernel density entropy estimator for smooth densities. *IEEE Trans. Inform. Theory* **45**, 1321–1326.

Eggermont, P. P. B. & LaRiccia, V. N. (2001). *Maximum penalized likelihood estimation, vol I, density estimation.* Springer, New York.

Efromovich, S. & Samarov, A. (2000). Adaptive estimation of the integral of squared regression derivatives. *Scand. J. Statist.* **27**, 335–351.

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

Klaassen, C. A. J. & Putter, H. (2002). Efficient estimation of Banach parameters in semiparametric models. Technical Report, Korteweg-de Vries Institute for Mathematics, University of Amsterdam. Available at http://preprint.beta.uva.nl/

Koul, H. L. & 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.

Kreiss, J.-P. (1990). Local asymptotic normality for autoregression with infinite order. *J. Statist. Plann. Inference* **26**, 185–219.

Kreiss, J.-P. (1991). Estimation of the distribution function of noise in stationary processes. *Metrika* **38**, 285–297.

Kutoyants, Yu. A. (1997). Some problems of nonparametric estimation by observations of ergodic diffusion process. *Statist. Probab. Lett.* **32**, 311–320.

Kutoyants, Yu. A. (1998). On density estimation by the observations of ergodic diffusion processes. In: Kabanov, Y. M., Rozovskii, B. L., Shiryaev, A. N. (Eds.), *Statistics and control of stochastic processes*, World Scientific, Singapore, pp. 253–274.

Kutoyants, Yu. A. (1999). Efficient density estimation for ergodic diffusion processes. *Stat. Inference Stoch. Process.* **1**, 131–155.

Levit, B. Y. (1975). Conditional estimation of linear functionals. *Problems Inform. Transmission* **11**, 39–54.

Mason, D. M. (2003). Representations for estimators of integral functionals of the density function. To appear in: *Austrian J. Statist.*

Müller, U. U., Schick, A. & Wefelmeyer, W. (2001). Improved estimators for constrained Markov chain models. *Statist. Probab. Lett.* **54**, 427-435.

Müller, U. U., Schick, A. & Wefelmeyer, W. (2002). Estimating linear functionals of the error distribution in nonparametric regression. To appear in: *J. Statist. Plann. Inference.*

Saavedra, A. & Cao, R. (1999a). Rate of convergence of a convolution-type estimator of the marginal density of an MA(1) process. *Stochastic Process. Appl.* **80**, 129–155.

Saavedra, A. & Cao, R. (1999b). A comparative study of two convolution-type estimators of the marginal density of moving average processes. *Comput. Statist.* **14**, 355–373.

Saavedra, A. & Cao, R. (2000). On the estimation of the marginal density of a moving average process.

*Canad. J. Statist.* **28**, 799–815.

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

Schick, A. & Wefelmeyer, W. (2002b). Efficient estimation in invertible linear processes. To appear in: *Math. Methods Statist.*

Schick, A. & Wefelmeyer, W. (2002c). Estimating invariant laws of linear processes by U-statistics. Technical Report, Department of Mathematical Sciences, Binghamton University. Available at http://math.binghamton.edu/anton/preprint.html.

Schick, A. & Wefelmeyer, W. (2002d). Functional convergence and optimality of plug-in estimators for stationary densities of moving average processes. Technical Report, Department of Mathematical Sciences, Binghamton University. Available at http://math.binghamton.edu/anton/preprint.html.

Tsybakov, A. B. & van der Meulen, E. C. (1996). Root-$n$ consistent estimators of entropy for densities with unbounded support. *Scand. J. Statist.* **23**, 75–83.

Veretennikov, A. Yu. (1999). On Castellana-Leadbetter's condition for diffusion density estimation. *Stat. Inference Stoch. Process.* **2**, 1–9.

Wefelmeyer, W. (1994). An efficient estimator for the expectation of a bounded function under the residual distribution of an autoregressive process. *Ann. Inst. Statist. Math.* **46**, 309–315.

Anton Schick, Binghamton University, Department of Mathematical Sciences, Binghamton, NY 13902-6000, USA, anton@math.binghamton.edu.

