JOURNAL OF TIME SERIES ANALYSIS
J. Time Ser. Anal. **35**: 4–15 (2014)
Published online 2 September 2013 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1111/jtsa.12044

ORIGINAL ARTICLE

NON-PARAMETRIC ESTIMATION UNDER STRONG DEPENDENCE

ZHIBIAO ZHAO[a,*,†], YIYUN ZHANG[b] AND RUNZE LI[a]

[a] *Penn State University*
[b] *Novartis Oncology*

We study non-parametric regression function estimation for models with strong dependence. Compared with short-range dependent models, long-range dependent models often result in slower convergence rates. We propose a simple differencing-sequence based non-parametric estimator that achieves the same convergence rate as if the data were independent. Simulation studies show that the proposed method has good finite sample performance.

Received 20 August 2012; Revised 18 June 2013; Accepted 9 July 2013

Keywords: Differencing; long-range dependence; non-parametric regression; short-range dependence; time series.
JEL. C14; C22.

1. INTRODUCTION

In the literature on non-parametric inference, considerable attention has been paid to independent or short-range dependent (SRD) processes. While being a reasonable assumption in some cases, the independence or SRD assumption may present a serious restriction in other applications. For example, long-range dependence (LRD) has been frequently observed in hydrology (Mandelbrot and Wallis, 1969), Internet communication traffic (Leland et al., 1994), and financial markets (Ding et al., 1993; Casas and Gao, 2008). For extensive exposition of LRD phenomena and their applications, we refer the reader to Beran (1992), Doukhan et al. (2003), and Robinson (2003).

Models with LRD often exhibit different behaviour from independent or SRD models, and hence, techniques developed for independent or SRD data may become less efficient or even fail in the presence of LRD. As shown in Mielniczuk and Wu (2004), the convergence rate of non-parametric estimation depends on the strength of dependence. For other contributions, see Hall and Hart (1990), Csörgő and Mielniczuk (1995, 1999), Robinson (1997), Masry and Mielniczuk (2001), and Yang (2001) for non-parametric estimation and Yajima (1991), Fox and Taqqu (1986), Giraitis and Surgailis (1990), Koul (1992), and Robinson and Hidalgo (1997) for parametric estimation in LRD models. As demonstrated by these works, estimations for LRD models usually have slower convergence rates that depend on the strength of dependence. However, it is generally a challenging problem to estimate the amount of dependence in non-parametric regression (Robinson, 1997).

In this article, we study non-parametric estimation for the following model:

$$Y_i = \mu(X_i) + g(i/n) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\{X_i\}$ are i.i.d. random covariates, $\{\varepsilon_i\}$ are stationary time-series noises with $E(\varepsilon_i) = 0$, $g(\cdot)$ is a time trend assumed to be Lipschitz continuous, and $\mu(\cdot)$ is a non-parametric function of the covariate X_i. By including both a time trend $g(\cdot)$ and a covariate function $\mu(\cdot)$, model (1) is more general and flexible than those studied in

* Correspondence to: Zhibiao Zhao, Department of Statistics, Penn State University, 326 Thomas Building, University Park, PA 16802, USA.
†E-mail: zuz13@stat.psu.edu

previous works where they considered either the time trend only $Y_i = g(i/n) + \varepsilon_i$ (Hall and Hart, 1990; Csörgő and Mielniczuk, 1995; Robinson, 1997; Beran and Feng, 2002) or the covariate function only $Y_i = \mu(X_i) + \varepsilon_i$ (Csörgő and Mielniczuk, 1999). While it is also important to draw inference about the overall time trend $g(\cdot)$ (Altman, 1990; Robinson, 1997; Wu and Zhao, 2007), our focus here is to estimate $\mu(\cdot)$, which quantifies the covariate effect.

We focus on the case that $\{\varepsilon_i\}$ exhibit strong dependence. We propose a simple differencing-sequence based non-parametric estimation method. It is shown that the proposed method can remove the LRD of $\{\varepsilon_i\}$ and thus can achieve the same convergence rate as if the data were independent. In contrast to existing works that often rely on a good estimate of the dependence (Robinson, 1997; Masry and Mielniczuk, 2001), the proposed method does not assume any knowledge about the amount or form of the dependence, and hence, it is non-parametric. Our simulation study shows that the proposed method delivers better finite sample performance than the existing method.

We introduce some notation. For sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \asymp b_n$ if $a_n/b_n \to 1$, $a_n = O(b_n)$ if $|a_n/b_n|$ is bounded, and $a_n \sim b_n$ if $|a_n/b_n|$ is bounded away from 0 and ∞.

2. KERNEL SMOOTHING ESTIMATION UNDER DEPENDENCE

In (1), the functions μ and g are identifiable only up to a constant: for any constant c, $\tilde{\mu} = \mu - c$ and $\tilde{g} = g + c$ also satisfy the same model. In practice, it often suffices to know a function up to a constant. For example, if $\mu(x)$ is the mean response of a patient taking dose x of a medicine, then the difference $\mu(x) - \mu(0)$ represents the treatment effect.

Throughout, we assume that we wish to estimate $\mu(\cdot)$ up to a constant at a fixed point x; see Remark 1 for discussions on the identifiability issue. Consider the popular Nadaraya–Watson kernel smoothing estimator

$$\hat{\mu}(x) = \frac{\sum_{i=1}^n Y_i K_i}{\sum_{i=1}^n K_i}, \quad \text{where} \quad K_i = K\left(\frac{x - X_i}{b_n}\right). \tag{2}$$

Here, $K(\cdot)$ is a kernel function and b_n is a bandwidth satisfying $b_n \to 0$ and $nb_n \to \infty$. Denote by $f_X(x)$ the density function of X_i. Then, we have the decomposition

$$\hat{\mu}(x) - \mu(x) = \omega_n(C_n + B_n + S_n) \quad \text{with} \quad \omega_n = \frac{nb_n f_X(x)}{\sum_{i=1}^n K_i}, \tag{3}$$

where

$$C_n = [nb_n f_X(x)]^{-1} \sum_{i=1}^n g(i/n)E(K_i), \tag{4}$$

$$B_n = [nb_n f_X(x)]^{-1} \sum_{i=1}^n [\mu(X_i) - \mu(x)]K_i, \tag{5}$$

$$S_n = [nb_n f_X(x)]^{-1} \sum_{i=1}^n \{\varepsilon_i K_i + g(i/n)[K_i - E(K_i)]\}. \tag{6}$$

By the well-known theory of non-parametric density estimation, under conditions A1–A4 in the Appendix, $\omega_n \to 1$. In (3), C_n is an unidentifiable constant, B_n is the bias, and S_n is the stochastic term that determines the asymptotic variance of $\hat{\mu}(x)$. We can show

$$C_n \to \int_0^1 g(t)\mathrm{d}t. \tag{7}$$

DOI: 10.1111/jtsa.12044

Consequently, $\hat{\mu}(x)$ is an estimator of $\mu(x) + \int_0^1 g(t)\mathrm{d}t$. Since both $\hat{\mu}(x)$ and our proposed estimator (Section 3) have the same bias B_n, we shall focus on S_n only.

