JOURNAL OF TIME SERIES ANALYSIS
J. Time Ser. Anal. **45**: 1006–1019 (2024)
Published online 17 May 2024 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1111/jtsa.12747

ORIGINAL ARTICLE

ESTIMATION FOR MARKOV CHAINS WITH PERIODICALLY MISSING OBSERVATIONS

URSULA U. MÜLLER[a] ANTON SCHICK[b] AND WOLFGANG WEFELMEYER[c]

[a]*Department of Statistics, Texas A&M University, College Station, TX, USA*
[b]*Department of Mathematics and Statistics, Binghamton University, Binghamton, NY, USA*
[c]*Abteilung Mathematik, Universität zu Köln, Köln, Germany*

When we observe a stationary time series with observations missing at periodic time points, we can still estimate its marginal distribution well, but the dependence structure of the time series may not be recoverable at all, or the usual estimators may have much larger variance than in the fully observed case. We show how non-parametric estimators can often be improved by adding unbiased estimators. We focus on a simple setting, first-order Markov chains on a finite state space, and an observation pattern in which a fixed number of consecutive observations is followed by an observation gap of fixed length, say workdays and weekends. The new estimators perform astonishingly well in some cases, as illustrated with simulations. The approach extends to continuous state space and to higher-order Markov chains.

Received 23 February 2023; Accepted 29 April 2024

Keywords: Markov chain; missing observations; empirical estimator; improved estimator.

MOS subject classification: 60J10; 62M05.

1. INTRODUCTION

The distribution of a (first-order) stationary Markov chain $X_0, X_1, \ldots$ on some state space S is determined by the (1-step) transition distribution $Q(x, \mathrm{d}y)$ and the corresponding (one-dimensional) marginal distribution $\pi(\mathrm{d}x)$. These are in turn determined by the joint distribution of (X_0, X_1). Our aim is to estimate this joint distribution from non-consecutive realizations of the Markov chain. We treat continuous and discrete state spaces in generality and elaborate details for the case when the state space is finite.

If the state space S is the real line and the transition distribution has a (Lebesgue) density $q(x, y)$, then the marginal distribution has a density, say $p(x)$, and both are determined by the joint density $p(x)q(x, y)$ of (X_0, X_1), which is in turn determined by expectations $E\big[f\big(X_0, X_1\big)\big]$ for sufficiently many (bounded) functions $f(x, y)$. On the other hand, if the state space is finite, say $S = \{1, \ldots, m\}$, then the transition distribution is a matrix $Q = (Q_{ij})_{i,j=1,\ldots,m}$, and the marginal distribution is a vector, say $\pi = (\pi_1, \ldots, \pi_m)^\top$, and both are determined by the matrix of joint probabilities $(\pi_i Q_{ij})_{i,j=1,\ldots,m}$. These are again determined by expectations $E\big[f\big(X_0, X_1\big)\big]$ for sufficiently many functions $f(x, y)$. Of course, for a discrete state space it would suffice to consider indicator functions $f = \mathbf{1}_{\{(x,y)\}}$ to identify the joint distribution of the chain. Even then, functions other than indicator functions are of interest in applications, for example covariances and joint moments.

Let $f(x, y)$ be a bounded function. The expectation $\theta = E[f(X_0, X_1)]$ of two consecutive observations can be estimated from observations $X_0, X_1, \ldots, X_n$ by the empirical estimator

$$\hat{\theta} = \frac{1}{n}\sum_{j=1}^{n} f(X_{j-1}, X_j).$$

*Correspondence to: Ursula U. Müller, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA.
Email: uschi@stat.tamu.edu

Under appropriate ergodicity and moment conditions, $\hat{\theta}$ is asymptotically normal. If the underlying model is non-parametric, $\hat{\theta}$ is also asymptotically efficient in the sense of a non-parametric version of the Hájek–Le Cam convolution theorem; see Penev (1991), Bickel (1993), and Greenwood and Wefelmeyer (1995).

We are interested in the case that observations are missing periodically. There is a large literature on regression models and time series with missing data. The emphasis is on interpolating plausible values for those that are missing, on calculating the loss of information through missing data, and on constructing consistent estimators; see, for example, Miller and Ferreiro (1984), Lu and Hui (2003), Bondon (2005), Pasanisi *et al.* (2012) and Efromovich (2014). Here we have a different goal, namely to obtain estimators that improve on the empirical estimator based on the observed pairs of consecutive realizations of the Markov chain.

Müller *et al.* (2008) discuss various missingness patterns without going into theoretical details, including a scenario with periodically observed data. An extreme case would be to have every other observation missing. Then we can estimate the 2-step transition distribution but we cannot identify the 1-step transition distribution. Here we focus on the most common periodic scenario in which the 1-step transition distribution is identifiable. We observe an initial value X_0. Then a gap of length a is followed by a block of $b+1$ consecutive realizations $X_a, \dots, X_{a+b}$, i.e. by b consecutive overlapping pairs $(X_a, X_{a+1}), (X_{a+1}, X_{a+2}), \dots, (X_{a+b-1}, X_{a+b})$. Then the pattern repeats itself,

$$X_0; X_a, X_{a+1}, \dots, X_{a+b};\ X_{2a+b}, \dots, X_{2a+2b};\ \dots;\ X_{(n-1)(a+b)+a}, \dots, X_{n(a+b)}.$$

Note that we consider n periods of length $a+b$ rather than n single realizations.

Our approach carries over immediately to more complicated periodic missingness patterns, as long as part of the observations are adjacent. It also carries over to random missingness patterns, as long as adjacent observations occur with positive probability.

Examples for such observation patterns are present and absent times during the day, and regular visits to remote places for data collection during the year. An obvious scenario with periodically missing observations are daily data that are only available on weekdays but not on weekends. Consider, for example, daily closing values of the stock market (of some index or some individual stock). The market is open only 5 days a week. Let us categorize a change in the closing value from one day to the next as positive, neutral or negative, where *positive* refers to an increase of at least p percent, *negative* to to a decrease of at least p percent, and *neutral* refers to the remaining case of no clear change. Here p is a small fixed number, say 0.2. We have four consecutive observations per week, the change from Monday to Tuesday, from Tuesday to Wednesday, from Wednesday to Thursday, and from Thursday to Friday. As markets are closed over the weekend, we treat the changes from Friday to Saturday, Saturday to Sunday, and Sunday to Monday as missing. In this scenario we have three states, gaps of length $a=4$ and blocks with $b+1=4$ observations. To estimate the probability of positive change on two consecutive days, we take f to be $f(x,y) = \mathbf{1}[x = y = \text{positive}]$.

