Comments

Priscilla E. Greenwood [1] **Anton Schick** [2] **Wolfgang Wefelmeyer**

In their thought-provoking essay, Bickel and Kwon (briefly, BK) touch on many important questions of semiparametric inference. We comment on only a few. Our Sections 1 to 4 concern BK's information calculus as applied to Markov chain models. In the first, we recall what BK call the traditional approach. The next two sections try to extract what we see as two essential points of the new information calculus in Markov chain models. The second of these points shows how to calculate efficient influence functions for Markov chains from corresponding bivariate i.i.d. models; this is particularly useful when the model and the parameter of interest are described in terms of the stationary law rather than the transition distribution. Section 4 is an aside on the converse: applying Markov chain results to bivariate i.i.d. models. Sections 5 to 7 discuss models more suited to the traditional approach: autoregression, conditional constraints, MCMC. Section 8 is on plugging kernel estimators into smooth functionals and into empirical estimators. Sections 9 and 10 briefly mention extensions of the traditional approach to continuous-time processes and to random fields.

1. The traditional approach. In order to illustrate the power of BK's approach, we compare it with the traditional approach, which we recall first. For a review see Wefelmeyer (1999). Let $X^{(n)} = (X_1, \dots, X_n)$ be observations from a stationary Markov chain with values in some state space E. (Here the state space may be arbitrary.) The natural parameter is the transition distribution, call it $q(x, dy)$. Let $\pi(dx)$, $b(dx, dy)$, and $P^{(n)}$ denote the laws of X_1, (X_1, X_2), and $X^{(n)}$, respectively. We have $b(dx, dy) = \pi(dx)q(x, dy)$. Consider (Hellinger differentiable) perturbations $q_{nh}(x, dy) \doteq q(x, dy)(1 + n^{-1/2}h(x, y))$ of q. For q_{nh} to be a transition distribution, we must restrict h to $\mathcal{H}_0 = \{h \in L_2(b) : q_x h = 0\}$, where $q_x h = \int h(x, y)q(x, dy)$ denotes conditional expectation. The space $\mathcal{H}_0$ is the tangent space of the full nonparametric model. It is well known that we have local asymptotic normality at q,

$$\log \frac{dP^{(n)}_{nh}}{dP^{(n)}} = n^{-1/2} \sum_{i=1}^{n-1} h(X_i, X_{i+1}) - \frac{1}{2} \int h^2 \, db + o_p(1). \tag{1}$$

(We do not need the stronger form of local asymptotic normality used in BK, with perturbations involving factors t_n converging to some t.)

Consider now a submodel. It is given by a subset of transition distributions. Its *tangent space* at q is a subset of $\mathcal{H}_0$, say $\mathcal{H}_0^s$, which we take to be linear. Consider a

[1]Supported in part by NSERC, Canada
[2]Supported in part by NSF Grant DMS 0072174

real-valued functional $\vartheta(q)$ on the submodel. Assume it is *differentiable* at q with respect to the inner product induced by local asymptotic normality, with *gradient* $g \in \mathcal{H}_0$,

$$n^{1/2}(\vartheta(q_{nh}) - \vartheta(q)) \to \int hg\, db \quad \text{for all } h \in \mathcal{H}_0^s. \tag{2}$$

The *canonical* gradient is the projection g_s of g onto (the closure of) $\mathcal{H}_0^s$.

An estimator $\hat{\vartheta}$ of $\vartheta(q)$ is *regular* at q with *limit* L if L is a random variable such that

$$n^{1/2}(\hat{\vartheta} - \vartheta(q_{nh})) \Rightarrow L \quad \text{under } P_{nh}^{(n)} \quad \text{for all } h \in \mathcal{H}_0^s. \tag{3}$$

The convolution theorem says that $L = \left(\int g_s^2\, db\right)^{1/2} \cdot N + M$ in distribution, with N standard Gaussian and M independent of N. This justifies calling a regular estimator $\hat{\vartheta}$ *efficient* for $\vartheta(q)$ if

$$n^{1/2}(\hat{\vartheta} - \vartheta(q)) \Rightarrow \left(\int g_s^2\, db\right)^{1/2} \cdot N \quad \text{under } P^{(n)}.$$

An estimator $\hat{\vartheta}$ is *asymptotically linear* at q with *influence function* f if $f \in \mathcal{H}_0$ and

$$n^{1/2}(\hat{\vartheta} - \vartheta(q)) = n^{-1/2} \sum_{i=1}^{n-1} f(X_i, X_{i+1}) + o_p(1). \tag{4}$$

It is well known that an asymptotically linear estimator is regular if and only if its influence function is a gradient, and that a (regular) estimator is efficient if and only if it is asymptotically linear with influence function equal to the canonical gradient.

Example 1. Let us illustrate the traditional approach with a simple example, estimating a linear functional $\vartheta(q) = \int k\, db$, with $k \in L_2(b)$, in the full nonparametric model, with tangent space $\mathcal{H}_0$. For $h \in \mathcal{H}_0$ let $b_{nh}(dx, dy) = \pi_{nh}(dx) q_{nh}(x, dy)$. By a perturbation expansion (see e.g. Kartashov (1985), (1996)) we have

$$n^{1/2} \left(\int w\, db_{nh} - \int w\, db \right) \to \int h \cdot Tw\, db \quad \text{for all } w \in L_2(b), \tag{5}$$

where the operator $T : L_2(b) \to \mathcal{H}_0$ is $Tw(x,y) = w(x,y) - q_x w + \sum_{j=1}^{\infty} (q_y^j w - q_x^{j+1} w)$. This operator is a projection, $T = T^2$. It can also be written, as in BK, $Tw(x,y) = \overline{w}(x,y) - \sum_{j=1}^{\infty} q_x^j \overline{w} + \sum_{j=1}^{\infty} q_y^j \overline{w}$, where $\overline{w}(x,y) = w(x,y) - \int w\, db$ denotes centering. Relation (5), applied to $w = k$, says that the functional $\int k\, db$ has canonical gradient Tk in the sense of (2).

By a martingale approximation we have, for $w \in L_2(b)$,

$$n^{-1/2} \sum_{i=1}^{n-1} (\overline{w}(X_i, X_{i+1}) - Tw(X_i, X_{i+1})) = o_p(1). \tag{6}$$

(Relation (6) is called martingale approximation because $Tw(X_i, X_{i+1})$ are martingale increments. This approximation has been found independently by many authors, e.g. Gordin (1969), Maigret (1978), Dürr and Goldstein (1986) and Greenwood and Wefelmeyer (1995). See also Bradley (1988a,b) and Meyn and Tweedie ((1993), Section 17.4). BK refer to Bickel (1993) and Künsch (1984).)

