1

Estimators For Partially Observed Markov Chains

Ursula U. Müller, Anton Schick and Wolfgang Wefelmeyer

Department of Statistics, Texas A&M University
Department of Mathematical Sciences, Binghamton University
Mathematisches Institut, Universität zu Köln

Abstract: Suppose we observe a discrete-time Markov chain at certain periodic or random time points only. Which observation patterns allow us to identify the transition distribution? In case we can identify it, how can we construct (good) estimators? We discuss these questions both for nonparametric models and for linear autoregression.

Keywords and phrases: Markov chain, linear autoregression, partial observation, periodic skipping, random skipping, empirical estimator, deconvolution.

1.1 Introduction

For Markov chains observed at certain periodic or random time points only, we discuss when one can identify the underlying transition distribution, and how one can construct estimators of linear functionals of the stationary distribution in nonparametric models, and of the innovation density in linear autoregression. By Markov chain we mean a Markov process in discrete time, with arbitrary state space.

In Section 1.2 we consider nonparametric estimators of linear functionals of the form $E[h(X_0, X_1)]$ of a real-valued first-order stationary Markov chain. We introduce different periodic and random partial observation patterns. If nothing is known about the structure of the transition distribution, consistent estimation of $E[h(X_0, X_1)]$ is, in general, impossible unless one occasionally sees adjacent pairs (X_{j-1}, X_j). We can use these pairs to construct an empirical estimator of $E[h(X_0, X_1)]$. In the simplest such situation, with every third of the realizations of the chain unobserved, we show how to use the information across the gaps for improving the empirical estimator. The approach carries over to the other observation patterns, and to higher-order Markov chains. In Section 1.3 we assume that the Markov chain is a first-order linear autoregressive process. In this case we can even treat observation patterns in which we never

see adjacent pairs, assuming only that we know the sign of the autoregression parameter. In the simplest such situation, only every second realization of the process is observed. We construct deconvolution-type estimators for the innovation density in this case. Again the approach carries over to more complicated observation patterns, and to higher-order linear autoregressive processes.

1.2 Nonparametric Estimators

Full observations. Let $X_0, \ldots, X_n$ be observations of a real-valued stationary and uniformly ergodic first-order Markov chain with transition distribution $Q(x, dy)$. We can identify Q from the stationary distribution of an adjacent pair (X_0, X_1), which in turn is identified from sufficiently many linear functionals $E[h(X_0, X_1)]$, for example from the distribution function $(s, t) \mapsto E[\mathbf{1}(X_0 \leq s, X_1 \leq t)]$ of (X_0, X_1). It suffices therefore to study estimation of such functionals. Let h be a bounded measurable function on $\mathbb{R}^2$. A natural estimator of $Eh = E[h(X_0, X_1)]$ is the empirical estimator

$$\mathbb{E}h = \frac{1}{n} \sum_{j=1}^{n} h(X_{j-1}, X_j).$$

It admits the martingale approximation

$$n^{1/2}(\mathbb{E}h - Eh) = n^{-1/2} \sum_{j=1}^{n} (Ah)(X_{j-1}, X_j) + o_p(1) \tag{1.1}$$

with

$$(Ah)(x, y) = h(x, y) - Q_x h + \sum_{k=1}^{\infty} (Q_y^k h - Q_x^{k+1} h),$$

where $Q_x h = \int h(x, y) Q(x, dy)$ and $Q_x^k h = \int Q_y h Q^{k-1}(x, dy)$ for $k = 2, 3, \ldots$. Hence by the martingale central limit theorem, $n^{1/2}(\mathbb{E}h - Eh)$ is asymptotically normal with variance $E[(Ah)^2(X_0, X_1)]$. See Meyn and Tweedie (1993), Chapter 17, for these results and for generalizations. In nonparametric models, with nothing known about the structure of the transition distribution, $\mathbb{E}h$ is efficient in the sense of Hájek and LeCam; see Penev (1991), Bickel (1993), Greenwood and Wefelmeyer (1995) and Bickel and Kwon (2001) for different proofs.