## Appendix 1: Auxiliary results

This appendix contains some lemmas which will be used in the proof of Theorem 1. Let $\varepsilon_t$, $t \in \mathbb{Z}$, be i.i.d. with distribution function $F$, mean $\mu$ and finite variance. Let $k_n(x,y)$ be $F \times F$-square-integrable. First we study the asymptotic behavior of the U-statistic

$$U_n = \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} k_n(\varepsilon_i, \varepsilon_j).$$

Denote its mean by

$$\nu_n = \int k_n dF \times F = \int k_n(y,z) dF(y) dF(z).$$

Set

$$k_{n1}(y) = \int k_n(y,z) dF(z) - \nu_n, \qquad k_{n2}(z) = \int k_n(y,z) dF(y) - \nu_n.$$

**Lemma 1**
*Suppose $\nu_n \to \nu$ for some $\nu$ in $\mathbb{R}$, and assume that*

$$\int k_n^2 dF \times F = o(n^2) \quad \text{and} \quad \int k_{ni}^2 dF = o(n), \quad i = 1, 2. \tag{8}$$

*Then* $U_n = \nu + o_p(1)$.

*Proof.* We have the Hoeffding decomposition $U_n = \nu_n + U_{n1} + U_{n2} + R_n$, where

$$\begin{aligned} U_{ni} &= \frac{1}{n}\sum_{j=1}^{n} k_{ni}(\varepsilon_j) - \nu_n, \\ R_n &= \frac{1}{n(n-1)}\sum_{i\neq j}\big(k_n(\varepsilon_i,\varepsilon_j) - k_{n1}(\varepsilon_i) - k_{n2}(\varepsilon_j) + \nu_n\big). \end{aligned}$$

We have $n(n-1)E[R_n^2] \le \int k_n^2 dF \times F = o(n^2)$ and $nE[U_{ni}^2] \le \int k_{ni}^2 dF = o(n)$. Hence $U_{ni} = o_p(1)$ and $R_n = o_p(1)$. The desired result is now immediate.

**Lemma 2**
*Suppose that*
(a) *there exist a* $\nu$ *in* $\mathbb{R}$ *such that* $n^{1/2}(\nu_n - \nu) \to 0$,
(b) *there are functions* $\gamma_1$ *and* $\gamma_2$ *in* $L_2(F)$ *such that* $\int (k_{ni} - \gamma_i)^2 dF \to 0$, $i = 1, 2$, *and*
(c) $\int k_n^2 dF \times F = o(n)$.
*Then*

$$n^{1/2}(U_n - \nu) = n^{-1/2}\sum_{j=1}^{n}\big(\gamma_1(\varepsilon_j) + \gamma_2(\varepsilon_j) - 2\nu\big) + o_p(1).$$

*Proof.* Consider the Hoeffding decomposition in the proof of Lemma 1. We have $n^{1/2}R_n = o_p(1)$ since $n(n-1)E[R_n^2] \le \int k_n^2 dF \times F = o(n)$ by (c). It follows from (a) and (b) that $\int \gamma_i dF = \nu$ and then that

$$n^{1/2}U_{ni} = n^{-1/2}\sum_{j=1}^{n}(\gamma_i(\varepsilon_j) - \nu) + o_p(1).$$

The above and (a) now yield the desired result.

Next we consider the statistic

$$W_n = \frac{1}{n(n-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{n} k_n(\varepsilon_i,\varepsilon_j)\sum_{s=1}^{\infty}\beta_s\varepsilon_{i-s} \quad \text{with} \quad \sum_{s=1}^{\infty}|\beta_s| < \infty.$$

**Lemma 3**
*Suppose* $\nu_n \to \nu$ *for some* $\nu$ *in* $\mathbb{R}$, *and*

$$\int k_n^2(y,z)(1 + y^2 + z^2)dF(y)dF(z) = o(n). \tag{9}$$

*Then*

$$W_n = \nu\mu\sum_{s=1}^{\infty}\beta_s + o_p(1).$$

*Proof.* We can write $W_n = \tilde{W}_n + U_n\mu\sum_{s=1}^{\infty}\beta_s$ with

$$\tilde{W}_n = \frac{1}{n(n-1)}\sum_{i\neq j} k_n(\varepsilon_i,\varepsilon_j)\sum_{s=1}^{\infty}\beta_s(\varepsilon_{i-s} - \mu).$$

