On Nonparametric Prediction of Linear Processes

By Jan Mielniczuk[1], Zhou Zhou[2] and Wei Biao Wu

Polish Academy of Sciences and University of Chicago

May 17, 2009

Abstract

We consider nonparametric prediction problem for both short- and long-range dependent linear processes. Asymptotic properties of local linear estimates are obtained and, for long-range dependent processes, an interesting dichotomous phenomenon is described: the limiting distribution depends on the interplay between the strength of the dependence and the magnitude of the bandwidth. A simulation study is carried out to assess the performance of the nonparametric prediction estimator.

Keywords. Dichotomy; Local linear prediction; long- and short-range dependence; linear process; subsampling.

1 Introduction

An important problem in the study of time series is prediction. Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary process with $\mathbb{E}(X_t^2) < \infty$. The classical Kolmogorov-Wiener theory concerns predicting future values by linear combinations of past values of the process. See Pourahmadi (2001) for an extensive treatment. If the underlying process is Gaussian, then the conditional expectation of a future value given the past values is a linear combination of the past values, and the linear predictor is thus indeed optimal with respect to mean squared error. For non-Gaussian processes, however, the linear relationship does not generally hold, and it appears quite difficult to find parametric forms of the predictor. In this case, one will obtain erroneous results if a linear predictor is used. To circumvent the difficulty, we can resort to nonparametric predictors such as the kernel-based Nadaraya-Watson or local linear estimators.

[1]This author is also with Warsaw University of Technology

[2]Corresponding author. *E-mail address:* zhouzhou@uchicago.edu.

In the paper we consider the estimation of the one-step ahead predictive function $g(x) = \mathbb{E}(X_{t+1}|X_t = x)$ based on observations $X_0, X_1, \ldots, X_n$ of the process. The function g is linear if (X_t) is Gaussian or if it is the linear AR(1) process $X_t = aX_{t-1} + \varepsilon_t$, where $|a| < 1$ and ε_t are independent and identically distributed (i.i.d.) with mean 0. There is a vast body of literature on the parallel problem of nonparametric estimation of conditional mean function in a regression setting when errors are independent, see for example Eubank (1988) or Härdle (1990). For nonparametric estimators of autoregression functions in specific models see Collomb and Härdle (1986), chapter 3 in Bosq (1996) and Wu and Huang (2006). Chapter 10 of Fan and Yao (2004) contains a general discussion of nonlinear prediction problems. For related problems, including estimation of conditional variances; see McKeague and Zhang (1994), Chen (1996), Phillips and Park (1998) and Robinson (1983). Various modeling strategies for linear least-squares prediction of long-memory or long-range dependent time series are discussed in Bhansali and Kokoszka (2004). Wu and Huang (2006) considered autoregressive function estimation for the case when $X_{t+1} = R(X_t, \varepsilon_{t+1})$ where (ε_t) is a sequence of i.i.d. innovations such that ε_{t+1} is independent of X_t. They proved that under mild weak dependence conditions the Nadaraya-Watson estimate of $g(x)$ is asymptotically normal with asymptotic variance $\mathrm{Var}(X_{t+1}|X_t = x)\kappa/f(x)$, where $\kappa = \int K^2(s)\,ds$ and f is the marginal density of X_t. This parallels analogous property of regression estimator in a random design heteroscedastic regression model with weakly dependent errors.

Kernel estimators in random design models with long-range dependent errors have been studied in several papers; see Csörgő and Mielniczuk (1999) and Mielniczuk and Wu (2004). Masry and Mielniczuk (1999) dealt with local linear estimators in the case of such errors. Assume that $\mathbb{E}(X_k) = 0$. Let $r(k) = \mathbb{E}(X_0 X_k)$ be the covariance function of the process (X_t). Generically speaking, the process (X_t) is said to be long-memory or long-range dependent if $r(k)$ is not summable:

$$\sum_{k\in\mathbb{Z}} r(k) = \infty. \tag{1}$$

It is known that in this case regression estimator exhibits dichotomous behavior for which correct normalization ensuring non-degenerate asymptotic distribution results from comparison of strength of dependence and size of a bandwidth. Here, we find that the same phenomenon holds for the prediction problem and determine normalizing constants and

asymptotic distributions for both parts of the dichotomy. We argue in Section 3 that for long-range dependent processes size of confidence intervals is mainly determined by the strength of dependence. In the paper we only consider one step ahead prediction based on the last available observation. It is possible to extend this to the case of m step predictors based on the lagged p values. Analogous asymptotic results can be similarly established. However, the conditions under which non-degenerate asymptotic law is obtained are rather complicated and the issue of curse of dimensionality emerges. So we do not pursue here.

The paper is structured as follows. Section 2 introduces the linear process model and the local linear estimate. Asymptotic properties of the estimate are discussed in Section 3, where both short- and long-range dependent processes are considered. Section 4 presents a simulation study concerning the performance of the nonparametric predictor and its dichotomous behavior for moderate size samples. Proofs are given in Section 5.

2 Preliminaries

A popular model for strong dependence or long-memory is linear processes (moving averages) with slowly decaying coefficients. Consider the one-sided infinite order moving average MA(∞) process

$$X_t = \sum_{i=0}^{\infty} a_i \varepsilon_{t-i}, \tag{2}$$

where $\varepsilon_i, i \in \mathbb{Z}$, are i.i.d. random variables with mean 0, $\mathbb{E}(\varepsilon_i^2) = \sigma^2 < \infty$ and coefficients $(a_i)_0^\infty$ are square summable. We assume without loss of generality that $a_0 = 1$. The strength of dependence of the process (X_t) is determined by the decay rate of (a_t). If $a_i = \ell(i) i^{-\beta}$, $i \in \mathbb{N}$, where $1/2 < \beta < 1$ and $\ell(\cdot)$ is slowly varying at ∞, routine calculations based on Karamata's theorem (Feller, 1971) imply that

$$r(i) = \mathrm{Cov}(X_0, X_i) \sim C_\beta \ell^2(i) i^{-(2\beta-1)} \mathbb{E}(\varepsilon_i^2), \text{ where } C_\beta = \int_0^\infty (x + x^2)^{-\beta}\, dx. \tag{3}$$

Here $a_n \sim b_n$ means $\lim_{n\to\infty} a_n/b_n \to 1$. Thus in this case sum of covariances diverge and we have (1). This is a case of long-range dependence (LRD) or long-memory which should be contrasted with the short-range dependence (SRD) case of summable covariances. Note that the frequently used models of long-range dependent sequences, namely fractionally

integrated ARIMA process and fractional Gaussian noise have representation (2). For a readable introduction to long-memory processes discussing MA(∞) expansions see Beran (1994).

Generally the function $g(x) = \mathbb{E}(X_{t+1}|X_t = x)$ is nonlinear in x even though (X_t) is a linear process. To estimate g, we shall apply the local linear method to the pairs $(X_0, X_1), \ldots, (X_{n-1}, X_n)$. Let K be a symmetric probability density, $K_b(\cdot) = K(\cdot/b)/b$ and $b_n > 0$ is a bandwidth sequence satisfying $b_n \to 0$ and $nb_n \to \infty$; moreover, $\mathcal{S}_{n,l}(x) = n^{-1}\sum_{i=0}^{n-1} K_{b_n}(X_i - x)\{(X_i - x)/b_n\}^l$. Let $(\hat{g}_n(x), \hat{h}_n(x))$ be defined as

$$(\hat{g}_n(x), \hat{h}_n(x)) = \text{argmin}_{\beta_0,\beta_1} \sum_{i=1}^{n} (X_i - \beta_0 - \beta_1(X_{i-1} - x))^2 K_{b_n}(X_{i-1} - x). \tag{4}$$

Then $\hat{g}_n(x)$ and $\hat{h}_n(x)$ are estimators of $g(x)$ and its derivative $g'(x)$, respectively. The bandwidth b_n determines the amount of smoothing employed by the local linear method. Recognizing that (4) is a weighted least squares regression problem one can represent $\hat{g}_n(x)$ as follows. For a column vector V, let $V[i]$ be the ith entry of V from the top. Define $\mathcal{S}_n(x) = ((\mathcal{S}_{n,0}(x), \mathcal{S}_{n,1}(x))^T, (\mathcal{S}_{n,1}(x), \mathcal{S}_{n,2}(x))^T)$, $\mathbf{X} = (\mathbf{1}_n, b_n^{-1}(X_0 - x, X_1 - x, \ldots, X_{n-1} - x)^T)$, where $\mathbf{1}_n$ is the column vector composed of n 1's. Let $\mathbf{W} = \text{diag}\{K_{b_n}(X_0 - x), \ldots, K_{b_n}(X_{n-1} - x)\}$ and $Y = (X_1, X_2, \ldots, X_n)^T$. Noting that $\mathcal{S}_n(x) = n^{-1}\mathbf{X^T W X}$, $\hat{g}_n(x) - g(x)$ can be be written as (Fan and Gijbels (1996))

$$\hat{g}_n(x) - g(x) = n^{-1}\{\mathcal{S}_n^{-1}(x)\mathbf{X}^T\mathbf{W}(Y - \mathbf{X}\theta_n(x))\}[1],$$

where $\theta_n(x) = (g(x), b_n g'(x))^T$. An alternative representation is

$$\hat{g}_n(x) = \frac{1}{n}\sum_{i=1}^{n} w_n(x, i)X_i, \tag{5}$$

where $w_n(x, i)$ are weights given by

$$w_n(x, i) = \frac{(\mathcal{S}_{n,2}(x) - \mathcal{S}_{n,1}(x)(X_{i-1} - x)/b_n)K_{b_n}(X_{i-1} - x)}{\mathcal{S}_{n,0}(x)\mathcal{S}_{n,2}(x) - \mathcal{S}_{n,1}^2(x)}.$$

Define

$$d_n(x) = \frac{1}{n}\{(\mathbb{E}\mathcal{S}_n(x))^{-1}\mathbb{E}[\mathbf{X}^T\mathbf{W}(Y - \mathbf{X}\theta_n(x))]\}[1]. \tag{6}$$

We shall see from Theorems 1 and 2 that $d_n(x)$ is the asymptotic bias of $\hat{g}_n(x)$.