Periodic skipping. Suppose now that we observe only some of the realizations, in a deterministic pattern that repeats itself periodically, say with period m. Specifically, in the first period we observe at k times $1 \leq i_1 < \ldots < i_k \leq m$

and then at times $m + i_1, \ldots$, $2m + i_1, \ldots$, for $n + 1$ periods, say. Then we observe up to time $(n+1)m$ and have $(n+1)k$ observations. Here it is understood that we *know* how many realizations we skip. We will consider below a pattern where we do not have this information. The *skip lengths* are

$$s_1 = i_2 - i_1, \ldots, s_{k-1} = i_k - i_{k-1},\ s_k = m + i_1 - i_k.$$

1. In the simplest case, some of the skip lengths are 1. For example, let $m = 3$, $k = 2$, $i_1 = 1$, $i_2 = 2$. Then every third realization is missing. A simple estimator of $E[h(X_0, X_1)]$ is the empirical estimator based on observed pairs (X_{3j-2}, X_{3j-1}) of successive realizations of the chain. Such an estimator does not use the information in the non-adjacent pairs (X_{3j-1}, X_{3j+1}), and we should be able to find better estimators, in the sense of smaller asymptotic variance (unless the observations happen to be independent). In the next subsection we describe how one could use the information in the non-adjacent pairs to improve on the empirical estimator.

2. Suppose that none of the skip lengths is 1, but they have no common divisor. Then we can represent 1 as a linear combination of skip lengths. Suppose, for example, that $m = 5$, $k = 2$, $i_1 = 1$, $i_2 = 3$. Then the skip lengths are $s_1 = 2$, $s_2 = 3$, and, since $1 = 3 - 2$, we can write $Q = Q^{-2}Q^3$. We can therefore identify Q from Q^2 and Q^3, which in turn can be estimated from the pairs (X_{5j+1}, X_{5j+3}) and (X_{5j-2}, X_{5j+1}), respectively. To estimate the inverse of a transition distribution, decompose the state space into a finite number of sets and invert the corresponding empirical transition matrix.

3. If the skip lengths have a common divisor, Q is not identifiable. Suppose, for example, that $m = 2$, $k = 1$, $i_1 = 1$. Then we skip every second realization. The remaining observations allow us to estimate Q^2, but this does not identify the root Q uniquely. In certain parametric and semiparametric models we can however still (nearly) identify Q, for example if the chain follows a first-order linear autoregressive model; see Section 1.3.

Observing two out of three. Suppose we observe (X_{3j-2}, X_{3j-1}) for $j = 1, \ldots, n$. A simple estimator for $E[h(X_0, X_1)]$ is the empirical estimator

$$\mathbb{E}h = \frac{1}{n}\sum_{j=1}^{n} h(X_{3j-2}, X_{3j-1}).$$

The information in the non-adjacent pairs (X_{3j-1}, X_{3j+1}) can be used as follows. Write (X, Y, Z) for $(X_{3j-1}, X_{3j}, X_{3j+1})$. We want to estimate $E[h(X, Y)]$. Introduce the conditional expecations

$$h_\ell(X, Z) = E(h(X, Y)|X, Z) \quad \text{and} \quad h_r(X, Z) = E(h(Y, Z)|X, Z).$$

We have

$$E[h_\ell(X, Z)] = E[h_r(X, Z)] = E[h(X, Y)].$$

If we knew h_ℓ and h_r, we could estimate $E[h(X,Y)]$ by empirical estimators

$$\frac{1}{n}\sum_{j=1}^{n} h_\ell(X_{3j-1}, X_{3j+1}) \quad \text{and} \quad \frac{1}{n}\sum_{j=1}^{n} h_r(X_{3j-1}, X_{3j+1})$$

or smoothed versions of these. We do not know h_ℓ and h_r and suggest to replace them by estimators as follows. Assume that the finite-dimensional stationary distributions of the chain have Lebesgue densities. Let p_1, p_2, p_3 denote the densities of X, (X,Y), (X,Y,Z), respectively. Write g for the density of (X,Z). Note that

$$g(x,z) = \int p_3(x,y,z)\,dy.$$

We have

$$h_\ell(x,z) = \frac{\int h(x,y)p_3(x,y,z)\,dy}{g(x,z)}.$$

Write

$$p_3(x,y,z) = \frac{p_2(x,y)p_2(y,z)}{p_1(y)}.$$

Estimate p_2 by a kernel estimator based on the adjacent pairs (X_{3j-2}, X_{3j-1}),

$$\hat{p}_2(x,y) = \frac{1}{n}\sum_{i=1}^{n} k_b(x - X_{3j-2})k_b(y - X_{3j-1}),$$

where $k_b(x) = k(x/b)/b$ with k a kernel and b a bandwidth. Estimate p_1 by

$$\hat{p}_1(y) = \frac{1}{2}\Big(\int \hat{p}_2(x,y)\,dx + \int \hat{p}_2(y,z)\,dz\Big).$$