The second moment of $\tilde{W}_n$ is

$$\frac{1}{n^2(n-1)^2}\sum_{s=1}^{\infty}\beta_s\sum_{s'=1}^{\infty}\beta_{s'}\sum_{i\neq j}\sum_{i'\neq j'}E[k_n(\varepsilon_i,\varepsilon_j)(\varepsilon_{i-s}-\mu)k_n(\varepsilon_{i'},\varepsilon_{j'})(\varepsilon_{i'-s'}-\mu)].$$

Note that the expectations appearing in this sum are zero if either $i-s$ does not equal one of $j$, $i'$, $j'$, $i'-s'$, or $i'-s'$ does not equal one of $j'$, $i$, $j$, $i-s$. Otherwise, each expectation is bounded by a multiple of the left-hand side of (9). Since, for fixed $s$ and $s'$, at most $7n(n-1)+n^2+n(n-1)^2 \le 7n^3$ terms are nonzero, the second moment of $\tilde{W}_n$ is of order $\frac{1}{n}(\sum_{s=1}^{\infty}|\beta_s|)^2 o(n) = o(1)$. This shows that $\tilde{W}_n = o_p(1)$, and the desired result follows since $U_n = \nu + o_p(1)$ by Lemma 1, which applies because (9) implies (8).

## Appendix 2: Proof of Theorem 1

To simplify notation, set $(n)_r = (n-r)(n-r-1)$, and write $\sum^*$ for the double sum extending over $i, j = r+1, \dots, n$ with $i \neq j$. We introduce

$$\begin{aligned}
\tilde{g}(x) &= \frac{1}{(n)_r}\sum\nolimits^* K_b(x-\varepsilon_i+\vartheta\varepsilon_j+(\hat{\vartheta}-\vartheta)(\varepsilon_j-\dot{\varepsilon}_i+\vartheta\dot{\varepsilon}_j)),\\
\overline{g}(x) &= \frac{1}{(n)_r}\sum\nolimits^* K_b(x-\varepsilon_i+\vartheta\varepsilon_j).
\end{aligned}$$

The desired expansion of $\hat{g}(x)$ follows if we show that

$$n^{1/2}(\hat{g}(x)-\tilde{g}(x)) = o_p(1), \tag{10}$$

$$n^{1/2}(\tilde{g}(x)-\overline{g}(x)) = \big(\dot{g}(x)-\mu g'(x)\big)n^{1/2}(\hat{\vartheta}-\vartheta)+o_p(1), \tag{11}$$

$$n^{1/2}(\overline{g}(x)-g(x)) = n^{-1/2}\sum_{j=1}^{n}\psi(\varepsilon_j)+o_p(1). \tag{12}$$

*Proof of (10).* The key to (10) is the following expansion:

$$\sum_{j=r+1}^{n}\big|\hat{\varepsilon}_j-\varepsilon_j-(\hat{\vartheta}-\vartheta)\dot{\varepsilon}_j\big| = O_p(1). \tag{13}$$

To this end, recall the representations for $\hat{\varepsilon}_j$, $\varepsilon_j$ and $\dot{\varepsilon}_j$. The left-hand side of (13) can be bounded by $T_1+T_2+|\hat{\vartheta}-\vartheta|T_3$ with

$$T_1 = \sum_{j=r+1}^{n}\sum_{s=0}^{r}|\hat{\vartheta}^s-\vartheta^s-(\hat{\vartheta}-\vartheta)s\vartheta^{s-1}||X_{j-s}|,$$

$$T_2 = \sum_{j=r+1}^{n}\Big|\sum_{s=r+1}^{\infty}\vartheta^s X_{j-s}\Big|, \qquad T_3 = \sum_{j=r+1}^{n}\Big|\sum_{s=r+1}^{\infty}s\vartheta^{s-1}X_{j-s}\Big|.$$

