Journal of Statistical Planning and Inference , 211, 90-106.

New robust confidence intervals for the mean under dependence

Martial Longla , Magda Peligrad

[a] Department of Mathematics, University of Mississippi, University, MS 38677, USA
[b] Department of Mathematical Sciences, University of Cincinnati, PO Box 210025, Cincinnati, OH 45221-0025, USA

A B S T R A C T

The goal of this paper is to indicate a new method for constructing normal confidence intervals for the mean, when the data is coming from stochastic structures with possibly long memory, especially when the dependence structure is not known or even the existence of the density function. More precisely we introduce a random smoothing suggested by the kernel estimators for the regression function. The normal confidence intervals are constructed under the sole condition that the sequence is ergodic and has finite second moments and a mild condition on the sample variance. Applications are presented for linear processes and reversible Markov chains with long memory.

1. Introduction

1.1. Motivation

Let us suppose that we have a stationary sequence $(Y_i)_{i\in Z}$ with finite variance ($\text{var}(Y_0) = \sigma_Y^2 < \infty$). Denote by $\mu_Y = EY_0$, the expected value of Y. Also, denote as usual the sample mean by $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. If, for instance, the sequence is ergodic, then by the Birkhoff ergodic theorem it is well-known that, $\lim_{n\to\infty} \bar{Y}_n = \mu_Y$ almost surely. If additional information on the dependence structure of $(Y_i)_{i\in Z}$ is available, such as martingale-like conditions or mixing conditions, we can derive a central limit theorem for $\sqrt{n}(\bar{Y}_n - \mu_Y)$, which naturally leads to the construction of confidence intervals for μ_Y based on the standard normal distribution. Without any other information on the dependence structure of $(Y_i)_{i\in Z}$, obviously, such a sequence might not obey the central limit theorem, and thus the sample mean becomes useless for confidence intervals. In this paper, we indicate a way to construct normal confidence intervals for μ_Y based on a smoothing method inspired by the Nadaraya–Watson estimators. These confidence intervals can be applied to a wide range of ergodic stochastic processes with finite second moment, therefore including processes with long-memory.

The procedure we propose is the following. The data $(Y_i)_{1\le i\le n}$ consists of a sample from a stationary sequence $(Y_i)_{i\in Z}$. Independently of $(Y_i)_{1\le i\le n}$ we generate a random sample $(X_i)_{1\le i\le n}$, from a distribution with bounded density $f(x)$, continuous at the origin, with $f(0) \neq 0$. In addition we select a sequence of positive constants called bandwidths $(h_n)_{n\in N}$ satisfying the condition

$$h_n \to 0 \text{ and } nh_n \to \infty \text{ as } n \to \infty, \tag{1}$$

and a kernel K that satisfies the following property

$$K \text{ is a symmetric bounded density function.} \tag{2}$$

Our estimator is

$$\hat{r}_n = \frac{1}{nE(K(X_1/h_n))} \sum_{i=1}^{n} Y_i K(\frac{1}{h_n} X_i). \tag{3}$$

We shall see that, under the condition

$$\bar{Y}_n \to \mu_Y \text{ in probability as } n \to \infty, \tag{4}$$

$\hat{r}_n$ is a consistent estimator of μ_Y. We mention that condition (4) is a very weak condition. By Lemma 7 in Appendix condition (4) is equivalent to apparently stronger condition

$$\text{var}(\bar{Y}_n) \to 0 \text{ as } n \to \infty. \tag{5}$$

If the sequence $(Y_i)_{i\in Z}$ is ergodic, by the Von Newman's mean ergodic theorem, (5) holds (see Theorem 8.1 in Eisner et al. (2015)). As a matter of fact the consistency result does not require the ergodicity of $(Y_i)_{i\in Z}$. Condition (5) also holds if, for instance, $\text{cov}(Y_0, Y_n) \to 0$.

1.2. Main results

We shall provide a rate of convergence via a central limit theorem, some confidence intervals for μ_Y, a functional central limit theorem. We also discuss the optimal bandwidth which minimizes the mean square error and provide the proper bandwidths for several examples. A simulation study is proposed for several models. To establish these results, we use the independence structure of the smoothing sequence $(X_i)_{i\in Z}$ that allows us not to restrict the dependence structure of $(Y_i)_{i\in Z}$ and also not to impose the existence of the density of Y. The closest idea to this one is the block-wise bootstrap. For instance, in the paper by Peligrad (1998), the central limit theorem for the mean is obtained via bootstrap smoothing, for a sequence that does not satisfy the CLT, but rather satisfies some restrictive mixing conditions.

1.3. Structure of the paper

The rest of this paper is organized as follows. In Section 2 we provide our results. In Section 3 we provide a brief discussion of the methods. In Section 4, we provide the proofs of our main results. In Section 5, we discuss the size of the optimal bandwidth and optimal kernel to be used in the confidence intervals. In Section 6 we provide several applications to processes with long memory. We show here that we can construct with our method confidence intervals for the mean of the population without estimating the variance or the memory parameter. In Section 7, we provide a simulation study to support our results. Simulations show that our confidence intervals, obtained without estimation of the memory parameter are statistically significant and very closely estimate the confidence intervals. Here we used the 90%, 95% and 99% confidence level for several models, including the Frechet copula family, the Clayton copula family and the ARFIMA$(0, d, 0)$ process. In Section 8 we have several remarks and conclusions. In Appendix, we give an auxiliary result.

2. Formulation of the main results

In the sequel we denote by $\Rightarrow$ the convergence in distribution. For positive sequences of numbers $a_n = O(b_n)$ means $\limsup_{n\to\infty} a_n/b_n < \infty$; $a_n = o(b_n)$ means $\lim_{n\to\infty} a_n/b_n = 0$. We use the notation $a_n \sim b_n$ for $\lim_{n\to\infty} a_n/b_n = 1$. Our first result is the consistency of $\hat{r}_n$.

Proposition 1. *Assume that $(Y_i)_{i\in Z}$ is a stationary sequence with finite second moments, satisfying (4). Also assume that conditions (1) and (2) are satisfied. Then, the following consistency result holds for the mean squared error:*

$$\text{MSE}(\hat{r}_n) := E(\hat{r}_n - \mu_Y)^2 \to 0 \text{ as } n \to \infty.$$

In order to establish the central limit theorem we need a stronger condition on the bandwidth sequence:

$$\sqrt{nh_n}(\bar{Y}_n - \mu_Y) \to^P 0, \tag{6}$$

which is implied by

$$nh_n \text{var}(\bar{Y}_n) \to 0. \tag{7}$$

Note that we can always find a sequence $(h_n)_{n\geq 1}$ satisfying both conditions (1) and (7), provided (5). In particular such a selection is possible for every stationary and ergodic sequence having finite second moments.

We establish the following theorem:

Theorem 2. *Assume that $(Y_i)_{i\in Z}$ is a stationary and ergodic sequence with finite second moments. Let $(h_n)_{n\geq 1}$ be a sequence of positive constants satisfying conditions (1) and (6). Also assume that K satisfies condition (2) and that $(X_i)_{i\in N}$ is an i.i.d. sequence of random variables, independent of $(Y_i)_{i\in Z}$, having a bounded density function $f(x)$, continuous at the origin, with $f(0)\neq 0$. Then we have*

$$\frac{\sqrt{nh_n}}{\sqrt{\overline{Y_n^2}}}(\hat{r}_n-\mu_Y)\Rightarrow N(0,\frac{1}{f(0)}\int K^2(x)dx).$$

where $\overline{Y_n^2}=\sum_{i=1}^n Y_i^2/n$ and $\hat{r}_n$ is defined by (3).

Based on Theorem 2 we can construct confidence intervals for the mean:

Corollary 3. *Under the conditions of Theorem 2, for $0<\alpha<1$, a $(1-\alpha)100\%$ confidence interval for μ_Y is*

$$\left(\hat{r}_n-z_{\alpha/2}\left(\frac{\overline{Y_n^2}\int K^2(x)dx}{nh_nf(0)}\right)^{1/2},\ \hat{r}_n+z_{\alpha/2}\left(\frac{\overline{Y_n^2}\int K^2(x)dx}{nh_nf(0)}\right)^{1/2}\right), \tag{8}$$

where $P(-z_{\alpha/2}<Z<z_{\alpha/2})=1-\alpha$ and Z is a standard normal variable.

Let us notice that, at no extra cost, our result can also be formulated as a functional CLT. If we consider the stochastic process

$$\hat{r}_n(t)=\frac{1}{nE(K(X_1/h_n))}\sum_{i=1}^{[nt]}Y_iK(\frac{1}{h_n}X_i),$$

from the proof of Theorem 2 and Donsker's theorem (see Theorem 8.2 in Billingsley, 1999) we obtain:

Corollary 4. *Under the conditions of Theorem 2 we have*

$$\sqrt{nh_n}(\hat{r}_n(t)-\mu_Y)/\sqrt{\overline{Y_n^2}}\Rightarrow\left(\frac{1}{f(0)}\int K^2(x)dx\right)^{1/2}W(t),$$

where $W(t)$ is the standard Brownian motion.