The empirical estimator for $\theta = E[f(X_0, X_1)]$ based on the nb observed adjacent pairs is

$$\hat{\theta}_0 = \frac{1}{nb} \sum_{j=1}^{n} \sum_{c \in B+(j-1)(a+b)} f(X_{c-1}, X_c),$$

with $B = \{a+1, \dots, a+b\}$. The estimator $\hat{\theta}_0$ has expectation θ, i.e. it is unbiased. We will suggest an unbiased modification with smaller asymptotic variance.

To motivate the modification, suppose first that we have i.i.d. observations $X_1, \dots, X_n$. The empirical estimator of the expectation $E[f(X_1)]$ of a (bounded) function f is the average $(1/n)\sum_{j=1}^{n} f(X_j)$. Assume that the observations are known to fulfill a linear constraint $E[h(X_1)] = 0$ for a known (bounded) function h. Then we get new unbiased estimators $H_f(c) = (1/n)\sum_{j=1}^{n} \big(f(X_j) - ch(X_j)\big)$ for $E[f(X_1)]$ and can determine the constant $c = c_*$ that minimizes the asymptotic variance of these estimators. This constant depends on the distribution of the observations but can be estimated consistently by an empirical estimator $\hat{c}_*$. The resulting plug-in estimator $H_f(\hat{c}_*)$ is then asymptotically equivalent to $H_f(c_*)$ and an asymptotically efficient estimator of $E[f(X_1)]$.

DOI: 10.1111/jtsa.12747

Suppose now that we have consecutive observations $X_0, \dots, X_n$ from a stationary Markov chain. We have the obvious linear constraint that the expectations $E[h(X_j)]$ for (bounded) functions h are equal for all j, so $E[h(X_{j-1}) - h(X_j)] = 0$. We can modify the empirical estimator $(1/n)\sum_{j=1}^n f(X_{j-1}, X_j)$ as in the i.i.d. case, but this will *not* improve it asymptotically because the modification telescopes to something negligible,

$$\frac{1}{n}\sum_{j=1}^{n}\left(f\left(X_{j-1},X_j\right)-h\left(X_{j-1}\right)+h\left(X_j\right)\right)=\frac{1}{n}\sum_{j=1}^{n}f\left(X_{j-1},X_j\right)+\frac{1}{n}\left(h\left(X_n\right)-h\left(X_0\right)\right).$$

Let us return to the pattern of periodic observations with gaps of length a and blocks of length $b+1$. The empirical estimator $\hat{\theta}_0$ of θ based on the observed adjacent pairs can be modified as

$$\hat{\theta}(h)=\frac{1}{nb}\sum_{j=1}^{n}\sum_{c\in B+(j-1)(a+b)}\left(f\left(X_{c-1},X_c\right)-h\left(X_{c-1}\right)+h\left(X_c\right)\right).$$

The modification telescopes to something non-negligible now,

$$\hat{\theta}(h)=\hat{\theta}_0-\frac{1}{n}\sum_{j=1}^{n}\left(h\left(X_{(j-1)(a+b)+a}\right)-h\left(X_{j(a+b)}\right)\right).$$

Since $E\left[h\left(X_{c-1}\right)-h\left(X_c\right)\right]=0$, we can and will assume in the following that $E[h(X_0)]=0$.

In Section 2 we calculate the asymptotic variance of $\hat{\theta}(h)$. For finite state space S we determine the function $h=h_*$ that minimizes the asymptotic variance. Since h_* depends on the unknown distribution of the Markov chain, it must be replaced by an estimator $\hat{h}_*$, resulting in the estimator $\hat{\theta}(\hat{h}_*)$. The variance reduction of $\hat{\theta}(\hat{h}_*)$ over $\hat{\theta}_0$ can be considerable. We illustrate this in Section 3 with simulations for gaps $a=2$ and blocks $b+1=2$. Section 4 contains auxiliary results needed in the proofs of Section 2.

2. RESULTS

We need to calculate the asymptotic variances of $\hat{\theta}_0$ and $\hat{\theta}(\hat{h}_*)$. We begin by recalling a martingale approximation for the empirical estimator $\hat{\theta}=(1/n)\sum_{j=1}^n f\left(X_{i-1},X_i\right)$ of the expectation $\theta=E\left[f\left(X_0,X_1\right)\right]$ of a bounded function $f: S\times S\to\mathbb{R}$ from *consecutive* observations $X_0, \dots, X_n$ of a stationary Markov chain on an arbitrary state space S. We assume that the Markov chain is positive Harris recurrent and geometrically ergodic with respect to the sup-norm. From Lemma 1 in Greenwood and Wefelmeyer (1995) we obtain the martingale approximation

$$\hat{\theta}-\theta=\frac{1}{n}\sum_{j=1}^{n}(Af)(X_{j-1},X_j)+O_P\left(\frac{\log n}{n}\right),$$

where A is a bounded linear operator mapping a function $f(x,y)$ into a function

$$(Af)(x,y)=f(x,y)-g_1(x)+\sum_{k=1}^{\infty}\left(g_k(y)-g_{k+1}(x)\right).$$

Here $g_k(X_0)=E(f(X_{k-1},X_k)|X_0)$ is the k-step conditional expectation of $f(X_{k-1},X_k)$ given X_0. We calculate $g_1(X_0)=E(f(X_0,X_1)|X_0)$ and

$$g_{k+1}(X_0)=E(f(X_k,X_{k+1})|X_0)=E(E(f(X_k,X_{k+1})|X_1)|X_0)=E(g_k(X_1)|X_0),\quad k>1.$$

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12747

This implies $E((Af)(X_0, X_1)|X_0) = 0$. Thus $(Af)(X_0, X_1)$ is a martingale increment, and the asymptotic variance of $n^{1/2}(\hat{\theta} - \theta)$ is

$$E\left[(Af)^2(X_0, X_1)\right] = E\left[f^2(X_0, X_1)\right] - \theta^2 + 2\sum_{k=1}^{\infty}\left(E\left[f(X_0, X_1)f(X_k, X_{k+1})\right] - \theta^2\right).$$

The martingale approximation was first proved by Gordin (1969). It was discovered independently by several other authors; see the references in section 1 of Greenwood *et al.* (2001).