By choice of $r$,

$$E[T_3] \le \sum_{j=r+1}^{n}\sum_{s=r+1}^{\infty}s|\vartheta|^{s-1}E[|X_0|] = E[|X_0|]O((n-r)r|\vartheta|^r) = o(1)$$

and, similarly, $E[T_2] = o(1)$. Since $\hat\vartheta$ is $n^{1/2}$-consistent, there is a constant $\rho < 1$ such that the probability of $|\hat\vartheta| > \rho$ tends to zero. On the event where $|\hat\vartheta| \le \rho$ we get

$$T_1 \le (\hat\vartheta - \vartheta)^2 \sum_{j=r+1}^{n} \sum_{s=0}^{r} s(s-1)\rho^{s-2}|X_{j-s}| \le (\hat\vartheta - \vartheta)^2 \sum_{s=0}^{\infty} s^2 \rho^{s-2} \sum_{j=1}^{n} |X_j| = O_p(1).$$

This proves (13).

The difference between the corresponding arguments of $K_b$ in $\hat g(x)$ and $\tilde g(x)$ is

$$\begin{aligned} & x - \hat\varepsilon_i + \hat\vartheta\hat\varepsilon_j - \big(x - \varepsilon_i + \vartheta\varepsilon_j + (\hat\vartheta - \vartheta)(\varepsilon_j - \dot\varepsilon_i + \vartheta\dot\varepsilon_j)\big) \\ &\quad = -\big(\hat\varepsilon_i - \varepsilon_i - (\hat\vartheta - \vartheta)\dot\varepsilon_i\big) + \hat\vartheta\big(\hat\varepsilon_j - \varepsilon_j - (\hat\vartheta - \vartheta)\dot\varepsilon_j\big) + (\hat\vartheta - \vartheta)^2\dot\varepsilon_j. \end{aligned}$$

Since $K$ has a bounded derivative, we can bound the absolute value of the left-hand side of (10) by

$$n^{1/2} b^{-2} \|K'\|_\infty \Big( (1 + |\hat\vartheta|) \frac{1}{n-r} \sum_{j=r+1}^{n} \big|\hat\varepsilon_j - \varepsilon_j - (\hat\vartheta - \vartheta)\dot\varepsilon_j\big| + (\hat\vartheta - \vartheta)^2 \frac{1}{n-r} \sum_{i=r+1}^{n} |\dot\varepsilon_i| \Big).$$

This bound converges to zero in probability by the $n^{1/2}$-consistency of $\hat\vartheta$, relation (13) and $n^{-1/2}b^{-2} \to 0$. This completes the proof of (10).

*Proof of (11).* Let

$$\begin{aligned} U &= \frac{1}{(n)_r} \sum\nolimits^* K_b'(x - \varepsilon_i + \vartheta\varepsilon_j)\varepsilon_j, \\ W_1 &= \frac{1}{(n)_r} \sum\nolimits^* K_b'(x - \varepsilon_i + \vartheta\varepsilon_j)\dot\varepsilon_i, \\ W_2 &= \frac{1}{(n)_r} \sum\nolimits^* K_b'(x - \varepsilon_i + \vartheta\varepsilon_j)\dot\varepsilon_j. \end{aligned}$$

Then a Taylor expansion and the $n^{1/2}$-consistency of $\hat\vartheta$ yield

$$\begin{aligned} S &= \Big| n^{1/2}(\tilde g(x) - \overline{g}(x)) - n^{1/2}(\hat\vartheta - \vartheta)(U - W_1 + \vartheta W_2) \Big| \\ &\le n^{1/2} b^{-3} \|K''\|_\infty \frac{1}{(n)_r} \sum\nolimits^* (\hat\vartheta - \vartheta)^2 (\varepsilon_j - \dot\varepsilon_i + \vartheta\dot\varepsilon_j)^2 = O_p(n^{-1/2}b^{-3}) = o_p(1). \end{aligned}$$