3. Discussion

Let us discuss now the relation of our estimator to the Nadaraya–Watson estimator. Given a sample $(X_i,Y_i)_{1\leq i\leq n}$ from a random vector (X,Y) on a probability space (Ω,K,P), the well-known Nadaraya–Watson estimator (see Nadaraya (1964) and Watson (1964), or pages 126–127 in Härdle (1991)) is defined by

$$\hat{m}_n(x)=\frac{1}{nh_n\hat{f}_n(x)}\sum_{i=1}^n Y_iK(\frac{1}{h_n}(X_i-x)),$$

where

$$\hat{f}_n(x)=\frac{1}{nh_n}\sum_{i=1}^n K(\frac{1}{h_n}(X_i-x)).$$

This estimator has been widely studied in the literature. For instance, when the vector (X,Y) has joint density $f(x,y)$, $\hat{m}_n(x)$ is used to estimate

$$E(Y|X=x)=r(x)=\int y[f(x,y)/f(x)]dy.$$

Under various smoothness assumptions on (X,Y) and various dependence assumptions on the process $(X_i,Y_i)_{i\in Z}$, the speed of convergence of $\hat{m}_n(x)$ to $r(x)$ was pointed out in numerous papers. The dependence structure considered in the literature is rather restrictive, of the weak dependence type, such as mixing conditions, function of mixing sequences, or martingale-like conditions. We mention for instance results in Bradley (1983), Collomb (1984), Peligrad (1992), Yoshihara (1994), Bosq (1996), Bosq et al. (1999), Laib and Louani (2010), Long and Qian (2013), and Hong and Linton (2016) among many others. Now, let us notice that if the variables $(X_i)_{i\in Z}$ are independent of $(Y_i)_{i\in Z}$, we have $E(Y|X)=E(Y)=\mu_Y$.

Since $\hat{f}_n(x)$ is a consistent estimator of $f(x)$, $\hat{m}_n(0)$ is asymptotically equivalent to (3). Actually this was the starting point of our paper. However we could not use any available results in the literature since our goal was to treat the case when no information is known about the dependence structure of the ergodic sequence $(Y_i)_{i\in Z}$ or the existence of the density of Y.

Let us also mention that our estimator is also formally related to the weighed estimators, so called W-estimators. (see sections 12.2.2 and 12.2.3 in Thode (2002)). These robust estimators are actually weighted M-estimators. The weighting could be based, for instance, on the Tukey's biweight function:

$$\omega(u) = (1-u^2)^2 I(|u| \le 1).$$

which can be normalized to be a kernel $K(u) = 15\omega(u)/16$. The W-estimators have the form

$$T_\omega = \frac{\sum_{i=1}^n Y_i K(X_{i,n})}{\sum_{i=1}^n K(X_{i,n})},$$

where $X_{i,n}$ is a certain function of the data, namely of $(Y_i)_{1\le i\le n}$. The rate of convergence of these robust estimators of location are theoretically difficult to study, unless $(Y_i)_{i\in Z}$ exhibit a known-type of dependence structure. Alternatively, we can use a different idea; we could construct the weighted estimators by taking (X_i) independent on (Y_i) and then set $X_{i,n} = X_i/h_n$. We obtain

$$T'_\omega = \frac{1}{nh_n\hat{f}_n(0)} \sum_{i=1}^n Y_i K(\frac{X_i}{h_n}),$$

We note that, T'_ω constructed above is asymptotically equivalent to (3).

4. Proofs of the main results

4.1. Proof of Proposition 1

For convenience, we shall drop the index n from the notation of h_n. Denote

$$V_{n,i} = h^{-1}K(X_i/h).$$

We shall compute first the bias

$$\text{Bias } \hat{r}_n = E(\hat{r}_n - \mu_Y) = \frac{\mu_Y}{E(V_{n,1})} E(\hat{f}_n(0)) - \mu_Y = 0.$$

This is an unbiased estimator with variance

$$\begin{aligned}
\text{var } \hat{r}_n &= E[\hat{r}_n - \mu_Y]^2 \\
&= E[\hat{r}_n - \frac{\mu_Y}{E(V_{n,1})}\hat{f}_n(0) + \frac{\mu_Y}{E(V_{n,1})}(\hat{f}_n(0) - E(\hat{f}_n(0)))]^2 \\
&= E[\hat{r}_n - \frac{\mu_Y}{E(V_{n,1})}\hat{f}_n(0)]^2 + \frac{\mu_Y^2}{E^2(V_{n,1})} \text{var } \hat{f}_n(0).
\end{aligned}$$

Simple computations based on stationarity show that the first term of the last sum becomes

$$\begin{aligned}
&E[\hat{r}_n - \frac{\mu_Y}{E(V_{n,1})}\hat{f}_n(0)]^2 = \frac{1}{E^2(V_{n,1})} E[\frac{1}{n}\sum_{i=1}^n (Y_i - \mu_Y)V_{n,i}]^2 \\
&= \frac{1}{E^2(V_{n,1})n^2}[n\sigma_Y^2 E(V_{n,1}^2) + 2\sum_{1\le i<j\le n} cov(Y_i, Y_j)(EV_{n,1})^2] \\
&= \frac{1}{E^2(V_{n,1})n^2}[n\sigma_Y^2 E(V_{n,i}^2) + (EV_{n,1})^2(\text{var } S_Y - n\sigma_Y^2)] \\
&= \frac{1}{E^2 V_{n,1} n^2}[n\sigma_Y^2 \text{var } V_{n,1} + (EV_{n,1})^2 \text{var } S_Y] \\
&= \frac{\text{var } \hat{f}_n(0)}{E^2(V_{n,1})}\sigma_Y^2 + \text{var } \bar{Y}_n.
\end{aligned}$$

Therefore, by combining these results, the mean squared error is

$$\begin{aligned} \text{MSE } \hat{r}_n &= E(\hat{r}_n - \mu_Y)^2 = \text{var } \hat{r}_n + [\text{Bias } \hat{r}_n]^2 \\ &= \frac{\text{var } \hat{f}_n(0)}{E^2(V_{n,1})} E(Y^2) + \text{var } \bar{Y}_n. \end{aligned} \tag{9}$$

It is known that $\hat{f}_n(0)$ is an asymptotically unbiased estimator of $f(0)$, provided the bandwidths h_n satisfies condition (1) and the kernel K satisfies condition (2) (see Parzen, 1962 or Härdle, 1991, page 59). We recall that for all $i \in N$,

$$\begin{aligned} \text{var } \hat{f}_n(0) &= \frac{1}{nh^2} \int K^2(\frac{t}{h}) f(t) dt - E(\hat{f}_n^2(0)) \\ &\le \frac{1}{nh} \int K^2(t) f(th) dt. \end{aligned}$$

By condition (5) it follows that

$$\lim_{n\to\infty} \sup \text{MSE } \hat{r}_n \le \lim_{n\to\infty} \text{var} \bar{Y}_n = 0.$$

4.2. *Proof of Theorem 2*

We condition on $(Y_i)_{i\in Z}$ and shall first find the limiting distribution of a related sequence of random variables under the regular conditional probability $P_Y^\omega(\cdot) = P(\cdot|(Y_i)_{i\in Z})(\omega)$. In the sequel E_Y^ω denotes the expected value with respect to P_Y^ω. We introduce the sequence of random variables

$$Z_{n,i} = \frac{1}{\sqrt{h}} \left(K(\frac{1}{h} X_i) - E(K(\frac{1}{h} X_i)) \right) Y_i = X_{n,i} Y_i, \tag{10}$$

where

$$X_{n,i} = \frac{1}{\sqrt{h}} \left[K(\frac{1}{h} X_i) - E(K(\frac{1}{h} X_i)) \right].$$

Note that, by independence of sequences $(Y_i)_{i\in Z}$ and $(X_i)_{i\in Z}$, for almost all ω, we have

$$E_Y^\omega(Z_{n,i}) = Y_i(\omega) E(X_{n,i}) = 0.$$

Denote

$$W_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{n,i} = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_{n,i} Y_i.$$

Let us find the limiting distribution of W_n under P_Y^ω, for almost all ω. We start by constructing Ω' such that, for all $\omega \in \Omega'$ the following convergences hold:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n Y_i^2(\omega) = E(Y^2), \tag{11}$$

and for all positive integer A,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n Y_i^2(\omega) I(|Y_i|(\omega) > A) = E[Y^2 I(|Y| > A)]. \tag{12}$$

This is possible because $(Y_i)_{i\in Z}$ is ergodic, so the convergences in (11) and (12) hold on sets of measure 1. We construct Ω' as a countable intersection of these sets, which will also have measure 1. Fix $\omega \in \Omega'$.

Under P_Y^ω, $(W_n)_{n\ge 1}$ becomes a sum of a triangular array of independent random variables. Therefore, in order to establish the CLT, we have to take care of the limiting variance and then verify the Lindeberg's condition. All the integrals below are taken over $R = (-\infty, \infty)$. First we recall that for all $i \in N$,

$$\text{var}(X_{n,i}) = \int K^2(t) f(th) dt - h(\int K(t) f(th) dt)^2.$$

So, by Bochner's theorem and condition (2),

$$\lim_{n\to\infty} \text{var}(X_{n,i}) = \lim_{n\to\infty} E(X_{n,i}^2) = f(0) \int K^2(u) du = C_1. \tag{13}$$

By independence of sequences $(Y_i)_{i\in Z}$ and $(X_i)_{i\in Z}$ and stationarity, we have

$$\sigma_n^2(\omega) = \text{var}_Y^\omega(W_n) = \frac{1}{n} \sum_{i=1}^n Y_i^2(\omega) \text{var}(X_{n,1}) = \overline{Y_n^2}(\omega) \text{var}(X_{n,1})$$

and therefore, by (11)

$$\lim_{n\to\infty} \sigma_n^2(\omega) = \lim_{n\to\infty} \frac{C_1}{n} \sum_{i=1}^{n} Y_i^2(\omega) = C_1 E(Y^2). \tag{14}$$

Let us establish now the Lindeberg's condition under P_Y^ω.