By the martingale approximation (6), applied to $w = k$, the empirical estimator $\hat{\vartheta} = \int k \, d\hat{b} = \frac{1}{n-1} \sum_{i=1}^{n-1} k(X_i, X_{i+1})$ satisfies

$$n^{1/2} \left(\int k \, d\hat{b} - \int k \, db \right) = n^{-1/2} \sum_{i=1}^{n-1} Tk(X_i, X_{i+1}) + o_p(1).$$

Hence the influence function, in the sense of (4), of the empirical estimator is Tk, the canonical gradient, and the estimator is regular and efficient by the two characterizations above.

2. An equivalence relation. The first point of BK on information calculus for Markov chains can be phrased as follows. Call $w, z \in L_2(b)$ *equivalent* if $Tw = Tz$. Then by the martingale approximation (6), $n^{-1/2} \sum_{i=1}^{n-1} (\overline{w}(X_i, X_{i+1}) - \overline{z}(X_i, X_{i+1})) = o_p(1)$. Now parametrize locally with equivalence classes in $L_2(b)$ rather than their representatives in $\mathcal{H}_0$. Then for $h = Tw$, local asymptotic normality (1) can be written

$$\log \frac{dP_{nh}^{(n)}}{dP^{(n)}} = n^{-1/2} \sum_{i=1}^{n-1} \overline{w}(X_i, X_{i+1}) - \frac{1}{2} \int (Tw)^2 \, db + o_p(1).$$

This is local asymptotic normality in the sense of Definition 1 of BK. Extend differentiability (2) of $\vartheta(q)$ correspondingly, calling m *gradient* of $\vartheta(q)$ if $m \in L_2(b)$ and

$$n^{1/2}(\vartheta(q_{nh}) - \vartheta(q)) \to \int h \cdot Tm \, db \quad \text{for all } h \in \mathcal{H}_0^s. \tag{7}$$

Any gradient m with Tm in (the closure of) $\mathcal{H}_0^s$ may then be called *canonical*. Extend asymptotic linearity (4) of $\hat{\vartheta}$, calling m *influence function* of $\hat{\vartheta}$ if $m \in L_2(b)$ and

$$n^{1/2}(\hat{\vartheta} - \vartheta(q)) = n^{-1/2} \sum_{i=1}^{n-1} \overline{m}(X_i, X_{i+1}) + o_p(1). \tag{8}$$

Then appropriate versions of the characterizations of regular and efficient estimators continue to hold.

Example 2. BK apply their approach in particular to the simple example above, estimating $\vartheta(q) = \int k \, db$ in the full nonparametric model, with tangent space $\mathcal{H}_0$. Write

$$n^{1/2} \left(\int k \, d\hat{b} - \int k \, db \right) = n^{-1/2} \sum_{i=1}^{n-1} \overline{k}(X_i, X_{i+1}) + o_p(1).$$

We have $Tk \in \mathcal{H}_0$. Hence k is a canonical gradient in the extended sense (7), and efficiency of the empirical estimator follows.

This proof is much shorter than the traditional one. Note, however, that the martingale approximation is also used here, namely for extending influence functions to equivalence classes. Efficiency of the empirical estimator was shown first by Penev (1990, 1991); he uses the perturbation expansion but circumvents the martingale approximation. Greenwood and Wefelmeyer (1995) show that the perturbation expansion follows from the martingale approximation.

3. From bivariate models to Markov chains. The second point of BK in their information calculus applied to Markov chains is that canonical gradients can be obtained as in bivariate models, with i.i.d. observations (X_i, Y_i). This is extremely useful, especially for models and functionals which are more easily described in terms of the joint law b than of the transition distribution q.

Parametrize the Markov chain by the law b of (X_1, X_2) rather than by q. Then b must have equal marginals π: $\int v(x)\, b(dx, dy) = \int v(y)\, b(dx, dy)$ for all $v \in L_2(\pi)$. Consider (Hellinger differentiable) perturbations $b_{nw}(dx, dy) \doteq b(dx, dy)(1 + n^{-1/2} w(x, y))$. For b_{nw} to be a probability measure, we must have $\int w\, db = 0$. Since b_{nw} must also have equal marginals, we get

$$\int v(x) w(x, y)\, b(dx, dy) = \int v(y) w(x, y)\, b(dx, dy) \quad \text{for all } v \in L_2(\pi).$$

Hence the tangent space at b, say $\mathcal{H}$, is defined by having the following orthogonal complement in $L_2(b)$: $\mathcal{H}^\perp = \{v(x) - v(y) : v \in L_2(\pi)\}$.

Locally, the parameters b and q are related as follows. To go from b to q, factor b_{nw} as $b_{nw}(dx, dy) = \pi_{nw}(dx) q_{nw}(x, dy)$. Then π_{nw} is perturbed as

$$\pi_{nw}(dx) = b_{nw}(dx, E) \doteq \pi(dx)(1 + n^{-1/2} q_x w). \tag{9}$$

Hence $q_{nw}(x, dy) \doteq q(x, dy)(1 + n^{-1/2} w_0(x, y))$, where $w_0(x, y) = w(x, y) - q_x w$ denotes conditional centering. In particular, $\mathcal{H}_0 = \{w_0 : w \in \mathcal{H}\}$. To go from q to b, start from a perturbation $q_{nh}(x, dy) \doteq q(x, dy)(1 + n^{-1/2} h(x, y))$ with $h \in \mathcal{H}_0$. Write $b_{nh}(dx, dy) = \pi_{nh}(dx) q_{nh}(x, dy)$. For $w \in L_2(b)$ and $h \in \mathcal{H}_0$ write

$$\int h \cdot Tw\, db = \int Sh \cdot w\, db, \tag{10}$$

with an operator $S : \mathcal{H}_0 \to \mathcal{H}$ which we may call the adjoint of T. We do not need the explicit form of S; see Greenwood and Wefelmeyer (1999) for it. From (10) and the perturbation expansion (5) we obtain the perturbation

$$b_{nh}(dx, dy) \doteq b(dx, dy)(1 + n^{-1/2} Sh(x, y)). \tag{11}$$

In particular, $\mathcal{H} = \{Sh : h \in \mathcal{H}_0\}$. An analogous local parameter change, between densities and hazard functions, is described in Ritov and Wellner (1988).