Remark 1. To ensure the identifiability of (1), we may assume $\mathbb{E}[\mu(X_i)] = 0$. Then, we can rewrite (1) as $Y_i = g(i/n) + \varepsilon_i^*$, where the new error $\varepsilon_i^* = \mu(X_i) + \varepsilon_i$ has zero mean. Thus, the time trend $g(\cdot)$ can be consistently estimated (see, e.g., the references cited in Section 1), denote the estimator by $\hat{g}(\cdot)$. Then, we can consider the detrended data $Y_i^* = Y_i - \hat{g}(i/n) = [g(i/n) - \hat{g}(i/n)] + \mu(X_i) + \varepsilon_i \approx \mu(X_i) + \varepsilon_i$. The covariate function $\mu(\cdot)$ can be estimated by applying (2) to (X_i, Y_i^*). All the aforementioned and subsequent analysis holds with $g(i/n)$ in (4) replaced by $g(i/n) - \hat{g}(i/n)$, and the resultant C_n is negligible.

For $\{\varepsilon_i\}$ in (1), denote by $\gamma_k = \mathrm{cov}(\varepsilon_i, \varepsilon_{i+k}), k \in \mathbb{Z}$,the autocovariance function. The process $\{\varepsilon_i\}$ is said to be SRD if $\sum_{k=1}^{\infty} |\gamma_k| < \infty$ or LRD if $\sum_{k=1}^{\infty} |\gamma_k| = \infty$. Theorem 1 studies the asymptotic variance of S_n.

Theorem 1. Assume that conditions A1–A5 in the Appendix hold. For S_n in (6),

$$\mathrm{var}(S_n) \asymp \frac{\gamma_0 + \int_0^1 g^2(t)\mathrm{d}t}{n b_n f_X(x)} \int_{\mathbb{R}} K^2(u)\mathrm{d}u + \frac{2\sum_{k=1}^{n}(n-k)\gamma_k}{n^2} := \Sigma_1 + \Sigma_2. \tag{8}$$

From (8), the asymptotic variance of S_n has two components: Σ_1 is the contribution from the aggregated noise level of $g(i/n) + \varepsilon_i$, and Σ_2 is the contribution from the dependence of $\{\varepsilon_i\}$. If $\{\varepsilon_i\}$ are independent so that $\gamma_k = 0, k \geq 1$, then $\Sigma_2 = 0$ and $\mathrm{var}(S_n) \asymp \Sigma_1$. For dependent processes, the relative magnitude of Σ_1 and Σ_2 depends on both the strength of dependence and the choice of bandwidth b_n. In particular, we have

$$\mathrm{var}(S_n) \asymp \begin{cases} \Sigma_1, & \text{if } b_n \sum_{k=1}^{n}(1-k/n)\gamma_k \to 0, \quad \text{small bandwidth;} \\ \Sigma_2, & \text{if } b_n \sum_{k=1}^{n}(1-k/n)\gamma_k \to \infty, \quad \text{large bandwidth;} \\ \Sigma_1 + \Sigma_2, & \text{if } 0 < c_1 < b_n \sum_{k=1}^{n}(1-k/n)\gamma_k < c_2 < \infty, \quad \text{borderline case.} \end{cases} \tag{9}$$

Therefore, for the small bandwidth case, $\mathrm{var}(S_n)$ achieves the same rate as if the data were independent. Clearly, the small bandwidth case covers all SRD processes for which $|b_n \sum_{k=1}^{n}(1-k/n)\gamma_k| \leq b_n \sum_{k=1}^{n}|\gamma_k| = O(b_n) \to 0$. For LRD processes under the large bandwidth scenario, the dependence term Σ_2 becomes the leading term. For the borderline case between small and large bandwidth, both Σ_1 and Σ_2 play non-negligible roles.

Csörgő and Mielniczuk (1999) studied the model $Y_i = \mu(X_i) + \varepsilon_i$ with i.i.d. covariates $\{X_i\}$. They assumed that $\varepsilon_i = G(X_i, Z_i)$ for some function $G(\cdot,\cdot)$ and an LRD Gaussian process $\{Z_i\}$ with $\mathrm{cov}(Z_i, Z_{i+k}) = L(k)/k^{\alpha}$ for some $\alpha \in (0,1)$ and a slowly varying function $L(\cdot)$. Their Theorems 1–3 showed that the Nadaraya–Watson kernel smoothing estimator of $\mu(\cdot)$ has different asymptotic behaviour, depending on whether the bandwidth is large, small, or intermediate. Therefore, the three-scenario phenomenon in (9) is in parallel to their results. However, in practice, it is difficult to determine whether a given bandwidth is small, large, or intermediate, and thus, it is desirable to have an estimator that has a unified convergence rate. Furthermore, our simulation study in Section 4 shows that too small or too large bandwidths lead to poor finite sample performance.

To see how Σ_2 in (8) depends on the dependence, we can easily obtain

$$\Sigma_2 \sim \begin{cases} \log(n)/n, & \text{if } \gamma_k \sim k^{-1}; \\ n^{-\lambda}, & \text{if } \gamma_k \sim k^{-\lambda}, \lambda \in (0,1); \\ \log^{-\varpi}(n), & \text{if } \gamma_k \sim \log^{-\varpi}(k), \varpi > 0. \end{cases} \tag{10}$$

DOI: 10.1111/jtsa.12044

We see that, in the presence of strong dependence, $\mathrm{var}(S_n)$ can decay very slowly.