Denote $\sigma_n(\omega) = \sqrt{\sigma_n^2(\omega)}$. We have to show that, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{\sigma_n(\omega)} \frac{1}{n} \sum_{i=1}^{n} E_Y^\omega[X_{n,i}^2 Y_i^2 I(|X_{n,i}Y_i| \ge \varepsilon\sigma_n(\omega)\sqrt{n})] = 0. \tag{15}$$

Now, by (14) there is $N(\omega)$ such that for all $n > N(\omega)$ we have $\sigma_n(\omega) \ge C_1 E(Y^2)/2$. By this remark, by independence of the two sequences (see Example 33.7 in Billingsley (1995)) and stationarity, we obtain

$$E_Y^\omega[X_{n,i}^2 Y_i^2 I(|X_{n,i}Y_i| \ge \varepsilon\sigma_n(\omega)\sqrt{n})] = Y_i^2(\omega) E_Y^\omega[X_{n,1}^2 I(|X_{n,1}Y_i(\omega)| \ge \varepsilon\sigma_n(\omega)\sqrt{n})].$$

It follows that, in order to show (15), we have to show instead

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Y_i^2(\omega) E_Y^\omega[X_{n,1}^2 I(|X_{n,1}Y_i(\omega)| \ge \varepsilon'\sqrt{n})] = 0,$$

where we denoted $\varepsilon' = \varepsilon C_1 E(Y^2)/2$. Denote the expression above:

$$G_n(\omega) = \frac{1}{n} E_Y^\omega[X_{n,1}^2 \sum_{i=1}^{n} Y_i^2(\omega) I(|X_{n,1}Y_i(\omega)| \ge \varepsilon'\sqrt{n})].$$

We shall decompose the sum in two parts. Let A be a positive integer and define the index sets

$$I_1(\omega) = (i : 1 \le i \le n,\ |Y_i|(\omega) \le A),$$

$$I_2(\omega) = (i : 1 \le i \le n,\ |Y_i|(\omega) > A).$$

Note $\{1, 2, \ldots, n\} = I_1(\omega) \cup I_2(\omega)$. We write $\sum_{i=1}^{n} = \sum_{i\in I_1(\omega)} + \sum_{i\in I_1(\omega)}$ and, by using the stationarity assumption, we shall upper bound G_n in the following way:

$$G_n(\omega) \le A^2 E[X_{n,1}^2 I(|X_{n,1}| \ge A^{-1}\varepsilon'\sqrt{n})] + E(X_{n,1}^2) \frac{1}{n} \sum_{i=1}^{n} Y_i^2(\omega) I(|Y_i|(\omega) > A). \tag{16}$$

Note that:

$$E[X_{n,1}^2 I(|X_{n,1}| \ge A^{-1}\varepsilon'\sqrt{n})] = \frac{1}{h} \int K^2(\frac{v}{h}) I(K(\frac{v}{h}) \ge A^{-1}\varepsilon'\sqrt{nh}) f(v) dv =$$

$$\int K^2(u) I(K(u) \ge A^{-1}\varepsilon'\sqrt{nh}) f(uh) du.$$

Since $nh \to \infty$ and K is bounded, this limit is 0 as $n \to \infty$. By passing to the limit in (16) with $n \to \infty$ and by using (13), we easily obtain

$$\limsup_{n\to\infty} G_n(\omega) = C_1 E[Y^2 I(|Y| > A)].$$

By letting $A \to \infty$, and using the fact that Y has finite second moment, we get

$$\lim_{n\to\infty} G_n(\omega) = 0.$$

Therefore, the Lindeberg's condition is satisfied under P_Y^ω. By all these considerations, we obtain that the following quenched central limit theorem holds: for any fixed $\omega \in \Omega'$

$$W_n \Rightarrow N(0, C_1 E(Y^2)) \text{ under } P_Y^\omega.$$

This quenched CLT is a stronger form of CLT. After representing it in terms of characteristic function we can integrate with respect to the measure P and we obtain the annealed CLT, namely

$$W_n \Rightarrow N(0, C_1 E(Y^2)) \text{ under } P. \tag{17}$$

Now recall that $Z_{n,i} = h^{-1/2} \left(K(\frac{1}{h}X_i) - E(K(\frac{1}{h}X_i))\right) Y_i$. Let us also note that by definition (3),

$$\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} Y_i K(\frac{1}{h}X_i) = \sqrt{nh} E(V_{n,1}) \hat{r}_n.$$

So we can rewrite

$$W_n = \sqrt{nh}\left(E(V_{n,1})\hat{r}_n - \frac{1}{n}\sum_{i=1}^{n} Y_i E(V_{n,1})\right) = \sqrt{nh}\left(E(V_{n,1})\hat{r}_n - \bar{Y}_n E(V_{n,1})\right) \tag{18}$$

Note that

$$\sqrt{nh}(\bar{Y}_n E(V_{n,1}) - E(V_{n,1})\mu_Y) = \sqrt{nh}(\bar{Y}_n - \mu_Y)E(V_{n,1}). \tag{19}$$

If we impose (7), then

$$\sqrt{nh}\left(\bar{Y}_n E(V_{n,1}) - E(V_{n,1})\mu_Y\right) \to^P 0 \tag{20}$$

and, by Theorem 25.2 in Billingsley (1995), we obtain

$$\sqrt{nh}E(V_{n,1})\left(\hat{r}_n - \mu_Y\right) \Rightarrow N(0, C_1 E(Y^2)).$$

or

$$\sqrt{nh}f(0)\left(\hat{r}_n - \mu_Y\right) \Rightarrow N(0, C_1 E(Y^2)).$$

By the ergodic theorem and Slutski's theorem we obtain the desired result. □

4.3. Alternative estimator

The provided estimator $\hat{r}_n$ is unbiased with an MSE that increases as h_n tends to 0. The selection of the best value of h_n is thus difficult through the MSE. The following estimator $\tilde{r}_n$ is not unbiased, but is asymptotically unbiased and the presence of the bias for finite samples allows to select the optimal bandwidths.

$$\tilde{r}_n = \frac{1}{nh_n f(0)}\sum_{i=1}^{n} Y_i K(\frac{X_i}{h_n}). \tag{21}$$

As for $\hat{r}_n$, it is easy to establish the following result.

Corollary 5. *Under the conditions of Theorem 2, the following CLT holds*

$$\frac{\sqrt{nh}}{\overline{Y_n^2}}(\tilde{r}_n - c\mu_Y) \Rightarrow N(0, \frac{1}{f(0)}\int K^2(x)dx), \tag{22}$$

where $c = E\hat{f}(0)/f(0)$.

The proof of this corollary repeats that of Theorem 2 replacing $E\hat{f}(0)$ by $f(0)$.

5. Bandwidth and kernel selection

The method we propose introduces new parameters, the bandwidth sequence $(h_n)_{n\geq 1}$ and the kernel. There is a vast literature on the selection of h_n for kernel estimation of the density and for the Nadaraya–Watson estimator of a regression, under independence or weak dependence assumptions. They can be found in books, such as in Section 5.1.2 in Härdle (1991) or in surveys, such as Jones et al. (1996). Our case deals with possible long range dependence for $(Y_i)_{i\in Z}$, but it benefits from the independence of $(Y_i)_{i\in Z}$ and $(X_i)_{i\in Z}$ and also from the fact that we know $f(x)$. Recall that we assume that $(Y_i)_{i\in Z}$ is ergodic. If we impose additional conditions on the smoothness of $f(x)$ and $K(x)$, namely $f(x)$ has a continuous and bounded second derivative and K satisfies condition (2) and $\int x^2K(x)dx < \infty$, we can analyze the global optimal bandwidth by optimizing the main part of the mean square error under the constraints (1) and (7). We shall see that this selection depends on the strength of dependence of $(Y_i)_{i\in Z}$. As a matter of fact we shall prove the following proposition.

Proposition 6. *Under the conditions above, the optimal data driven bandwidth, which minimizes the mean squared error of* $\tilde{r}_n$ *is*

$$h_o = [\frac{f(0)B\overline{Y_n^2}}{n(f''(0)A)^2(\bar{Y}_n)^2}]^{1/5} \text{ provided that } \mathrm{var}(\bar{Y}_n) = o(n^{-4/5}) \text{ and } \mu_Y \neq 0. \tag{23}$$

where

$$A = \int x^2K(x)dx \text{ and } B = \int K^2(x)dx. \tag{24}$$

The condition $var(\bar{Y}_n) = o(n^{-4/5})$ is necessary for the CLT to hold. If this condition does not hold, this proposition defines a range of h_n for which our theorem is not applicable. It provides a hint for the selection of h_n. In such a case, knowing that the MSE decreases prior to h_0, we need to select h_n as close as possible to h_o, ensuring that $nh_n var(\bar{Y}) \to 0$ and $nh_n \to \infty$.