The local constant analogue of (5) is the Nadaraya-Watson estimator of $g(x)$ defined in Remark 4. It is well-known that (Fan and Gijbels (1996)) local linear estimator has several advantages over the local constant estimator such as alleviation of bias at boundary points. In Section 3 we shall consider asymptotic properties of $\hat{g}_n(x) - g(x)$.

Many previous asymptotic results on nonparametric estimation of autoregressive function relied on Markovian assumptions of the underlying time series models such as $X_{t+1} = R(X_t, \varepsilon_{t+1})$ from which it follows that $\mathbb{E}(X_{t+1}|\ldots, \varepsilon_{t-1}, \varepsilon_t) = \mathbb{E}(X_{t+1}|X_t)$, see for example Theorem 3.7 in Bosq (1996). Here, for model (2) the latter equality fails. Thus technical aspect of all derivations differs much from previous techniques employed for this problem.

3 Asymptotic Theory

Denote by f_ε the density of ε_1 and let $\kappa = \int K^2(s)\, ds$. Recall $a_0 = 1$. Let $Y_t = X_{t+1} - \varepsilon_{t+1}$ and $Z_t = X_{t+2} - \varepsilon_{t+2} - a_1\varepsilon_{t+1}$. Then both Y_t and Z_t are $\mathcal{F}_t = (\ldots, \varepsilon_{t-1}, \varepsilon_t)$-measurable. Assume throughout the paper that the kernel K is symmetric, compactly supported and bounded. Let $N(\mu, \sigma^2)$ denote a normal distribution with mean μ and variance σ^2 and $\xrightarrow{\mathcal{D}}$ the weak convergence. Asymptotic properties of $\hat{g}_n$ under short- and long-range dependence are presented in Theorems 1 and 2, respectively.

Theorem 1. *Assume that (i) $\sum_{i=0}^{\infty} |a_i| < \infty$; (ii) f_ε is Lipschitz continuous; (iii) g is twice continuous differentiable at x and $f(x) \neq 0$; (iv) $b_n \to 0$ and $nb_n \to \infty$; and (v) $\mathbb{E}|\varepsilon_i|^q < \infty$ for some $q > 2$. Then*

$$(nb_n)^{1/2}[\hat{g}_n(x) - g(x) - d_n(x)] \xrightarrow{\mathcal{D}} N[0,\, \kappa(v_1^2(x) + v_2^2(x))/f^2(x)], \tag{7}$$

where $v_1^2(x) = \sigma^2 f(x)$ and $v_2^2(x) = \mathbb{E}[(Z_{t-1} + a_1(x - Y_{t-1}) - g(x))^2 f_\varepsilon(x - Y_{t-1})]$.

The result can be better understood by writing (X_t) in the form

$$X_{t+1} = g(X_t) + \varepsilon_{t+1} + Y_t - g(X_t) := g(X_t) + e_{1,t+1} + e_{2,t+1}, \tag{8}$$

where $e_{1,t+1} = \varepsilon_{t+1}$ and $e_{2,t+1} = Y_t - g(X_t)$. Then $\mathbb{E}(e_{1,t+1}|X_t) = \mathbb{E}(e_{2,t+1}|X_t) = 0$, moreover, $e_{1,t+1}$ and $e_{2,t+1}$ are uncorrelated. It can be seen from the proof of Theorem 1 that the decomposition of the asymptotic variance in (7) corresponds to decomposition of the error in (8). Actually, $(nb_n)^{-1}\kappa\sigma^2/f(x)$ is the usual form of the asymptotic variance

of the local linear estimator when homoscedastic errors are independent from the random regressors. The form of the second summand of the asymptotic variance in (7) is more complicated due to the $MA(\infty)$ structure of (X_t). Note that the asymptotic variance is *not* equal $\kappa \mathrm{Var}(X_{t+1}|X_t = x)/f(x)$ as in the case of random design regression estimation or in the autoregressive model considered by Wu and Huang (2006).

Remark 1. It is a routine exercise to deal with the asymptotic bias $d_n(x)$, which does not involve the dependence structure of (X_t). Elementary calculations show that

$$d_n(x) = C_B(x) b_n^2 + o(b_n^2),$$

where $C_B(x) = \mu_2 g''(x)/2$ and $\mu_2 = \int s^2 K(s)\, ds$. So the central limit theorem in (7) holds with $g_n(x)$ therein replaced by $g(x)$ if $nb_n^5 \to 0$. In Section 4.1, a bias corrected estimator is proposed. $\diamondsuit$

It is worthwhile to mention that the central limit theorem (7) holds under the natural condition on the bandwidth (iv) and the natural short-range dependence condition (i), without any additional constraints on the decay rates of (a_i) and (b_n).

For long-range dependent processes, the asymptotic behavior of $\hat{g}_n(x)$ has a more complicated nature. To this end, we introduce

$$J(y, z) = [z + a_1(x - y) - g(x)] f_\varepsilon(x - y).$$

Let $\sigma_n = \| \sum_{t=1}^n X_t \|$, where $\| \cdot \|$ denotes $\mathcal{L}^2$ norm, namely, $\|X\| = (\mathbb{E}X^2)^{1/2}$. Recall (3) for the definition of C_β. For long-range dependent processes with $a_i = \ell(i) i^{-\beta}$, where $1/2 < \beta < 1$ and $\ell(\cdot)$ is slowly varying, by Karamata's theorem we have

$$\sigma_n^2 \sim D_\beta n^{2-(2\beta-1)} \ell^2(n) \mathbb{E}(\varepsilon_1^2), \text{ where } D_\beta = \frac{C_\beta}{(2 - 2\beta)(3/2 - \beta)}.$$

Let $\mathbf{1}_{2\times 2}$ be the 2×2 unit matrix consisting of ones and

$$J_\infty(y, z) = \mathbb{E}[J(y + Y_t, z + Z_t)] \text{ and } J'_\infty(0) = [\frac{\partial}{\partial y} J_\infty(y, z), \frac{\partial}{\partial z} J_\infty(y, z)]\Big|_{(y,z)=(0,0)}. \tag{9}$$

Under suitable regularity conditions,

$$J'_\infty(0) = [-a_1 f(x) - \mathbb{E}\{f'_\varepsilon(x - Y_t)(Z_t + a_1(x - Y_t) - g(x))\},\ f(x)].$$

Theorem 2. *Assume that (i)* $\sup_u\{f_\varepsilon(u) + |f'_\varepsilon(u)| + |f''_\varepsilon(u)|\} < \infty$, $\mathbb{E}(|\varepsilon_i|^q) < \infty$ *for some* $q > 2$*; (ii)* $a_i = \ell(i)i^{-\beta}$ *where* $1/2 < \beta < 1$ *and* ℓ *is slowly varying; (iii)* $b_n \to 0$ *and* $nb_n \to \infty$*; and (iv) g is continuous at x and* $f(x) \neq 0$*. Then [a] under* $\sigma_n/n = o((nb_n)^{-1/2})$*, we have (7); [b] under* $(nb_n)^{-1/2} = o(\sigma_n/n)$*, we have*

$$\frac{n}{\sigma_n}[\hat{g}_n(x) - g(x) - d_n(x)] \xrightarrow{\mathcal{D}} N[0,\, s^2(x)],\ \textit{where}\ s^2(x) = \frac{J'_\infty(0)\mathbf{1}_{2\times 2}J'_\infty(0)^T}{f^2(x)}. \tag{10}$$

Remark 2. Theorems 1 and 2 allow for multivariate extensions. Let $x_1 < x_2 < \ldots < x_k$ and $W_n = [\hat{g}_n(x_1) - g_n(x_1) - d_n(x_1), \ldots, \hat{g}_n(x_k) - g(x_k) - d_n(x_k)]$. Then under assumptions of Theorem 1, $(nb_n)^{1/2}W_n \xrightarrow{\mathcal{D}} (\gamma_1 Z_1, \ldots, \gamma_k Z_k)$, where Z_i are independent $N(0,1)$ and $\gamma_i^2 = \kappa(v_1^2(x_i) + v_2^2(x_i))/f^2(x_i)$. For a proof of the latter claim, we can use the Cramér-Wold device. The details are omitted since they involve no essential extra difficulties. The same holds true in case [a] of Theorem 2 whereas for the case [b] we have $(n/\sigma_n)W_n \xrightarrow{\mathcal{D}} Z_1(s(x_1), \ldots, s(x_k))$. Thus for large bandwidths satisfying [b], coordinates of the asymptotic law only differ by multiplicative constants. ♢

Remark 3. As noted in Remark 1, the bias $d_n(x) = C_B(x)b_n^2 + o(b_n^2)$. Thus the first part of the dichotomy remains valid when $nb_n^5 \to c_0 \neq 0$ with an asymptotic mean equal to $c_0 C_B(x)$ instead of 0. For such b_n, if $9/10 < \beta$, then $\sigma_n/n = o((nb_n)^{-1/2})$. However, if $1/2 < \beta < 9/10$, then the latter is violated. ♢