Example 1 (Linear processes)**.** For $\{a_j\}$ satisfying $\sum_{j=0}^{\infty} a_j^2 < \infty$, define the linear process $\varepsilon_i = \sum_{j=0}^{\infty} a_j \xi_{i-j}$, where $\{\xi_i\}$ are i.i.d. innovations with $E(\xi_i) = 0$ and $E(\xi_i^2) = 1$. Then, $\gamma_k = \sum_{j=0}^{\infty} a_j a_{j+k}$. If $a_j \sim j^{-\lambda}$ with $\lambda > 1/2$, then simple calculations show that

$$\gamma_k \sim \begin{cases} k^{-\lambda}, & \text{if } \lambda > 1; \\ k^{-1} \log k, & \text{if } \lambda = 1; \\ k^{-(2\lambda-1)}, & \text{if } 1/2 < \lambda < 1. \end{cases}$$

The two cases $\lambda > 1$ and $1/2 < \lambda \leq 1$ correspond to the SRD and LRD cases, respectively. For the case a_j decaying at the rate of $j^{-1/2}$ up to a logarithm factor (to ensure $\sum_{j=0}^{\infty} a_j^2 < \infty$), by the proof in the Appendix,

$$\gamma_k \sim \log^{-\varpi}(k), \quad \text{if } a_j \sim j^{-1/2} \log^{-(1+\varpi)/2}(j) \text{ for some } \varpi > 0. \tag{11}$$

Then, $\mathrm{var}(S_n)$ can be calculated using (10).

Example 2 (Fractionally integrated process)**.** Denote by B the backshift operator. For $d \in (-1/2, 1/2)$, consider the pure fractionally integrated process $I(d)$: $(1-B)^d \varepsilon_i = \xi_i$ for i.i.d. innovations $\{\xi_i\}$ with $E(\xi_i) = 0$ and $E(\xi_i^2) = 1$. Then, the representation $\varepsilon_i = \sum_{j=0}^{\infty} a_j \xi_{i-j}$ holds with $a_j \sim j^{d-1}$. Depending on the value of d, the process can exhibit quite different behaviour. If $d \in (0, 1/2)$, then $\{\varepsilon_i\}$ is LRD. If $d \in (-1/2, 0)$, then we have anti-persistence: the autocovariances are summable but decay at a polynomial rate slower than the exponential rate of causal ARMA models. See Granger (1980), Granger and Joyeux (1980), and Lo (1991) for more discussions.

For $\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q \in \mathbb{R}$, define $\alpha(z) = 1 - \sum_{i=1}^{p} \alpha_i z^i$ and $\beta(z) = 1 + \sum_{i=1}^{q} \beta_i z^i$. The aforementioned properties continue to hold for general FARIMA(p, d, q): $\alpha(B)(1-B)^d \varepsilon_i = \beta(B)\xi_i$, provided that the polynomial $\alpha(\cdot)$ has all its roots outside the unit circle.

Example 3 (Near unit-root models)**.** Near unit-root models have been extensively studied (Phillips, 1988). Consider the near unit-root model $\varepsilon_i = \lambda_n \varepsilon_{i-1} + \sqrt{1-\lambda_n^2}\xi_i$, where $\lambda_n < 1$ but $\lambda_n \to 1$, and $\{\xi_i\}$ are i.i.d. innovations with $E(\xi_i) = 0$ and $E(\xi_i^2) = 1$. Then, $\gamma_k = \lambda_n^k$. Write $\lambda_n = 1 - \delta_n$ with $\delta_n \to 0$. We can show that $\sum_{k=1}^{n}(n-k)\gamma_k \sim n/\delta_n$ if $\liminf_{n\to\infty} n\delta_n > 0$ and $\sum_{k=1}^{n}(n-k)\gamma_k \sim n^2$ if $n\delta_n \to 0$. In these cases, as $\lambda_n \to 1$, the dependence becomes increasingly stronger and $\mathrm{var}(S_n)$ can be very large. As a result, the estimator $\hat{\mu}(x)$ in (2) may perform poorly.

Example 4 (Fractional Gaussian noises)**.** Fractional Gaussian noise (Mandelbrot and Van Ness, 1968) with index $\alpha \in (0, 1)$ is a centred Gaussian process with autocovariance $\gamma_k = [(k+1)^{2\alpha} - 2k^{2\alpha} + (k-1)^{2\alpha}]/2, k \geq 0$. Then, $\sum_{k=1}^{\infty} |\gamma_k| < \infty$ for $\alpha \in (0, 1/2]$ and $\sum_{k=1}^{\infty} |\gamma_k| = \infty$ for $\alpha \in (1/2, 1)$, leading to SRD and LRD, respectively.

3. NON-PARAMETRIC ESTIMATION VIA DIFFERENCING SEQUENCE

For ε_i in (1), define the differencing sequence:

$$\eta_i = \varepsilon_i - \varepsilon_{i-1}. \tag{12}$$

For the $I(d)$ model $(1-B)^d \varepsilon_i = \xi_i, d \in (-1/2, 1/2)$, in Example 2, $\eta_i = (1-B)^{-(d-1)}\varepsilon_i$. By Granger (1980), $\mathrm{cov}(\eta_i, \eta_{i+k}) \sim k^{2d-3}, k \in \mathbb{N}$. Thus, $\sum_{k=0}^{\infty} |\mathrm{cov}(\eta_0, \eta_k)| = O(1)$.

DOI: 10.1111/jtsa.12044

In general, the autocovariances of $\{\eta_i\}$ are given by (recall that $\gamma_k = \mathrm{cov}(\varepsilon_i, \varepsilon_{i+k})$)

$$\tilde{\gamma}_k = \mathrm{cov}(\eta_i, \eta_{i+k}) = (\gamma_k - \gamma_{k+1}) - (\gamma_{k-1} - \gamma_k). \tag{13}$$

Then, $\sum_{k=1}^{\infty} |\tilde{\gamma}_k| \le 2\sum_{k=0}^{\infty} |\gamma_k - \gamma_{k+1}|$. If γ_k is non-increasing, then $\sum_{k=1}^{\infty} |\tilde{\gamma}_k| \le 2\gamma_0$. A special example is the near unit-root model in Example 3 with $\gamma_k = \lambda_n^k, \lambda_n \in (0,1)$. For the fractional Gaussian noise in Example 4, we can show $\gamma_k - \gamma_{k-1} = O(k^{2\alpha-3})$ and thus $\sum_{k=1}^{\infty} |\tilde{\gamma}_k| = O(1)$ for $\alpha \in (0,1)$. As another example, let $\gamma_k = k^{-\lambda}L(k)$ for some $\lambda \in (0,1)$ and a slowly varying function $L(\cdot)$ at ∞. Assume that $L(1+k)/L(k) = 1 + O(1/k)$. Then, we can show $\gamma_k - \gamma_{k+1} = O[k^{-(1+\lambda)}L(k)]$ and thus $\sum_{k=1}^{\infty} |\tilde{\gamma}_k| < \infty$. Therefore, for the models in Examples 1–4, the differencing sequence $\{\eta_i\}$ is SRD.