Then we can estimate p_3 by

$$\hat{p}_3(x,y,z) = \frac{\hat{p}_2(x,y)\hat{p}_2(y,z)}{\hat{p}_1(y)}$$

and g by

$$\bar{g}(x,z) = \int \hat{p}_3(x,y,z)\,dy.$$

We arrive at the following estimator for h_ℓ,

$$\hat{h}_\ell(x,z) = \frac{\int h(x,y)\hat{p}_3(x,y,z)\,dy}{\bar{g}(x,z)}.$$

Rather than looking at the empirical estimator $(1/n)\sum_{j=1}^{n} \hat{h}_\ell(X_{3j-1}, X_{3j+1})$, it is technically convenient to look at the smoothed version

$$\mathbb{E}_\ell h = \int \hat{h}_\ell(x,z)\hat{g}(x,z)\,dx\,dz,$$

where $\hat{g}$ is a kernel estimator of g based on non-adjacent pairs (X_{3j-1}, X_{3j+1}),

$$\hat{g}(x,z) = \frac{1}{n}\sum_{j=1}^{n} k_b(x - X_{3j-1})k_b(y - X_{3j+1}).$$

Similarly,

$$\hat{h}_r(x,z) = \frac{\int h(y,z)\hat{p}_3(x,y,z)\,dy}{\bar{g}(x,z)}$$

and

$$\mathbb{E}_r h = \int \hat{h}_r(x,z)\hat{g}(x,z)\,dx\,dz.$$

Under appropriate conditions, the three estimators $\mathbb{E}h$, $\mathbb{E}_\ell h$ and $\mathbb{E}_r h$ can be shown to be asymptotically normal. We can take linear combinations of them to obtain estimators with smaller asymptotic variance than the empirical estimator $\mathbb{E}h$. The best weights are expressed in terms of the variances and covariances of the three estimators. They depend on the unknown distribution but can be estimated empirically. Consider for example the empirical estimator $\mathbb{E}h = (1/n)\sum_{j=1}^{n} h(X_{3j-2}, X_{3j-1})$ based on the observations (X_{3j-2}, X_{3j-1}), $j = 1, \ldots, n$. The observations follow a Markov chain with transition distribution of (X_{3j+1}, X_{3j+2}) given $(X_{3j-2}, X_{3j-1}) = (v, w)$ not depending on v and defined by

$$R(w, dy, dz) = Q^2 \otimes Q(w, dy, dz) = Q^2(w, dy)Q(y, dz).$$

We can apply the martingale approximation (1.1) to obtain

$$n^{1/2}(\mathbb{E}h - Eh) = n^{-1/2}\sum_{j=1}^{n}(Bh)(X_{3j-1}, X_{3j+1}, X_{3j+2}) + o_p(1)$$

with

$$(Bh)(w,y,z) = h(y,z) - R_w h + \sum_{k=1}^{\infty}(R_z^k h - R_w^{k+1}h).$$

By the martingale central limit theorem, $\mathbb{E}h$ is asymptotically normal with variance $E[(Bh)^2(X_2, X_4, X_5)]$ of the form

$$Eh^2 - (Eh)^2 + 2\sum_{k=1}^{\infty} E[(h(X_1, X_2) - Eh)h(X_{3k+1}, X_{3k+2})].$$

This variance can be estimated empirically, by

$$\mathbb{E}h^2 - (\mathbb{E}h)^2 + 2\sum_{k=1}^{m(n)} \frac{1}{n-k}\sum_{j=1}^{n-k}(h(X_{3j-2}, X_{3j-1}) - \mathbb{E}h)h(X_{3(j+k)-2}, X_{3(j+k)-1})$$

with $m(n)$ slowly increasing to infinity. Compare Müller *et al.* (2001). Similar martingale approximations can be obtained for $\mathbb{E}h_\ell$ and $\mathbb{E}h_r$, and their variances and the covariances of the three estimators can be estimated similarly as the variance of $\mathbb{E}h$.

Random skipping. Suppose that, after an observation at time j, we make the next observation at time $j+s$ with probability a_s. Then the skip lengths are i.i.d. random variables S_i, $i = 0, 1, \ldots$, with values in $\mathbb{N}$ and distribution given by $A(\{s\}) = a_s$, $s \in \mathbb{N}$. Set $T_0 = 0$ and $T_j = \sum_{i=0}^{j-1} S_i$, and write $Y_j = X_{T_j}$. Suppose we observe the pairs (S_j, Y_j) for $j = 0, \ldots, n$, say. They form a Markov chain with transition distribution

$$R(x, ds, dy) = A(ds)Q^s(x, dy).$$

Let N_s denote the observed number of skip lengths $S_j = s$. We can estimate a_s by N_s/n. Estimation of $E[h(X_0, X_1)]$ is similar to the case of periodic skipping considered above. In particular, if a_1 is positive, a simple estimator of $E[h(X_0, X_1)]$ is the empirical estimator

$$\frac{1}{N_1} \sum_{S_j=1} h(Y_j, Y_{j+1}).$$

The information in the pairs (Y_j, Y_{j+1}) with skip lengths $S_j = 2, 3, \ldots$ can be exploited similarly as for periodic skipping.