Now consider a submodel described by some set of joint laws of (X_1, X_2). Its tangent space at b is a subset of $\mathcal{H}$, say $\mathcal{H}^s$, which we take to be linear. Consider a real-valued functional $\vartheta(b)$ on the submodel. Call it *differentiable* with *gradient* m if $m \in L_2(b)$ and

$$n^{1/2}(\vartheta(b_{nw}) - \vartheta(b)) \to \int wm\, db \quad \text{for all } w \in \mathcal{H}^s. \tag{12}$$

The *canonical* gradient is the projection m_s of m onto (the closure of) $\mathcal{H}^s$. Writing $w = Sh$ and using (10), we can characterize m_s as the function in (the closure of) $\mathcal{H}^s$ which fulfills

$$0 = \int Sh \cdot (m - m_s)\, db = \int h \cdot (Tm - Tm_s)\, db \quad \text{for all } h \in \mathcal{H}_0.$$

This means that Tm_s is the canonical gradient, in the traditional sense (2), in the Markov chain model. This is essentially Theorem 1 in Greenwood and Wefelmeyer (1999a). BK's second point is the following interpretation of this result. Suppose we have i.i.d. observations (X_i, Y_i). Consider a bivariate model of distributions $b(dx, dy)$, with equal marginals, and a real-valued differentiable functional $\vartheta(b)$ on this model. Calculate its canonical gradient m_s in the sense of (12). The canonical gradient of the corresponding Markov chain model is then Tm_s. Hence, using BK's first point, any function $z \in L_2(b)$ with $Tz = Tm_s$, in particular m_s itself, is a canonical gradient and efficient influence function in their extended sense.

Example 3. BK illustrate their second point with their Example 3b, a Markov chain model with *known* marginal distribution π. Then we must have $\pi_{nw}(dx) = \pi(dx)$ and hence $q_x w = 0$ by (9), and similarly $\pi(dy) = \pi_{nw}(dy) \doteq \pi(dy)(1 + n^{-1/2} q_y^- w)$. Here q^- is the transition distribution of the *reversed* chain, defined by $\pi(dx)q(x, dy) = \pi(dy)q^-(y, dx)$, and $q_y^- w = \int q^-(y, dx) w(x, y)$ is the conditional expectation under q^-, acting on the *first* argument of w. Hence the tangent space is $\mathcal{H}^s = \{w \in L_2^0(b) : qw = q^- w = 0\}$. Following BK, for $w \in L_2^0(b)$ we can write $qw = q^- w = 0$ as $\int (u(x) + v(y)) w(x, y)\, b(dx, dy) = 0$ for all $u, v \in L_2(\pi)$. In words: w is orthogonal to functions of the form $u(x) + v(y)$. Now let $\vartheta(b)$ be differentiable, in the sense (12) of the bivariate model, with gradient $m \in L_2(b)$, say. As BK note, the *canonical* gradient in this sense can be obtained from Bickel, Klaassen, Ritov and Wellner ((1998), p. 440) as $m_s = m - \mathrm{ACE}(m)$, where $\mathrm{ACE}(m)$ is the projection of m onto the space of functions $u(x) + v(y)$. For $w \in \mathcal{H}^s$ we have $qw = 0$, i.e. $w(x, y) = w(x, y) - q_x w = w_0(x, y)$. The model is therefore degenerate: $\mathcal{H}^s = \mathcal{H}_0^s$. Hence m_s is also the traditional canonical gradient and efficient influence function in the sense (2).

Example 4. We agree that the Markov chain model with known marginals is possibly unrealistic. It is, however, not, as BK suggest, the model considered by Kessler, Schick and Wefelmeyer (2001). The latter assume not that the marginal is fixed but that it belongs to some parametric family π_ϑ, with ϑ one-dimensional, and they construct an

efficient estimator for ϑ. The justification for such models comes from financial time series in which the marginal can be modeled more convincingly than the dynamics, especially when one has discrete observations from a continuous-time process. The efficient estimator is a complicated one-step improvement. We will consider elsewhere the possibility of finding a conceptually simpler estimator using BK's approach.

Example 5. Here is another application of BK's approach. Greenwood and Wefelmeyer (1999a) consider the model of all *reversible* Markov chains. This means that b is symmetric, $b(dx, dy) = b(dy, dx)$, or equivalently, $q = q^-$. They prove that the symmetrized empirical estimator

$$\hat{\vartheta}_s = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (k(X_i, X_{i+1}) + k(X_{i+1}, X_i))$$

is efficient for $\int k\, db$. The proof is also based on their Theorem 1 used above. For a more elegant version of this proof we follow BK and parametrize with b. The tangent space of the bivariate model is $\mathcal{H}^s = \{w \in L_2^0(b) : w(x, y) = w(y, x)\}$. Consider a real-valued functional $\vartheta(b)$ which is differentiable, in the sense (12), with gradient $m \in L_2^0(b)$. The *canonical* gradient in the bivariate model is the symmetrized $m_s(x, y) = \frac{1}{2}(m(x, y) + m(y, x))$. Hence the canonical gradient in the Markov chain model is Tm_s. Hence, by BK's first point, m_s is also an efficient influence function in the Markov chain model. This proves that if $\hat{\vartheta}$ is asymptotically linear with influence function m in the Markov chain model, then its symmetrization $\hat{\vartheta}_s = \frac{1}{2}(\hat{\vartheta}(X_1, \ldots, X_n) + \hat{\vartheta}(X_n, \ldots, X_1))$ is regular and efficient in the model of all reversible Markov chains. In particular, the symmetrized *empirical* estimator is efficient.