Proof. We start from the computation of MSE$(\tilde{r}_n)$ similar to those leading to formula (9), to obtain

$$\mathrm{MSE}(\tilde{r}_n) = \frac{1}{f^2(0)}[E(Y^2)var(\hat{f}(0)) + \mu_Y^2 Bias^2(\hat{f}(0)) + E^2(V_{n,1})var(\bar{Y}_n)]. \tag{25}$$

When $f(x)$ has a continuous and bounded second derivative, according to formula (2.3.2) in Härdle (1991), we have

$$\text{Bias } \hat{f}_n(0) = \frac{h^2}{2} f''(0)A + o(h^2) \text{ as } n \to \infty.$$

Also by formula (2.3.3) in the same book

$$\text{var } \hat{f}_n(0) = \frac{1}{nh} Bf(0) + o(\frac{1}{nh}) \text{ as } n \to \infty.$$

Under condition (7),

$$(EV_{n,1})^2 \text{var } \bar{Y}_n = (1 + o(1))\text{var } \bar{Y}_n = o(\frac{1}{nh_n}).$$

It follows that

$$\text{MSE } \tilde{r}_n = \frac{1}{f^2(0)}[\frac{E(Y^2)}{nh} Bf(0) + \frac{h^4}{4}\mu_Y^2 (f''(0)A)^2 + o(\frac{1}{nh_n}) + o(h^4)]. \tag{26}$$

In order to minimize it, we set 0 the derivative with respect to h of the main part and obtain

$$h_{o'} = [\frac{f(0)BE(Y^2)}{n(f''(0)A)^2 \mu_Y^2}]^{1/5},$$

provided $\mu_Y \neq 0$. Since the optimal $h_{o'}$ depends on the unknown parameters $E(Y^2)$ and $\mu_Y^2 \neq 0$, we shall replace them by plug in estimators which are consistent because of the ergodicity of $(Y_n)_{n>0}$ to obtain (23). Finally, the size of h_o combined to condition (7) leads to the restriction $\text{var}(\bar{Y}_n) = o(nh_o)^{-1} = o(n^{-4/5})$. □

If $\liminf_{n\to\infty} n^{4/5}\text{var}(\bar{Y}_n) \neq 0$ this implies that $\liminf_{n\to\infty} nh_{o'}\text{var}(\bar{Y}_n) \neq 0$, and we need to select an $h < h_0$ in order for (7) to be satisfied. Note that the MSE is decreasing for $h < h_{o'}$. Therefore, in this case, the MSE is minimized when h_n satisfies $h_n = \varepsilon_n[n\text{var}(\bar{Y}_n)]^{-1}$, with ε_n converging very slowly to 0. But such a selection would be difficult to estimate when no information is available on the dependence of $(Y_n)_{n\in Z}$. In fact, when one can estimate the variance of partial sums, ε_n and h_n can be chosen to satisfy the following conditions:

$$\varepsilon_n \to 0, \quad var(\bar{Y}_n) = o(\varepsilon_n), \quad h_n = \frac{\varepsilon_n}{nvar(\bar{Y}_n)} \to 0. \tag{27}$$

The most important thing in formula (27) is the rate of convergence of $var(\bar{Y}_n)$. When estimated, this rate can replace $var(\bar{Y}_n)$.

The mean squared error given in formula (26) depends on the kernel estimator via the quantities A and B, given in (24). Epanechnikov suggested to compare the performance of kernels satisfying the restriction

$$A = \int x^2 K(x)dx = 1.$$

For this case, studying kernel density estimators, he found that the MSE is optimized when B is minimized, and this happens for the so called Epanechnikov kernel, namely $K(u) = (3/4)(1 - u^2)I(|u| \leq 1)$. However, other kernels are also giving very good results with a minimal loss of efficiency and therefore the success of the method does not depend too much on the kernel selection for kernel density estimation. Other kernels that are often used are the quadratic kernel $(K(u) = (15/16)(1 - u^2)^2 I(|u| \leq 1))$ and the standard normal kernel.

5.1. Kernels, densities and efficiency

We provide here few tables containing quantities that are needed for use of various kernels and distribution functions for X.

In Table 1 we have values of A and B for each of the provided kernels. Using these values, we will propose a study of the impact of kernels and densities on the proposed estimator of the mean and the subsequent confidence intervals.

We use Table 1 to obtain Table 2 of various optimal bandwidths for a set of kernels and densities of X. This Table shows that the quartic kernel in combination with the $\chi^2(2)$ for $1 + X$ provides the largest optimal bandwidth. These

Table 1
Values of A and B for various kernels.

Kernel	Gaussian	Epanechnikov	Uniform	Quartic
$A\|B$	$1\|1/2\sqrt{\pi}$	$1/5\|3/5$	$1/3\|1/2$	$1/7\|5/7$

Table 2
Optimal bandwidths.

	Density of X		
Kernel	Gaussian	$\chi^2(2)$	Cauchy
Gaussian	$(\frac{\overline{y_n^2}}{n\sqrt{2}\bar{y}_n^2})^{1/5}$	$(\frac{16\sqrt{e}\overline{y_n^2}}{n\sqrt{\pi}\bar{y}_n^2})^{1/5}$	$(\frac{\sqrt{\pi}\overline{y_n^2}}{8n\bar{y}_n^2})^{1/5}$
Epanechnikov	$(\frac{15\sqrt{2\pi}\overline{y_n^2}}{n\bar{y}_n^2})^{1/5}$	$(\frac{480\sqrt{e}\overline{y_n^2}}{n\bar{y}_n^2})^{1/5}$	$(\frac{15\pi\overline{y_n^2}}{4n\bar{y}_n^2})^{1/5}$
Uniform	$(\frac{9\sqrt{\pi}\overline{y_n^2}}{n\sqrt{2}\bar{y}_n^2})^{1/5}$	$(\frac{144\sqrt{e}\overline{y_n^2}}{n\bar{y}_n^2})^{1/5}$	$(\frac{9\pi\overline{y_n^2}}{8n\bar{y}_n^2})^{1/5}$
Quartic	$(\frac{35\sqrt{2\pi}\overline{y_n^2}}{n\bar{y}_n^2})^{1/5}$	$(\frac{1120\sqrt{e}\overline{y_n^2}}{n\bar{y}_n^2})^{1/5}$	$(\frac{35\pi\overline{y_n^2}}{4n\bar{y}_n^2})^{1/5}$

Table 3
Asymptotic variance when the kernel is known.

Kernel	Gaussian	Epanechnikov	Uniform	Quartic
Efficiency	$.363D_1$	$.349D_1$	$.370D_1$	$.351D_1$

Table 4
Asymptotic variance when the Distribution of X is known.

Density of X	Gaussian	$\chi^2(2)-1$	Cauchy
Efficiency	$.48D_2$	$.55D_2$	$.53D_2$

bandwidths when used for estimation also show that the theoretical asymptotic variance for the quartic kernel and $\chi^2(2)$ density for $1+X$ is not the smallest. In fact, a larger optimal bandwidth does not necessarily imply a shorter confidence interval. This is because the size of the confidence interval depends also on $f(0)$, A and B, considered separately from h. To have a procedure with shorter confidence intervals, we will look for the scenario with the lowest asymptotic variance. We call asymptotic variance the quantity

$$var(Case) = \frac{B\mu_{Y^2}}{f(0)nh_n}.$$

A set of kernel K and density of X that provides a smaller value of $var(Case)$, defines a smaller confidence interval. So, we will call this quantity efficiency as well. Using the optimal bandwidths for h_n in $var(Case)$ leads to the following tables.

In Table 3 of theoretical asymptotic variances, D_1 is a constant that depends on μ_Y, μ_{Y^2}, $f(0)$ and $f''(0)$. The form of the constant is irrelevant at this point. This table shows that when the optimal bandwidths are used, the quartic kernel is the most efficient among the considered kernels in terms of variance minimization along with the Epanechnikov kernel. The Gaussian kernel performs a little better than the uniform kernel. Overall, the difference in variance is relatively small compared to the smallest, these are (97%, 100%, 95% and 100%).

In Table 4 we provide the asymptotic variance for 3 distributions of X. Here, D_2 is a constant that depends only on A and B. The choice is made to have a centered bell-shaped distribution, a distribution with heavy tail and a distribution that does not have a mean. The result shows that the Normal distribution seems to be a better distribution for X (smallest asymptotic variance for a fixed kernel). Entries of this table are asymptotic variances when for the same data and same kernel we use various distributions for X.

Combining Tables 3 and 4, using the special form of $var(Case)$, we obtain the following Table 5 by multiplying the coefficients in each of the cases.

In Table 5, G is a constant that depends only on μ_Y and μ_Y^2, if we consider that the sample size is known. We can see that for the proposed distributions, the normal distribution for X outperforms others in all cases. This fact motivates the definition of asymptotic relative efficiency of estimators of the provided form as

$$e = var(Case1)/var(Case2),$$

where $var(Case1)$ is the variance of the estimator when the kernel is Gaussian and the distribution of X is standard normal; and $var(Case2)$ is the variance of the considered alternative. We obtain the following table of asymptotic relative efficiencies.

Table 5
Asymptotic variances for given kernel and density of X using the optimal bandwidths.