"Skipping at random". In the previous subsection we have assumed that the skip lengths are independent of the Markov chain. It is however conceivable that the skip lengths depend on the previous state. Let $A(x, ds)$ denote the skip length distribution out of state x. Then we observe pairs (S_j, Y_j) for $j = 0, \ldots, n$ with transition distribution

$$R(x, ds, dy) = A(x, ds)Q^s(x, dy).$$

This factorization is analogous to the factorization $Q(x, dy)A(x, y, ds)$ of the transition distribution of a Markov renewal process; for efficient estimation in semiparametric models of the corresponding semi-Markov process see Greenwood *et al.* (2004). The name "skipping at random" is chosen because of the similarity with responses "missing at random" in regression models; for efficient semiparametric estimation see Müller *et al.* (2006). Recent monographs treating missing data are Little and Rubin (2002), van der Laan and Robins (2003) and Tsiatis (2006). Random skipping as considered above, with A not depending on x, would correspond to "missing totally at random". We can estimate $a_s(x) = A(x, \{s\})$ by the kernel estimator

$$\hat{a}_s(x) = \frac{\sum_{i=1}^n k_b(x - Y_i)\mathbf{1}(S_i = s)}{\sum_{i=1}^n k_b(x - Y_i)},$$

where $k_b(x) = k(x/b)/b$ with k a kernel and b a bandwidth. Again, if $a_1(x) = A(x, \{1\})$ is positive with positive probability, a simple estimator of the expecation $E[h(X_0, X_1)]$ can be based on the observed pairs of successive observations:

$$\sum_{j=0}^{n-1} \frac{\mathbf{1}(S_j = 1)}{\hat{a}_1(Y_j)} h(Y_j, Y_{j+1}).$$

Again, the information in the pairs (Y_j, Y_{j+1}) with skip lengths $S_j = 2, 3, \ldots$ can be exploited similarly as for periodic skipping.

1.3 Linear Autoregression

Full observations. Let $X_0, \ldots, X_n$ be observations from a stationary first-order autoregressive linear model

$$X_j = \vartheta X_{j-1} + \varepsilon_j \tag{1.2}$$

with $|\vartheta| < 1$ and i.i.d. innovations ε_j that have mean zero, finite variance and density f. This is a first-order Markov chain with transition distribution $Q(x, dy) = f(y - \vartheta x)\, dy$, parametrized by ϑ and f. A simple estimator for ϑ is the least squares estimator

$$\bar{\vartheta} = \frac{\sum_{j=1}^n X_{j-1} X_j}{\sum_{j=1}^n X_{j-1}^2}.$$

We can use it to estimate the innovation ε_j by the residual $\bar{\varepsilon}_j = X_j - \bar{\vartheta} X_{j-1}$. An estimator for the innovation density f is the residual-based kernel estimator

$$\hat{f}(x) = \frac{1}{n} \sum_{j=1}^n k_b(x - \bar{\varepsilon}_j)$$

where $k_b(x) = k(x/b)/b$ with k a kernel and b a bandwidth.

Observing one out of two. As mentioned in Section 1.2, the transition distribution of a Markov chain is not identifiable if observations are skipped periodically with skip lenghts having a common divisor. In the simplest such case, only every second of the realizations of the chain is observed. The situation is much better for autoregression (1.2). Then the transition distribution is still identifiable, up to the sign of ϑ. To see this, suppose that we observe $X_0, X_2, \ldots, X_{2n}$ and write

$$X_{2j} = \vartheta^2 X_{2j-2} + \eta_{2j} \tag{1.3}$$

with

$$\eta_{2j} = \varepsilon_{2j} + \vartheta\varepsilon_{2j-1}. \tag{1.4}$$

The X_{2j} follow again a first-order linear autoregesssive model, now with autoregression parameter $s = \vartheta^2$ and innovation η_{2j}. The non-uniqueness of the square-root of the two-step transition distribution Q^2 reduces to the non-uniqueness of the square-root of ϑ^2. Let us assume that we know the sign of ϑ, say ϑ is positive. This knowledge is realistic in many applications. We can estimate ϑ^2 by the least squares estimator

$$\hat{s} = \frac{\sum_{j=1}^n X_{2j-2}X_{2j}}{\sum_{j=1}^n X_{2j-2}^2}.$$

Then $\hat{\vartheta} = \hat{s}^{1/2}$ estimates ϑ.

It remains to estimate f. We introduce three different approaches. All are solutions of certain deconvolution problems. Write $\varphi_Y(t) = E[\exp(itY)]$ for the characteristic function of a random variable Y.