We apply the martingale approximation to the pattern in the Introduction, with n periods of gaps of length a followed by $b + 1$ consecutive observations. Denote the jth block of consecutive observations by

$$Z_j = \left(X_{(j-1)(a+b)+a}, \ldots, X_{(j-1)(a+b)+a+b}\right), \quad j = 1, \ldots, n,$$

and set

$$Y_j = X_{j(a+b)}, \quad j = 0, \ldots, n.$$

Now consider a bounded function g from S^{b+1} into $\mathbb{R}$. Then we have the identity

$$\frac{1}{n}\sum_{j=1}^{n} g(Z_j) = \frac{1}{n}\sum_{j=1}^{n}\left(g(Z_j) - g_0(Y_{j-1})\right) + \frac{1}{n}\sum_{j=1}^{n} g_0(Y_j) + \frac{1}{n}\left(g_0(Y_0) - g_0(Y_n)\right),$$

with

$$g_0\left(Y_{j-1}\right) = E\left(g(Z_j)|Y_{j-1}\right).$$

Since $Y_0, Y_1, \ldots, Y_n$ is a Markov chain with transition distribution Q^{a+b}, the $(a + b)$-step transition distribution of the original chain, we have the martingale representation

$$\frac{1}{n}\sum_{j=1}^{n} g_0(Y_j) = E[g_0(Y_0)] + \frac{1}{n}\sum_{j=1}^{n}\sum_{k=0}^{\infty}\left(g_k(Y_j) - g_{k+1}(Y_{j-1})\right) + O_P\left(\frac{\log n}{n}\right),$$

with

$$g_{k+1}(Y_{j-1}) = E\left(g_k(Y_j)|Y_{j-1}\right), \quad k = 0, 1, \ldots.$$

We compute

$$g_k(y) = \int Q^{k(a+b)}(y, \mathrm{d}x)g_0(x), \quad k = 1, 2, \ldots.$$

The above shows that the expansion

$$\frac{1}{n}\sum_{j=1}^{n} g(Z_j) = E[g(Z_1)] + \frac{1}{n}\sum_{j=1}^{n} B_j(g) + O_P\left(\frac{\log n}{n}\right),$$

holds with

$$B_j(g) = g(Z_j) - g_0\left(Y_{j-1}\right) + \sum_{k=0}^{\infty}\left(g_k(Y_j) - g_{k+1}\left(Y_{j-1}\right)\right), \quad j = 1, \ldots, n.$$

Note that $B_1(g), \dots, B_n(g)$ form a martingale difference array for the filtration $\mathcal{F}_j, j = 0, \dots, n$, where

$$\mathcal{F}_0 = \sigma(Y_0),\ \mathcal{F}_1 = \sigma(Y_0, Z_1),\ \mathcal{F}_2 = \sigma\big(Y_0, Z_1, Z_2\big),\ \dots, \mathcal{F}_n = \sigma\big(Y_0, Z_1, \dots, Z_n\big).$$

Indeed, $B_j(g)$ is $\mathcal{F}_j$-measurable and $E\big(B_j(g)|\mathcal{F}_{j-1}\big) = E\big(B_j(g)|Y_{j-1}\big) = 0$.

Let us now look at two special cases. Note that the estimator $\hat{\theta}_0$ can be expressed as

$$\hat{\theta}_0 = \frac{1}{n}\sum_{j=1}^{n} \phi(Z_j),$$

with

$$\phi(z) = \frac{1}{b}\sum_{i=1}^{b} f(z_{i-1}, z_i), \quad z = (z_0, \dots, z_b) \in S^{b+1}.$$

Then we obtain

$$\hat{\theta}_0 - \theta = \frac{1}{n}\sum_{j=1}^{n} B_j(\phi) + O_P\left(\frac{\log n}{n}\right).$$

Here the role of the g_k is played by ϕ_k which is calculated as

$$\phi_k(y) = \int M_k(y, \mathrm{d}x) f_*(x),$$

with

$$M_k(y, \mathrm{d}x) = \frac{1}{b}\sum_{j=a}^{a+b-1} Q^{k(a+b)+j}(y, \mathrm{d}x) \quad \text{and} \quad f_*(x) = \int Q(x, \mathrm{d}y) f(x, y).$$

Next, we look at the case $g = \Delta h$, where

$$\Delta h(z) = h(z_b) - h(z_0), \quad z = (z_0, \dots, z_b) \in S^{b+1},$$

and h is a bounded function from S to $\mathbb{R}$. Then we obtain

$$\frac{1}{n}\sum_{j=1}^{n} \big(h\big(X_{j(a+b)}\big) - h\big(X_{j(a+b)-b}\big)\big) = \frac{1}{n}\sum_{j=1}^{n} B_j(\Delta h) + O_P\left(\frac{\log n}{n}\right).$$

For the choice $g = \Delta h$, the role of g_k is played by

$$(\Delta h)_k(x) = \int Q^{(k+1)(a+b)}(x, \mathrm{d}y) h(y) - \int Q^{k(a+b)+a}(x, \mathrm{d}y) h(y), \quad k = 0, 1, \dots.$$

From this we derive

$$\hat{\theta}(h) - \theta = \frac{1}{n}\sum_{j=1}^{n} \big(B_j(\phi) - B_j(\Delta h)\big) + O_P\left(\frac{\log n}{n}\right).$$

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12747

The asymptotic variance of $\hat{\theta}(h)$ is then

$$E\big[\big(B_1(\phi) - B_1(\Delta h)\big)^2\big].$$

We use the minimizer h_* of the asymptotic variance with respect to h with $E[h(X_0)] = 0$. This optimal choice depends on the distribution of the Markov chain. Thus we will replace h_* by an estimator.

From now on we work with a finite state space $S = \{1, \dots, m\}$. Then a function h with mean zero can be written as a step function

$$h = \sum_{i=1}^{m} h_i \mathbf{1}_{\{i\}} = h_1 \mathbf{1}_{\{1\}} + \cdots + h_m \mathbf{1}_{\{m\}},$$

with coefficients $h_1, \dots, h_m$ satisfying $\sum_{i=1}^m h_i \pi_i = 0$ and with $\mathbf{1}_{\{i\}}$ denoting the indicator of the singleton set $\{i\}$, $i = 1, \dots, m$. For such h, we have $B_1(\Delta h) = \sum_{i=1}^n h_i B_1(\Delta \mathbf{1}_{\{i\}})$. This shows that the optimal h_* is obtained by minimizing

$$E\left[\left(B_1(\phi) - \sum_{i=1}^{m} h_i B_1\big(\Delta \mathbf{1}_{\{i\}}\big)\right)^2\right],$$

subject to the constraint $\sum_{j=1}^m h_i \pi_i = 0$. With the optimal h_* we associate the estimator

$$\hat{\theta}(h_*) = \frac{1}{n} \sum_{j=1}^{n} \left(\phi(Z_j) - \sum_{i=1}^{m} h_{*,i} B_j\big(\Delta \mathbf{1}_{\{i\}}\big)\right).$$