Remark 2. While differencing can alleviate the strength of dependence, it may also introduce some undesirable issues, such as the non-invertibility. To appreciate the benefit of differencing, consider the near unit-root model in Example 3 with $\lambda_n = 0.95$, then $\gamma_1 = 0.95$, suggesting strong dependence. For the differenced data, by (13), $\tilde{\gamma}_1 = -(1-\lambda_n)^2 = -0.0025 \approx 0$. Our simulation study in Section 4 suggests that estimators based on differenced data tend to have better finite sample performance for highly correlated data.

Motivated by the aforementioned discussion, we propose a differencing-sequence based non-parametric estimator of $\mu(x)$ (up to a constant). For Y_i in (1), define the difference $\tilde{Y}_i = Y_i - Y_{i-1}$. In (1), assume that $g(\cdot)$ is Lipschitz continuous. Then,

$$\tilde{Y}_i = Y_i - Y_{i-1} = O(1/n) + \mu(X_i) - \mu(X_{i-1}) + \eta_i, \quad \text{where} \quad \eta_i = \varepsilon_i - \varepsilon_{i-1}. \tag{14}$$

Because $\{X_i\}$ are i.i.d. and independent of $\{\eta_i\}$, $\mathbb{E}(\tilde{Y}_i | X_i = x) = \mu(x) - \mathbb{E}[\mu(X_0)] + O(1/n)$. Therefore, as in (2), consider the Nadaraya–Watson kernel smoothing estimator

$$\tilde{\mu}(x) = \frac{\sum_{i=1}^n \tilde{Y}_i K_i}{\sum_{i=1}^n K_i}, \quad \text{where} \quad K_i = K\left(\frac{x - X_i}{b_n}\right). \tag{15}$$

Similar to (3), we have the decomposition

$$\tilde{\mu}(x) - \mu(x) = O(1/n) + \omega_n(\tilde{C}_n + B_n + \tilde{S}_n), \tag{16}$$

where ω_n and B_n are defined as in (3) and (5),

$$\tilde{C}_n = [n b_n f_X(x)]^{-1} \sum_{i=1}^n E[\mu(X_{i-1})K_i], \tag{17}$$

$$\tilde{S}_n = [n b_n f_X(x)]^{-1} \sum_{i=1}^n \{[\eta_i - \mu(X_{i-1})]K_i + \mathbb{E}[\mu(X_{i-1})K_i]\}. \tag{18}$$

As in the decomposition (3), $\tilde{C}_n$ is an unidentifiable constant, B_n is the bias, and $\tilde{S}_n$ is the stochastic term that determines the asymptotic variance of $\tilde{\mu}(x)$. We can show

$$\tilde{C}_n \to \mathbb{E}[\mu(X_0)]. \tag{19}$$

Consequently, $\tilde{\mu}(x)$ is an estimator of $\mu(x) - \mathbb{E}[\mu(X_0)]$. In particular, under the identifiability condition $\mathbb{E}[\mu(X_i)] = 0$ in Remark 1, $\tilde{\mu}(x)$ is an estimator of $\mu(x)$.

DOI: 10.1111/jtsa.12044

Theorem 2. Assume that conditions A1–A5 in the Appendix hold and that $\mathbb{E}[\mu^2(X_0)] < \infty$. Further assume that $\sum_{k=0}^{\infty} |\gamma_k - \gamma_{k+1}| < \infty$ so that $\{\eta_i\}$ is SRD. Then,

$$\operatorname{var}(\tilde{S}_n) \asymp \frac{2(\gamma_0 - \gamma_1) + E[\mu^2(X_0)]}{n b_n f_X(x)} \int_{\mathbb{R}} K^2(u)\mathrm{d}u. \tag{20}$$

By Theorem 2, the proposed differencing-sequence based estimator $\tilde{\mu}(x)$ achieves the same convergence rate $\sqrt{nb_n}$ as if the data were independent. Unlike $\operatorname{var}(S_n)$ in (8) that depends on γ_k at all lags, $\operatorname{var}(\tilde{S}_n)$ depends on $\gamma_0 - \gamma_1$ only. Define the relative variance

$$\mathrm{RV} = \frac{\operatorname{var}(\tilde{S}_n)}{\operatorname{var}(S_n)}. \tag{21}$$

By Theorems 1 and 2, we can easily obtain Corollary 1.

Corollary 1. Assume the same conditions in Theorems 1 and 2.

(i) If $\{\varepsilon_i\}$ is SRD, then

$$\mathrm{RV} \to \frac{2(\gamma_0 - \gamma_1) + E[\mu^2(X_0)]}{\gamma_0 + \int_0^1 g^2(t)\mathrm{d}t}.$$

(ii) If $\{\varepsilon_i\}$ is LRD and $b_n \sum_{k=1}^{n}(1 - k/n)\gamma_k \to \infty$, then RV$\to 0$.

By Corollary 1, for LRD processes, under appropriate conditions on b_n, the differencing-sequence based estimator $\tilde{\mu}(x)$ has a relative variance converging to zero. For SRD processes, the relative variance depends on γ_0, γ_1, $E[\mu^2(X_0)]$, and $\int_0^1 g^2(t)\mathrm{d}t$. Since the goal is to estimate $\mu(\cdot)$, we can view $\mu(\cdot)$ and $g(\cdot)$ as the 'signal' and 'noise', respectively. By Corollary 1(i), as the noise $\int_0^1 g^2(t)$ increases, the relative performance of $\tilde{\mu}(x)$ becomes increasingly better. To see the effect of γ_0 and γ_1, consider the special case $\mu(\cdot) = g(\cdot) \equiv 0$, then RV $= 2(1-\rho_1)$, where $\rho_k = \gamma_k/\gamma_0$ is the autocorrelation function. For strongly dependent data (for example, the near unit-root model in Example 3), we often have $\rho_1 \geq 1/2$, which implies RV ≤ 1. We illustrate this phenomenon via simulations.