1. The most straightforward estimator for f uses only the autoregressive representation (1.2), which implies

$$\varphi_X(t) = \varphi_{\vartheta X}(t)\varphi_\varepsilon(t) = \varphi_X(\vartheta t)\varphi_\varepsilon(t).$$

Estimate φ_X by the empirical characteristic function

$$\hat{\varphi}_X(t) = \frac{1}{n}\sum_{j=1}^n \exp(itX_{2j}).$$

An estimator for φ_ε is then given by

$$\hat{\varphi}_{\varepsilon,1}(t) = \frac{\hat{\varphi}_X(t)}{\hat{\varphi}_X(\hat{\vartheta}t)}.$$

Let K be a kernel, φ_K its characteristic function, and b a bandwidth that tends to zero as n tends to infinity. By Fourier inversion we arrive at an estimator for f,

$$\hat{f}_1(x) = \frac{1}{2\pi}\int \exp(-itx)\varphi_K(bt)\hat{\varphi}_{\varepsilon,1}(t)\,dt.$$

2. Another estimator for f uses only the moving average representation (1.4) of the η_{2j}. It is based on the approach of Belomestny (2003); see also Belomestny and Prokhorov (2003) and Belomestny (2005). Belomestny considers i.i.d. random variables Y_1 and Y_2 and estimates their density on the basis of i.i.d. observations distributed as $\vartheta Y_1 + Y_2$. The moving average representation (1.4) is of this form, but we do not know ϑ and do not observe the η_{2j} and must

replace them by an estimator $\hat{\vartheta}$ and residuals $\hat{\eta}_{2j} = X_{2j} - \hat{s}X_{2j-2}$. From (1.4) we obtain

$$\varphi_\eta(t) = \varphi_\varepsilon(t)\varphi_{\vartheta\varepsilon}(t) = \varphi_\varepsilon(t)\varphi_\varepsilon(\vartheta t).$$

Iteratively solving for φ_ε we arrive at the representation

$$\varphi_\varepsilon(t) = \frac{\varphi_\eta(t)}{\varphi_\varepsilon(\vartheta t)} = \prod_{r=0}^{\infty} \frac{\varphi_\eta(\vartheta^{2r}t)}{\varphi_\eta(\vartheta^{2r+1}t)}.$$

Estimate φ_η by the residual-based empirical characteristic function

$$\hat{\varphi}_\eta(t) = \frac{1}{n}\sum_{j=1}^{n} \exp(it\hat{\eta}_{2j}).$$

An estimator for φ_ε is then given by

$$\hat{\varphi}_{\varepsilon,2}(t) = \prod_{r=0}^{N} \frac{\hat{\varphi}_\eta(\hat{s}^r t)}{\hat{\varphi}_\eta(\hat{\vartheta}^{2r+1}t)}$$

with N tending to infinity. By Fourier inversion we arrive at a second estimator for f,

$$\hat{f}_2(x) = \frac{1}{2\pi}\int \exp(-itx)\varphi_K(bt)\hat{\varphi}_{\varepsilon,2}(t)\,dt.$$

3. A third estimator for f uses (1.2) together with the autoregression representation (1.3) of the observations X_{2j}. They give

$$\varphi_X(t) = \varphi_X(\vartheta t)\varphi_\varepsilon(t) \quad \text{and} \quad \varphi_X(t) = \varphi_X(\vartheta^2 t)\varphi_\eta(t)$$

and hence

$$\varphi_\varepsilon(t) = \frac{\varphi_X(\vartheta^2 t)\varphi_\eta(t)}{\varphi_X(\vartheta t)}.$$

An estimator for φ_ε is therefore given by

$$\hat{\varphi}_{\varepsilon,3}(x) = \frac{\hat{\varphi}_X(\hat{s}t)\hat{\varphi}_\eta(t)}{\hat{\varphi}_X(\hat{\vartheta}t)}.$$

By Fourier inversion we arrive at a third estimator for f,

$$\hat{f}_3(x) = \frac{1}{2\pi}\int \exp(-itx)\varphi_K(bt)\hat{\varphi}_{\varepsilon,3}(t)\,dt.$$

The estimator $\hat{f}_1$ is the easiest to calculate. However, the representation of φ_ε as a ratio $\varphi_X/\varphi_{\vartheta X}$ does not lead to a good estimator of φ_ε and f. It is comparable with the usual deconvolution estimators treated in the literature;

see Fan (1991) for their convergence rates, which can be very slow. The estimators $\hat{f}_2$ and $\hat{f}_3$ do not have this disadvantage, at least not to the same extent. This is easier to explain for $\hat{f}_3$, which is based on the representation of φ_ε as $\varphi_{\vartheta^2 X}\varphi_\eta/\varphi_{\vartheta X}$ whose tail behavior is governed by the numerator. Of course, $\hat{\varphi}_{\varepsilon,3}$ and $\hat{f}_3$ are preferable because they are simpler than $\hat{\varphi}_{\varepsilon,2}$ and $\hat{f}_2$. Apart from this, $\hat{f}_2$ and $\hat{f}_3$ have similar convergence rates.