Kernel	Density of X		
	Gaussian	$\chi^2(2)-1$	Cauchy
Gaussian	$.17G$	$.20G$	$.19G$
Epanechnikov	$.17G$	$.19G$	$.18G$
Uniform	$.18G$	$.20G$	$.19G$
Quartic	$.17G$	$.19G$	$.18G$

Table 6
Relative efficiencies using the optimal bandwidths.

Kernel	Density of X		
	Gaussian	$\chi^2(2)-1$	Cauchy
Gaussian	100%	85%	89%
Epanechnikov	100%	89%	94%
Uniform	94%	85%	89%
Quartic	100%	89%	94%

Table 7
Asymptotic variances for given kernel and density of X without optimal bandwidths.

Kernel	Density of X		
	Gaussian	$\chi^2(2)-1$	Cauchy
Gaussian	$.707G_1$	$.930G_1$	$.886G_1$
Epanechnikov	$1.504G_1$	$1.978G_1$	$1.885G_1$
Uniform	$1.253G_1$	$1.648G_1$	$1.571G_1$
Quartic	$1.790G_1$	$2.355G_1$	$2.244G_1$

Table 8
Relative efficiencies when not using the optimal bandwidths.

Kernel	Density of X		
	Gaussian	$\chi^2(2)-1$	Cauchy
Gaussian	100%	76%	80%
Epanechnikov	47%	36%	38%
Uniform	56%	43%	45%
Quartic	40%	30%	32%

It is clear from Table 6 that when the optimal bandwidths are used, in order to minimize the size of the confidence interval, one needs to use the Gaussian, quartic or Epanechnikov kernel and the standard normal distribution for X. This result is valid only for the considered kernels and distributions. Also, one can notice that the lack of mean for the Cauchy distribution does not stop it from outperforming the shifted $\chi^2(2)$ distribution on all considered kernels. When h_n is not the optimal bandwidth in the confidence interval, minimizing the asymptotic variance (the size of the confidence interval) over kernels and densities of X presents a different pattern as shown below.

It can be seen from Table 7 that the Gaussian kernel combined with the standard normal distribution for X performs better than any other proposed scenario. In this table, G_1 is a constant that depends on h_n, μ_Y and μ_Y^2. The relative efficiencies can be obtained easily.

Table 8 shows that when the optimal bandwidths is not used, the Gaussian kernel and the standard normal outperform any other considered scenario by at least 20% of the length of the confidence interval. The Gaussian kernel is indisputably by far better that any of the used popular kernels. Differences in lengths of the confidence intervals for the considered distributions of X are not too high when the Gaussian kernel is used. We recommend for this reason to use the normal kernel when the optimal bandwidth is not used.

6. Applications to stationary sequences with long memory

Condition (4) used in our theorems is satisfied if the sequence $(Y_n)_{n\in Z}$ is ergodic in the ergodic theoretical sense (see for instance (Krengel, 1985). Convergence of $\overline{Y_n^2}$ to the second moment of Y is also a consequence of ergodicity. There are numerous sequences which are ergodic. For instance, given $(\xi_n)_n$ a sequence of i.i.d. and $f : R^N \to R$ a Borel function, then $Y_k = f(\ldots.\xi_{k-2,}\xi_{k-1}, \xi_k)$ is an ergodic sequence. Other examples are the countable Markov chains which are irreducible and aperiodic. Also general space Markov chains which are Harris recurrent and aperiodic. A Gaussian sequence with spectral density is a function of i.i.d., and therefore ergodic. For other examples we refer to the book of Bradley (2007).

Example 1. Polynomial restriction on the covariance structure.

Let us recall that our confidence intervals are obtained for stationary ergodic sequences with finite second moments. No other restriction on the dependence structure and on the distribution of Y is assumed other that $nh_n var(\bar{Y}_n) \to 0$. In particular, when $|\text{cov}(Y_0, Y_k)| \sim C(k^{-\alpha})$ as $k \to \infty$ for $0 < \alpha < 1$, the covariances are not summable and $(Y_k)_{k\in Z}$ has long memory. Note first that we have

$$nh_n \text{var}(\bar{Y}_n) \le 2h_n \sum_{k=0}^{n} |\text{cov}(Y_0 Y_k)| = O(h_n n^{-\alpha+1}) \text{ as } n \to \infty. \tag{28}$$

Therefore condition (7) of Theorem 2 holds, if using formula (27), we select

$$\varepsilon_n = \frac{\log n}{n^\alpha} \quad \text{and} \quad h_n = \frac{\log n}{n}.$$

The CLT applies with $h_n = \log n/n$. Notice that for this example, the rate α is not necessary for the choice of h_n.

Example 2. Logarithmic restriction on the covariance structure.

If we consider $|\text{cov}(Y_0, Y_k)| \sim C(\log k)^{-1}$, then the above computations lead to

$$nvar(\bar{Y}_n) \le 2 \sum_{k=0}^{n} |\text{cov}(Y_0 Y_k)| = O(n/\log n).$$

For this example, applying formula (27) leads to the possibility

$$\varepsilon_n = \frac{\log\log n}{\log n}, \quad h_n = \frac{\log\log n}{n}.$$

This, with $h_n = \log\log n/n$, Theorem 2 can be applied. In this case, the rate of convergence to 0 of the MSE is slower than $(\log n)^{-1}$ as $n \to \infty$. This shows that when the memory is very long the rates of convergence can be rather slow, therefore a very large sample size might be necessary.

Example 3. Long memory linear processes.

Let $(\xi_j)_{j\in Z}$ be an i.i.d. sequence of random variables, centered with finite second moments. Let $(a_j)_{j\in Z}$ be a sequence of constants. We consider the linear process

$$Y_k = \sum_{j=-\infty}^{\infty} a_{k-j}\xi_j. \tag{29}$$

Denote $S_n = \sum_{k=1}^{n} Y_k$. If $\sum_{i\in\mathbb{Z}} a_i^2 < \infty$, Y_k in (29) is well defined a.s. and in L_2. Note that, being a function of i.i.d., the sequence $(Y_k)_{k\in Z}$ is ergodic and our results apply. To find the condition satisfied by h_n we start by writing $S_n = \sum_{i=-\infty}^{\infty} b_{ni}\xi_i$ with $b_{ni} = a_{1-i} + \cdots + a_{n-i}$. Using this notation we have $\text{var}(S_n) = \text{var}(\xi_0^2)\sum_i b_{ni}^2$. Then $\text{var}(\sqrt{nh}(S_n/n)) = h_n n^{-1}\sum_i b_{ni}^2$. If we select h_n to satisfy (1) and $h_n n^{-1}\sum_i b_{ni}^2 \to 0$, then the conclusion of Theorem 2 holds.

As a particular case, we consider the important case of causal long-memory processes with

$$a_i = l(i+1)(1+i)^{-\alpha}, \; i \ge 0, \text{ with } 1/2 < \alpha < 1, \text{ and } a_i = 0 \text{ otherwise.}$$

Here $l(\cdot)$ is a slowly varying function at infinite. These processes have long memory because $\sum_{j\ge 0} |a_j| = \infty$.

For this case, $var(\bar{Y}_n) \sim \kappa_\alpha n^{1-2\alpha}\ell^2(n)$ (see for instance Relations (12) in Wang et al. (2001)), where κ_α is a positive constant depending on α. Theorem 2 can be applied as soon as $h_n = o((n^{2(1-\alpha)}\ell^2(n))^{-1})$ as $n \to \infty$, provided that $nh_n \to \infty$.

This example covers the ARFIMA $(0, d, 0)$ processes (cf. Granger and Joyeux, 1980; Hosking, 1981), which play an important role in financial time series modeling and applications. As a special case, let $0 < d < 1/2$ and B be the backward shift operator with $B\varepsilon_k = \varepsilon_{k-1}$,

$$X_k = (1-B)^{-d}\xi_k = \sum_{i\ge 0} a_i \xi_{k-i}, \text{ where } a_i = \frac{\Gamma(i+d)}{\Gamma(d)\Gamma(i+1)}.$$

Notice that here, $\lim_{n\to\infty} a_n/n^{d-1} = 1/\Gamma(d)$. Thus, for this case, $var(\bar{Y}_n) \sim \kappa_d n^{2d-1}$ and the CLT applies when $h_n = o(n^{-2d})$, provided that $nh_n \to \infty$. The CLT holds for instance if

$$h_n = \frac{\ln n}{n}.$$

So, for applications, one does not need to know the value of d. We will provide a simulation study that shows that using the same h_n across all values of d does not affect our procedure.

Example 4. A long memory reversible Markov chain.

For a nonlinear example we would like to mention an example given in Zhao et al. (2010), describing a stationary and ergodic reversible Markov chain, which does not satisfy the CLT. This is their Example 2. Let $1 < \alpha < 2$. One starts with a measurable function $p : R \to (0, 1)$, $p(x) = e^{-1/|x|}I(|x| > 1)$ and a probability measure υ such that for $|x| > 1$,

$$\upsilon(x) = \frac{[1 - p(x)]dx}{2\gamma_\alpha |x|^\alpha} \text{ where } \gamma_\alpha = \int_0^1 y^{\alpha-2}(1 - e^{-y})dy.$$

We define now a stationary and reversible Markov chain, $(X_n)_{n\in Z}$, with transition operator:

$$Q(x, A) = p(x)\delta_x(A) + (1 - p(x))\upsilon(A),$$

where δ_x denotes the Dirac measure. This Markov chain is stationary, reversible and ergodic with invariant distribution

$$\pi(dx) = (\alpha - 1)dx/(2|x|^\alpha) \text{ for } |x| > 1.$$

Zhao et al. (2010) showed that $S_n = \sum_{i=1}^n \text{sign}(X_i)$ does not satisfy the central limit theorem under any normalization. They showed in fact $n^{-1/\alpha}S_n \to Z$, where Z has a symmetric stable distribution. This implies that $\text{var}(\bar{Y}_n) \sim cn^{-2+2/\alpha}$. For statistical inference in this case, one would need to look for properties of Z. With the proposed Theorem 2, we just need to select $h_n = n^{-1/\alpha}$ when α is known or just take in general for this case h_n as in the above example.