4. A SIMULATION STUDY

We carry out a small simulation study to examine the performance of our proposed differencing-sequence based estimator $\tilde{\mu}(\cdot)$ in (15) to that of the direct kernel smoothing estimator $\hat{\mu}(\cdot)$ in (2). By the discussions in Sections 2 and 3, both estimators are estimates of $\mu(\cdot)$ up to some constant: $\int_0^1 g(t)\mathrm{d}t$ for $\hat{\mu}(\cdot)$ and $E[\mu(X_0)]$ for $\tilde{\mu}(\cdot)$. To make a sensible comparison, we choose $g(\cdot)$ and $\mu(\cdot)$ satisfying $\int_0^1 g(t)\mathrm{d}t = 0$ and $E[\mu(X_0)] = 0$ so that both $\hat{\mu}(\cdot)$ and $\tilde{\mu}(\cdot)$ are consistent estimators of $\mu(\cdot)$.

For $\tilde{\mu}$ in (15), its mean integrated squared error (MISE) is computed as follows:

(i) Simulate n pairs of observations (X_i, Y_i) from a given model.
(ii) On the basis of the differences $\tilde{Y}_i$ in (14), obtain the estimator $\tilde{\mu}$ in (15).
(iii) Compute $D_n = \int_{\ell_2}^{\ell_1} [\tilde{\mu}(x) - \mu(x)]^2 \mathrm{d}x$, $[\ell_1, \ell_2]$ is the interval over which $\mu(\cdot)$ is estimated.
(iv) Repeat (i)–(iii) 1000 times, the MISE of $\tilde{\mu}$ is the average of those 1000 $D_n's$.

DOI: 10.1111/jtsa.12044

Similarly, we can compute the MISE of $\hat{\mu}$ in (2).

In our simulation, we consider the following model:

$$Y_i = \mu(X_i) + g(i/n) + \sigma\varepsilon_i, \quad X_i : \text{ uniform } [0,1], \quad i = 1, \ldots, n, \tag{22}$$

where $\mu(x) = \sin(2\pi x)$, $g(t) = \cos(2\pi t)$, and $\sigma > 0$ controls the noise level. In the aforementioned procedure, the MISE is calculated by approximating the integral $\int_{\ell_2}^{\ell_1}$ using 99 evenly spaced grid points on $[\ell_1, \ell_2] = [0.01, 0.99]$.

For the errors $\{\varepsilon_i\}$, we consider two models

$$\text{Model I:} \quad \varepsilon_i = \sum_{j=0}^{\infty} a_j \xi_{i-j}, \quad a_j = \frac{(j+1)^{-\lambda}}{\sqrt{\sum_{j=0}^{\infty}(j+1)^{-2\lambda}}}, \quad \lambda > 1/2. \tag{23}$$

$$\text{Model II:} \quad \varepsilon_i = \theta\varepsilon_{i-1} + \sqrt{1-\theta^2}\xi_i, \quad \theta \in [0,1). \tag{24}$$

Here, $\{\xi_i\}$ are i.i.d. standard normal random variables. For both models, the coefficients are properly normalized so that $\text{var}(\varepsilon_i) = 1$. For model I, by Example 1, as λ decreases, the strength of dependence increases. With $\lambda = 1$ determining the boundary, $\lambda > 1$ and $\lambda \in (0.5, 1]$ correspond to SRD and LRD, respectively. We consider four choices of $\lambda = 3/2, 1, 2/3, 0.51$. Then, $\lambda = 3/2$ results in SRD, whereas the other three choices lead to LRD. Furthermore, the strength of the LRD for $\lambda = 1, 2/3, 0.5$, increases from mild, intermediate, to strong dependence.

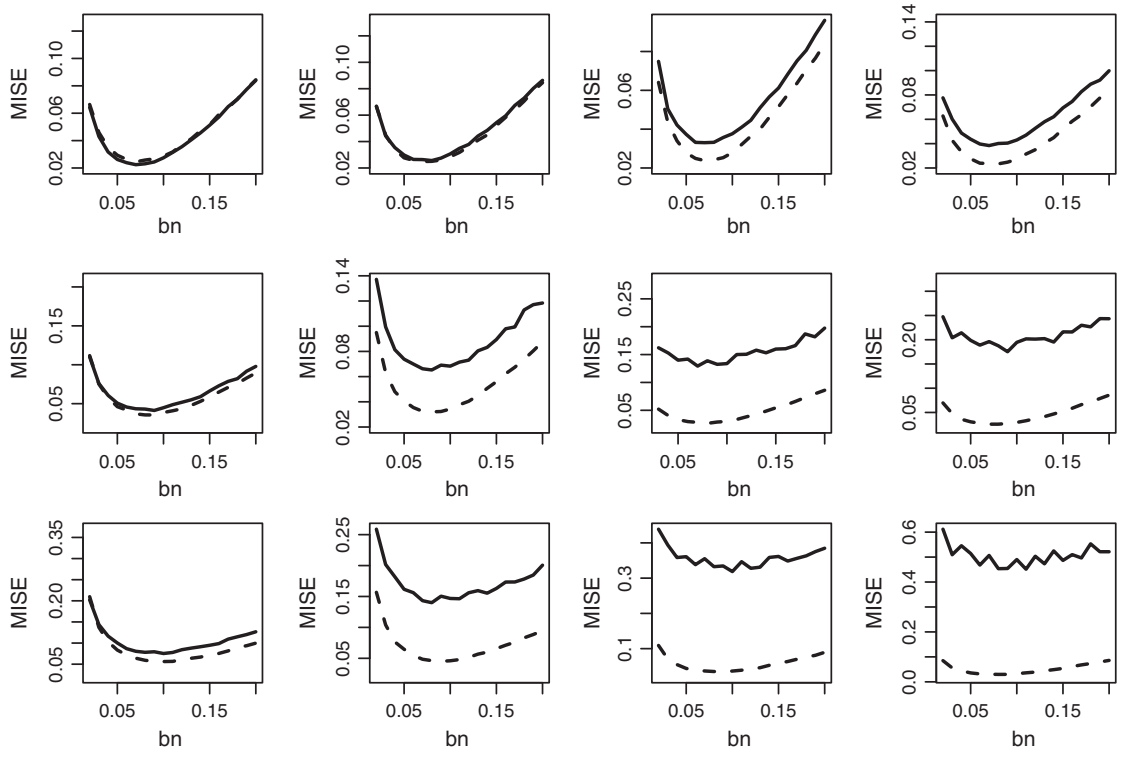

Figure 1. MISE comparison for model I in (23) as a function of bandwidth b_n (horizontal axis). Sample size $n = 150$. Solid and dashed curves are the MISEs for $\hat{\mu}$ and $\tilde{\mu}$, respectively. Top row (from left to right): $\sigma = 0.2$ and $\lambda = 3/2$, 1, 2/3, 0.51, respectively; middle row (from left to right): $\sigma = 0.6$ and $\lambda = 3/2$, 1, 2/3, 0.51, respectively; bottom row (from left to right): $\sigma = 1.0$ and $\lambda = 3/2$, 1, 2/3, 0.51, respectively