Let g denote the density of the innovation η_{2j}. Paradoxically, it can be estimated at a better rate than the density of the innovation ε_j of the fully observed time series. From (1.4) we have the representation

$$g(y) = \int f(y - \vartheta x) f(x)\, dx$$

and can estimate g by the plug-in estimator

$$\hat{g}(y) = \int \hat{f}(y - \hat{\vartheta} x) \hat{f}(x)\, dx,$$

where $\hat{f}$ is $\hat{f}_2$ or $\hat{f}_3$. The estimator $\hat{g}$ can be root-n consistent; compare Frees (1994), Schick and Wefelmeyer (2004a, 2004b, 2006) and Giné and Mason (2006) for related results.

Higher lags. Versions of the three estimators $\hat{f}_1$, $\hat{f}_2$, $\hat{f}_3$ can also be constructed if we observe the AR(1) process each k-th time only. We have seen that an AR(1) process, observed every second time, is again AR(1), with different innovation distribution and autoregression parameter. If we observe the process each k-th time only, we also have an AR(1) process

$$X_{kj} = \vartheta^k X_{k(j-1)} + \eta_{kj} \tag{1.5}$$

with innovations

$$\eta_{kj} = \sum_{i=0}^{k-1} \vartheta^i \varepsilon_{kj-i}. \tag{1.6}$$

Suppose we observe $X_0, X_k, \ldots, X_{kn}$. Then ϑ^k can be estimated by the least squares estimator

$$\hat{s} = \frac{\sum_{j=1}^n X_{k(j-1)} X_{kj}}{\sum_{j=1}^n X_{k(j-1)}^2}.$$

If k is even, we cannot identify the sign of ϑ and will again assume that we know ϑ to be positive. Then $\hat{\vartheta} = \hat{s}^{1/k}$ estimates ϑ.

A version of $\hat{f}_1$ is obtained by using again the Fourier inverse of

$$\hat{\varphi}_{\varepsilon,1}(t) = \frac{\hat{\varphi}_X(t)}{\hat{\varphi}_X(\hat{\vartheta} t)},$$

now with the empirical characteristic function

$$\hat{\varphi}_X(t) = \frac{1}{n}\sum_{j=1}^{n} \exp(itX_{kj}).$$

In view of (1.5) and (1.6) we obtain

$$\varphi_X(t) = \varphi_\eta(t)\varphi_X(\vartheta^k t) \tag{1.7}$$

and then, by the representation (1.2),

$$\varphi_\varepsilon(t) = \frac{\varphi_X(t)}{\varphi_X(\vartheta t)} = \frac{\varphi_\eta(t)\varphi_X(\vartheta^k t)}{\varphi_X(\vartheta t)}. \tag{1.8}$$

A version of $\hat{\varphi}_{\varepsilon,3}$ is therefore

$$\hat{\varphi}_{\varepsilon,3}(t) = \frac{\hat{\varphi}_\eta(t)\hat{\varphi}_X(\hat{s}t)}{\hat{\varphi}_X(\hat{\vartheta}t)},$$

now with empirical characteristic function

$$\hat{\varphi}_\eta(t) = \frac{1}{n}\sum_{j=1}^{n} \exp(it\hat{\eta}_{kj})$$

based on residuals $\hat{\eta}_{kj} = X_{kj} - \hat{s}X_{k(j-1)}$. An estimator for f is now obtained by Fourier inversion of $\hat{\varphi}_{\varepsilon,3}$.

For a version of the second estimator, $\hat{f}_2$, we apply (1.7) repeatedly to (1.8) and obtain

$$\varphi_\varepsilon(t) = \frac{\varphi_\eta(t)\varphi_X(\vartheta^k t)}{\varphi_\eta(\vartheta t)\varphi_X(\vartheta^{k+1} t)} = \frac{\varphi_\eta(t)}{\varphi_\eta(\vartheta t)}\prod_{r=1}^{\infty}\frac{\varphi_\eta(\vartheta^{kr} t)}{\varphi_\eta(\vartheta^{kr+1} t)}.$$

From this we obtain a version of $\hat{\varphi}_{\varepsilon,2}$ and hence of $\hat{f}_2$.