Remark 4. Both theorems are also valid for Nadaraya-Watson estimator of g defined as $\tilde{g}_n(x) = \sum_{i=1} X_i K_{b_n}(x - X_{i-1}) / \sum_{i=1} K_{b_n}(x - X_{i-1}) =: v_n(x)/w_n(x)$ with the centering term $g(x) + d_n(x)$ replaced by $g_n(x) = \mathbb{E}v_n(x)/\mathbb{E}w_n(x)$. Furthermore, if f and g have two continuous derivatives in a neighborhood of x, then it is easy to check that $g_n(x) - g(x) = \bar{C}_B(x)b_n^2 + o(b_n^2)$, where $\bar{C}_B(x) = \mu_2((fg)^{(2)}(x) - gf^{(2)}(x))/[2f(x)]$. Thus in the both results for Nadaraya-Watson estimator $g_n(x)$ can be replaced by $g(x)$ provided $nb_n^5 \to 0$. ♢

Remark 5. Both presented results are valid for $X'_t = \mu + X_t$, where μ is the mean and (X_t) is as in (2) under the same conditions for (X'_t) as those assumed for (X_t). Let $g_{X'}(\cdot)$ and $\hat{g}_{n,X'}(\cdot)$ denote the one-step ahead predictive function and its local linear estimator for the process (X'_t), respectively. The claim is justified by noting that $g_{X'}(x) = \mu + g_X(x - \mu)$, $\hat{g}_{n,X'}(x) = \mu + \hat{g}_{n,X}(x - \mu)$, and that asymptotic variances in (7) and (10) calculated at $x - \mu$ for the process (X'_t) and at x for the process (X_t), coincide. ♢

Remark 6. In our long-range dependence model, we assume that (X_t) is a linear process of form (2). The linearity assumption is crucial for Theorem 2. Robinson (1991) showed that, for long-range dependent processes which are functionals of Gaussian processes, the limiting distribution for the kernel density estimate may be non-Gaussian. See Csörgő and Mielniczuk (1995a, 1995b) for parallel papers on regression. $\diamondsuit$

Theorem 2 describes the interesting dichotomous phenomenon: if the bandwidth b_n is small such that $\sigma_n/n = o((nb_n)^{-1/2})$, then the asymptotic distribution of $\hat{g}_n(x) - g_n(x)$ is same as the one obtained under short-range dependence. On the other hand, for larger bandwidths, one has a central limit theorem (10) with a different normalizing constant and a different asymptotic variance: both quantities are changed. If $b_n = n^{-\alpha}\ell_1(n)$, where $\alpha \in (0,1)$ and ℓ_1 is slowly varying, let the triangles $T_1 = \{(\alpha,\beta) \in (0,1) \times (1/2,1) : 2-2\beta < \alpha\}$ and $T_2 = \{(\alpha,\beta) \in (0,1) \times (1/2,1) : 2-2\beta > \alpha\}$. Then $\sigma_n/n = o((nb_n)^{-1/2})$ holds if $(\alpha,\beta) \in T_1$, while $(nb_n)^{-1/2} = o(\sigma_n/n)$ if $(\alpha,\beta) \in T_2$. For the boundary case when $2-2\beta = \alpha$, we conjecture that the limiting distribution is the convolution of distributions in (7) and (10). One can refer to Theorem 4 in Wu and Mielniczuk (2002) for a similar result on kernel density estimation for linear processes.

From Theorem 2 we conjecture that the Asymptotic Mean Squared Error (MSE) of $\hat{g}_n(x)$ for LRD sequences satisfies

$$\mathrm{MSE}(g_n(x)) \sim \frac{v^2(x)}{nb_n} + C_B^2(x)b_n^4 + \frac{\sigma_n^2}{n^2}s^2(x),$$

where $v^2(x)$ is the limiting variance in Theorem 1 and $C_B(x)$ is defined in Remark 1. Hall and Hart (1990) and Yang (2001) proved results on such behavior of $MSE(x)$ for kernel density estimators and regression estimators in a random design model for LRD observations. Thus in the second case of dichotomy when $(nb_n)^{-1/2} = o(\sigma_n/n)$, the main term of the variance of $\hat{g}_n(x)$ does not depend on b_n. In other words, minimization of the above expression resulting in $b_n = C_{opt}n^{-1/5}$ yields $MSE(x)$ of order $\max(n^{-4/5}, \sigma_n^2/n^2)$. As noted in Remark 3 this order equals $n^{-4/5}$ for $\beta > 9/10$ when 'light' LRD occurs, but equals σ_n^2/n^2 when $\beta < 9/10$. This has important consequences for the related statistical inference such as construction of confidence intervals, as it implies that the minimal asymptotic variance of $\hat{g}_n(x)$ for $\beta < 9/10$ equals $s^2(x)\sigma_n^2/n^2(1+o(1))$ and is attained for $b_n = Cn^{-1/5}$ with *any* $C > 0$. Note also that the constant $C_{opt} = (v^2(x)/4C_B^2(x))^{1/5}$ for the MSE

optimal bandwidth $b_n = C_{opt} n^{-1/5}$ depends in an involved way on parameters of the linear process and it is not clear how to estimate it. Here we shall propose to use the generalized cross-validation (GCV) (Wahba (1977) and Craven and Wahba (1979)) method. It works as follows: write the local linear estimate $\hat{g}_n(x)$ as $\hat{g}_{n,b}(x)$ to emphasize its dependence on the bandwidth b. By (5), we can write the predicted values

$$(\hat{g}_{n,b}(X_0), \ldots, \hat{g}_{n,b}(X_{n-1}))^T = H(b)Y, \tag{11}$$

where $H(b)$ is the $n \times n$ smoothing matrix. Recall $Y = (X_1, \ldots, X_n)^T$. The generalized cross-validation criterion chooses the bandwidth which minimizes

$$\mathrm{GCV}(b) = \frac{n^{-1} \sum_{i=1}^n (X_i - \hat{g}_{n,b}(X_{i-1}))^2}{\{1 - \mathrm{trace}[H(b)]/n\}^2}. \tag{12}$$

Roughly speaking, the optimal bandwidth under the GCV criterion balances goodness of fit measured in the numerator of (12) and model complexity measured in the denominator. This criterion has various favorable properties such as avoiding estimating nuisance parameters of the model and ease of implementation. See also Golub et al (1979), Li (1985) and Wahba (1990) for more discussions on GCV. Asymptotic performance of GCV under long-memory requires further study. In the next section we propose a method of estimating the variance of $\hat{g}_n(\cdot)$ with GCV bandwidth based on sub-sampling.

4 Simulation studies

If the process (X_t) is a stationary Gaussian time series, then the predictive function is linear: $g(x) = \alpha_0 + \alpha_1 x$, with $\alpha_1 = \rho$ and $\alpha_0 = \mathbb{E}X_1(1-\rho)$, where $\rho = r(1)/r(0)$ is the correlation coefficient between X_0 and X_1. Given the data $X_0, \ldots, X_n$, the regression coefficients α_0 and α_1 can be estimated, for example, by the classical linear regression methods. As mentioned in previous sections, for non-Gaussian processes $g(x)$ is generally nonlinear in x. For SRD processes, we shall conduct in Section 4.1 a simulation study to assess the performance of the nonparametric estimates. In Section 4.2 we shall illustrate the dichotomous phenomenon described in Theorem 2.

4.1 Performance of nonparametric predictors

For the SRD case, we consider the MA(1) model:

$$X_i = \varepsilon_i + 2\varepsilon_{i-1}, \tag{13}$$

where ϵ_i are i.i.d. innovations with density $f_\varepsilon(u) = \sqrt{2}\pi^{-1}(1+u^4)^{-1}$. In this case we are able to find an explicit form of $g(x) = \mathbb{E}(X_{i+1}|X_i = x)$. Since ε_i has mean 0, $g(x) = 2\mathbb{E}(\varepsilon_i|X_i = x)$. Note that X_t has the marginal density $\int_{\mathbb{R}} f_\varepsilon(t)f_\varepsilon(x/2-t/2)dt$. Elementary calculations show that the conditional mean

$$g(x) = 2\frac{\int_{\mathbb{R}} t f_\varepsilon(x/2-t/2)dt}{\int_{\mathbb{R}} f_\varepsilon(t)f_\varepsilon(x/2-t/2)dt} = \frac{2x(153+64x^2+x^4)}{945+192x^2+9x^4}. \tag{14}$$

The best linear predictor of X_{i+1} given X_i assumes the form $\alpha_0+\alpha_1 X_i$ with the parameters α_0 and α_1 minimizing $\mathbb{E}[X_{i+1}-(\alpha_0+\alpha_1 X_i)]^2$. It is easily seen that $\alpha_0 = 0$ and $\alpha_1 = \rho = 2/5$. The function g and the linear function $y = 2x/5$ are plotted in Figure 1.