Example 6. Müller, Schick and Wefelmeyer (2001b) consider the nonparametric Markov chain model with linear constraint $\int z\, db = 0$ for some d-dimensional vector $z \in L_2(b)^d$. They construct efficient estimators for linear functionals $\int k\, db$, following the traditional approach and Levit (1975), who considers the i.i.d. case. The canonical gradient is $Tk - c_*^\top Tz$ with $c_* = \left(\int Tz \cdot Tz^\top\, db\right)^{-1} \int Tz \cdot Tk\, db$. Let us derive this result using BK's approach. Parametrize by b. We have $n^{1/2} \left(\int z\, db_{nw} - \int z\, db\right) \to \int zw\, db$. Because of the constraints $\int z\, db = \int z\, db_{nw} = 0$ we must have $\int zw\, db = 0$. Hence the tangent space of the corresponding bivariate model is $\mathcal{H}^s = \{w \in \mathcal{H} : \int zw\, db = 0\}$. By Example 2, k is a gradient of $\int k\, db$ in the extended sense (7). Write $w_\mathcal{H}$ for the projection of a function $w \in L_2(b)$ onto $\mathcal{H}$. In the bivariate model, because of the constraint $\int z\, db = 0$, all functions $k - a^\top z$ with $a \in \mathbf{R}$ are gradients, and hence all functions $k_\mathcal{H} - a^\top z_\mathcal{H}$ are gradients in $\mathcal{H}$. The canonical gradient is the projection of any of these gradients onto $\mathcal{H}^s$. It must minimize $\int (k_\mathcal{H} - a^\top z_\mathcal{H})^2\, db$ in a. The minimizing value of a is $a_* = \left(\int z_\mathcal{H} z_\mathcal{H}^\top\, db\right)^{-1} \int z_\mathcal{H} k_\mathcal{H}\, db$. By BK's second point, $k_\mathcal{H} - a_*^\top z_\mathcal{H}$ is also a canonical gradient in the constrained Markov chain model, in their extended sense (7).

Of course, $k_\mathcal{H} - a_*^\top z_\mathcal{H}$ must be equivalent to the traditional canonical gradient $Tk - c_*^\top Tz$ in the bivariate model. This follows from two observations.

1. If $w \in L_2(b)$ and $w_{\mathcal{H}}$ is its projection onto $\mathcal{H}$, then $w - w_{\mathcal{H}}$ is in $\mathcal{H}^{\perp}$, i.e., of the form $v(x) - v(y)$. Such functions are annihilated by T. Hence w and $w_{\mathcal{H}}$ are equivalent: $Tw = Tw_{\mathcal{H}}$. In particular, $k_{\mathcal{H}} - a_*^{\top} z_{\mathcal{H}}$ is equivalent to $Tk - a_*^{\top} Tz$.

2. The operator ST is a projection onto $\mathcal{H}$. Hence we obtain, using (10),

$$\int Tw \cdot Tm \, db = \int w \cdot STm \, db = \int wm \, db \quad \text{for all } w, m \in \mathcal{H}.$$

In particular, $a_* = c_*$.

An efficient estimator for $\int k \, db$ under the constraint $\int z \, db = 0$ is the improved empirical estimator

$$\frac{1}{n-1} \sum_{i=1}^{n-1} (k(X_i, X_{i+1}) - \hat{a}_*^{\top} z(X_i, X_{i+1})).$$

It requires a consistent estimator $\hat{a}_*$ for a_*. Such an estimator is constructed in Müller, Schick and Wefelmeyer (2001b). It is based on an explicit representation of a_*. Calculating a_* requires calculating the projections of z and k onto $\mathcal{H}$. Example 7 shows how projections $w_{\mathcal{H}}$ of functions $w \in L_2(b)$ onto $\mathcal{H}$ are obtained, via the traditional approach, as $w_{\mathcal{H}} = STw$. One checks that by (10) this gives again $a_* = c_*$.

This example shows that even if the model and functional of interest are in terms of the joint law b rather than the transition distribution q, the traditional approach is not necessarily more awkward than the approach via the bivariate model. One reason is the following. The traditional approach parametrizes by q and uses an unpleasant local parameter space $\mathcal{H}_0$, equipped however with the natural norm $\int w^2 \, db$. If we introduce equivalence classes as suggested in BK's first point, then we end up with a simple local parameter space $L_2(b)$, but now equipped with the unpleasant semi-norm $\int (Tw)^2 \, db$. On the other hand, if we parametrize by b as suggested in BK's second point, then we end up with the natural norm but with an unpleasant local parameter space $\mathcal{H}$.

4. From Markov chains to bivariate models. We have seen in Section 3 how canonical gradients in Markov chain models can be obtained from canonical gradients in bivariate models. The converse is of course also possible and, more surprisingly, sometimes useful.

Consider a Markov chain model described by some set of transition distributions. Its tangent space at q is a subset $\mathcal{H}_0^s$ of $\mathcal{H}_0$, taken to be linear. Let $\vartheta(q)$ be a real-valued functional which is differentiable, in the (traditional) sense (2), with canonical gradient $g_s \in \mathcal{H}_0^s$. Set $h = Tw$ and use (10) to rewrite differentiability (2) as

$$n^{1/2}(\vartheta(q_{nh}) - \vartheta(q)) \to \int Tw \cdot g_s \, db = \int w \cdot Sg_s \, db \quad \text{for all } w \in \mathcal{H}^s.$$

This is differentiability in the sense (12) of the bivariate model. Hence Sg_s is the canonical gradient of $\vartheta(q)$, viewed as functional on the bivariate model.

Example 7. This is already useful in the simplest example, estimating the linear functional $\vartheta(q) = \int k\,db$, with $k \in L_2(b)$, in the full nonparametric Markov chain model. Its (canonical) gradient is $g = Tk$. The corresponding bivariate model is the model with equal marginals. It follows that $Sg = STk$ is the canonical gradient in this model. The explicit form of ST can be obtained from results for Markov chain models, see Greenwood and Wefelmeyer (1999a). An efficient estimator in the bivariate i.i.d. model with equal marginals is constructed in Peng and Schick (2001). It does not use the explicit form of the canonical gradient.

5. Regression and autoregression. An important class of Markov chain models are autoregressive models $X_{i+1} = r(X_i) + \varepsilon_{i+1}$, where the innovations ε_i are i.i.d. with mean zero and finite variance σ^2 and have an absolutely continuous and positive density f with finite Fisher information $J = \int \ell^2\,dF$ for location, where $\ell = -f'/f$ and F is the distribution function of f. For convenience we consider only first-order autoregression. For the model to be ergodic, the autoregression function r must satisfy some growth conditions; see e.g. Bhattacharya and Lee (1995). BK consider the nonparametric model, with r unknown. Submodels are the linear model, with $r(x) = \vartheta x$, and nonlinear models with parametric families $r_\vartheta(x)$ of autoregression functions. Here it suggests itself to follow the traditional approach and describe the model by its transition distribution $q(x, dy) = f(y - r(x))\,dy$.