Since h_* is unknown, we must work with an estimator $\hat{h}_*$ of h_* and use

$$\hat{\theta}(\hat{h}_*) = \frac{1}{n} \sum_{j=1}^{n} \left(\phi(Z_j) - \sum_{i=1}^{m} \hat{h}_{*,i} B_j\big(\Delta \mathbf{1}_{\{i\}}\big)\right).$$

Suppose that $\hat{h}_*$ is a consistent estimator of h_* in the sense that $\hat{h}_{*,i}$ converges to $h_{*,i}$ in probability for $i = 1, \dots, m$. Then the following rates

$$\frac{1}{n} \sum_{j=1}^{n} B_j\big(\Delta \mathbf{1}_{\{i\}}\big) = O_P\big(n^{-1/2}\big), \quad i = 1, \dots, m,$$

imply the asymptotic equivalence of the plug-in estimator $\hat{\theta}(\hat{h}_*)$ and the estimator $\hat{\theta}(h_*)$, i.e.

$$\hat{\theta}(\hat{h}_*) = \hat{\theta}(h_*) + o_P\big(n^{-1/2}\big).$$

Let us now address how to construct a consistent estimator of h_*. We shall take $\hat{h}_*$ to be the minimizer of

$$\frac{1}{n} \sum_{j=1}^{n} \left(\hat{B}_j(\phi) - \sum_{i=1}^{m} h_i \hat{B}_j\big(\Delta \mathbf{1}_{\{i\}}\big)\right)^2,$$

subject to the constraint $\sum_{i=1}^m h_i\hat\pi_i = 0$. Here

$$\hat\pi_i = \frac{1}{1+n(b+1)}\sum_{\ell\in O}\mathbf{1}[X_\ell = i],$$

is the empirical estimator of π_i, $i = 1, \dots, m$, with $O = \{0, a, \dots, a+b, \dots, n(a+b)\}$ denoting the indices of the observations $X_0, X_a, \dots, X_{a+b}, \dots, X_{n(a+b)}$, and

$$\hat B_j(g) = g(Z_j) - \hat g_0\big(Y_{j-1}\big) + \sum_{k=1}^{r_n}\big(\hat g_k\big(Y_j\big) - \hat g_{k+1}\big(Y_{j-1}\big)\big),$$

is an estimator of $B_j(g)$. Here r_n is a positive integer allowed to grow to infinity slowly, and

$$\hat g_0(y) = \sum_{(z_0,\dots,z_b)\in S^{b+1}}\hat Q(y,z_0)\hat Q(z_0,z_1)\cdots\hat Q(z_{b-1},z_b)g(z_0,\dots,z_b),$$

and

$$\hat g_k(y) = \sum_{s\in S}\hat Q^k(y,s)\hat g_0(s), \quad k = 1, 2, \dots, r_n+1,$$

are plug-in estimators of $g_0(y), \dots, g_{r_n+1}(y)$ obtained be replacing the unknown transition matrix Q by the empirical transition matrix $\hat Q$ defined by

$$\hat Q(x,y) = \frac{\sum_{\ell:\{\ell-1,\ell\}\in O}\mathbf{1}[X_{\ell-1}=x, X_\ell = y]}{\sum_{y'=1}^m\sum_{\ell:\{\ell-1,\ell\}\in O}\mathbf{1}[X_{\ell-1}=x, X_\ell = y']}.$$

To state our theorem we also introduce the functions

$$\psi_i = \mathbf{1}_{\{i\}} - \frac{\pi_i}{\pi_m}\mathbf{1}_{\{m\}}, \quad i = 1, \dots, m-1.$$

Theorem 1. Suppose $r_n \to \infty$ and $n^{-1/4}r_n \to 0$ and the $(m-1)\times(m-1)$ matrix D with entries

$$D_{ik} = E[B_1(\Delta\psi_i)B_1(\Delta\psi_k)], \quad i,k = 1, \dots, m-1,$$

is invertible. Then $\hat h_*$ is a consistent estimator of h_*.

Proof. The constraint $\sum_{i=1}^m h_i\pi_i = 0$ gives $h_m = -\sum_{i=1}^{m-1}\pi_i h_i/\pi_m$. This lets us write a function h with mean $E[h(Y_0)] = 0$ as $h = h_1\psi_1 + \cdots + h_{m-1}\psi_{m-1}$. Thus the optimal h_* can be expressed as

$$h_* = \sum_{i=1}^{m-1}\beta_i^*\psi_i = \sum_{i=1}^{m-1}\beta_i^*\mathbf{1}_{\{i\}} - \sum_{i=1}^{m-1}\frac{\beta_i^*\pi_i}{\pi_m}\mathbf{1}_{\{m\}},$$

where $\beta_1^*, \dots, \beta_{m-1}^*$ are the minimizers of

$$E\left[\left(B_1(\phi) - \sum_{i=1}^{m-1}\beta_i B_1(\Delta\psi_i)\right)^2\right].$$

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12747

The minimizing vector is given by

$$\beta^* = \begin{bmatrix} \beta_1^* \\ \vdots \\ \beta_{m-1}^* \end{bmatrix} = D^{-1}R \quad \text{with} \quad R = \begin{bmatrix} E[B_1(\Delta\psi_1)B_1(\phi)] \\ \vdots \\ E[B_1(\Delta\psi_{m-1})B_1(\phi)] \end{bmatrix}.$$

In a similar fashion we can show that

$$\hat{h}_* = \sum_{i=1}^{m-1} \hat{\beta}_i \hat{\psi}_i = \sum_{i=1}^{m-1} \hat{\beta}_i \mathbf{1}_{\{i\}} - \sum_{i=1}^{m-1} \frac{\hat{\beta}_i \hat{\pi}_i}{\hat{\pi}_m} \mathbf{1}_{\{m\}},$$

where

$$\hat{\psi}_i = \mathbf{1}_{\{i\}} - \frac{\hat{\pi}_i}{\hat{\pi}_m} \mathbf{1}_{\{m\}}, \quad i = 1, \ldots, m-1,$$

and $\hat{\beta}_1^*, \ldots, \hat{\beta}_{m-1}^*$ are minimizers of

$$\frac{1}{n} \sum_{j=1}^{n} \left(\hat{B}_j(\phi) - \sum_{i=1}^{m-1} \beta_i \hat{B}_1(\Delta\hat{\psi}_i) \right)^2.$$