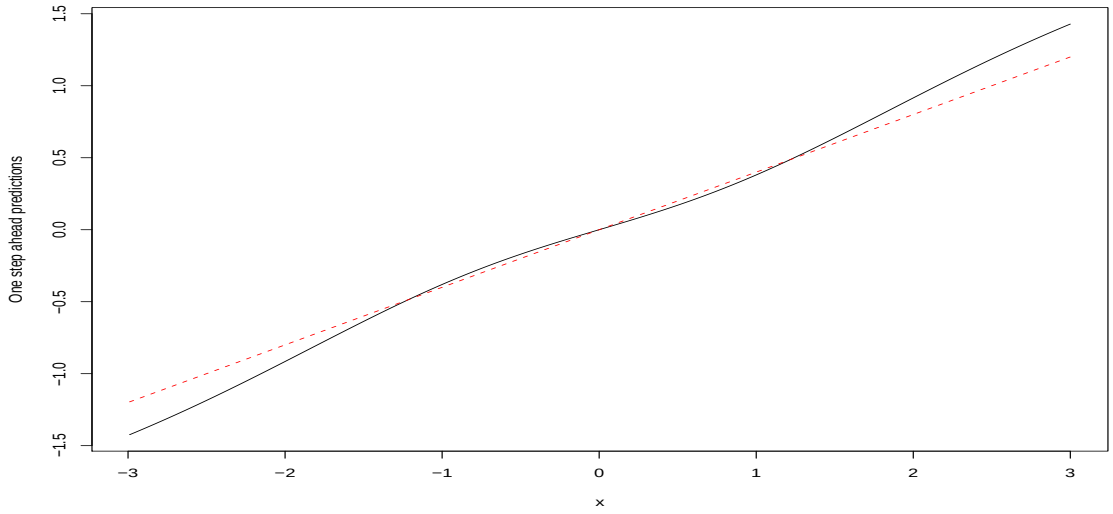

Figure 1: One step ahead predicted means for model (13). The solid line stands for the function $g(x) = \mathbb{E}(X_{i+1}|X_i = x)$ given in (14). The dashed line is the linear approximation $g_*(x) = \alpha_0 + \alpha_1 x = 2x/5$.

In our simulation study we choose two levels of n: $n = 500$ and $n = 1000$. The bandwidth b is chosen by the generalized cross-validation and K is the Epanechnikov

kernel $\tilde{K}(u) = 3\max(1-u^2, 0)/4$. As mentioned in Remark 1, the bias is of the form $g_{n,b_n}(x) - g(x) = C_B(x)b_n^2 + o(b_n^2)$ and $C_B(x)$ depends on unknown parameters which are not easily estimated. Following Wu and Zhao (2007), we apply the simple jackknife-type bias corrected estimate

$$\tilde{g}_{n,b_n}(x) = 2\hat{g}_{n,b_n}(x) - \hat{g}_{n,\sqrt{2}b_n}(x). \tag{15}$$

Then the bias of $\tilde{g}_{n,b_n}(x)$ is of the higher order $o(b_n^2)$ than the bias of $\hat{g}_{n,b_n}(x)$. The above estimate is equivalent to using the 4th order kernel $\tilde{K}^*(u) = 2\tilde{K}(u) - \tilde{K}(u/\sqrt{2})/\sqrt{2}$.

To apply Theorem 1, we need to estimate the asymptotic variance $(v_1^2(x)+v_2^2(x))/f^2(x)$ in (7). The variance estimation problem is generally not easy. A popular way is to use the subsampling technique (Politis, Romano and Wolf, 1999). For a chosen block size h, one divides the series $X_1, \ldots, X_n$ into $(n-h+1)$ consecutive blocks of size h, namely $\{(X_j, X_{j+1}, \ldots, X_{j+h-1})\}_{j=1}^{n-h+1}$. For each block, the bias corrected estimate $\tilde{g}(x)$ is calculated. Then the asymptotic variance in (7) is estimated as the sample variance of these bias corrected estimates. Regarding the choice of block size h, we suggest using the minimum volatility method proposed in Chapter 9 of Politis, Romano and Wolf (1999). The idea behind this approach is that, if a block size is in a reasonable range, then confidence intervals for the conditional mean constructed by the above sub-sampling technique should be stable when considered as a function of block size. See also Chapter 9 in the latter book for a more detailed description. Hence one could first propose a grid of possible block sizes and then choose one which minimizes the volatility of the end points of the confidence intervals near this size. More precisely, let the grid of possible block sizes be $\{h_1, \ldots, h_M\}$ and confidence intervals constructed by those block sizes be $\{(I_{1,l}, I_{1,u}), \ldots, (I_{M,l}, I_{M,u})\}$. For each block size h_i, calculate standard deviations of the sequences $(I_{i-3,l}, I_{i-2,l}, \ldots, I_{i+3,l})$ and $(I_{i-3,u}, I_{i-2,u}, \ldots, I_{i+3,u})$. Choose the block size which minimizes sum of those two standard deviations. In our simulations, this block size selector performs reasonably well and it is also found that the estimated variances are not sensitive to the choice of block size as long as this block size is not very different from the one chosen by the minimum volatility method.

In our experiments, we generate 10^4 sequences each having a length n. Then we obtain 10^4 bias corrected estimates $\tilde{g}(x)$. We consider $x = 2$. Then the true value $g(2) = 1700/1857 \approx 0.915455$ and the linear predictor gives the wrong value $\alpha_0 + 2\alpha_1 = 0.8$.

We estimate the mean and the variance of $\tilde{g}(2)$ by the sample mean and the sample variance of the simulated estimators. The estimated means and variances for $\tilde{g}(2)$ with $n = 500$ and 1000 are given in Table 1, and the histograms of $\tilde{g}(2)$ are displayed in Figure 2. The histograms suggest that the claim of approximate normality of $\tilde{g}$ is plausible. Figure 3 shows the histograms of the estimated variances of $\tilde{g}(2)$ of the 10^4 simulated sequences using the subsampling technique. Compared to the 'true' variance of $\tilde{g}(2)$ given in Table 1, we see from the histograms that the estimated variances are centered around the 'true' variance with small variability. Actually, the sample variances of the 10^4 estimated variances for $n = 1000$ and $n = 500$ are 8×10^{-4} and 5×10^{-3} respectively. Therefore the subsampling variance estimator with minimum volatility block size selector seems plausible in this experiment. For each realization, with the estimated variance of $\tilde{g}(2)$, we construct a 95% confidence interval for $g(2)$. Thus we have 10^4 confidence intervals in total. In Table 1, $\hat{p}$ gives the simulated coverage probability of these 10^4 confidence intervals that contain the true mean $g(2) \approx 0.915455$ and $\hat{l}$ gives the median half length of the confidence intervals. Observe that $\hat{p}$ is very close to the pre-assigned nominal level .95 and the length of the confidence intervals is moderate.

n	$n = 500$	$n = 1000$
$g(2)$	0.915455	0.915455
$\hat{\mathbb{E}}\tilde{g}_{n,b_n}(2)$	0.918588	0.917172
$\hat{\mathrm{Var}}[\tilde{g}_{n,b_n}(2)]$	0.055822	0.049586
$\hat{p}$	0.9429	0.9457
$\hat{l}$	0.52616	0.44578

Table 1. The estimated means and variances for $n = 500$ and 1000, respectively. $\hat{p}$: the simulated coverage probability of the .95 confidence intervals that contain the true mean $g(2) \approx 0.915455$. $\hat{l}$: median half length of the confidence intervals.

4.2 A simulation for the dichotomy

Theorem 2 describes the dichotomous phenomenon: for small bandwidths, the estimate performs as if the data were independent; while for large bandwidths, both the normalization and the limiting variance of the estimate are changed. In this section we shall design a simulation study to confirm this assertion for medium sample sizes.

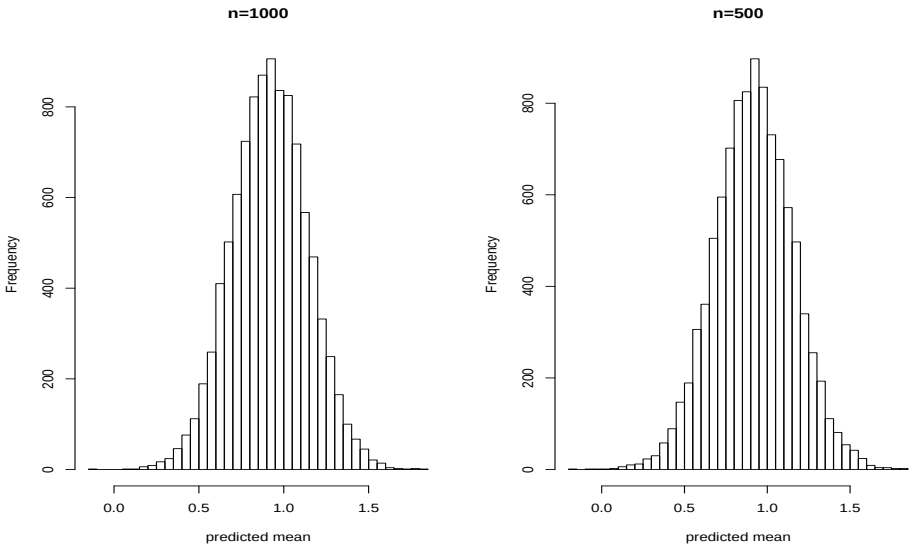

Figure 2: Histograms of 10^4 bias corrected estimates $\tilde{g}(2)$. Left panel: $n = 1000$. Right panel: $n = 500$.