The information calculus of Section 3 would suggest looking at the bivariate i.i.d. model described by the joint law $b(dx, dy) = \pi(dx)q(x, dy)$ of (X_1, X_2). Perturbation of q would, however, result in a complicated perturbation of π, see (11), and in a complicated tangent space of the bivariate model.

Nevertheless, it pays to look at an i.i.d. model analogous to the Markov chain model, namely regression $Y_i = r(X_i) + \varepsilon_i$, with ε_i as before, and i.i.d. covariates X_i, independent of the ε_i, with known law $c(dx)$, say. The joint law of (X_1, Y_1) is $c(dx)f(y - r(x))\,dy$. Tangent spaces and gradients for autoregression are therefore the same as for regression. Schick (1993) considers functionals of (c, r); for extensions to heteroscedastic regression see Schick (1994).

Following the traditional approach to autoregression, see Koul and Schick (1997), consider (Hellinger differentiable) perturbations $f_{nv} \doteq f(1 + n^{-1/2}v)$. Since the innovations are assumed to have mean zero, the local parameters v must be in the orthogonal complement V in $L_2(F)$ of the polynomials of degree at most one,

$$V = \{v \in L_2(F) : \int v(\varepsilon)\,dF(\varepsilon) = \int \varepsilon v(\varepsilon)\,dF(\varepsilon) = 0\}.$$

The model also specifies a family of autoregression functions. Consider (π-square-differentiable) perturbations $r_{nu} \doteq r + n^{-1/2}u$. The model restricts u to some subset of $L_2(\pi)$, say U, which we take to be (closed and) linear. The transition density determined by f_{nv} and r_{nu} is $f_{nv}(y - r_{nu}(x)) \doteq f(\varepsilon)\big(1 + n^{-1/2}(u(x)\ell(\varepsilon) + v(\varepsilon))\big)$. Hence the tangent space of the autoregressive model is $\mathcal{H}_0(U) = \{u(x)\ell(\varepsilon) + v(\varepsilon) : u \in U, v \in V\}$.

The tangent space is the sum of the tangent spaces $\{u(x)\ell(\varepsilon) : u \in U\}$ for known f, and $\{v(\varepsilon) : v \in V\}$ for known r. It is well known that one can estimate (all smooth functionals of) f and r adaptively with respect to each other if and only if these two spaces are orthogonal.

Example 8. Schick and Wefelmeyer (2001b) obtain efficient estimators for $\int a\,dF$ when the autoregression functions are restricted to a parametric family r_ϑ. For simplicity, we take ϑ one-dimensional here. Then U is the linear span $[\dot r_\vartheta]$ of the derivative of r_ϑ with respect to ϑ, and the tangent space is $\mathcal{H}_0([\dot r_\vartheta]) = \{t\dot r_\vartheta(x)\ell(\varepsilon) + v(\varepsilon) : t \in \mathbf{R}, v \in V\}$. Unless $\int \dot r_\vartheta\,d\pi = 0$, the tangent space is not an orthogonal sum, and f and r cannot be estimated adaptively with respect to each other. A natural estimator of $\int a\,dF$ is the empirical estimator $\frac{1}{n-1}\sum_{i=1}^{n-1} a(\hat\varepsilon_{i+1})$ based on estimated innovations $\hat\varepsilon_{i+1} = X_{i+1} - r_{\hat\vartheta}(X_i)$. It can be improved using that the innovations have mean zero,

$$\hat A = \frac{1}{n-1}\sum_{i=1}^{n-1}(a(\hat\varepsilon_{i+1}) - \hat c\hat\varepsilon_{i+1}), \tag{13}$$

with $\hat c$ a consistent estimator for the optimal constant

$$c = \sigma^{-2}\int \varepsilon a(\varepsilon)\,dF(\varepsilon). \tag{14}$$

An obvious choice is $\hat c = \sum_{i=1}^{n-1}\hat\varepsilon_{i+1}a(\hat\varepsilon_{i+1})\big/\sum_{i=1}^{n-1}\hat\varepsilon_{i+1}^2$. The influence function of $\hat A$ requires some notation, and we do not give it here. In the non-adaptive situation, with $\int \dot r_\vartheta\,d\pi$ not zero, for $\hat A$ to be efficient we must estimate $\varepsilon_{i+1} = X_{i+1} - r_\vartheta(X_i)$ using an *efficient* estimator for ϑ.

Plug-in of finite-dimensional estimators in not necessarily adaptive situations is studied in Klaassen and Putter (1997, 2000) for i.i.d. models, and more generally in Müller, Schick and Wefelmeyer (2001a).

Example 9. In their Example 3a, BK consider estimating $\int a\,dF$ in the nonparametric autoregressive model, with r unknown except for mean zero. Then $U = L_2(\pi)$, and the tangent space is $\mathcal{H}_0(L_2(\pi)) = \{u(x)\ell(\varepsilon) + v(\varepsilon) : u \in L_2(\pi), v \in V\}$. This is not an orthogonal sum. Hence f and r cannot be estimated adaptively with respect to each other. (BK state that the tangent space equals that with Gaussian innovation distribution with known variance, their (3.30), and later that it contains all functions $v(\varepsilon)$ with $v \in L_2(\pi)$. These statements are not consistent with each other and with the tangent space obtained here.) The canonical gradient for $\int a\,dF$ is the same as in the corresponding regression model, Müller, Schick and Wefelmeyer (2001c), namely $\overline{a}(\varepsilon) - \int a\ell\,dF\cdot\varepsilon$. One can show that the empirical estimator $\frac{1}{n-1}\sum_{i=1}^{n-1} a(\hat\varepsilon_{i+1})$ based on estimated innovations $\hat\varepsilon_{i+1} = X_{i+1} - \hat r(X_i)$ has this influence function. To check that this function is indeed in the tangent space $\mathcal{H}_0(L_2(\pi))$, rewrite it as

$$\overline{a}(\varepsilon) - \int a\ell\,dF\cdot\varepsilon = -\sigma^2\int a\ell_V\,dF\cdot\ell(\varepsilon) + a_V(\varepsilon) + \sigma^2\int a\ell_V\,dF\cdot\ell_V(\varepsilon),$$

where a_V and ℓ_V are the projections of a and ℓ onto V, $a_V(\varepsilon) = \overline{a}(\varepsilon) - c\varepsilon$, $\ell_V(\varepsilon) = \ell(\varepsilon) - \sigma^{-2}\varepsilon$. We note that in this non-adaptive model, the canonical gradient for *known* regression function r is indeed different: It is just the projection a_V of a onto V, and an efficient estimator is the *improved* empirical estimator $\frac{1}{n-1}\sum_{i=1}^{n-1}(a(\varepsilon_{i+1}) - \hat{c}\varepsilon_{i+1})$ based on *true* innovations. Compare also Example 8 on parametric autoregression functions r_ϑ.