The minimizing vector satisfies

$$\hat{\beta}^* = \begin{bmatrix} \hat{\beta}_1^* \\ \vdots \\ \hat{\beta}_{m-1}^* \end{bmatrix} = \hat{D}^{-1}\hat{R} \quad \text{with} \quad \hat{R} = \frac{1}{n} \sum_{j=1}^{n} \begin{bmatrix} \hat{B}_j(\Delta\hat{\psi}_1)\hat{B}_j(\phi)] \\ \vdots \\ \hat{B}_j(\Delta\hat{\psi}_{m-1})\hat{B}_j(\phi) \end{bmatrix},$$

on the event where the matrix $\hat{D}$ with entries

$$\hat{D}_{i,k} = \frac{1}{n} \sum_{j=1}^{n} \hat{B}_j(\Delta\hat{\psi}_i)\hat{B}_j(\Delta\hat{\psi}_k), \quad i, k = 1, \ldots, m-1,$$

is invertible.

We shall show in Lemma 2 of Section 4 that

$$\max_{1 \le j \le n} |\hat{B}_j(g) - B_j(g)| = o_P(1),$$

for every function g from S^{b+1} into $\mathbb{R}$. Using this with $g = \phi$ and $g = \Delta\mathbf{1}_{\{i\}}$, $i = 1, \ldots, m$, and the fact that $\hat{\pi}_i$ is a consistent estimator of π_i for $i = 1, \ldots, m$, one derives

$$\begin{aligned} \frac{1}{n} \sum_{j=1}^{n} \hat{B}_j(\Delta\hat{\psi}_i)\hat{B}_j(\phi) &= \frac{1}{n} \sum_{j=1}^{n} \hat{B}_j(\Delta\mathbf{1}_{\{i\}})\hat{B}_j(\phi) - \frac{\hat{\pi}_i}{\hat{\pi}_m} \frac{1}{n} \sum_{j=1}^{n} \hat{B}_j(\Delta\mathbf{1}_{\{m\}})\hat{B}_j(\phi) \\ &= \frac{1}{n} \sum_{j=1}^{n} B_j(\Delta\mathbf{1}_{\{i\}})B_j(\phi) - \frac{\pi_i}{\pi_m} \frac{1}{n} \sum_{j=1}^{n} B_j(\Delta\mathbf{1}_{\{m\}})B_j(\phi) + o_P(1) \\ &= \frac{1}{n} \sum_{j=1}^{n} B_j(\Delta\psi_i)B_j(\phi) + o_P(1), \end{aligned}$$

and thus

$$\frac{1}{n}\sum_{j=1}^{n}\hat{B}_j(\Delta\hat{\psi}_i)\hat{B}_j(\phi) = E[B_1(\Delta\psi_i)B_1(\phi)] + o_P(1),$$

for $i = 1, \dots, m-1$. Similarly, one obtains

$$\frac{1}{n}\sum_{j=1}^{n}\hat{B}_j(\Delta\hat{\psi}_i)\hat{B}_j(\Delta\hat{\psi}_k) = E[B_1(\Delta\psi_i)B_1(\Delta\psi_k)] + o_P(1),$$

for $i, k = 1, \dots, m-1$. We conclude that $\hat{R}$ is a consistent estimator of R and $\hat{D}$ is a consistent estimator of D. Since D is invertible, we conclude that $\hat{\beta}^*$ is a consistent estimator of β^*, and this implies that $\hat{h}_*$ is a consistent estimator of h_*. ■

3. SIMULATIONS

We study the behavior of our estimator with simulations for a simple scenario with state space $S = \{1, 2, 3, 4\}$, and with $a = 2$ and $b = 1$, so a gap of length $a = 2$ is followed by a block of $b + 1 = 2$ observations, and we observe $X_0, X_2, X_3, X_5, X_6, \dots, X_{3n}$. We compare our estimator $\hat{\theta}(\hat{h}_*)$ with the empirical estimator $\hat{\theta}_0$, for various functions $f(x, y)$, transition matrices M, and $n = 50{,}100, 150{,}200$. The simulations are based on 4000 iterations. The tables contain n times the simulated mean square errors (MSE) of the estimators and the relative efficiency $\mathrm{RE} = \mathrm{MSE}(\hat{\theta}_0)/\mathrm{MSE}(\hat{\theta}(h_*))$. The values were computed with R and are rounded to five and three decimal places respectively. A value $\mathrm{RE} > 1$ indicates the superiority of our approach.

We begin with the three transition matrices M_1, M_2 and M_3 given below. The stationary distribution for each matrix is the uniform distribution, $\pi = (1/4, 1/4, 1/4, 1/4)$.

$$M_1 = \begin{bmatrix} 1/8 & 1/8 & 1/4 & 1/2 \\ 1/8 & 1/4 & 1/2 & 1/8 \\ 1/4 & 1/2 & 1/8 & 1/8 \\ 1/2 & 1/8 & 1/8 & 1/4 \end{bmatrix} \quad M_2 = \begin{bmatrix} 1/2 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/2 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/2 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/2 \end{bmatrix} \quad M_3 = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}$$

The MSEs of our estimator and the empirical estimator as well as the relative efficiency are given in Table I.

We now examine the behavior of our estimator if the stationary distribution of the chain is *not* the uniform distribution. We consider the transition matrices M_4, M_5 and M_6 given below, with stationary distributions $\pi = (0.2, 0.2, 0.266 \dots, 0.33 \dots)$, $(0.2, 0.2, 0.3, 0.3)$ and $(0.1, 0.2, 0.3, 0.4)$ respectively. The last matrix, M_6, corresponds to the special case that the data are i.i.d.

$$M_4 = \begin{bmatrix} 0.2 & 0.2 & 0.2 & 0.4 \\ 0.2 & 0.2 & 0.2 & 0.4 \\ 0.2 & 0.2 & 0.2 & 0.4 \\ 0.2 & 0.2 & 0.4 & 0.2 \end{bmatrix} \quad M_5 = \begin{bmatrix} 0.2 & 0.2 & 0.2 & 0.4 \\ 0.2 & 0.2 & 0.4 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.4 \\ 0.2 & 0.2 & 0.4 & 0.2 \end{bmatrix} \quad M_6 = \begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.4 \\ 0.1 & 0.2 & 0.3 & 0.4 \\ 0.1 & 0.2 & 0.3 & 0.4 \\ 0.1 & 0.2 & 0.3 & 0.4 \end{bmatrix}$$

The simulation results for these matrices are provided in Table II.