This is the only place where we need that $nb^6 \to \infty$. Let us now show that we can obtain $S = o_p(1)$ under $nb^4 \to \infty$ if we also require $f$ to have a finite third moment and $K$ to possess a bounded third derivative. In this case we use a higher order Taylor expansion to obtain the bound

$$S \le \frac{1}{(n)_r} \sum\nolimits^* \Big( n^{1/2}(\hat\vartheta - \vartheta)^2 V_{i,j}^2 |K_b''(x - \varepsilon_i + \vartheta\varepsilon_j)| + n^{1/2}|\hat\vartheta - \vartheta|^3 b^{-4} \|K'''\|_\infty V_{i,j}^3 \Big)$$

with $V_{i,j} = |\varepsilon_j - \dot\varepsilon_i + \vartheta\dot\varepsilon_j|$. In view of the $n^{1/2}$-consistency of $\hat\vartheta$ we obtain $S = O(n^{-1/2}b^{-2})$ if we show that that $E[V_{i,j}^3] < C$ and $D_{i,j} = E[V_{i,j}^2 |K_b''(x - \varepsilon_i + \vartheta\varepsilon_j)|] \le Db^{-2}$ for positive constants $C$ and $D$ and all $i \ne j$. The former is easy. For the latter recall that $K$ has compact support. This implies that $K''$ has compact support and shows that $|K_b''| \le ab^{-3}\mathbf{1}_{[-Lb,Lb]}$ for finite constants $a$ and $L$. For $j < i$, $\varepsilon_i$ is independent of $(V_{i,j}, \varepsilon_j)$ and we get

$$D_{i,j} = E\Big[ V_{i,j}^2 \int |K_b''(x - y + \vartheta\varepsilon_j)| f(y)\, dy \Big] \le 2ab^{-2}L\|f\|_\infty E[V_{i,j}^2].$$

For $j > i$, $\varepsilon_j$ is independent of $(\varepsilon_i, \dot\varepsilon_i, \dot\varepsilon_j)$, and we get

$$D_{i,j} = E\Big[\int (y - \dot\varepsilon_i + \vartheta\dot\varepsilon_j)^2 |K_b''(x - \varepsilon_i + \vartheta y)| f(y)\, dy\Big] \le 6ab^{-2}L\|f\|_\infty E[(|x-\varepsilon_i| + Lb)^2 + \dot\varepsilon_i^2 + \dot\varepsilon_j^2].$$

This establishes the desired bound $D_{i,j} \le Db^{-2}$.

To deal with $U$, $W_1$ and $W_2$, we will utilize Lemmas 1 and 3. For this we study

$$\nu_{n,a} = \iint K_b'(x - y + \vartheta z) z^a f(y) f(z)\, dydz = \frac{1}{b}\iint K'(u) f(x + \vartheta z - bu) z^a f(z)\, dudz.$$

for $a = 0, 1$. We have $\int K'(u)\, du = 0$ and $\int uK'(u)\, du = -1$. By a Taylor expansion

$$\nu_{n,a} = \int f'(x + \vartheta z) z^a f(z)\, dz - \iint uK'(u) \int_0^1 \big(f'(x + \vartheta z - tbu) - f'(x + \vartheta z)\big)\, dt\, z^a f(z)\, dudz.$$

Since $f'$ is bounded and continuous, an application of the Lebesgue dominated convergence theorem gives $\nu_{n,a} \to \int f'(x + \vartheta z) z^a f(z)\, dz$. The limit is $g'(x)$ for $a = 0$, and $\dot g(x)$ for $a = 1$. An application of Lemma 1 with $k_n(y, z) = K_b'(x - y + \vartheta z)z$ yields $U = \dot g(x) + o_p(1)$; while an application of Lemma 3 with $k_n(y, z) = K_b'(x - y + \vartheta z)$, together with the representation (2), yields $W_1 = \mu g'(x)/(1-\vartheta) + o_p(1)$ and $W_2 = \mu g'(x)/(1-\vartheta) + o_p(1)$. Hence $U - W_1 + \vartheta W_2 = \dot g(x) - \mu g'(x) + o_p(1)$, and (11) is proved.