DOI: 10.1111/jtsa.12044

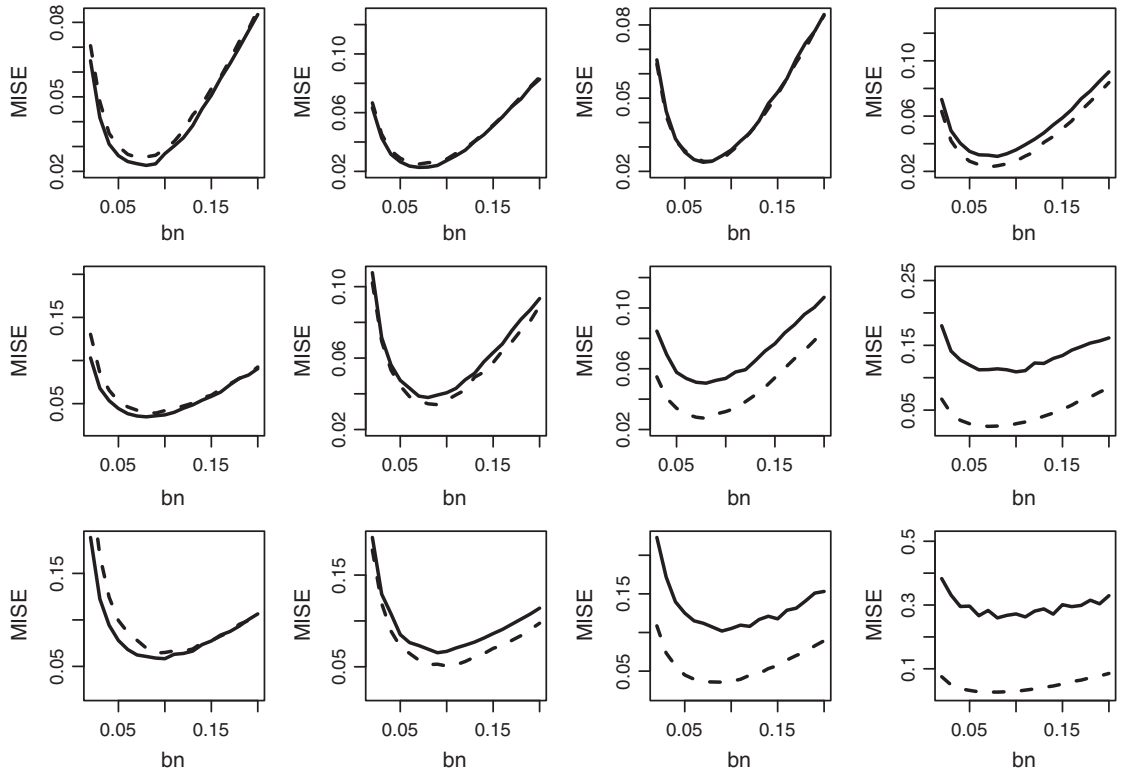

Figure 2. MISE comparison for model II in (24) as a function of bandwidth b_n (horizontal axis). Sample size $n = 150$. Solid and dashed curves are the MISEs for $\hat{\mu}$ and $\tilde{\mu}$, respectively. Top row (from left to right): $\sigma = 0.2$ and $\theta = 0.2, 0.5, 0.8, 0.95$, respectively; middle row (from left to right): $\sigma = 0.6$ and $\theta = 0.2, 0.5, 0.8, 0.95$, respectively; bottom row (from left to right): $\sigma = 1.0$ and $\theta = 0.2, 0.5, 0.8, 0.95$, respectively

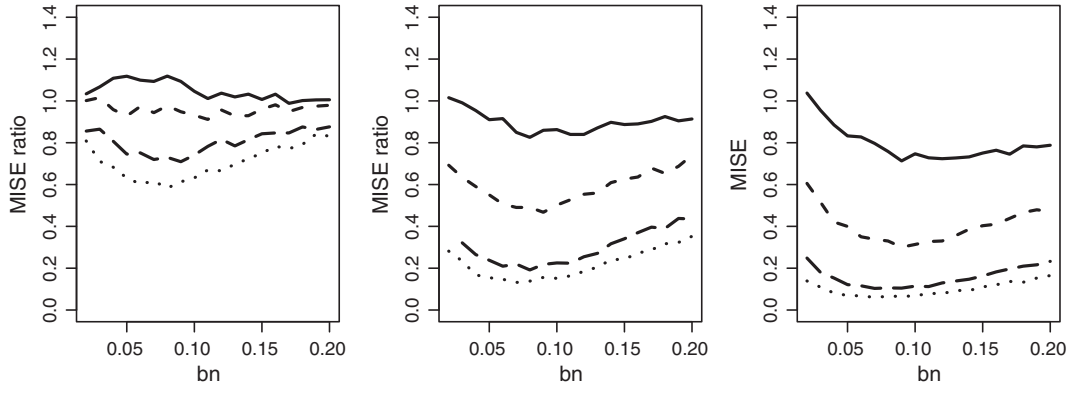

Figure 3. MISE ratio $\mathrm{MISE}(\tilde{\mu})/\mathrm{MISE}(\hat{\mu})$ for model I in (23) as a function of bandwidth b_n (horizontal axis). Sample size $n = 150$. From left to right: $\sigma = 0.2, 0.6, 1.0$, respectively. In each plot, solid, dashed, long-dash, and dotted curves correspond to $\lambda = 3/2, 1, 2/3, 0.51$, respectively