Why are the relative efficiencies so different for different functions f? For observations $X_0, \dots, X_n$ from a fully observed non-parametric Markov chain, the empirical estimator $(1/n)\sum_{j=1}^{n} f(X_{j-1}, X_j)$ is asymptotically efficient for $\theta = E[f(X_0, X_1)]$. For observations $X_0, X_2, X_3, X_5, X_6, \dots, X_{3n}$ considered here, with every third realization of the Markov chain missing, the estimator $\hat{\theta}_0 = (1/n)\sum_{j=1}^{n} f(X_{3j-1}, X_{3j})$ uses only half the available pairs of observations, ignoring the pairs $(X_{3(j-1)}, X_{3j-1})$ that are separated by a gap. One might conjecture that the asymptotic variance of

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12747

Table I. The table entries are n times the mean squared errors for the empirical estimator $\hat{\theta}_0$ and for our estimator $\hat{\theta}(\hat{h}_*)$ for various functions f and transition matrices M when $n = 50$

		nMSE		RE			
M	$f(x,y)$	$\hat{\theta}_0$	$\hat{\theta}(\hat{h}_*)$	$n = 50$	100	150	200
M_1	$\mathbf{1}[x > y]$	0.20856	0.07046	2.960	3.144	3.080	3.455
	x/y	1.21480	0.26822	4.865	4.865	4.705	5.217
	$4x/(x+y)$	0.48158	0.0080	602.621	605.333	636.142	665.854
	$\log(x/(1+y))$	0.43177	0.01329	32.490	35.773	36.774	36.059
	$\max(x,y)$	0.67036	0.77102	0.869	0.934	0.962	0.975
	$xy/4$	0.89172	1.00182	0.890	0.943	0.960	0.975
M_2	$\mathbf{1}[x > y]$	0.18680	0.11233	1.663	1.812	1.982	1.940
	x/y	0.63671	0.15127	4.209	4.649	4.986	4.872
	$4x/(x+y)$	0.28913	0.00069	416.483	418.912	440.513	437.387
	$\log(x/(1+y))$	0.26070	0.02970	8.777	9.675	9.475	9.750
	$\max(x,y)$	1.18596	1.49045	0.796	0.912	0.933	0.960
	$xy/4$	1.64064	1.94167	0.845	0.934	0.953	0.970
M_3	$\mathbf{1}[x > y]$	0.26130	0.08334	3.135	3.471	3.579	3.489
	x/y	0.88804	0.15536	5.716	6.345	6.497	6.611
	$4x/(x+y)$	0.45836	0.00128	358.570	384.219	387.795	417.381
	$\log(x/(1+y))$	0.39884	0.01840	21.680	23.287	24.864	24.081
	$\max(x,y)$	0.84995	0.92632	0.918	1.026	1.042	1.047
	$xy/4$	1.12398	1.19860	0.938	1.038	1.066	1.085

Note: The last four columns provide the relative efficiencies for samples sizes $n = 50$,100, 150 and 200.

Table II. The table entries are n times the mean squared errors ($n = 50$) and the relative efficiency as in Table I, now with transition matrices M_4, M_5 and M_6

		nMSE		RE			
M	$f(x,y)$	$\hat{\theta}_0$	$\hat{\theta}(\hat{h}_*)$	$n = 50$	100	150	200
M_4	$\mathbf{1}[x > y]$	0.24842	0.08772	2.832	3.126	3.139	2.983
	x/y	0.96055	0.18451	5.206	5.624	5.959	5.755
	$4x/(x+y)$	0.43647	0.00098	444.776	475.583	450.609	470.466
	$\log(x/(1+y))$	0.37619	0.01790	21.020	25.812	25.138	24.734
	$\max(x,y)$	0.72452	0.83773	0.865	0.935	0.963	0.976
	$xy/4$	1.16930	1.37596	0.850	0.926	0.939	0.961
M_5	$\mathbf{1}[x > y]$	0.23344	0.08370	2.789	3.052	3.069	2.997
	x/y	0.89833	0.18706	4.802	5.437	5.828	6.093
	$4x/(x+y)$	0.41521	0.00099	419.700	464.1001	471.879	459.925
	$\log(x/(1+y))$	0.36790	0.01765	20.839	24.349	24.002	26.072
	$\max(x,y)$	0.73561	0.85871	0.856	0.939	0.949	0.970
	$xy/4$	1.06267	1.26272	0.842	0.917	0.955	0.962
M_6	$\mathbf{1}[x > y]$	0.23204	0.09357	2.480	2.755	2.806	2.743
	x/y	0.66162	0.14203	4.658	5.405	5.398	5.854
	$4x/(x+y)$	0.31495	0.00964	32.671	285.273	589.983	546.436
	$\log(x/(1+y))$	0.26491	0.01332	19.895	23.901	25.144	23.842
	$\max(x,y)$	0.46617	0.54032	0.863	0.940	0.954	0.970
	$xy/4$	1.17293	1.49311	0.786	0.882	0.920	0.939

DOI: 10.1111/jtsa.12747 *Journal of Time Series Analysis* published by John Wiley & Sons Ltd.

$\hat{\theta}_0$ is therefore approximately twice the variance of an efficient estimator that exploits the information provided by both types of pairs, with and without gap, and that a modified estimator

$$\hat{\theta}(h) = \frac{1}{n}\sum_{j=1}^{n}\left(f(X_{3j-1}, X_{3j}) - h(X_{3j-1}) + h(X_{3j})\right),$$

will have about half the variance of $\hat{\theta}_0$ at best.

This conjecture is false. Indeed, our simulations show that the improvement of $\hat{\theta}(h)$ over $\hat{\theta}_0$ can be considerably greater, especially for functions f that are close to antisymmetric, and in particular, close to an antisymmetric function of the form $f(x,y) = g(x) - g(y)$. If f has exactly this form, for the Markov chain without gaps, the empirical estimator is a telescoping function $(1/n)\sum_{j=1}^{n}(g(X_{j-1}) - g(X_j)) = (1/n)(g(X_0) - g(X_n))$, and the asymptotic variance of the standardized empirical estimator is zero. The variance will be close to zero if f is close to this form. On the other hand, for the Markov chain with gaps, $\hat{\theta}_0$ is close to $(1/n)\sum_{j=1}^{n}(g(X_{3j-1}) - g(X_{3j}))$, which does not telescope. However the improved estimator $\hat{\theta}(h)$ with $h = g$ is close to

$$\frac{1}{n}\sum_{i=1}^{n}\left(g(X_{3j-1}) - g(X_{3j}) - g(X_{3j-1}) + g(X_{3j})\right) = 0,$$

an extreme variance reduction over $\hat{\theta}_0$.