Higher order autoregression. Generalizations of our results to higher-order autoregression are not straightforward. In general we lose the Markov property. Consider an AR(2) process

$$X_j = \vartheta_1 X_{j-1} + \vartheta_2 X_{j-2} + \varepsilon_j,$$

with innovations ε_j as before. Assume that the polynomial $1 - \vartheta_1 z - \vartheta_2 z^2$ does not have zeroes on the closed complex unit disk. Suppose we observe the process at even times only. We have

$$X_{2j} = \vartheta_1 X_{2j-1} + \vartheta_2 X_{2j-2} + \varepsilon_{2j}.$$

Replacing X_{2j-1} by its AR(2) representation, we obtain

$$X_{2j} = (\vartheta_1^2 + \vartheta_2)X_{2j-2} + \vartheta_1\vartheta_2 X_{2j-3} + \varepsilon_{2j} + \vartheta_1\varepsilon_{2j-1}. \tag{1.9}$$

Iterating this for odd-numbered indices, we arrive at an ARMA(∞,∞) representation for X_{2j},

$$X_{2j} = (\vartheta_1^2 + \vartheta_2)X_{2j-2} + \sum_{i=1}^{\infty} \vartheta_1^2\vartheta_2^i X_{2j-2i-2} + \varepsilon_{2j} + \vartheta_1\varepsilon_{2j-1} + \sum_{i=1}^{\infty} \vartheta_1\vartheta_2^i \varepsilon_{2j-2i-1}.$$

If we replace all X_{2j-i} by their AR(2) representations, we arrive at an AR(∞) representation for X_{2j}.

A simpler representation is obtained if we subtract

$$\vartheta_2 X_{2j-2} = \vartheta_2(\vartheta_1 X_{2j-3} + \vartheta_2 X_{2j-4} + \varepsilon_{2j-2})$$

from (1.9). This gives the ARMA(2,2) representation

$$X_{2j} - (\vartheta_1^2 + 2\vartheta_2)X_{2j-2} + \vartheta_2^2 X_{2j-4} = \varepsilon_{2j} + \vartheta_1\varepsilon_{2j-1} - \vartheta_2\varepsilon_{2j-2}.$$

The parameters are identifiable if we know their signs.

Such a representation can be obtained for arbitrary ARMA(p,q) processes observed at even times. Introduce polynomials $\varrho(z) = 1 + \varrho_1 z + \ldots + \varrho_p z^p$ and $\varphi(z) = 1 + \varphi_1 z + \ldots + \varphi_q z^q$. Assume that ϱ does not vanish on the closed complex unit disk. Define the backshift operator by $BX_j = X_{j-1}$. Consider the ARMA(p,q) process

$$\varrho(B)X_j = \varphi(B)\varepsilon_j,$$

with ε_j as before. Let $\varrho_1^*, \ldots, \varrho_p^*$ denote the zeroes of ϱ. They lie outside the unit disk. Factor ϱ as

$$\varrho(z) = \prod_{i=1}^{p}(z - \varrho_i^*).$$

Introduce the polynomials

$$\varrho_2(z) = \prod_{i=1}^{p}(z - \varrho_i^{*2}), \quad \varrho_+(z) = \prod_{i=1}^{p}(z + \varrho_i^*).$$

We can write

$$\varrho_2(z^2) = \prod_{i=1}^{p}(z^2 - \varrho_i^{*2}) = \prod_{i=1}^{p}(z + \varrho_i^*)(z - \varrho_i^*) = \varrho_+(z)\varrho(z)$$

and obtain an ARMA(p,$p+q$) representation for the ARMA(p,q) process observed at even times only:

$$\varrho_2(B^2)X_{2j} = \varrho_+(B)\varphi(B)\varepsilon_{2j}.$$

We *retain* a Markovian representation if we have observations in blocks of length at least equal to the order of the process. For example, suppose we do not see every third observation of the AR(2) process, so our observations are (X_{3j-2}, X_{3j-1}) for $j = 1, \ldots, n$, say. Then we can write

$$X_{3j+1} = (\vartheta_1^2 + \vartheta_2)X_{3j-1} + \vartheta_1\vartheta_2 X_{3j-2} + \varepsilon_{3j+1} + \vartheta_1\varepsilon_{3j}$$

and

$$X_{3j+2} = \vartheta_1 X_{3j+1} + \vartheta_1\vartheta_2 X_{3j-1} + \vartheta_2^2 X_{3j-2} + \varepsilon_{3j+2} + \vartheta_2\varepsilon_{3j}.$$

This means that the observations (X_{3j-2}, X_{3j-1}) follow an alternating autoregressive process, with orders alternating between 2 and 3, and independent innovations $\eta_{3j+1} = \varepsilon_{3j+1} + \vartheta_1\varepsilon_{3j}$, $j = 1, \ldots, n$, and $\eta_{3j+2} = \varepsilon_{3j+2} + \vartheta_2\varepsilon_{3j}$, $j = 1, \ldots, n$, respectively. Note however that for fixed j the innovations η_{3j+1} and η_{3j+2} depend on each other. The observations (X_{3j-2}, X_{3j-1}) can also be viewed as a two-dimensional autoregressive process of order 3.