DOI: 10.1111/jtsa.12044

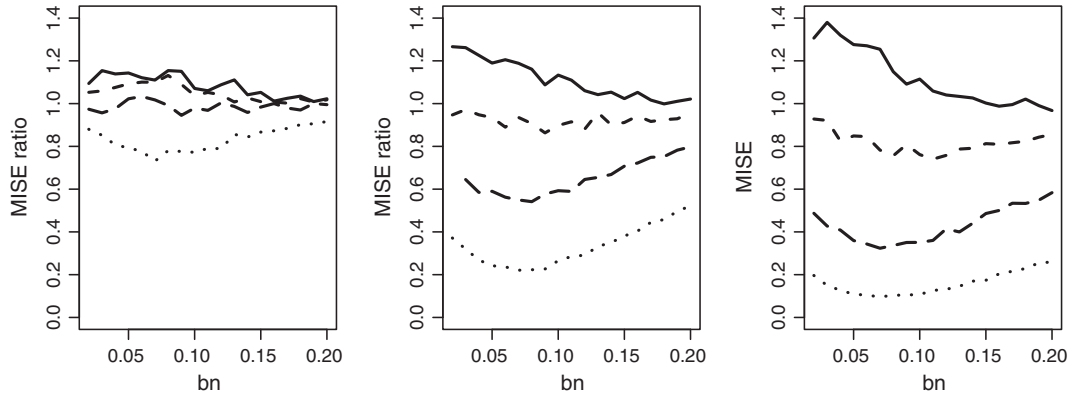

Figure 4. MISE ratio MISE($\tilde{\mu}$)/MISE($\hat{\mu}$) for model II in (24) as a function of bandwidth b_n (horizontal axis). Sample size $n = 150$. From left to right: $\sigma = 0.2, 0.6, 1.0$, respectively. In each plot, solid, dashed, long-dash, and dotted curves correspond to $\theta = 0.2, 0.5, 0.8, 0.95$, respectively

For model II, we consider $\theta = 0.2, 0.5, 0.8, 0.95$, where the two cases $\theta = 0.8$ and $\theta = 0.95$ are used to examine the performance under the near unit-root model setting (Example 3).

We use sample size $n = 150$ to evaluate the MISE across different choices of bandwidth $b_n = 0.01, 0.02, \ldots, 0.20$ and noise level $\sigma = 0.2, 0.6, 1.0$. In our simulation, we adopt the local linear estimator (Fan and Gijbels, 1996) to reduce the boundary effect. The function `locpoly` under the package `KernSmooth` in software `R` fits local polynomial regression.

Figures 1 and 2 present the MISE for models I and II, respectively. From Figure 1, $\tilde{\mu}$ performs uniformly better than $\hat{\mu}$ in the presence of intermediate or strong dependence ($\lambda = 2/3, 0.51$) and high noise level ($\sigma = 0.6, 1.0$). Furthermore, the two estimators have comparable performance under weak dependence ($\lambda = 3/2, 1$) or low noise level ($\sigma = 0.2$). As the dependence increases from $\lambda = 3/2$ to $\lambda = 0.51$ (from left to right in Figure 1), $\tilde{\mu}$ tends to be increasingly better. As the noise level increases from $\sigma = 0.2$ to $\sigma = 1.0$ (from top to bottom in Figure 1), the dependence plays an increasingly more significant role, and $\tilde{\mu}$ tends to be increasingly better. Figure 3 presents the MISE ratio MISE($\tilde{\mu}$)/MISE($\hat{\mu}$). We observe that, as the dependence increases (from top to bottom curves in each plot of Figure 3) or as the noise level increases (from left to right plots in Figure 3), the ratio tends to decrease. This is in good agreement with our theoretical result. For model II, Figures 2 and 4 exhibit similar patterns. In summary, we conclude that the proposed estimator delivers overall superior performance.

ACKNOWLEDGEMENTS

We are grateful to two anonymous referees for their insightful comments that have significantly improved an earlier draft. Zhao's research was supported by a National Institute on Drug Abuse (NIDA) grant P50-DA10075-15. Li's research was supported by NIDA grants R21-DA024260 and P50-DA10075-15 and a National Natural Science Foundation of China grant 11028103. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the NIH.

REFERENCES

Altman, NS. (1990) Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association* **85**: 749–759.

Beran, J. (1992) Statistical methods for data with long-range dependence. *Statistical Science* **7**: 404–416.

DOI: 10.1111/jtsa.12044

Beran, J, Feng, Y. (2002) Local polynomial fitting with long-memory, short-memory and antipersistent errors. *Annals of the Institute of Statistical Mathematics* **54**: 291–311.
Casas, I, Gao, J. (2008) Econometric estimation in long-range dependent volatility models: theory and practice. *Journal of Econometrics* **147**: 72–83.
Csörgő, S, Mielniczuk, J. (1995) Nonparametric regression under long-range dependent normal errors. *Annals of Statistics* **23**: 1000–1014.
Csörgő, S, Mielniczuk, J. (1999) Random-design regression under long-range dependence. *Bernoulli* **5**: 209–224.
Ding, Z, Granger, CWJ and Engle, RF. (1993) A long memory property of stock market returns and a new model. *Journal of Empirical Finance* **1**: 83–106.
Doukhan, P, Oppenheim, G and Taqqu, MS. (2003) *Theory and Applications of Long-Range Dependence*. Boston: Birkhäuser Press.
Fan, J, Gijbels, I. (1996) *Local Polynomial Modeling and Its Applications*. London: Chapman & Hall.
Fox, R, Taqqu, MS. (1986) Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *Annals of Statistics* **14**: 517–532.
Giraitis, L, Surgailis, D. (1990) A central limit theorem for quadratic forms in strongly dependent linear variables and its application to asymptotic normality of Whittle's estimate. *Probability Theory and Related Fields* **86**: 87–104.
Granger, CWJ. (1980) Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics* **14**: 227–238.
Granger, CWJ, Joyeux, R. (1980) An introduction to long range time series models and fractional differencing. *Journal of Time Series Analysis* **1**: 15–29.
Hall, PG, Hart, JD. (1990) Nonparametric regression with long-range dependence. *Stochastic Processes and their Applications* **36**: 339–351.
Koul, HL. (1992) M-estimators in linear models with long range dependent errors. *Statistics & Probability Letters* **14**: 153–164.
Leland, WE, Taqqu, MS, Willinger, W and Wilson, DV. (1994) On the self-similar nature of Ethernet traffic. *IEEE/ACM Transactions on Networking* **2**: 1–15.
Lo, AW. (1991) Long-term memory in stock market prices. *Econometrica* **59**: 1279–1313.
Mandelbrot, BB, Van Ness, JW. (1968) Fractional Brownian motion, fractional noises and applications. *SIAM Review* **10**: 422–437.
Mandelbrot, BB, Wallis, JR. (1969) Some long-run properties of geophysical records. *Water Resources Research* **5**: 321–340.
Masry, E, Mielniczuk, J. (2001) Local linear regression estimation for time series with long-range dependence. *Stochastic Processes and their Applications* **82**: 173–193.
Mielniczuk, J, Wu, WB. (2004) On random-design model with dependent errors. *Statistica Sinica* **14**: 1105–1126.
Phillips, PCB. (1988) Regression theory for near-integrated time series. *Econometrica* **56**: 1021–1043.
Robinson, PM. (1997) Large-sample inference for nonparametric regression with dependent errors. *Annals of Statistics* **25**: 2054–2083.
Robinson, PM. (2003). *Long memory time series*. In *Time Series with Long Memory*, Robinson, PM (ed.), Advanced texts in econometrics. Oxford University Press: Oxford, pp. 1–48.
Robinson, PM, Hidalgo, FJ. (1997) Time series regression with long-range dependence. *Annals of Statistics* **25**: 77–104.
Wu, WB, Zhao, Z. (2007) Inference of trends in time series. *Journal of the Royal Statistical Society: Series B* **69**: 391–410.
Yajima, Y. (1991) Asymptotic properties of LSE in a regression model with long memory stationary errors. *Annals of Statistics* **19**: 158–177.
Yang, Y. (2001) Nonparametric regression and prediction with dependent errors. *Bernoulli* **7**: 633–655.