*Proof of (12).* We rely on Lemma 2 to derive (12). We begin by studying

$$\nu_n = \iint K_b(x - y + \vartheta z) f(y) f(z)\, dydz = \iint K(u) f(x + \vartheta z - bu) f(z)\, dudz.$$

We will use twice

$$f(a + h) - f(a) - hf'(a) = h^2 \iint_{0<s<t<1} f''(a + sth)\, dsdt,$$

together with the properties of $K$ and a substitution, to obtain

$$\begin{aligned}
\nu_n - g(x) &= b^2 \iint u^2 K(u) \iint_{0<s<t<1} f''(x + \vartheta z - stbu)\, dsdtdu f(z)\, dz \\
&= b^2 \iint u^2 K(u) \iint_{0<s<t<1} f''(v) f_\vartheta(v - x + stbu)\, dsdtdudv \\
&= b^4 \iint u^4 K(u) f''(v) \iint_{0<s<t<1} s^2t^2 \iint_{0<p<q<1} f_\vartheta''(v - x + pqstbu)\, dpdqdsdtdudv.
\end{aligned}$$

Since $f''$ is bounded and integrable,

$$\nu_n = g(x) + O(b^4) = g(x) + o(n^{-1/2}).$$

This is (a) of Lemma 2 with $\nu = g(x)$. We have

$$\begin{aligned}
k_{n1}(y) &= \int K_b(x - y + \vartheta z)\, f(z)\, dz = \int f_\vartheta(y - x + bu)\, K(u)\, du, \\
k_{n2}(z) &= \int K_b(x - y + \vartheta z)\, f(y)\, dy = \int f(x + \vartheta z - bu)\, K(u)\, du.
\end{aligned}$$

Since $f'$ is bounded by Condition 2, relation (b) of Lemma 2 holds with $\gamma_1(y) = f_\vartheta(y - x)$ and $\gamma_2(y) = f(x + \vartheta y)$. An application of Lemma 2 yields (12).

## Appendix 3: Proof of Theorem 2

Since $f$ has finite Fisher information, $f$ is bounded, and its almost everywhere derivative $f'$ is in $L_1$. We will repeatedly use the fact that for $\varphi \in L_1$ and $a_n \to a \neq 0$,

$$\int |\varphi(a_n y) - \varphi(ay)|\, dy \to 0. \tag{14}$$

Since Hellinger differentiability implies $L_1$-differentiability, we have

$$n^{1/2} \int \left| f_{nh}(y) - f(y) - n^{-1/2} h(y) f(y) \right| dy \to 0. \tag{15}$$

Using the properties of $f$, we get

$$n^{1/2} \int \big( f(x + \vartheta_{nc} y) - f(x + \vartheta y) \big) f(y)\, dy \to c \int y f'(x + \vartheta y) f(y)\, dy. \tag{16}$$

Since $f$ is bounded and continuous,

$$\int f(x + \vartheta_{nc} y) h(y) f(y)\, dy \to \int f(x + \vartheta y) h(y) f(y)\, dy. \tag{17}$$

Since $f$ is bounded, (15) gives

$$n^{1/2} \int f(x + \vartheta_{nc} y) \Big( f_{nh}(y) - f(y) - n^{-1/2} h(y) f(y) \Big)\, dy \to 0. \tag{18}$$

Combining (17) and (18), we get

$$n^{1/2} \int f(x + \vartheta_{nc} y)(f_{nh}(y) - f(y))\, dy \to \int f(x + \vartheta y) h(y) f(y)\, dy. \tag{19}$$

Since $\|f_{nh} - f\|_\infty \to 0$,

$$\int (hf)(x + \vartheta_{nc} y)(f_{nh}(y) - f(y))\, dy \to 0. \tag{20}$$

Also, $\|f_{nh}\|_\infty$ is bounded. Thus the substitution $u = x + \vartheta_{nc} y$ and (15) give

$$n^{1/2} \int \Big( f_{nh}(x + \vartheta_{nc} y) - f(x + \vartheta_{nc} y) - n^{-1/2} (hf)(x + \vartheta_{nc} y) \Big) f_{nh}(y)\, dy \to 0. \tag{21}$$