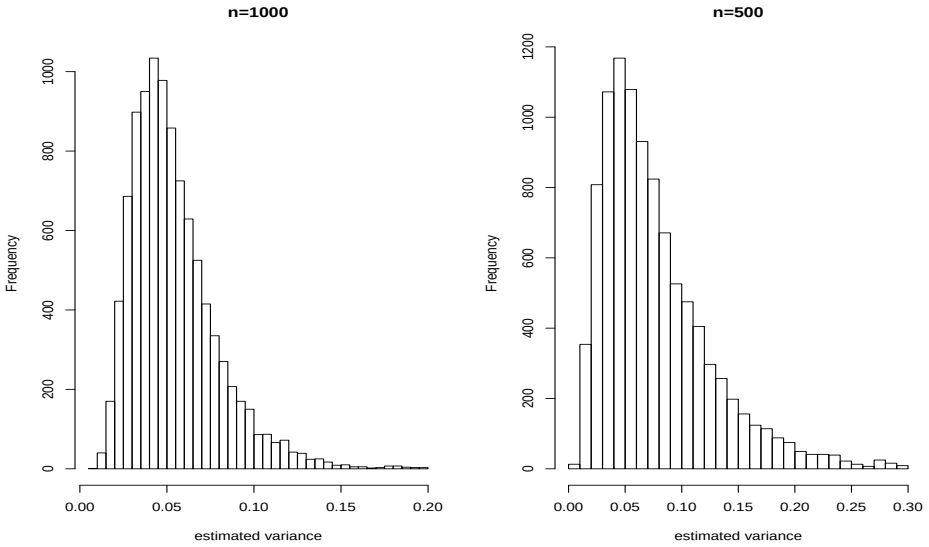

Figure 3: Histograms of 10^4 estimated variances of $\tilde{g}(2)$ using the subsampling technique. Left panel: $n = 1000$. Right panel: $n = 500$.

The considered innovations ε_k are i.i.d. with the common distribution being a mixture of normal distribution $\frac{1}{2}N(0,1)+\frac{1}{2}N(0,1.25)$. Let $a_i = (i+1)^{-0.8}$, $i \geq 0$. Then the process (X_t) is long-range dependent. With the convolution structure in (X_t), we can employ the powerful FFT (fast Fourier transforms) algorithm which uses circular embedding; see Wu et al (2004). Using a version of the algorithm described there, we can quickly generate $m = 5000$ sequences and each of them has length $n = 1000$. For each sequence, we apply the local linear estimate (5) to estimate $g(0)$, $g(0.5)$, $g(1)$ and $g(1.5)$ with the Epanechnikov kernel. Let $b = l/200$, $l = 1, 2, \ldots, 200$. For each b, the asymptotic variances of $\hat{g}_{n,b}(0.5i)$, $i = 0, 1, 2, 3$ are estimated by the sample variances of the m estimates. Figure 4 shows the estimated variances of the local linear estimates plotted against respective bandwidths. Theorem 2 asserts that, for relatively large b_n, the variance of $\hat{g}_{n,b_n}(x)$ is proportional to $(\sigma_n/n)^2$ which is independent of b_n. The latter fact is illustrated in Figure 4: the variances are relatively stable for larger bandwidths. Figure 4 also supports the claim that, for small b, the variances are proportional to $(nb)^{-1}$.

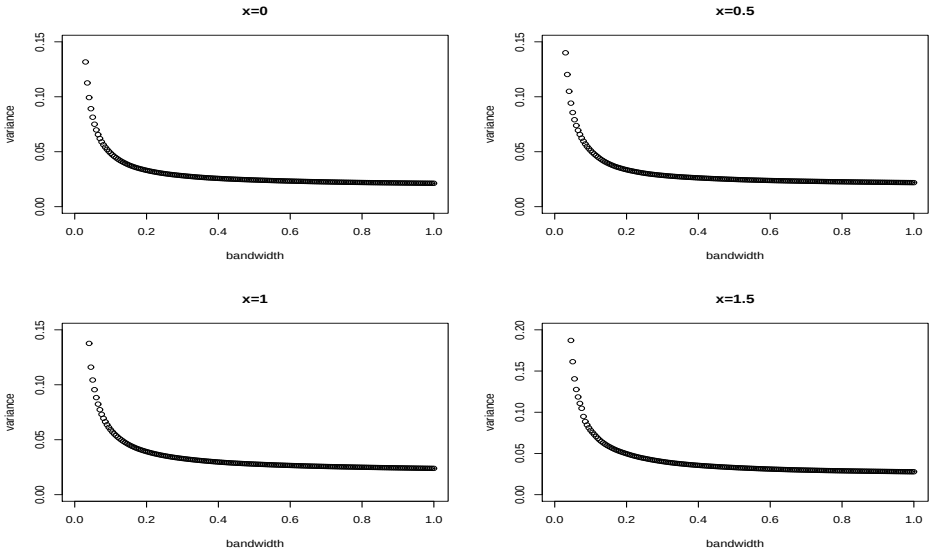

Figure 4: Estimated variances of the local linear estimate at $x = 0, 0.5, 1$ and 1.5 with respect to different bandwidths.

5 Proofs

In this section we shall prove Theorems 1 and 2. Let $\hat{h}_n(x)$ be the local linear estimator of the derivative $g'(x)$. Following equation (3.5) in Fan & Gijbels (1996), we have

$$(\hat{g}_n(x), b_n\hat{h}_n(x))^T = \mathcal{S}_n^{-1}(x)\frac{1}{n}\mathbf{X}^T\mathbf{W}Y. \tag{16}$$

Recall $\theta_n(x) = (g(x), b_n g'(x))^T$. Write

$$\hat{\theta}_n(x) = (\hat{g}_n(x), b_n\hat{h}_n(x))^T - \theta_n(x),\ \hat{\nu}_n(x) = \frac{1}{n}\mathbf{X}^T\mathbf{W}[Y - \mathbf{X}\theta_n(x)],$$

$D_n(x) = \hat{\theta}_n(x) - [\mathbb{E}\mathcal{S}_n(x)]^{-1}\mathbb{E}\hat{\nu}_n(x)$ and $\mathcal{S} := \mathrm{diag}(1, \mu_2)$, where $\mu_2 = \int_{-1}^{1} x^2K(x)dx$.

Assumptions of Proposition 1 below imply that $\mathcal{S}_n(x)$ is a weakly consistent estimate of $f(x)\mathcal{S}$ (Wu and Mielniczuk (2002)). Thus, when $f(x) \neq 0$, in order to investigate asymptotic laws of $D_n(x)$ given (2), it suffices to study laws of $\mathcal{S}_n(x)D_n(x)$, which in view of $\mathcal{S}_n(x) = n^{-1}\mathbf{X^T}\mathbf{W}\mathbf{X}$ is equal to

$$\begin{aligned}\mathcal{S}_n(x)D_n(x) &= \hat{\nu}_n(x) - \mathcal{S}_n(x)[\mathbb{E}\mathcal{S}_n(x)]^{-1}\mathbb{E}\hat{\nu}_n(x)\\ &= [\hat{\nu}_n(x) - \mathbb{E}\hat{\nu}_n(x)] - \{(\mathcal{S}_n(x) - \mathbb{E}\mathcal{S}_n(x))[\mathbb{E}\mathcal{S}_n(x)]^{-1}\mathbb{E}\hat{\nu}_n(x)\}. \end{aligned}\tag{17}$$

By the Cramér-Wold device, in order to investigate the asymptotic behavior of $\hat{\nu}_n(x)$, it is sufficient to study that of $L_n(x) := (c_1, c_2)\hat{\nu}_n(x)$ for all $(c_1, c_2) \in \mathbb{R}^2$ such that $c_1^2 + c_2^2 = 1$. Note that

$$L_n(x) = \frac{1}{n}\sum_{t=0}^{n-1}[c_1 + c_2(X_t - x)/b_n](X_{t+1} - g(x) - g'(x)(X_t - x))K_{b_n}(x - X_t).$$

Recall that $Y_t = X_{t+1} - \varepsilon_{t+1}$ and let $\mathcal{F}_t$ be a σ-field generated by $(\ldots, \varepsilon_{t-1}, \varepsilon_t)$. Moreover,

$$\begin{aligned}m_{t,n}(x) &= (Y_t - g(x) - g'(x)(X_t - x))(c_1 + c_2(X_t - x)/b)K_b(x - X_t)\\ &+ \varepsilon_t(c_1 + c_2(X_{t-1} - x)/b)K_b(x - X_{t-1}).\end{aligned}$$

Write

$$L_n(x) - \mathbb{E}L_n(x) = M_n(x) + N_n(x) + R_n(x), \tag{18}$$

where

$$M_n(x) = n^{-1}\sum_{t=0}^{n-1}[m_{t,n}(x) - \mathbb{E}(m_{t,n}(x)|\mathcal{F}_{t-1})],$$

$$N_n(x) \;=\; n^{-1}\sum_{t=0}^{n-1}[\mathbb{E}(m_{t,n}(x)|\mathcal{F}_{t-1}) - \mathbb{E}m_{t,n}(x)]$$

and the third term $R_n = (nb_n)^{-1}\{\varepsilon_n(c_1+c_2(X_n-x)/b_n)K[(x-X_{n-1})/b_n] - \varepsilon_0(c_1+c_2(X_{-1}-x)/b_n)K[(x-X_{-1})/b_n]\}$ satisfying

$$\begin{aligned}\|R_n\|^2 &\le 4(nb_n)^{-2}\mathbb{E}\{\varepsilon_0^2(c_1+c_2(X_{-1}-x)/b_n)^2K^2[(x-X_{-1})/b_n]\}\\ &= O(b_n n^{-2})\int K^2(u)(c_1+c_2u)^2 f(x-b_nu)du = O(b_n^{-1}n^{-2}) \qquad (19)\end{aligned}$$

since f is bounded.

In Section 5.1 we shall show that $(nb_n)^{1/2}M_n(x)$ is asymptotically normal without any conditions on the decay rate of (a_i). The second term $N_n(x)$ is also asymptotically normal, however, with different normalizing sequences when the process is long-range dependent and it is negligible in the opposite case of short-range dependence. The last term $R_n(x)$ is asymptotically negligible in view of (19).