Suppose now that f is close to *symmetric* rather than antisymmetric, so $f(x,y) = f(y,x)$. Suppose also that the Markov chain is close to *reversible*, so the joint distributions of (X_1, X_m) and (X_m, X_1) are close for $m = 2, 3, \ldots$. Then $\hat{\theta}(h)$ will not improve much over $\hat{\theta}_0$, whatever h: Subtracting $h(x) - h(y)$ from $f(x,y)$ will change $f(x,y)$ by roughly the same amount, but with the opposite sign, when (x,y) is replaced by the reflected point (y,x). Also, $P[(X_1, X_m) = (x,y)]$ is about $P[(X_1, X_m) = (y,x)]$. Hence $\hat{\theta}(h)$ and therefore $\hat{\theta}(\hat{h}_*)$ will have about the same asymptotic variance as $\hat{\theta}_0$. This is illustrated by the simulations for approximately symmetric f.

The additional randomness introduced by estimating a small optimal correction $h(x) - h(y)$ with $\hat{h}(x) = \hat{\beta}^\top \mathbf{1}[x = \cdot]$ leads occasionally to a slight variance *increase* of $\hat{\theta}(\hat{h}_*)$ over $\hat{\theta}_0$.

4. TECHNICAL DETAILS

Let S be a finite set with at least two elements. For a positive integer k, we let $\mathcal{G}_k$ denote the vector space of all functions from S^k to $\mathbb{R}$ and set

$$\|g\|_\infty = \max_{x \in S^k} |g(x)|, \quad g \in \mathcal{G}_k.$$

We abbreviate $\mathcal{G}_1$ by $\mathcal{G}$ and $\mathcal{G}_2$ by $\mathcal{M}$. On $\mathcal{M}$ we also introduce the norm

$$\|M\| = \max_{s \in S}\sum_{t \in S}|M(s,t)|, \quad M \in \mathcal{M}.$$

We let $\mathcal{M}_1 = \{M \in \mathcal{M} : \|M\| \le 1\}$ denote the unit ball in $\mathcal{M}$ for this norm. For $M_1, \ldots, M_k$ in $\mathcal{M}$, we define a linear operator $M_1 \otimes \cdots \otimes M_k$ from $\mathcal{G}_k$ into $\mathcal{G}$ which maps $g \in \mathcal{G}_k$ to the function $(M_1 \otimes \cdots \otimes M_k)g$ defined by

$$(M_1 \otimes \cdots \otimes M_k)g(s_0) = \sum_{s_1, \ldots, s_k \in S} M_1(s_0, s_1)M_2(s_1, s_2)\cdots M_k(s_{k-1}, s_k)g(s_1, \ldots, s_k), \quad s_0 \in S.$$

If $M_1 = \cdots = M_k$, we abbreviate $M_1 \otimes \cdots \otimes M_k$ by $M^{\otimes k}$. It is easy to verify the inequality

$$\|(M_1 \otimes \cdots \otimes M_k)g\|_\infty \le \prod_{j=1}^{k}\|M_k\|\,\|g\|_\infty, \quad g \in \mathcal{G}_k.$$

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12747

For $N_1, \dots, N_k, M_1, \dots, M_k$ in $\mathscr{M}_1$, we have the inequality

$$\|(N_1 \otimes \cdots \otimes N_k)g - (M_1 \otimes \cdots \otimes M_k)g\|_\infty \le \sum_{j=1}^{k} \|N_j - M_j\| \|g\|_\infty, \quad g \in \mathscr{G}_k. \tag{4.1}$$

This is verified by successively replacing the factors N_i by M_i in the product $N_1 \otimes \cdots \otimes N_k$ and using the triangle inequality.

For M_1 and M_2 in $\mathscr{M}$, we let $M_1 \odot M_2$ denote the matrix product of M_1 and M_2 defined by

$$(M_1 \odot M_2)(s,t) = \sum_{u \in S} M_1(s,u) M_2(u,t), \quad s,t \in \mathbb{R}.$$

We write M^k for the k-fold matrix product $M \odot \cdots \odot M$ of M. We have

$$\|M \odot N\| \le \|M\| \|N\|, \quad M, N \in \mathscr{M} \tag{4.2}$$

and thus

$$\|M^k\| \le \|M\|^k.$$

Lemma 1. Consider M, N in $\mathscr{M}$ and a positive integer k. Then we have the following inequalities:

$$\|(M+N)^k\| \le \sum_{j=0}^{k} \binom{k}{j} \|M\|^j \|N\|^{k-j}, \tag{4.3}$$

$$\|N^k - M^k\| \le \sum_{j=0}^{k-1} \binom{k}{j} \|M\|^j \|N - M\|^{k-j} \tag{4.4}$$

and, if M in $\mathscr{M}_1$,

$$\|N^k - M^k\| \le k \|N - M\| \exp(k \|N - M\|). \tag{4.5}$$

Proof. For a subset I of $\{1, \dots, k\}$, set

$$T(j, I) = M \mathbf{1}(j \in I) + N \mathbf{1}(j \notin I), \quad j = 1, \dots, k.$$

Inequality (4.3) is a consequence of the identity

$$(M+N)^k = \sum_{I \subset \{1, \dots, k\}} T(1, I) \odot \cdots \odot T(k, I)$$

and the inequality (4.2), while inequality (4.4) follows from the identity with the role of N played by $N - M$ and (4.2).

To prove (4.5) we use (4.4) with $\|M\| \le 1$ and the inequality

$$\binom{k}{j} \le \frac{k^j}{j!} \le \frac{k^j}{(j-1)!}, \quad 1 \le j \le k.$$

This and the substitution $j = k - i$ give

$$\|N^k - M^k\| \le \sum_{i=0}^{k-1} \binom{k}{i} \|N - M\|^{k-i} = \sum_{j=1}^{k} \binom{k}{j} \|N - M\|^j \le \sum_{j=1}^{k} \frac{k^j}{(j-1)!} \|N - M\|^j.$$

Factoring out one term $k\|N - M\|$, extending the summation to infinity and using the Taylor expansion of the exponential function yields the desired result (4.5). ∎

Lemma 2. Suppose $r_n \to \infty$ and $r_n = o(n^{1/4})$. Then we have

$$\max_{1\le j\le n} |\hat{B}_j(g) - B_j(g)| = o_P(1),$$

for every g in $\mathcal{G}_{b+1}$.