In both cases described above we have obtained ARMA(p,q) representations for the partially observed process. Such representations can again be used to construct estimators for the innovation density. Consider an ARMA(2,2) process of the form

$$X_j + aX_{j-1} + bX_{j-2} = \varepsilon_j + c\varepsilon_{j-1} + d\varepsilon_{j-2} = \eta_j.$$

To construct an estimator analogous to $\hat{f}_2$, write

$$\varphi_\varepsilon(t) = \frac{\varphi_\eta(t)}{\varphi_\varepsilon(ct)\varphi_\varepsilon(dt)},$$

replace $\varphi_\varepsilon(ct)$ and $\varphi_\varepsilon(dt)$ by such ratios to obtain

$$\varphi_\varepsilon(t) = \frac{\varphi_\eta(t)\varphi_\varepsilon(c^2t)\varphi_\varepsilon(d^2t)\varphi_\varepsilon^2(cdt)}{\varphi_\eta(ct)\varphi_\eta(dt)},$$

and iterate these steps to obtain an infinite product in terms of φ_η. An estimator for φ_η can be based on residuals $\hat{\eta}_j = X_j + \hat{a}X_{j-1} + \hat{b}X_{j-2}$.

Acknowledgment. Anton Schick was supported in part by NSF Grant DMS 0405791. We thank an Editor and a Referee for their suggestions.

References

1. Belomestny, D. (2003). Rates of convergence for constrained deconvolution problem, Technical Report, arXiv:math.ST/0306237.

2. Belomestny, D. (2005). Reconstruction of a general distribution from the distribution of some statistics, *Theory of Probability and Its Applications*, **49**, 1–15.

3. Belomestny, D. and Prokhorov, A. V. (2003). On the problem of reconstructing the general distribution from the distribution of a linear statistic, *Moscow University Mathematics Bulletin*, **58**, 1–6.

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

5. Bickel, P. J. and Kwon, J. (2001). Inference for semiparametric models: Some questions and an answer (with discussion), *Statistica Sinica*, **11**, 863–960.

6. Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems, *Annals of Statistics*, **19**, 1257–1272.

7. Frees, E. W. (1994). Estimating densities of functions of observations, *Journal of the American Statistical Association*, **89**, 517–525.

8. Giné, E. and Mason, D. (2006). On local U-statistic processes and the estimation of densities of functions of several variables, To appear in *Annals of Statistics*.

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

10. Greenwood, P. E., Müller, U. U. and Wefelmeyer, W. (2004). Efficient estimation for semiparametric semi-Markov processes, *Communications in Statistics Theory and Methods*, **33**, 419–435.

11. Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*, Second Edition, Wiley Series in Probability and Statistics, Wiley-Interscience, Hoboken, NJ.

12. Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*, Communications and Control Engineering Series, Springer-Verlag, London.

13. Müller, U. U., Schick, A. and Wefelmeyer, W. (2001). Improved estimators for constrained Markov chain models, *Statistics & Probability Letters*, **54**, 427–435.

14. Müller, U. U., Schick, A. and Wefelmeyer, W. (2006). Imputing responses that are not missing, In *Probability, Statistics and Modelling in Public Health* (Eds., N. Nikulin, D. Commenges, and C. Huber), pp. 350-363, Springer-Verlag, New York.

15. Penev, S. (1991). Efficient estimation of the stationary distribution for exponentially ergodic Markov chains, *Journal of Statistical Planning and Inference*, **27**, 105–123.

16. Schick, A. and Wefelmeyer, W. (2004a). Root n consistent and optimal density estimators for moving average processes, *Scandinavian Journal of Statistics*, **31**, 63–78.

17. Schick, A. and Wefelmeyer, W. (2004b). Functional convergence and optimality of plug-in estimators for stationary densities of moving average processes. *Bernoulli*, **10**, 889–917.

18. Schick, A. and Wefelmeyer, W. (2006). Root-n consistent density estimators of convolutions in weighted L_1-norms, To appear in *Journal of Statistical Planning and Inference*.

19. Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*, Springer Series in Statistics, Springer-Verlag, New York.

20. Van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*, Springer Series in Statistics, Springer-Verlag, New York.