These results are not consistent with the statements of BK that the empirical estimators with true and estimated innovations are asymptotically equivalent, that their influence function is $a(\varepsilon)$, and that this function is in the tangent space, which would imply that $\frac{1}{n-1}\sum_{i=1}^{n-1} a(X_{i+1} - \hat{r}(X_i))$ is adaptive with respect to r.

Example 10. BK ascribe their statements about $\frac{1}{n-1}\sum_{i=1}^{n-1} a(X_{i+1} - \hat{r}(X_i))$ in nonparametric autoregression to Wefelmeyer (1994). But the latter treats only *linear* autoregression $X_{i+1} = \vartheta X_i + \varepsilon_{i+1}$, and proves that the *improved* empirical estimator $\hat{A}$, now with innovations estimated by $\hat{\varepsilon}_{i+1} = X_{i+1} - \hat{\vartheta}X_i$, is efficient. Linear autoregression is a special case of the nonlinear model above, with $r_\vartheta(x) = \vartheta x$ and $\dot{r}_\vartheta(x) = x$. The tangent space is therefore $\mathcal{H}_0^s = \{tx\ell(\varepsilon) + v(\varepsilon) : t \in \mathbf{R}, v \in V\}$. Since the innovations have mean zero, so has the stationary law π. This implies that the tangent space is an orthogonal sum, and ϑ and f can be estimated adaptively with respect to each other. In particular, $\hat{A}$ is efficient for $\int a\,dF$ even when an inefficient estimator $\hat{\vartheta}$ is used in the estimated innovations $\hat{\varepsilon}_{i+1} = X_{i+1} - \hat{\vartheta}X_i$.

Example 11. Another adaptive example is nonparametric autoregression with innovations that are *symmetric about zero*. The tangent space is $\mathcal{H}_0^s = \{u(x)\ell(\varepsilon) + v(\varepsilon) : u \in L_2(\pi), v \in L_2(F)$ symmetric about zero$\}$. Here $\ell(\varepsilon) = -\ell(-\varepsilon)$. Hence $\int v\ell\,dF = 0$ for all v that are symmetric about zero, and the tangent space is an orthogonal sum. Koshevnik (1996) shows that the symmetrized empirical distribution function based on estimated innovations is efficient.

Example 12. Kwon (2000) and BK also consider estimating $\int r(x)\lambda(x)\,dx$ in the nonparametric autoregression model with mean zero innovations. Here λ is known and has compact support. They suggest that an efficient estimator is obtained by plugging in a suitable (kernel) estimator $\hat{r}$ for r. Kwon (2000) shows that the estimator $\int \hat{r}(x)\lambda(x)\,dx$ has influence function $\varepsilon\lambda(x)/f(x)$. From Schick ((1993), (3.5)), the canonical gradient of $\int r(x)\lambda(x)\,dx$ is obtained as

$$\left(\frac{\lambda(x)}{f(x)} - \int \lambda(y)\,dy\right)\frac{\ell(\varepsilon)}{J} + \int \lambda(y)\,dy \cdot \varepsilon,$$

with J the Fisher information for location of the innovation distribution. This is the influence function of BK's estimator only if the innovation distribution is Gaussian, so their estimator is efficient only if the true innovation distribution happens to be Gaussian.

The traditional approach has also been used in more complicated autoregressive models. For example, Schick (1999a) treats the semiparametric model $X_{i+1} = \vartheta X_i + r(X_{i-1}) + \varepsilon_{i+1}$. Maercker (1997) and Schick (2001) treat the heteroscedastic autoregressive model $X_{i+1} = \vartheta X_i + s(X_i)\varepsilon_{i+1}$ with symmetric errors, while Schick (1999b) considers it with arbitrary errors. Efficient estimation in invertible linear processes is treated in Schick and Wefelmeyer (2001c).

6. Conditional constraints. Another class of submodels described through transition distributions rather than joint laws are models with constraints $E(v_\vartheta(X_1, X_2)|X_1) = 0$ for some d-dimensional vector $v_\vartheta \in L_2(b)$. These comprise quasi-likelihood models, with parametric models for conditional mean and variance of the Markov chain:

$$\begin{aligned} E(X_2|X_1) &= r_\vartheta(X_1), \\ E((X_2 - r_\vartheta(X_1))^2|X_1) &= s^2_\vartheta(X_1). \end{aligned}$$

Here $v_\vartheta(x, y)$ has components $y - r_\vartheta(x)$ and $(y - r_\vartheta(x))^2 - s^2_\vartheta(x)$. The quasi-maximum-likelihood estimator solves an estimating equation of the form

$$\sum_{i=1}^{n-1} w_\vartheta(X_i, X_{i+1})\big(X_{i+1} - r_\vartheta(X_i)\big) = 0,$$

with weights w_ϑ chosen to minimize the asymptotic variance. It does not use the information in the specification of the conditional variance and is not efficient. Efficient estimating equations are constructed in Wefelmeyer (1996). For similar regression models with i.i.d. observations, quite different efficient estimators are introduced in Li (2000) and (2001). Efficient estimation of invariant laws in such models is discussed in Schick and Wefelmeyer (1999).

7. MCMC. A third class of submodels described by transition distributions are Monte Carlo Markov chains. Here one starts with a distribution $\pi(dx)$ which is in principle known, and constructs a transition distribution $q(x, dy)$ with π as invariant law. Then one runs the corresponding Markov chain and approximates, e.g., $\int a(x)\pi(dx)$ by the empirical estimator $\frac{1}{n}\sum_{i=1}^n a(X_i)$. Greenwood, McKeague and Wefelmeyer (1998) calculate the information in the knowledge that a Gibbs sampler was used. A review is Greenwood and Wefelmeyer (2001).