5.1 Asymptotic normality of M_n

Note that $M_n(x)$ is a sum of nth row of a triangular array of martingale differences with respect to $\mathcal{F}_t$. We shall apply the martingale central limit theory to prove the following proposition. Recall Theorem 1 for the definition of $v_1^2(x)$ and $v_2^2(x)$.

Proposition 1. *Assume that f_ε is Lipschitz continuous, K is compactly supported and bounded, $\mathbb{E}|\varepsilon_i|^q < \infty$ for some $q > 2$ and $b_n \to 0$, $nb_n \to \infty$. Then*

$$(nb_n)^{1/2}M_n(x) \xrightarrow{\mathcal{D}} N[0, [v_1^2(x)+v_2^2(x)](\kappa c_1^2 + \lambda c_2^2)], \qquad (20)$$

where $\lambda = \int u^2K^2(u)du$.

Remark 7. If $a_i = 0$ for $i > 0$ then $X_t = \varepsilon_t$ is i.i.d., $g(x) = N_n(x) = 0$ and it follows that asymptotic variance of $\hat{g}_n(x)$ in this case is $\sigma^2\kappa/f_\varepsilon(x)$. $\diamondsuit$

Proof of Proposition 1. Let $V_t(x) = Y_t - g(x) - g'(x)(X_t - x)$. As $M_n(x)$ is a sum of martingale differences it suffices to check conditions of martingale CLT (cf Chow and Teicher (1988)). Consider first convergence of conditional variance. We will prove that

$$\frac{1}{nb_n}\sum_{t=0}^{n-1}\mathbb{E}\left(V_t^2(x)(c_1+c_2\frac{X_t-x}{b_n})^2K^2\left(\frac{x-X_t}{b_n}\right)|\mathcal{F}_{t-1}\right) \xrightarrow{\mathbb{P}} v_2^2(x)(\kappa c_1^2+\lambda c_2^2), \qquad (21)$$

$$\frac{1}{nb_n}\sum_{t=0}^{n-1}\left\{\mathbb{E}\left(V_t(x)(c_1+c_2\frac{X_t-x}{b_n})K\left(\frac{x-X_t}{b_n}\right)|\mathcal{F}_{t-1}\right)\right\}^2=o_{\mathbb{P}}(1), \tag{22}$$

$$\frac{1}{nb_n}\sum_{t=0}^{n-1}\mathbb{E}\left(\varepsilon_t^2(c_1+c_2\frac{X_t-x}{b_n})^2K^2\left(\frac{x-X_{t-1}}{b_n}\right)|\mathcal{F}_{t-1}\right)\xrightarrow{\mathbb{P}}v_1^2(x)(\kappa c_1^2+\lambda c_2^2), \tag{23}$$

$$\begin{aligned}&\sum_{t=0}^{n-1}\mathbb{E}\left(\varepsilon_tV_t(x)(c_1+c_2\frac{X_t-x}{b_n})(c_1+c_2\frac{X_{t-1}-x}{b_n})K\left(\frac{x-X_t}{b_n}\right)K\left(\frac{x-X_{t-1}}{b_n}\right)|\mathcal{F}_{t-1}\right)\\&=o_{\mathbb{P}}(nb_n).\end{aligned} \tag{24}$$

Let $A_t(u)=[Z_{t-1}+a_1u-g(x)-g'(x)(Y_{t-1}+u-x)]^2f_\varepsilon(u)$ and

$$B_t(y,r)=|Z_{t-1}+a_1(y-Y_{t-1})-g(x)-(y-x)g'(x)|^rf_\varepsilon(y-Y_{t-1}).$$

Write $B_t(y):=B_t(y,2)$. Consider first (21) and observe that

$$\begin{aligned}&b_n^{-1}\mathbb{E}\left(V_t^2(x)(c_1+c_2(X_t-x)/b_n)^2K^2\left(\frac{x-X_t}{b_n}\right)|\mathcal{F}_{t-1}\right)\\&=b_n^{-1}\int A_t(u)(c_1+c_2(Y_{t-1}+u-x)/b_n)^2K^2\left(\frac{x-Y_{t-1}-u}{b_n}\right)du\\&=\int B_t(x-b_nv)(c_1+c_2v)^2K^2(v)\,dv.\end{aligned} \tag{25}$$

Thus the left hand side of (21) equals $\int(c_1+c_2v)^2K^2(v)\bar{H}_n(x-b_nv)\,dv$ with

$$\bar{H}_n(y)=n^{-1}\sum_{t=0}^{n-1}B_t(y).$$

As $B_t(x)$ is ergodic, in order to prove (21) in view of ergodic theorem and symmetry of K, it suffices to show that

$$\sup_{|\delta|\le\delta_n}|\bar{H}_n(x+\delta)-\bar{H}_n(x)|\to 0 \tag{26}$$

in probability when $\delta_n\to 0$. Observe that $|B_t(x+\delta)-B_t(x)|$ is bounded by

$$\begin{aligned}&[Z_{t-1}+a_1(x+\delta-Y_{t-1})-g(x)-\delta g'(x)]^2|f_\varepsilon(x-Y_{t-1}+\delta)-f_\varepsilon(x-Y_{t-1})|+\\&|[Z_{t-1}+a_1(x+\delta-Y_{t-1})-g(x)-\delta g'(x)]^2-[Z_{t-1}+a_1(x-Y_{t-1})-g(x)]^2|f_\varepsilon(x-Y_{t-1})\end{aligned}$$

$$
\begin{aligned}
&\leq 2L\delta([Z_{t-1} - g(x) - a_1 Y_{t-1}]^2 + [a_1(x+\delta) - \delta g'(x)]^2) \\
&+ C\delta[|Z_{t-1}| + |a_1(x - Y_{t-1})| + |a_1\delta| + |g(x)| + |\delta g'(x)|],
\end{aligned}
$$

where L is the Lipschitz constant of f_ε, as Lipschitz continuity of a density f_ε implies that it is bounded. ¿From this using ergodicity again (26) easily follows. In order to prove (22) observe that its LHS can be written as

$$
\begin{aligned}
&\frac{b_n}{n}\sum_{t=0}^{n-1}\Big\{\int K(v)B_t^{1/2}(x - b_n v)(c_1 + c_2 v) f_\varepsilon^{1/2}(x - Y_{t-1} - b_n v)\, dv\Big\}^2 \\
\leq\ &\frac{b_n}{n}\sum_{t=0}^{n-1}\int K(v)B_t(x - b_n v)(c_1 + c_2 v)^2\, dv \int K(v) f_\varepsilon(x - Y_{t-1} - b_n v)\, dv \leq \\
\leq\ &Cb_n \int \bar{H}_n(x - b_n v)(c_1 + c_2 v)^2 K(v)\, dv \int K(v)\, dv
\end{aligned}
$$

using boundedness of f_ε again. It follows from the proof of (21) that the bound tends to 0. Proofs of (23) and (24) are similar and thus we omit the details.

It remains to verify the Lindeberg condition. Let $\Xi_t = V_t(x)(c_1 + c_2(X_t - x)/b_n)K_{b_n}(x - X_t)$. By assumptions on K, similarly as (25), we have for some function $C_0(x)$ that

$$
\begin{aligned}
\mathbb{E}(|\Xi_t|^q|\mathcal{F}_{t-1}) &= O(b_n^{1-q})\int_{\mathbb{R}} B_t(x - b_n v, q)\{(|c_1| + |c_2 v|)|K(v)|\}^q dv \\
&= O(b_n^{1-q})[|Z_{t-1}|^q + |Y_{t-1}|^q + C_0(x)].
\end{aligned}
$$

Since $\mathbb{E}|\varepsilon_i|^q < \infty$, $q > 2$, we have $\mathbb{E}|\Xi_t|^q = O(b_n^{1-q})$. We can similarly obtain that $\mathbb{E}\{|\varepsilon_t(c_1 + c_2(X_{t-1} - x)/b_n)K_{b_n}(x - X_{t-1})|^q\} = O(b_n^{1-q})$. Hence $n\mathbb{E}|m_{t,n}(x)\sqrt{b_n/n}|^q = n(b_n/n)^{q/2}O(b_n^{1-q}) = o(1)$ since $q/2 > 1$, $nb_n \to \infty$. The Lindeberg condition follows. $\diamondsuit$

5.2 Proof of Theorem 1

By extension of Theorem 1 in Wu and Mielniczuk (2002), under (i) and (ii), we have $\sqrt{nb_n}[\mathcal{S}_n(x) - \mathbb{E}\mathcal{S}_n(x)] = O_{\mathbb{P}}(1)$. By Remark 1, $\mathbb{E}\hat{\nu}_n(x) \to 0$. Furthermore, it is easy to see that $\mathbb{E}\mathcal{S}_n(x) \to f(x)\mathcal{S}$. Note that the last matrix is positive definite. So the second term in (17) is of order $o_{\mathbb{P}}[(nb_n)^{-1/2}]$. Therefore, by Proposition 1 and (19), it remains to show that the second term $N_n(x)$ in (18) satisfies $\|N_n(x)\| = O(n^{-1/2})$ since $b_n \to 0$. To this end, let

$$
Q_n(y) = \sum_{t=0}^{n-1} B_t(y) \tag{27}
$$

and $H_n(y) = Q_n(y) - \mathbb{E}Q_n(y)$. Then $nN_n(x) = \int H_n(x - b_n v)(c_1 + c_2 v)K(v)dv$.