7. Simulation study

7.1. Generalities for the confidence interval for the mean

Recalling that for the normal kernel and the standard normal distribution for X the estimator of the mean of the sequence $(Y_i, 1 \le i \le n)$ is $\tilde{r}_n = \dfrac{1}{nh}\displaystyle\sum_{i=1}^n Y_i exp(-\frac{1}{2}(\frac{X_i}{h})^2)$. Using the quartic kernel and the standard normal distribution for X, the estimator is

$$\tilde{r}_n = \frac{15\sqrt{2\pi}}{16nh}\sum_{i=1}^n Y_i(1 - (\frac{X_i}{h})^2)^2 I(|X_i| < h).$$

For a sample of observations $\{Y_i.1 \le i \le n\}$ with unquantified dependence, based on Corollary 5, a $(1-\alpha)100\%$ confidence interval for the mean μ_Y is

$$\left(\hat{r}_n - \left(\frac{\overline{Y_n^2}\int K^2(x)dx}{nhf(0)}\right)^{1/2} z_{\alpha/2},\ \hat{r}_n + \left(\frac{\overline{Y_n^2}\int K^2(x)dx}{nhf(0)}\right)^{1/2} z_{\alpha/2}\right).$$

Using the standard normal distribution for X we obtain

$$\left(\hat{r}_n - \left(\frac{\overline{Y_n^2}}{nh\sqrt{2}}\right)^{1/2} z_{\alpha/2},\ \hat{r}_n + \left(\frac{\overline{Y_n^2}}{nh\sqrt{2}}\right)^{1/2} z_{\alpha/2}\right), \quad \hat{r}_n = \tilde{r}_n\sqrt{1+h^2}. \tag{30}$$

7.2. Using $\tilde{r}_n$ to estimate μ for some reversible Markov chain

Longla (2015) proposed conditions for mixing properties of mixtures of copulas that generate reversible Markov chains. A class of copulas for such Markov chains was the Frechet family of copulas $C(x, y) = aW(x, y) + (1 - a)M(x, y)$, for $0 \le a \le 1$. This copula is the joint distribution of a bivariate random variable (U, V) with uniform marginals on $(0, 1)$. It generates reversible Markov chains with any initial distribution (see Longla and Peligrad (2012), Longla (2013) or Longla (2015) for more). Technically, any sample from any Markov chain generated by this copula will be a string made of two values X_0 and $1 - X_0$ with changes depending on the value of a. The number a is typically the probability to obtain $1 - x$ after obtaining x for the previous sample point. So, the larger a, the more flips we will have in the sample. Our aim here is to apply the results of the study to the population mean and compare the performances of various estimators.

Table 9(a)–(d) indicates some sample results from the Markov chains with distributions given in the headings and having transition probabilities defined by the Frechet copula with parameter a. The sample size in this table is given in thousands. The sequence of X observations is from the standard normal distribution and we use the Gaussian kernel. To generate observations from these stationary Markov chains, if F is the cumulative distribution of Y, we generate a Markov chain $(U_i, 1 \le i \le n)$ with uniform distribution as marginals and Frechet copula for transitions, then set $Y_i = F^{-1}(U_i)$ for $i = 1, \ldots, n$. The estimators $\tilde{r}_n$ and $\bar{Y}$ are applied to the same data set with the same sample $(X_i, 1 \le i \le n)$ from the standard normal distribution. Here, the optimal bandwidths are used and provided pretty accurate estimates for the mean with reasonable variances for large samples. To obtain coverage probabilities of our confidence intervals, we generate

Table 9
Applications to the Frechet family of copulas.

(a) $a = .3$ and $Y = N(50, 1)$					(b) $a = .5$ and $Y = Normal(6, 1)$				
$n(10^3)$	20	10	5	.1		20	10	5	.1
$\tilde{r}_n$	49.2	49.18	50.79	50.19	$\tilde{r}_n$	5.91	5.89	5.71	6.33
$\sigma_{\tilde{r}_n}$	.83	1.09	1.44	6.89	$\sigma_{\tilde{r}_n}$	.10	.13	.17	.83
$\bar{y}$	49.99	49.99	50.01	49.95	$\bar{y}$	6.01	6.00	5.99	6
σ_y	.30	.62	1.33	1.10	σ_y	.50	.22	1.52	.51
(c) $a = .7$ and $Y = N(200, 4)$					(d) $a = .7$ and $Y = Normal(-5, 4)$				
	20	10	5	.1		20	10	5	.1
$\tilde{r}_n$	199.14	194.88	199.86	211.35	$\tilde{r}_n$	−5.00	−5.01	−5.07	−4.68
$\sigma_{\tilde{r}_n}$	3.31	4.37	5.77	27.61	$\sigma_{\tilde{r}_n}$	.08	.11	.15	.72
$\bar{y}$	199.96	200.00	199.98	200.15	$\bar{y}$	−4.99	−5.00	−4.99	−4.82
σ_y	1.40	2.71	3.17	3.72	σ_y	.84	.90	.96	2.30

Table 10
Coverage probabilities of confidence intervals.

	a=.3	$\mu = 200$	a=.7	$\mu = -5$	$a = .5$	$\mu = 50$
n	20,000	10,000	5000	1000	500	100
99%	99/100	96/100	98/88	97/91	100/99	100/100
95%	96/100	94/100	98/83	95/86	95/90	100/95
90%	94/100	91/100	91/78	92/79	90/87	95/93

Table 11
Coverage probabilities of confidence intervals for $\hat{r}_n$ with modification.

	$a = .3$	$\mu = 100$	$a = .3$	$\mu = 10$	$a = .3$	$\mu = 10$
n	4,000	3,000	4,000	3,000	10,000	20,000
99%	99.9/99.8	99.9/99.6	99.8/99.7	99.9/99.8	99.6/99.5	99.3/99.3
95%	97.9/97.8	96.9/96.6	97.4/97.2	97.8/97.4	97.3/97.3	96.2/96.2
90%	95.0/94.9	94.0/93.4	93.2/93.1	94.2/94.1	93.2/92.9	92.8/92.5

one hundred samples of the described Markov chain with the normal distribution with mean μ and variance $\sigma^2 = 4$ as invariant distribution. We have selected 3 values of a for the analysis in Table 10. The value $a = .5$ is the case when half of the sample is expected to have value $1 - X_0$. The other two cases $a = .3$ and $a = .7$ are closer to the extremes of the copula. One can see that large sample confidence intervals are pretty accurate. For instance, Table 10 indicates that when the sample size is $n = 20,000$, $a = .3$ and $\mu = 200$, we find that 99, 96 and 94 of the 99%, 95% and 90% confidence intervals cover the true mean respectively (while inaccurately the standard interval $\bar{Y}_n \pm z_\alpha S/\sqrt{n}$ covers 100 times the true mean). Each of the entries of Table 10 is a set of two proportions (p_1/p_2), where p_1 is the percentage of intervals that cover the true mean when $\tilde{r}_n$ is used and p_2 is the proportion of confidence intervals that cover the mean when $\bar{Y}_n$ is used.

In Table 11, we have coverage probabilities of two confidence intervals. The performance is compared on the same data sets. Entries of Table 11 are percentages of confidence intervals covering the true mean of the process after 1000 simulations with $\hat{r}_n$ and a modified version of the above interval by dividing the length of the interval by $\sqrt{1+h^2}$. The outputs show that the modified intervals are more accurate on moderate samples, and the performances of both intervals are identical for larger samples ($n \leq 10{,}000$). In any of the cases, the test of the confidence level is statistically significant.

7.3. *Estimation of μ for Markov chain with Clayton copula*

Copulas are bivariate distributions that are used to capture the strength of the dependence between random variables. When using a copula to model the dependence for a bivariate random variable with uniform marginals, the conditional distribution for transitions is the derivative with respect to the first variable of the copula (see Nelsen (2006) or Longla and Peligrad (2012)). Thus, to obtain the data, we generate a Markov chain with uniform marginals and Clayton copula for transition probabilities $(Z_i, 1 \leq i \leq n)$. This is done using the Clayton copula and its derivative

$$C(u, v) = (u^{-\alpha} + v^{-\alpha} - 1)^{-1/\alpha}, \quad C_u(u, v) = u^{-\alpha-1}(u^{-\alpha} + v^{-\alpha} - 1)^{-1/\alpha-1}.$$

Knowing the previous value u_0 of the Markov chain, the following is obtained by generating a value from the distribution $C_u(u_0, v)$ (see Nelsen (2006) for more). An independent observation v_i is generated from the uniform distribution. Then $Z_i = (u_0^{-\alpha}(v_i^{-\alpha/(\alpha+1)} - 1) + 1)^{-1/\alpha}$. We then set $(Y_i = F^{-1}(Z_i), 1 \leq i \leq n)$, where F is the cumulative distribution of the invariant distribution of the generated Markov chain. The estimators use the same sample of X values from the standard normal distribution.