APPENDIX: ASSUMPTIONS AND PROOFS

Recall that $f_X(x)$ is the density function of X_i and x is a fixed point. We impose the following regularity conditions that are commonly used in non-parametric estimation.

A1 $X_1, \ldots, X_n$ are i.i.d. random variables independent of $\varepsilon_1, \ldots, \varepsilon_n$.
A2 $f_X(\cdot)$ is twice continuously differentiable in a neighborhood of x, and $f_X(x) > 0$.
A3 The kernel function K has a bounded support, is symmetric and differentiable with bounded derivative, and integrates to one.
A4 The bandwidth b_n satisfies the natural condition $b_n \to 0$ and $nb_n \to \infty$.
A5 $g(\cdot)$ is Lipschitz continuous.

DOI: 10.1111/jtsa.12044

Proof of Theorems 1 and 2

We only give the proof for Theorem 1 since Theorem 2 can be similarly treated. By the symmetry of K, $\int_{\mathbb{R}} uK(u)\mathrm{d}u = 0$. Under assumptions A1–A4, elementary calculations show that

$$E(K_i) \asymp b_n f_X(x), \qquad E(K_i^2) \asymp b_n f_X(x) \int_{\mathbb{R}} K^2(u)\mathrm{d}u. \tag{A.1}$$

Let $e_i = [\varepsilon_i + g(i/n)]K_i - g(i/n)E(K_i)$. Since $\{X_i\}$ are i.i.d. and independent of $\{\varepsilon_i\}$, using (A.1), we can show

$$\begin{aligned} \mathrm{var}(e_i) &= \gamma_0 E(K_i^2) + g^2(i/n)\mathrm{var}(K_i) \asymp b_n f_X(x)[\gamma_0 + g^2(i/n)] \int_{\mathbb{R}} K^2(u)\mathrm{d}u, \\ \mathrm{cov}(e_i, e_j) &= \gamma_{j-i} E(K_i)E(K_j) \asymp b_n^2 f_X^2(x)\gamma_{j-i}. \end{aligned}$$

Finally, (8) easily follows from $\mathrm{var}(S_n) = \sum_{i=1}^n \mathrm{var}(e_i) + 2\sum_{1\le i<j\le n} \mathrm{cov}(e_i, e_j)$. QED

Proof of (11)

Denote by $\lfloor z \rfloor$ the largest integer not exceeding z. Note the decomposition

$$\begin{aligned} \sum_{j=1}^{\infty} a_j a_{j+k} &\sim \left\{ \sum_{j=k+1}^{\infty} + \sum_{j=\lfloor\sqrt{k}\rfloor}^{k} + \sum_{j=2}^{\lfloor\sqrt{k}\rfloor-1} \right\} \frac{1}{\sqrt{j(j+k)\log^{1+\varpi}(j)\log^{1+\varpi}(j+k)}} \\ &:= I_1 + I_2 + I_3. \end{aligned}$$

We consider the three terms separately. First, $I_3 \le \sum_{j=2}^{\lfloor\sqrt{k}\rfloor-1}(jk)^{-1/2} = O(k^{-1/4})$. Notice that, for $\lfloor\sqrt{k}\rfloor \le j \le k$, we have $j + k \sim k$ and $\log(j) \sim \log(j+k) = O[\log(k)]$. Thus, $I_2 \sim \log^{-(1+\varpi)}(k) \sum_{j=\lfloor\sqrt{k}\rfloor}^{k}(jk)^{-1/2} \sim \log^{-(1+\varpi)}(k)$. For I_1, notice that

$$\frac{1}{\sqrt{j(j+k)\log^{1+\varpi}(j)\log^{1+\varpi}(j+k)}} \sim \frac{1}{j\log^{1+\varpi}(j)} \sim \frac{1}{2^s k \log^{1+\varpi}(2^s k)}$$

uniformly over $2^s k + 1 \le j \le 2^{s+1}k$ and $s \ge 0$. Therefore,

$$\begin{aligned} I_1 = \sum_{s=0}^{\infty} \sum_{j=2^s k+1}^{2^{(s+1)}k} &\sim \sum_{s=0}^{\infty} \sum_{j=2^s k+1}^{2^{s+1}k} \frac{1}{2^s k \log^{1+\varpi}(2^s k)} \\ &= \sum_{s=0}^{\infty} \frac{1}{[\log(k) + s\log(2)]^{1+\varpi}} \\ &= \left\{ \sum_{s=\lfloor\log(k)\rfloor+1}^{\infty} + \sum_{s=0}^{\lfloor\log(k)\rfloor} \right\} \frac{1}{[\log(k) + s\log(2)]^{1+\varpi}} \sim \log^{-\varpi}(k), \end{aligned}$$

  J. Time Ser. Anal. **35**: 4–15 (2014)
DOI: 10.1111/jtsa.12044

where the last assertion follows from

$$\sum_{s=0}^{\lfloor \log(k) \rfloor} \frac{1}{[\log(k) + s\log(2)]^{1+\varpi}} \sim \sum_{s=0}^{\lfloor \log(k) \rfloor} \frac{1}{\log^{1+\varpi}(k)} \sim \log^{-\varpi}(k),$$
$$\sum_{s=\lfloor \log(k) \rfloor + 1}^{\infty} \frac{1}{[\log(k) + s\log(2)]^{1+\varpi}} \sim \sum_{s=\lfloor \log(k) \rfloor + 1}^{\infty} \frac{1}{s^{1+\varpi}} \sim \log^{-\varpi}(k).$$

This completes the proof. □

J. Time Ser. Anal. **35**: 4–15 (2014)
DOI: 10.1111/jtsa.12044