8. Plug-in estimators. As BK point out, $n^{1/2}$-consistent and even efficient estimators can often be obtained by plugging density estimators or regression function estimators into smooth functionals or into "empirical estimators" involving such functions. BK's estimators for $\int r(x)\lambda(x)\,dx$ and $\int a\,dF$ in nonparametric autoregression are examples of plug-in into a smooth functional and into an empirical estimator.

For i.i.d. observations with density f, smooth functionals of f can be estimated efficiently by plugging in (undersmoothed) kernel estimators; see Abramson and Goldstein (1991), Goldstein and Messer (1992) and Goldstein and Khas'minskii (1995).

For expectations of functions of more than two arguments, e.g. $E\psi(X_1, X_2, X_3)$, the empirical estimator based on Markov chain observations is not efficient in the nonparametric Markov chain model. Writing $E\psi(X_1, X_2, X_3) = \int \psi(x, y, z)\, b(dx, dy)q(y, dz)$, one sees that for discrete state space a better estimator is obtained by replacing b and q by their empirical estimators. For general state space, Schick and Wefelmeyer (2001a) construct a complicated efficient estimator as one-step improvement of the empirical estimator. Bickel (1993) has suggested a conceptually simpler estimator, using the empirical estimator for b as before, and plugging in a nonparametric estimator $\hat{q}$ for the transition density. Kwon (2000) treats a modification of this idea, writing the density of the joint law of (X_1, X_2, X_3) as $p(x, y)p(y, z)/g(y)$ with g and p the densities of X_1 and (X_1, X_2), respectively, and replacing these densities by kernel estimators.

Example 13. Here is another application of plug-in. For moving average processes $X_{i+1} = \varepsilon_{i+1} - \vartheta\varepsilon_i$, the density $g(x)$ of X_{i+1} can be written as convolution of the density f of ε_{i+1} and the density of $\vartheta\varepsilon_i$, i.e., $g(x) = \int f(x+\vartheta y)f(y)dy$. Saavedra and Cao (1999) and (2000) propose plugging in (undersmoothed) kernel estimators $\hat{f}(z) = \frac{1}{n}\sum_{i=1}^n K_c(z-\hat{\varepsilon}_i)$, where $K_c(u) = K(u/c)/c$ and $\hat{\varepsilon}_i$ are estimated innovations. They obtain $n^{1/2}$-consistency of their estimator $\int \hat{f}(x+\hat{\vartheta}y)\hat{f}(y)dy$. Schick and Wefelmeyer (2001e) propose the asymptotically equivalent, but simpler, U-statistic

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

and prove that it is efficient. The estimator can be written (approximately) as the plug-in estimator $\frac{1}{n}\sum_{i=1}^n \hat{f}(x + \hat{\vartheta}\hat{\varepsilon}_i)$.

We note that estimators based on U-statistics have many applications in semiparametric inference. For example, U-statistics with *fixed* kernel are used in Schick and Wefelmeyer (2001d) to estimate expectations under the stationary law of invertible linear processes.

9. Continuous-time processes. The traditional approach generalizes immediately to continuous-time processes X_t, $t \geq 0$, observed on an increasing time interval $[0, n]$, say. For counting processes, the intensity plays the role of the transition distribution as natural parameter; diffusion processes $X_t = r(X_t)dt + s(X_t)dB_t$ are parametrized by drift r and diffusion coefficient s. More generally, semimartingales are parametrized by their predictable characteristics; Jacod and Shiryaev (1987) is the standard reference for structure theory and limit theorems. Other types of asymptotics are also possible. For counting processes we may let the intensity increase. For diffusion processes, we may let the diffusion coefficient decrease, see Kutoyants (1994). In survival analysis one usually considers an increasing number of paths; a comprehensive reference including non- and semiparametric efficiency results is Andersen, Borgan, Gill and Keiding (1993).

Efficient plug-in estimators for the stationary density of diffusion processes are obtained in Kutoyants (1997), (1998) and (1999). Empirical estimators are shown to

be efficient in nonparametric Markov step process and semi-Markov process models by Greenwood and Wefelmeyer (1994a) and (1996), and in nonparametric multivariate point process models by Greenwood and Wefelmeyer (1994b). It seems possible to use versions of BK's approach in such models.

10. Random fields. The traditional approach also generalizes to homogeneous random fields on lattices, where the transition distribution is replaced by the local characteristic, the conditional distribution at a site given the rest of the configuration. For random fields with local interactions, Greenwood and Wefelmeyer (1999b) determine which empirical estimators are efficient.

References

Abramson, I. and Goldstein, L. (1991). Efficient nonparametric testing by functional estimation. *J. Theoret. Probab.* **4**, 137–159.

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

Bhattacharya, R. and Lee, C. (1996). On geometric ergodicity of nonlinear autoregressive models. *Statist. Probab. Lett.* **22**, 311–315.

Bickel, P. J. (1993). Estimation in semiparametric models. In *Multivariate Analysis: Future Directions* (C. R. Rao, ed.), 55–73. North-Holland, Amsterdam.

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

Bradley, R. C. (1988a). On a theorem of Gordin. *Stochastics* **24**, 357–392.

Bradley, R. C. (1988b). On some results of M. I. Gordin: A clarification of a misunderstanding. *J. Theoret. Probab.* **1**, 115–119.

Dürr, D. and Goldstein, S. (1986). Remarks on the central limit theorem for weakly dependent random variables. In *Stochastic Processes — Mathematics and Physics* (S. Albeverio, P. Blanchard and L. Streit, eds.), 104–118. Lecture Notes in Mathematics 1158, Springer, Berlin.

Goldstein, L. and Khas'minskii, R. (1995). On efficient estimation of smooth functionals. *Theory Probab. Appl.* **40**, 151–156.

Goldstein, L. and Messer, K. (1992). Optimal plug-in estimators for nonparametric functional estimation. *Ann. Statist.* **20**, 1306–1328.

Gordin, M. I. (1969). The central limit theorem for stationary processes. *Soviet Math. Dokl.*, **10**, 1174–1176.

Greenwood, P. E., McKeague, I. W. and Wefelmeyer, W. (1998). Information bounds for Gibbs samplers. *Ann. Statist.* **26**, 2128–2156.

Greenwood, P. E. and Wefelmeyer, W. (1994a). Nonparametric estimators for Markov step processes. *Stochastic Process. Appl.* **52**, 1–16.