Table 12
Number of 95% confidence intervals covering the true $\mu = 30$ for $\alpha = 3$.

n	20,000	10,000	5000	1000	500
Gaussian and h_0	93	93	94	94	96
Gaussian and $h = n^{-.19}$	91	93	93	92	96
Epanechnikov and h_0	88	92	92	91	90
Epanechnikov and $h = n^{-.19}$	89	89	94	92	91

Table 13
Percentage of 95% confidence intervals covering the true $\mu = 30$ for $\alpha = 3$.

n	20,000	10,000	5000	1000	500
Gaussian and h_0	91	95.5	94.1	94.7	93.5
Gaussian and $h = n^{-.19}$	89	94.9	93.9	94.1	93.5
Epanechnikov and h_0	86	87.4	85.9	85.0	85.6
Epanechnikov and $h = n^{-.19}$	89	90	90.1	88.8	89.8

Table 14
Proportions of confidence intervals covering the true mean using h_0.

Sample size	7000	5000	3000	2000	1000	500
$\tilde{r}_n$, Gaussian kernel, $d = .1$	95.2	95.2	95.9	96.3	96.5	96.6
$\tilde{r}_n$, Gaussian kernel, $d = .3$	95.0	95.6	95.1	96.2	96.0	96.7
$\tilde{r}_n$, Gaussian kernel, $d = .4$	95.3	95.3	96.0	96.3	96.3	97.1
$\tilde{r}_n$, Gaussian kernel, $d = .6$	95.2	94.8	95.2	96.4	94.8	97.9

Entries of Table 12 are coverage probabilities of 95% confidence intervals using the indicated kernel and bandwidths applied to the same data set. The data $(Y_i, i = 1, \ldots, n)$ is simulated from the Clayton copula with $\alpha = 3$ and invariant distribution $T(2) + 30$ (T distribution with 2 degrees of freedom shifted to a mean of 30). This is done by generating a reversible Markov chain $(Z_i, = 1, \ldots, n)$ from the Clayton copula the Uniform distribution as invariant marginal distribution. Then, the SAS function TINV is used to find $Y = 30 + \text{TINV}(Z, 2)$ for each of the Z observations generated via the Clayton copula and the uniform distribution. The fourth row for example means that using $\tilde{r}_n$, 89 of the 100 computed 95% confidence intervals cover $\mu = 30$.

In Table 13, we have results of simulations for Y following $\chi^2(30)$ using the copula with $\alpha = 3$ and standard normal distribution for X. It can be seen from Table 13 that $\tilde{r}_n$ defined via the Epanechnikov kernel performs poorly for confidence intervals for the sample size up to 20,000, but the Gaussian kernel agrees more with the 95% confidence intervals under the optimal bandwidths. Table 13 also shows that moving away from the optimal bandwidths reduces accuracy when the Gaussian kernel is used. The opposite is true when the Epanechnikov kernel is used. The percentage is computed for 1000 simulations of samples of the given sizes.

7.4. Comparisons and example of simulation with Gaussian innovations

The ARFIMA$(0, d, 0)$ model that we have considered in this paper is very popular in the literature. For this model, when innovations are independent and normally distributed, it follows that the stationary distribution of the model is Gaussian. It has been shown (see Beran (1989), Theorem 2.2. Beran (1959)) that the variance of the sample mean satisfies $n^{2-2H} var(\bar{Y}) \to K$ where K is a function of $H = d + 1/2$. Also, by the comment after Corollary 2.1 in Beran (1959), for any stationary square integrable self-similar sequence $(Y_i)_{i\in N}$, $var(\bar{Y}) = \sigma^2 n^{2H-2}$. Moreover, if the sequence is Gaussian with mean μ and variance σ^2, then $n^{1-H}\sigma^{-1}(\bar{Y} - \mu)$ has a standard normal distribution. As seen here, even in the Gaussian case, inference highly relies on estimators of H and their distributions. With our proposed strategy, one just needs to define a range of d, that can be obtained using one of the methods described in Chapter 4 of Beran (1959), such as the variance plot. The variance plot plots $log(var(mean\ of\ subsample\ of\ size\ k))$ against $log(subsample\ size\ k)$ for integers $2 \le k \le n/2$. This plot must exhibit a line with slope $2H - 2 = 2d - 1$. It is clear that the only issue we have appears when d is close to or equal to $1/2$. For $d = 1/2$, $nhvar(\bar{Y}) \to \infty$ is impossible. But, for any value of d one can choose $h_n = \ln n/n$. Table 14 below indicates coverage proportions for our estimator in the case when the Y observations are generated from the ARFIMA$(0, d, 0)$ shifted to mean $\mu = 50$ after 1000 samples of the indicated size n, using the optimal bandwidths. Table 15 is obtained in the same scenario as Table 14, but using $h_n = \ln n/n$. The results are striking even for $d = .6 > .5$. All estimates are statistically significant. This result is obtained without need to estimate σ^2 or d.

8. Conclusion and remarks

In this paper we propose a method for constructing confidence intervals for the mean or for testing statistical hypotheses for the mean of a dependent stationary sequence with finite second moment. The method is robust in the

Table 15
Proportions of confidence intervals covering the true mean using $h_n = \ln n/n$.

Sample size	7000	5000	3000	2000	1000	500
$\tilde{r}_n$, Gaussian kernel, $d = .1$	94.4	95.0	94.9	96.2	96.0	95.6
$\tilde{r}_n$, Gaussian kernel, $d = .3$	94.9	95.5	95.6	94.3	95.8	95.9
$\tilde{r}_n$, Gaussian kernel, $d = .4$	94.7	95.4	95.5	95.4	94.7	95.2
$\tilde{r}_n$, Gaussian kernel, $d = .6$	94.8	94.3	95.2	94.7	95.8	95.8

sense that we do not impose a specific restriction on the dependence structure of the sequence except for the ergodicity and stationarity. The estimators we propose are $\hat{r}_n$, defined by (3) and $\tilde{r}_n$ leading to the confidence intervals defined by (30). For simulations, it is convenient to use a kernel K following a standard normal distribution and to generate $(X_i;\ 1 \le i \le n)$ also from a standard normal variable. For this choice of f and K, we obtain $f(0) = 1/\sqrt{2\pi}$, $\int K^2(x)dx = 1/(2\sqrt{\pi})$, and the $(1-\alpha)100\%$ confidence intervals for μ_Y become

$$\left(\hat{r}_n - z_{\alpha/2}\left(\frac{\overline{Y_n^2}}{\sqrt{2}h_n n}\right)^{1/2},\ \hat{r}_n + z_{\alpha/2}\left(\frac{\overline{Y_n^2}}{\sqrt{2}h_n n}\right)^{1/2}\right) \tag{31}$$

and

$$\left(\tilde{r}_n - z_{\alpha/2}\left(\frac{\overline{Y_n^2}}{\sqrt{2}h_n n}\right)^{1/2},\ \tilde{r}_n + z_{\alpha/2}\left(\frac{\overline{Y_n^2}}{\sqrt{2}h_n n}\right)^{1/2}\right) \tag{32}$$

where

$$\tilde{r}_n = \frac{1}{nh_n}\sum_{i=1}^{n} Y_i \exp[-\frac{1}{2}(\frac{1}{h_n}X_i)^2] = c\hat{r}_n, \quad c = \frac{1}{\sqrt{1+h^2}}.$$

The size of the confidence interval depends on $\mathrm{var}(\bar{Y}_n)$ via condition (7), which restricts the size of nh_n. The larger $\mathrm{var}(\bar{Y}_n)$, the larger the size of the interval. Also, for the standard normal distribution of X and the Gaussian kernel, $\int x^2K(x)dx = 1$ and $f''(0) = -1/\sqrt{2\pi}$. Thus, the plug in estimator of the optimal bandwidth for estimating μ_Y is

$$h_o = (\frac{\overline{Y_n^2}}{n\sqrt{2}\bar{Y}_n^2})^{1/5} \text{ provided } \mathrm{var}(\bar{Y}_n) = o(n^{-0.8}). \tag{33}$$

The obtained optimal bandwidths minimize the mean squared error for large but finite samples. This is very different from minimizing the asymptotic variance, which would imply minimizing the length of the confidence interval as in the cited case of Epanechnikov. The MSE finds a balance between the bias and the variance of the estimator. After minimizing the MSE over h_n, we have minimized the size of the interval (via the asymptotic variance of the estimator) by looking for the kernel that provides the smallest interval from a given set of kernels (Gaussian, quartic, Epanechnikov, Uniform). We found out that the Gaussian, the quartic and Epanechnikov kernels perform best when X has the standard normal distribution. We also found that the standard normal distribution for X is better than the Cauchy distribution and the shifted $\chi^2(2)$.