Let $(\varepsilon_i')_{i\in\mathbb{Z}}$ be an i.i.d. copy of $(\varepsilon_i)_{i\in\mathbb{Z}}$ and define

$$\begin{aligned} Y_t' &= Y_t - a_{t+1}\varepsilon_0 + a_{t+1}\varepsilon_0', \\ Z_t' &= Y_t - a_{t+2}\varepsilon_0 + a_{t+2}\varepsilon_0', \\ U_t &= [a_1(y - Y_t) + Z_t - g(x) - g'(x)(y-x)]f_\varepsilon(y - Y_t), \\ U_t' &= [a_1(y - Y_t') + Z_t' - g(x) - g'(x)(y-x)]f_\varepsilon(y - Y_t'). \end{aligned}$$

Since f_ε is bounded and Lipschitz continuous, elementary calculations show that there exists a constant C, independent of t, x and y such that

$$\|U_t - U_t'\| \le C(|a_{t+1}| + |a_{t+2}|)[|g(x)| + |y| + 1 + |g'(x)||y - x|].$$

Note that $\mathbb{E}(U_t'|\mathcal{F}_0) = \mathbb{E}(U_t|\mathcal{F}_{-1})$. By Jensen's inequality $\|\mathcal{P}_0 U_t\| \le \|U_t - U_t'\|$. By Theorem 1 in Wu (2007) and (i),

$$\|H_n(y)\| \le \sqrt{n}\sum_{t=0}^{\infty}\|\mathcal{P}_0 U_t\| \le C\sqrt{n}[|g(x)| + |y| + 1 + |g'(x)||y - x|].$$

Using Schwarz's inequality we have

$$n\|N_n(x)\| \le \int \|H_n(x - b_n v)\||c_1 + c_2 v||K(v)|dv = O(\sqrt{n}),$$

Hence we have

$$(nb_n)^{1/2}[\hat{\theta}_n(x) - (\mathbb{E}\mathcal{S}_n(x))^{-1}\mathbb{E}\hat{\nu}_n(x)] \xrightarrow{\mathcal{D}} N(0, (v_1^2(x) + v_2^2(x))\mathcal{S}^{-2}\mathcal{S}^*/f^2(x)), \tag{28}$$

where $\mathcal{S}^* = \mathrm{diag}(\kappa, \lambda)$. Clearly (28) implies Theorem 1. ♢

5.3 Proof of Theorem 2

The proof of Theorem 2 follows from Theorem 1 and the following Theorem 3. Let $\mathbf{W}_t = \sum_{i=0}^{\infty}\mathbf{d}_i\varepsilon_{t-i} = (Y_t, Z_t)^T$, where $\mathbf{d}_i = (a_{i+1}, a_{i+2})^T$, $\mathbf{S}_n = \sum_{t=0}^{n-1}\mathbf{W}_t$.

Theorem 3. *Assume that (i) and (ii) of Theorem 2 are satisfied. Then we have*

$$\frac{n}{\sigma_n}N_n(x) - c_1\frac{1}{\sigma_n}J_\infty'(0)\mathbf{S}_n \xrightarrow{\mathbb{P}} 0 \tag{29}$$

and

$$\frac{n}{\sigma_n}N_n(x) \xrightarrow{\mathcal{D}} N(0, c_1^2 J_\infty'(0)\mathbf{1}J_\infty'(0)^{\mathbf{T}}). \tag{30}$$

Proof of Theorem 3. Let $H_n(y)$ be defined in the proof of Theorem 1; see (27). Reasoning as previously and using symmetry of kernel K we have

$$nN_n(x) = \int H_n(x - b_n v)(c_1 + c_2 v)K(v)\, dv = c_1 H_n(x) + o_{\mathbb{P}}(\sigma_n).$$

The first statement follows from Proposition 3 below as it implies that $H_n(x) - J'_\infty(0)\mathbf{S}_n = o_{\mathbb{P}}(\sigma_n)$. The second statement follows from the first and the fact that $\sigma_n^{-1}\mathbf{S}_n \xrightarrow{\mathcal{D}} N_2(0, \mathbf{1})$ (cf Lemma 8 in Mielniczuk and Wu (2004)). Let $\underline{\mathbf{W}}_{n,k} = \mathbb{E}(\mathbf{W}_n|\mathcal{F}_k)$ and $\bar{\mathbf{W}}_{n,k} = \mathbf{W}_n - \underline{\mathbf{W}}_{n,k}$. Moreover, let $J_n(\mathbf{u}) = \mathbb{E}J(\mathbf{u} + \bar{\mathbf{W}}_{n,0})$.

Proposition 2. *Let (ε_t) be an i.i.d. sequence with mean 0, $\mathbb{E}|\varepsilon_t|^q < \infty$ for some $q > 2$, and assume that the density function $f_\varepsilon(\cdot)$ satisfies*

$$\sup_u \{|f''_\varepsilon(u)| + |f'_\varepsilon(u)| + f_\varepsilon(u)\} < \infty. \tag{31}$$

Then

$$\|J_n(\underline{\mathbf{W}}_{n,0}) - J_{n+1}(\underline{\mathbf{W}}_{n,-1}) - J'_\infty(0)\mathbf{d}_n\varepsilon_0\| = O(n^{-\beta'}) \tag{32}$$

for any $\beta' \in (\beta, \beta_0)$, where $\beta_0 = \min\{2\beta, q\beta/2, \beta + (2\beta - 1)/(2p)\}$ and $p = q/(q-2)$.

To prove Proposition 2, we need the following lemma.

Lemma 1. *For any $n \in \mathbb{N}$, J_n is twice differentiable. Furthermore, there exists a constant $C < \infty$, such that for any $\mathbf{u} \in \mathbb{R}^2$ and $n \in \mathbb{N}$,*

$$|J_n(\mathbf{u})| + |L_{J'_n}(\mathbf{u})| + |J'_n(\mathbf{u})| \le C(1 + |\mathbf{u}|), \tag{33}$$

where $L_g(\mathbf{u})$ is the local Lipschitz constant of the function g:

$$L_g(\mathbf{u}) = \sup_{\mathbf{y} \neq \mathbf{u}:\, |\mathbf{y}-\mathbf{u}| \le 1} \frac{|g(\mathbf{y}) - g(\mathbf{u})|}{|\mathbf{y} - \mathbf{u}|}.$$

Proof of Lemma 1. Let C be a finite generic constant the value of which may vary from line to line. Recall that $J_n(\mathbf{u}) = \mathbb{E}J(\mathbf{u} + \bar{\mathbf{W}}_{n,0})$. Using the form of L Lemma 1 follows easily by simple calculations. ♢

Proof of Proposition 2. Let $U = J_{n+1}(\underline{\mathbf{W}}_{n,0}) - J_{n+1}(\underline{\mathbf{W}}_{n,-1}) - J'_{n+1}(\underline{\mathbf{W}}_{n,-1})\mathbf{d}_n\varepsilon_0$. Let $\delta = \mathbf{d}_n\varepsilon_0$. Then

$$\begin{aligned}
\mathbb{E}(|U|^2)/2 &\leq \mathbb{E}(|UI_{|\delta|\leq 1}|^2) + \mathbb{E}(|UI_{|\delta|>1}|^2) \\
&\leq \mathbb{E}(|L_{J'_{n+1}}(\underline{\mathbf{W}}_{n,-1})|\delta|^2 I_{|\delta|\leq 1}|^2) + 3\mathbb{E}(|J_{n+1}(\underline{\mathbf{W}}_{n,0})|I_{|\delta|>1})^2 \\
&+ 3\mathbb{E}(|J_{n+1}(\underline{\mathbf{W}}_{n,-1})|I_{|\delta|>1})^2 + 3\mathbb{E}(|J'_{n+1}(\underline{\mathbf{W}}_{n,-1})||\delta|I_{|\delta|>1})^2] \\
&:= I_n + II_n + III_n + IV_n
\end{aligned}$$

By Lemma 1, and the fact that $\underline{\mathbf{W}}_{n,-1}$ and δ are independent, we have $I_n = O(|\mathbf{d}_n|^{q_0})$, where $q_0 = \min(q, 4)$. Similarly, $III_n + IV_n = O(|\mathbf{d}_n|^q)$. For II_n, again, by Lemma 1,

$$\begin{aligned}
II_n/3 &= \mathbb{E}(|J_{n+1}(\underline{\mathbf{W}}_{n,-1} + \delta)|^2 I_{|\delta>1|}) \leq C\mathbb{E}(1 + |\underline{\mathbf{W}}_{n,-1}|)^2\mathbb{E}[(1 + |\delta|)^2 I_{|\delta|>1}] \\
&\leq C\mathbb{E}[(1 + |\delta|)^2 I_{|\delta|>1}] \leq 2C\mathbb{E}[|\delta|^2 I_{|\delta|>1}] \leq 2C\mathbb{E}|\delta|^q = O(|\mathbf{d}_n|^q)
\end{aligned}$$

Thus, $\|U\| = O(|\mathbf{d}_n|^{q_0/2})$. Denote $J_n(\underline{\mathbf{W}}_{n,0} + \mathbf{d}_n\varepsilon'_0) - J_n(\underline{\mathbf{W}}_{n,0}) - J'_n(\underline{\mathbf{W}}_{n,0})\mathbf{d}_n\varepsilon'_0$ by V, where $\{\varepsilon'_i, i \in \mathbb{Z}\}$ is an i.i.d. copy of $\{\varepsilon_i, i \in \mathbb{Z}\}$. Similarly as U, $\|V\| = O(|\mathbf{d}_n|^{q_0/2})$. Hence we can get that $\|J_{n+1}(\underline{\mathbf{W}}_{n,0}) - J_n(\underline{\mathbf{W}}_{n,0})\| = O(|\mathbf{d}_n|^{q_0/2})$ in view of $J_{n+1}(\mathbf{u}) - J_n(\mathbf{u}) = \mathbb{E}[J_n(\mathbf{u} + \mathbf{d}_n\varepsilon'_0) - J_n(\mathbf{u}) - \mathbf{d}_n\varepsilon'_0 J'_n(\mathbf{u})]$, and Jensen's inequality. To finish the proof of the theorem, it suffices to show that

$$\|J'_{n+1}(\underline{\mathbf{W}}_{n,-1}) - J'_\infty(0)\| = O(r_n^{1/(2p)}), \text{ where } r_n = \sum_{i+1}^{\infty} |\mathbf{d}_i|^2. \tag{34}$$