Greenwood, P. E. and Wefelmeyer, W. (1994b). Optimality properties of empirical estimators for multivariate point processes. *J. Multivariate Anal.* **49**, 202–217.

Greenwood, P. E. and Wefelmeyer, W. (1995). Efficiency of empirical estimators for Markov chains. *Ann. Statist.* **23**, 132–143.

Greenwood, P. E. and Wefelmeyer, W. (1996). Empirical estimators for semi-Markov processes. *Math. Methods Statist.* **5**, 299–315.

Greenwood, P. E. and Wefelmeyer, W. (1999a). Reversible Markov chains and optimality of symmetrized empirical estimators. *Bernoulli* **5**, 109–123.

Greenwood, P. E. and Wefelmeyer, W. (1999b). Characterizing efficient empirical estimators for local interaction Gibbs fields. *Stat. Inference Stoch. Process.* **2**, 119–134.

Greenwood, P. E. and Wefelmeyer, W. (2001). Empirical estimators based on MCMC data. To appear in *Handbook of Statistics* **21** (D. Shanbhag, ed.). Elsevier, Amsterdam.

Jacod, J. and Shiryaev, A. N. (1987). *Limit Theorems for Stochastic Processes.* Grundlehren der Mathematischen Wissenschaften 288, Springer, Berlin.

Kartashov, N. V. (1985). Criteria for uniform ergodicity and strong stability of Markov chains with a common phase space. *Theory Probab. Math. Statist.* **30**, 71–89.

Kartashov, N. V. (1996). *Strong Stable Markov Chains.* VSP, Utrecht.

Kessler, M., Schick, A. and Wefelmeyer, W. (2001). The information in the marginal law of a Markov chain. *Bernoulli* **7**, 243–266.

Klaassen, C. A. J. and Putter, H. (1997). Efficient estimation of the error distribution in a semiparametric linear model. In *Contemporary Multivariate Analysis and Its Applications* (K. T. Fang and F. J. Hickernell, eds.). Hong Kong Baptist University.

Klaassen, C. A. J. and Putter, H. (2000). Efficient estimation of Banach parameters in semiparametric models. Technical Report, Department of Mathematics, University of Amsterdam.

Koshevnik, Y. (1996). Semiparametric estimation of a symmetric error distribution from regression models. *Publ. Inst. Statist. Univ. Paris* **40**, 77–91.

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

Künsch, H. R. (1984). Infinitesimal robustness for autoregressive processes. *Ann. Statist.* **12**, 843–863.

Kutoyants, Yu. A. (1994). *Identification of Dynamical Systems with Small Noise.* Mathematics and its Applications 300, Kluwer, Dordrecht.

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 *Statistics and Control of Stochastic Processes* (Y. M. Kabanov, B. L. Rozovskii and A. N. Shiryaev, eds.), 253–274. World Scientific, Singapore.

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

Kwon, J. (2000). *Calculus of Statistical Efficiency in a General Setting; Kernel Plug-in Estimation for Markov Chains; Hidden Markov Modeling of Freeway Traffic.* Ph.D. Dissertation, Department of Statistics, University of California at Berkeley. `http://www.stat.berkeley.edu/users/kwon/index.html`

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

Li, B. (2000). Nonparametric estimating equations based on a penalized information criterion. *Canad. J. Statist.* **28**, 621–639.

Li, B. (2001). On quasilikelihood equations with nonparametric weights. To appear in *Scand. J. Statist.*

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

Maigret, N. (1978). Théorème de limite centrale fonctionnel pour une chaîne de Markov récurrente au sens de Harris et positive. *Ann. Inst. H. Poincaré Probab. Statist.* **14**, 425–440.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability.* Springer, London.

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

Müller, U. U., Schick, A. and Wefelmeyer, W. (2001b). Improved estimators for constrained Markov chain models. To appear in *Statist. Probab. Lett.*

Müller, U. U., Schick, A. and Wefelmeyer, W. (2001c). Estimating linear functionals of the error distribution in nonparametric regression. Technical Report, Department of Mathematical Sciences, Binghamton University.
`http://math.binghamton.edu/anton/preprint.html`

Penev, S. (1990). Convolution theorem for estimating the stationary distribution of Markov chains. *C. R. Acad. Bulgare Sci.* **43**, 29–32.

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

Peng, H. and Schick, A. (2001). Efficient estimation of linear functionals of a bivariate distribution with equal but unknown marginals: The least squares approach. Technical Report, Department of Mathematical Sciences, Binghamton University.

Ritov, Y. and Wellner, J. A. (1988). Censoring, martingales, and the Cox model. In *Statistical Inference from Stochastic Processes* (N. U. Prabhu, ed.), 191-219. Contemporary Mathematics 80, American Mathematical Society, Providence, Rhode Island.

Saavedra, A. and Cao, R. (1999). 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. and Cao, R. (2000). On the estimation of the marginal density of a moving average process. *Canad. J. Statist.* **28**, 799–815.

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

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

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

Schick, A. (1999b). Efficient estimation in a semiparametric heteroscedastic autoregressive model. Technical Report, Department of Mathematical Sciences, Binghamton University. `http://math.binghamton.edu/anton/preprint.html`

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

Schick, A. and Wefelmeyer, W. (1999). Efficient estimation of invariant distributions of some semiparametric Markov chain models. *Math. Methods Statist.* **8**, 119–134.

Schick, A. and Wefelmeyer, W. (2001a). Estimating joint distributions of Markov chains. To appear in *Stat. Inference Stoch. Process.*

Schick, A. and Wefelmeyer, W. (2001b). Estimating the innovation distribution in nonlinear autoregressive models. To appear in *Ann. Inst. Statist. Math.*

Schick, A. and Wefelmeyer, W. (2001c). Efficient estimation in invertible linear processes. Technical Report, Department of Mathematical Sciences, Binghamton University. `http://math.binghamton.edu/anton/preprint.html`

Schick, A. and Wefelmeyer, W. (2001d). Estimating invariant laws of linear processes by U-statistics. Technical Report, Department of Mathematical Sciences, Binghamton University. `http://math.binghamton.edu/anton/preprint.html`

Schick, A. and Wefelmeyer, W. (2001e). Root n consistent and optimal density estimators for moving average processes. Technical Report, Department of Mathematical Sciences, Binghamton University.
`http://math.binghamton.edu/anton/preprint.html`

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.

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

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.