In the case when $\mu = 0$, we have suggested to shift the data $(Y_i, i = 1 \cdots n)$ by a constant c. This shift by itself implies the reasonable assumption that a gain could be obtained by minimizing the MSE over (h, c). But, simple calculations show that the MSE cannot be minimized over (h, c) simultaneously. The minimization procedure leads to $c = 0$ as critical point. In general, for $\mu \neq 0$, the minimization procedure leads to $c = -\mu$, but for no possible h. Thus, the shift contributes to lower the bias when equal to $-\mu$, but in this case does not guarantee minimization of the asymptotic variance. A weaker condition on the sequence of Y observations implies the main result of this paper. Namely, convergence to the second moment of the average of sample squares. This follows from the proof of our result. Given that the boundary is defined for the bandwidths sequence h_n by $nh_n var(\bar{Y}) \to 0$, for implementation purposes it is good to consider estimating the variance of partial sums in some cases. Several papers in the literature address this problem. It is mostly done through estimation of the spectral density as in McElroy and Politis (2014), where taper-based estimates of the spectral density utilizing a fixed-b asymptotic framework are provided. Approximations of the variance of partial sums can also be found in Deligiannidis and Utev (2013) for weakly dependent stationary random sequences. In the case of ρ-mixing sequences, Peligrad and Shao (1995) recall an earlier result of Peligrad (1982)

$$var(S_n) \le 8000exp(2\sum_{i=1}^{[\ln n]} \rho(2^i))n\sigma^2,$$

and show that for square integrable ρ-mixing sequences with $\sum_{i=1}^{\infty} \rho(2^i) < \infty$, the CLT holds when $var(S_n) \to \infty$. Moreover, $var(S_n)/n \to \sigma^2$. They mention that the difficulty in using this CLT is in approximating the value of σ^2. Our result applies to such sequences without additional work to find σ^2 and with any h_n satisfying $nh_n \to \infty$.

Our results are asymptotic. We have conducted a numerical study to test the performance of the confidence intervals based on formula (31) on finite sample sizes. We have constructed confidence intervals based on samples from an ARFIMA(0, d, 0) with innovations $(\xi_j)_{j\in Z}$. In our simulations we vary the size of d, which controls the dependence strength, and accordingly the size of h_n. Since the second moment of Y is important we also vary the distribution of Y by considering various distributions for the innovations. In all the situations, for relatively large sample size, our methods returned reliable results.

Based on standard normal innovations we simulated an ARFIMA(0, .09, 0) sequence (Y'_n) and set $Y_n = 3 + Y'_n$. For a sample size $n = 100$, and using optimal bandwidth, we found that a 95% confidence for μ_Y is (2.9, 3.49), while for $n = 1000$ the 95% confidence for μ_Y is (2.81, 3.04). Constructed confidence intervals for long memory processes have shown that this procedure is very reliable and allows to have inference ignoring the memory parameter. This is very important, given that estimation of the memory parameter is cumbersome.

Acknowledgments

The authors thank reviewers for valuable inputs. The first author was supported by the College of Liberal Arts Summer grant and the second author was partially supported by the National Science Foundation (NSF), USA grant DMS-1811373 as well as a grant from the Taft Research Center.

Appendix

The following lemma might be known. However we could not find it in the literature.

Lemma 7. *Assume that $(Y_i)_{i\in Z}$ is a stationary sequence with finite second moments satisfying*

$$\bar{Y}_n \to E(Y_0) \text{ in probability as } n \to \infty. \tag{34}$$

Then

$$var(\bar{Y}_n) \to 0 \text{ as } n \to \infty.$$

Proof. Let us center the sequence at expectations. By stationarity, $E(\bar{Y}_n^2) < \infty$ and therefore (34) implies that $E|\bar{Y}_n| \to 0$. Note that

$$E(S_n^2) = nE(Y_0^2) + 2\sum_{i=1}^{n}\sum_{j=1}^{i-1} E(Y_0 Y_j)$$
$$= nE(Y_0^2) + 2EY_0[\sum_{i=1}^{n}\sum_{j=1}^{i-1} Y_j].$$

For all $A > 0$, by writing $|Y_0| = |Y_0|I[|Y_0| \leq A) + |Y_0|I(|Y_0| > A)$, and using the stationarity we obtain

$$\frac{E(S_n^2)}{n^2} \leq \frac{2A}{n}\sum_{i=1}^{n} E|\bar{Y}_i| + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=0}^{i-1} E|Y_0|I(|Y_0| > A)|Y_j|.$$

The first term on the right hand side consists of a Cesàro sum of a sequence which is convergent to 0. As for the second term, we apply first the Cauchy–Schwarz inequality and note that it is upper bounded by $\sqrt{E(Y_0^2 I(|Y_0| > A))}$ which is convergent to 0 as $A \to \infty$. The result follows.

References

Beran, J., 1959. Statistics for Long-Memory Processes. Chapman & Hall.
Beran, J., 1989. A test of location for data with slowly decaying serial correlations. Biometrika 76, 261–269.
Billingsley, P., 1995. Probability and Measure, third ed. John Willey & Sons.
Billingsley, P., 1999. Convergence of Probability Measures, second ed. John Willey & Sons.
Bosq, D., 1996. Nonparametric Statistics for Stochastic Processes: Estimation and Prediction. Springer-Verlag, New York.
Bosq, D., Merlevède, F., Peligrad, M., 1999. Asymptotic normality for density kernel estimators in discrete and continuous time. J. Multivariate Anal. 68, 79–95.
Bradley, R., 1983. Asymptotic normality of some kernel-type estimators of probability density. Statist. Probab. Lett. 1, 295–300.
Bradley, R.C., 2007. Introduction To Strong Mixing Conditions 1, 2, 3. Kendrick Press, Heber City, UT.
Collomb, G., 1984. Propriétés de convergence presque complète du prédicteur à noyau. Z. Wahrscheinlichkeitstheor. Verwandte Geb. 66, 441–460.
Deligiannidis, G., Utev, S., 2013. Variance of partial sums of stationary sequences. Ann. Probab. 41, 3606–3616.
Eisner, F., Farkas, B., Haase, M., Nagel, R., 2015. Operator theoretic aspects of ergodic theory. In: Graduate Texts in Mathematics, vol. 272, Springer.
Granger, C.W., Joyeux, R., 1980. An introduction to long-memory time series models and fractional differencing. J. Time Ser. Anal. 1, 15–29.
Härdle, W., 1991. Smoothing Techniques with Implementation in S. Springer-Veralg.
Hong, S.Y., Linton, O., 2016. Asymptotic properties of a nadaraya–watson type estimator for regression functions of infinite order. arXiv:1604.06380.

Hosking, J.R.M., 1981. Fractional differencing. Biometrika 68, 165–176.
Jones, M., Marron, J., Sheather, S., 1996. A brief survey of bandwidth selection for density estimation. J. Amer. Statist. Assoc. 91 (433), 401–407.
Krengel, U., 1985. Ergodic Theorems. De Gruyter Studies in Mathematics, Berlin.
Laib, N., Louani, D., 2010. Nonparametric kernel regression estimation for functional stationary ergodic data: Asymptotic properties. J. Multivariate Anal. 101, 2266–2281.
Long, H., Qian, L., 2013. Nadaraya–Watson estimator for stochastic processes driven by stable Lévy motions. Electron. J. Stat. 7, 1387–1418.
Longla, M., 2013. Remarks on the speed of convergence of mixing coefficients and applications. Statist. Probab. Lett. 83 (10), 2439–2445.
Longla, M., 2015. On mixtures of copulas and mixing coefficients. J. Multivariate Anal. 139, 259–265.
Longla, M., Peligrad, M., 2012. Some aspects of modeling dependence in copula-based Markov chains. J. Multivariate Anal. 111, 234–240.
McElroy, T.S., Politis, D.N., 2014. Spectral density and spectral distribution inference for long memory time series via fixed-b asymptotics. J. Econometrics 182, 211–225.
Nadaraya, E.A., 1964. On estimating regression. Theory Probab. Appl. 9, 141–142.
Nelsen, R.B., 2006. An Introduction To Copulas. Springer, New York.
Parzen, E., 1962. On estimation of a probability density function and mode. Ann. Math. Statist. 33 (3), 1065–1076.
Peligrad, M., 1982. Invariance principles for mixing sequences of random variables. Ann. Probab. 10, 968–981.
Peligrad, M., 1992. Properties of uniform consistency of the kernel estimators of density and of regression functions under dependence assumptions. Stoch. Stoch. Rep. 40, 147–168.
Peligrad, M., 1998. On the blockwise bootstrap for empirical processes for stationary sequences. Ann. Probab. 26, 877–901.
Peligrad, M., Shao, Q.-M., 1995. Estimation of the variance for partial sums for *rho*-mixing random variables. J. Multivariate Anal. 52, 140–157.
Thode, H., 2002. Testing for Normality. Marcel Dekker, New York.
Wang, Q., Lin, X.-Y., Gulati, C.M., 2001. Asymptotics for moving average processes with dependent innovations. Statist. Probab. Lett. 54, 347–356.
Watson, G.S., 1964. Smooth regression analysis. Sankhya A 26, 359–372.
Yoshihara, K., 1994. Weakly Dependent Stochastic Sequences and their Applications: Volume IV: Curve Estimation Based on Weakly Dependent Data. Sanseido, Tokyo, Japan.
Zhao, O., Woodroofe, M., Volný, D., 2010. A central limit theorem for reversible processes with nonlinear growth of variance. J. Appl. Probab. 47, 1195–1202.