It is easy to see that $J'_\infty(0) = \mathbb{E}(J'_{n+1}(\underline{\mathbf{W}}_{n,-1}))$ for any n. Let $\underline{\mathbf{W}}^*_{n,-1} = \sum_{i+1}^{\infty} \mathbf{d}_i\varepsilon'_{n-i}$, $\Delta = \underline{\mathbf{W}}_{n,-1} - \underline{\mathbf{W}}^*_{n,-1}$. We have

$$\begin{aligned}
\|J'_{n+1}(\underline{\mathbf{W}}_{n,-1}) - J'_\infty(0)\| &= \|\mathbb{E}[J'_{n+1}(\underline{\mathbf{W}}_{n,-1}) - J'_{n+1}(\underline{\mathbf{W}}^*_{n,-1})]|\varepsilon_{-1}, \varepsilon_{-2}, \dots\| \\
&\leq \|J'_{n+1}(\underline{\mathbf{W}}_{n,-1}) - J'_{n+1}(\underline{\mathbf{W}}^*_{n,-1})\| \\
&\leq \|L_{J_{n+1}}(\underline{\mathbf{W}}^*_{n,-1})|\Delta|I_{|\Delta|\leq 1}\| \\
&\quad + \|J'_{n+1}(\underline{\mathbf{W}}^*_{n,-1})I_{|\Delta|>1}\| + \|J'_{n+1}(\underline{\mathbf{W}}_{n,-1})I_{|\Delta|>1}\| \\
&:= I'_n + II'_n + III'_n.
\end{aligned}$$

By Lemma 1 and Hölder's inequality,

$$\begin{aligned}
(I'_n)^2 &\leq (\mathbb{E}[L^q_{J'_{n+1}}(\underline{\mathbf{W}}^*_{n,-1})])^{2/q}(\mathbb{E}[|\Delta|^{2p} I_{|\Delta|\leq 1}])^{1/p} \\
&\leq C(\mathbb{E}[|\Delta|^{2p} I_{|\Delta|\leq 1}])^{1/p} \leq C(\mathbb{E}[|\Delta|^2 I_{|\Delta|\leq 1}])^{1/p} = O(r_n^{1/p})
\end{aligned}$$

Similarly, it can be shown that $II'_n + III'_n = O(r_n^{1/(2p)})$. Thus (32) holds. ♢

Proposition 3. *Let* $\mathbf{S}_n = \sum_{i=0}^{n-1} \mathbf{W}_i$. *We have* $\|H_n(x) - J'_\infty(0)\mathbf{S}_n\| = O(n^{3/2-\beta'})$ *for* β' *defined in Proposition 1.*

Proof of Proposition 3 . The proof is analogous to the proof of Theorem 2 in Wu (2003). Let $T_i = J(\mathbf{W}_i) - J'_\infty(0)\mathbf{W}_i$, $\theta_i = \|\mathcal{P}_0 T_i\|$, and $\Theta_i = \sum_{j=0}^{i} \theta_j$. Then

$$\|H_n(x) - J'_\infty(0)\mathbf{S}_n\|^2 = \sum_{j=-\infty}^{n-1} \|\mathcal{P}_j[H_n(x) - J'_\infty(0)\mathbf{S}_n]\|^2 \le \sum_{j=-\infty}^{n-1} [\sum_{i=0}^{n-1} \theta_{i-j}]^2$$

$$= \sum_{j=0}^{n-1}[\sum_{i=0}^{n-1} \theta_{i-j}]^2 + \sum_{j=-n}^{-1} [\sum_{i=0}^{n-1} \theta_{i-j}]^2 + \sum_{j=-\infty}^{-n-1} [\sum_{i=0}^{n-1} \theta_{i-j}]^2 := I_n^* + II_n^* + III_n^*$$

By Proposition 1, we have that $\theta_n = O(n^{-\beta'})$, $\Theta_n = O(n^{1-\beta'})$. Thus, $I_n^* \le n[\sum_{i=0}^{n-1} \theta_i]^2 = O(n\Theta_{n-1}^2) = O(n^{3-2\beta'})$. Similarly, $II_n^* + III_n^* = O(n^{3-2\beta'})$. Thus Proposition 3 holds. $\diamondsuit$

Acknowledgements. The authors are grateful to the referee for his/her valuable comments. The third-named author was supported in part by the NSF grant DMS-0478704 and the Taft fellowship at the University of Cincinnati.

References

Beran, J. (1994) *Statistics for Long Memory Processes*, Chapman and Hall, New York.

Bhansali R. and P. Kokoszka (2004) Prediction of long memory time series: a tutorial review. In: *Processes with Long-Range Correlations* (eds. G. Rangarajan and M. Ding), 3-21, Springer, New York.

Bosq, D. (1996). *Nonparametric Statistics for Stochastic Processes*, Springer, New York.

Csörgő, S. and J. Mielniczuk (1995a). Distant long-range dependent sums with application to regression estimation. *Stochastic Processes and their Applications* **59**, 143–155.

Csörgő, S. and J. Mielniczuk (1995b). Close short-range dependent sums and regression estimation. *Acta Scientiarum Mathematicarum (Szeged)* **60**, 177–196.

Csörgő, S. and J. Mielniczuk (1999) Random design regression under long-range dependence. *Bernoulli* **5**, 209-224.

Chen, R. (1996) A nonparametric multi-step prediction estimator in Markovian structures. *Statistica Sinica* **6**, 603-615.

Chow, Y.S. and H. Teicher (1988) *Probability Theory*, 2nd ed. Springer, New York.

Collomb, G. and W. Härdle (1986) Strong uniform convergence rates in robust nonparametric time series analysis and prediction. *Stochastic Processes and their Applications* **23**, 77-89.

Craven, P. and G. Wahba (1979) Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, 377-403.

Eubank, R. (1988) *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, Boston.

Fan, J. and I. Gijbels (1996) *Local Polynomial Modelling and Its Applications*, Chapman and Hall.

Fan, J. and Q. Yao (2003) *Nonlinear Time Series*, Springer, New York.

Feller, W. (1971) *An Introduction to Probability Theory and its Applications II.* Wiley, New York.

Golub, G., M. Heath and G. Wahba (1979) Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics* **21** 215-224.

Hall, P. and Hart, J. (1990) Convergence rates in density estimation for data from infinite-order moving average processes. Probability Theory and Related Fields **87**, 253-274.

Härdle, W. (1990) *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.

Li, K.C. (1985). From Stein's unbiased risk estimates to the method of generalized cross-validation. *The Annals of Statistics* **13** 1352-1377.

Masry, E. and Mielniczuk, J. (1999) Local linear regression estimation for time series with long-range dependence. *Stochastic Processes and Their Applications* **82**, 173-194.

McKeague, I. and M.I. Zhang (1994) Identification of nonlinear time series from first order cumulative characteristics. *The Annals of Statistics* **22** 495–514.

Mielniczuk, J. and W. B. Wu (2004) On random-design regression model with dependent errors. *Statistica Sinica* **14**, 1105-126.

Phillips, P. and J. Park (1998) Nonstationary density estimation and kernel autoregression, Cowles Foundation discussion paper 1181, Yale University.

Politis, D. N., Romano, J. P. and M. Wolf (1999) *Subsampling.* Springer, New York.

Pourahmadi, M. (2001) *Foundations of Time Series Analysis and Prediction Theory.* Wiley, New York.

Robinson, P. M. (1983) Nonparametric estimators for time series. *Journal of Time Series Analysis* **4** 185–207.

Robinson, P. M. (1991) Nonparametric function estimation for long memory time series. In: Barnett, W. A., Powell, J. and Tauchen, G. E. (eds.) Nonparametric and semiparametric methods in econometrics and statistics - proceedings of the fifth international symposium in. Cambridge University Press, Cambridge, pp. 437-458.

Wahba, G. (1977) A survey of some smoothing problems and the method of generalized cross-validation for solving them. In *Applications of Statistics* (P.R. Krishnaiah, ed.). North-Holland, Amsterdam, pp. 507-523.

Wahba, G. (1990) *Spline Models for Observational Data*. SIAM, Philadelphia. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59.

Wu, W. B. and J. Mielniczuk (2002) Kernel density estimation for linear processes *The Annals of Statistics* **30**, 1441-1459.

Wu, W. B. (2003) Additive functionals of infinite-variance moving averages. *Statistica Sinica* **13**, 1259-1267.

Wu, W. B. (2007) Strong invariance principles for dependent random variables. *Annals of Probability*, **35**, 2294-2320.

Wu, W. B. and Y. Huang (2006) Kernel estimation of time series: asymptotic theory, manuscript.

Wu, W. B., G. Michailidis and D. Zhang (2004) Simulating sample paths of linear fractional stable motion. *IEEE Transactions on Information Theory* **50** 1086-1096.

Wu W. B. and Z. Zhao (2007) Inference of Trends in Time Series. *Journal of the Royal Statistical Society, Series B.* **69** 429-446.

Yang, Y. (2001) Nonparametric regression with dependent errors. *Bernoulli* **7**, 633-655.