Finally, since $f$ is bounded and $\vartheta_{nc} \to \vartheta$, we obtain from (14), applied with $\varphi = hf$,

$$\begin{aligned} &\left| \int \big( (hf)(x + \vartheta_{nc} y) - (hf)(x + \vartheta y) \big) f(y)\, dy \right| \qquad (22) \\ &\quad \le \|f\|_\infty \int \big| (hf)(x + \vartheta_{nc} y) - (hf)(x + \vartheta y) f(y) \big| dy \to 0. \end{aligned}$$

Combining (20), (21) and (22),

$$n^{1/2} \int \big( f_{nh}(x + \vartheta_{nc} y) - f(x + \vartheta_{nc} y) \big) f_{nh}(y)\, dy \to \int h(x + \vartheta y) f(x + \vartheta y) f(y)\, dy. \tag{23}$$

Now Theorem 2 follows from (16), (19) and (23).

Table 1: $10^5 \times$ MSE of the estimators $\overline{g}$ and $\hat{g}$, for selected arguments $x$, parameter values $\vartheta$, bandwidths $b$ and sample sizes $n$.

| $x = 0$ | | $n = 30$ | | | | | $n = 60$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\vartheta = .3$ | $b$ | .8 | .9 | 1.0 | 1.1 | 1.2 | .7 | .8 | .9 | 1.0 | 1.1 |
| | $\overline{g}$ | 295 | 224 | 189 | 187 | 217 | 211 | 150 | 115 | 106 | 124 |
| | $\hat{g}$ | 273 | 205 | 169 | 165 | 192 | 183 | 131 | 100 | 92 | 110 |
| $\vartheta = .7$ | $b$ | .6 | .7 | .8 | .9 | 1.0 | .5 | .6 | .7 | .8 | .9 |
| | $\overline{g}$ | 687 | 582 | 520 | 492 | 494 | 491 | 416 | 375 | 360 | 368 |
| | $\hat{g}$ | 332 | 302 | 286 | 287 | 308 | 198 | 193 | 196 | 209 | 235 |
| $x = .5$ | | $n = 30$ | | | | | $n = 60$ | | | | |
| $\vartheta = .3$ | $b$ | .9 | 1.0 | 1.1 | 1.2 | 1.3 | .8 | .9 | 1.0 | 1.1 | 1.2 |
| | $\overline{g}$ | 199 | 159 | 140 | 143 | 165 | 133 | 99 | 82 | 81 | 97 |
| | $\hat{g}$ | 177 | 141 | 124 | 127 | 149 | 113 | 86 | 71 | 71 | 87 |
| $\vartheta = .7$ | $b$ | .7 | .8 | .9 | 1.0 | 1.1 | .6 | .7 | .8 | .9 | 1.0 |
| | $\overline{g}$ | 434 | 372 | 337 | 324 | 330 | 308 | 265 | 242 | 236 | 244 |
| | $\hat{g}$ | 208 | 193 | 188 | 193 | 211 | 130 | 127 | 131 | 141 | 161 |
| $x = 1$ | | $n = 30$ | | | | | $n = 60$ | | | | |
| $\vartheta = .3$ | $b$ | 1.3 | 1.4 | 1.5 | 1.6 | 1.7 | 1.2 | 1.3 | 1.4 | 1.5 | 1.6 |
| | $\overline{g}$ | 62 | 51 | 45 | 44 | 49 | 38 | 30 | 25 | 24 | 27 |
| | $\hat{g}$ | 57 | 47 | 41 | 41 | 45 | 34 | 27 | 23 | 22 | 25 |
| $\vartheta = .7$ | $b$ | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 | 1.0 | 1.1 | 1.2 | 1.3 | 1.4 |
| | $\overline{g}$ | 90 | 77 | 70 | 68 | 71 | 58 | 49 | 45 | 44 | 47 |
| | $\hat{g}$ | 43 | 39 | 39 | 40 | 46 | 24 | 23 | 23 | 26 | 31 |