Proof. Fix $g \in \mathcal{G}_{b+1}$. The difference $\hat{B}_j(g) - B_j(g)$ equals $T_j - U_j + V_j$, where

$$T_j = \sum_{k=0}^{r_n} \big(\hat{g}_k(Y_j) - g_k(Y_j)\big) = \sum_{k=0}^{r_n} \Big((\hat{Q}^{k(a+b)} \otimes \hat{Q}^a \otimes \hat{Q}^{\otimes b})g - (Q^{k(a+b)} \otimes Q^a \otimes Q^{\otimes b})g\Big)\big(Y_j\big),$$

$$U_j = \sum_{k=0}^{r_n+1} \big(\hat{g}_k\big(Y_{j-1}\big) - g_k\big(Y_{j-1}\big)\big) = \sum_{k=0}^{r_n+1} \Big(\hat{Q}^{k(a+b)} \otimes \hat{Q}^a \otimes \hat{Q}^{\otimes b} g - Q^{k(a+b)} \otimes Q^a \otimes Q^{\otimes b} g\Big)\big(Y_{j-1}\big),$$

and

$$V_j = \sum_{k=r_n+1}^{\infty} \big(Q^{k(a+b)} g_0\big(Y_j\big) - Q^{(k+1)(a+b)} g_0\big(Y_{j-1}\big)\big)\,.$$

Here we used the identities $g_0 = \big(Q^a \otimes Q^{\otimes b}\big)g$ and $\hat{g}_0 = \big(\hat{Q}^a \otimes \hat{Q}^{\otimes b}\big)g$.

Note the identities $\|Q\| = 1$ and $\|\hat{Q}\| = 1$. With the help of (4.1) and (4.5) we derive

$$\begin{aligned}
\max_{1\le j\le n} |T_j| &\le \sum_{k=0}^{r_n} \|\hat{Q}^{k(a+b)} \otimes \hat{Q}^a \otimes \hat{Q}^{\otimes b} g - Q^{k(a+b)} \otimes Q^a \otimes Q^{\otimes b} g\|_\infty \\
&\le \|g\|_\infty \sum_{k=0}^{r_n} \Big(\|\hat{Q}^{k(a+b)} - Q^{k(a+b)}\| + \|\hat{Q}^a - Q^a\| + b\|\hat{Q} - Q\|\Big) \\
&\le \|g\|_\infty \sum_{k=0}^{r_n} (k+1)(a+b)\|\hat{Q} - Q\| \exp\big(r_n(a+b)\|\hat{Q} - Q\|\big) \\
&\le \|g\|_\infty (r_n+1)^2(a+b)\|\hat{Q} - Q\| \exp\big(r_n(a+b)\|\hat{Q} - Q\|\big)\,,
\end{aligned}$$

and similarly

$$\max_{1\le j\le n} |U_j| \le (r_n+2)^2(a+b)\|\hat{Q} - Q\| \exp\big((r_n+1)(a+b)\|\hat{Q} - Q\|\big)\,.$$

As $r_n = o(n^{1/4})$ and $\|\hat{Q} - Q\| = O_P(n^{-1/2})$, these maxima converge in probability to zero.

Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12747

Finally, using inequality (2.1) in Greenwood and Wefelmeyer (1995), for some $0 < \alpha < 1$,

$$\begin{aligned}\max_{1\le j\le n}|V_j| &\le \sum_{k=r_n+1}^{\infty} \|Q^{k(a+b)}g_0 - Q^{(k+1)(a+b)}g_0\|_\infty \\ &\le \sum_{k=r_n+1}^{\infty} \left(\|Q^{k(a+b)}g_0 - E[g_0(Y_0)]\|_\infty + \|Q^{(k+1)(a+b)}g_0 - E[g_0(Y_0)]\|_\infty\right) \\ &\le 2\|g\|_\infty \sum_{k=r_n+1}^{\infty} \left(\alpha^{(a+b)k} + \alpha^{(a+b)(k+1)}\right) = o_P(1).\end{aligned}$$

In the last step we used the fact that the L_1 distance of two probability measures P_1 and P_2 equals twice their total variation distance,

$$\sup_t \left|\int t\, dP_1 - \int t\, dP_2\right| = 2\sup_A |P_1(A) - P_2(A)|.$$

Here the first supremum is over all measurable functions t satisfying $|t| \le 1$ and the second supremum over all sets A in the common domain of the measures. ■

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

Bickel PJ. 1993. *Estimation in semiparametric models*. In *Multivariate Analysis: Future Directions*, Rao CR (ed.). North-Holland, Amsterdam; 55–73.

Bondon P. 2005. Influence of missing values on the prediction of a stationary time series. *Journal of Time Series Analysis* **26**:519–525.

Efromovich S. 2014. Efficient non-parametric estimation of the spectral density in the presence of missing observations. *Journal of Time Series Analysis* **35**:407–427.

Gordin MI. 1969. The central limit theorem for stationary processes. *Soviet Mathematics Doklady* **10**:1174–1176.

Greenwood PE, Wefelmeyer W. 1995. Efficiency of empirical estimators for Markov chains. *Annals of Statistics* **23**:132–143.

Greenwood PE, Schick A, Wefelmeyer W. 2001. Comment on: Inference for semiparametric models: some questions and an answer. By Peter J. Bickel and Jaimyoung Kwon. *Statistica Sinica* **11**:892–906.

Lu Z, Hui YV. 2003. L_1 linear interpolator for missing values in time series. *Annals of the Institute of Statistical Mathematics* **55**:197–216.

Miller RB, Ferreiro O. 1984. *A strategy to complete a time series with missing observations*. In *Time Series Analysis of Irregularly Observed Data*. Lecture Notes in Statistics, Vol. 25, Parzen E (ed.). Springer, New York; 251–275.

Müller UU, Schick A, Wefelmeyer W. 2008. *Estimators for partially observed Markov chains*. In *Statistical Models and Methods for Biomedical and Technical Systems*, Vonta F, Nikulin M, Limnios N, Huber-Carol C (eds.). Birkhäuser, Boston; 419–433.

Pasanisi A, Fu S, Bousquet N. 2012. Estimating discrete Markov models from various incomplete data schemes. *Computational Statistics & Data Analysis* **56**:2609–2625.

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

DOI: 10.1111/jtsa.12747 *Journal of Time Series Analysis* published by John Wiley & Sons Ltd.