Local Linear Regression for Data with AR Errors

Runze Li and Yan Li

Department of Statistics and The Methodology Center

Pennsylvania State University

University Park, PA 16802-2111

April 9, 2008

Abstract

In many statistical applications, data are collected over time, and they are likely correlated. In this paper, we investigate how to incorporate the correlation information into the local linear regression. Under the assumption that the error process is an auto-regressive (AR) process, a new estimation procedure is proposed for the nonparametric regression by using local linear regression method and the profile least squares techniques. We further propose the SCAD penalized profile least squares method to determine the order of AR process. Extensive Monte Carlo simulation studies are conducted to examine the finite sample performance of the proposed procedure, and to compare the performance of the proposed procedures with the existing one. From our empirical studies, the newly proposed procedures can dramatically improve the accuracy of naive local linear regression with working-independent error structure. We illustrate the proposed methodology by an analysis of real data set.

Key Words: Auto-regressive error, local linear regression, partially linear model, profile least squares, SCAD.

1 Introduction

When data are correlated, it is of great interest to improve efficiency of parameter estimates by including the correlation information into estimation procedures. This issue has been well studied in the longitudinal or panel data. The generalized method of moments (GMM, Hansen, 1982), the generalized estimating equation (GEE, Liang and Zeger, 1986; Zeger and Liang, 1986) and quadratic inference function (QIF, Qu, Li and Lindsay, 2000) are well-known methods to incorporate the correlation information into estimation procedure for parametric regression models with longitudinal data. Lin and Carroll (2000) showed that kernel GEE, a direct estimation of the parametric GEE, fails to incorporate the correlation information into the kernel estimate for the nonparametric function of clustered/longitudinal data. Wang (2003) proposed the marginal kernel method for longitudinal data. The marginal kernel method achieves its efficiency by incorporating the true correlation structure. Fan, Huang and Li (2007) proposed the idea of minimizing generalized variance (MGV) to improve the efficiency of estimates of nonparametric regression function under the context of longitudinal data using working independence.

Beyond the setting of longitudinal data, many authors have studied the topic of nonparametric regression with correlated errors. A good review on this topic is given in Opsomer, Wang and Yang (2001), in which attentions have been paid to a nonparametric regression model with fixed designs:

$$y_t = m(x_t) + \varepsilon_t, \tag{1.1}$$

where $x_t = t/n$ or $x_t = (t - 0.5)/n$, and ε_t is a correlated error. Opsomer, Wang and Yang (2001) discussed the problems with correlation and illustrate the failure of inclusion the correlation between the errors may yield an undesirable results. Pioneer works in this topic are Altman (1990) and Hart (1991). Both Altman (1990) and Hart (1991) assumed that the correlation between ε_t and ε_s was of the form $\rho_n(|t - s|)$, and addressed the issue how to select a bandwidth adjusting the correlation structure $\rho_n(|t - s|)$, which is assumed

to be known or required to estimate based on the observed data.

In this paper, we consider the situation in which x_t is a random design. More specifically, it is assumed that (x_t, y_t), $t = 1, 2, \cdots$, is a sequence of strictly stationary random vectors. Thus, (x_t, y_t), $t = 1, 2, \cdots$, are identically distributed. In this paper, we are interested in that when the data are correlated, how to incorporate the correlation into the local linear regression estimation procedure. This issue has been studied in Xiao, Linton, Carroll and Mammen (2003), in which the authors proposed a new estimation procedure based on a pre-whitening transformation of the dependent variable that must be estimated from the data. They also established the asymptotic distribution of their estimator under weak conditions, while they didn't address the critical issues related to practical implementation, such as the selection of smoothing parameter of their nonparametric regression. They assumed that the error process was is an invertible linear process, i.e., a moving average (MA) process with order infinite, while they did not discuss how to determine the order of the error process. In this paper, it is assumed that the error process is an AR process of order d. We propose a new estimation procedure for the regression using the profile least squares techniques, and study the asymptotic property of the resulting estimate. We discuss how to select the bandwidth in the local linear regression. We further propose the penalized profile least squares with the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) to determine the order of the AR process. Monte Carlo simulation studies are conducted to examine the finite sample performance. Our simulation results show that the proposed estimation procedure is much more efficient than the conventional local linear regression method when the error is highly correlated. The efficiency gain can be achieved in moderate-sized samples.

The remainder of this paper is organized as follows. In Section 2, we propose a new estimation procedure, and discuss the issues related to practical implementation. Section 3 presents numerical comparison and analysis of a real data example. Regularity conditions and technical proofs are given in Section 4. Some discussions and final concluding remarks are given in Section 5.

2 A new estimation procedure

Suppose that $(x_t, y_t), t = 1, \cdots, n$ be a random sample from the nonparametric regression model

$$y_t = m(x_t) + \varepsilon_t, \tag{2.1}$$

where error process ε_t is a correlated random error with mean zero. Throughout this paper, it is assumed that the error process ε_t is independent of the covariate process x_t. Altman (1990) and Hart (1991) proposed kernel regression estimation for $m(\cdot)$ under the 'fixed design' case, i.e., either $x_t = t/n$ or $x_t = (t - 0.5)/n$ and $\text{cov}(\varepsilon_t, \varepsilon_{t+k}) = \sigma^2 \rho_n(|k|)$, and further studied how to choose the bandwidth with adjusting for the correlation. Xiao, *et al* (2003) proposed a local polynomial estimate for the regression function with x_t being a 'random design' and following a non-degenerate distribution, and the residual process is stationary, mean zero and has an invertible linear process representation. That is, it can be represented as a moving average with infinite order (MA(∞)). Throughout this paper, x_t is a random design and that ε_t is an autoregressive (AR) series

$$\varepsilon_t = \beta_1 \varepsilon_{t-1} + \cdots + \beta_d \varepsilon_{t-d} + \eta_t,$$

where η_t is independent and identically distributed random error with mean zero and variance σ^2. The order d can be large, and the selection of the order d will be discussed in next section. If the values for ε_t were available, then we would work on the following partially linear model

$$y_t = m(x_t) + \beta_1 \varepsilon_{t-1} + \cdots + \beta_d \varepsilon_{t-d} + \eta_t.$$

In practice, ε_t is not available, but it may be estimated by $\widehat{\varepsilon_t} = y_t - \widehat{m}_I(x_t)$, where $\widehat{m}_I(\cdot)$ is a local linear estimate of $m(\cdot)$ based on (2.1) without considering the AR error structure. We will address the issue of bandwidth selection for $\widehat{m}_I(\cdot)$ in next section.

Replacing ε_t's with $\widehat{\varepsilon_t}$'s, we have

$$y_t = m(x_t) + \mathbf{e}_t^T \boldsymbol{\beta} + \eta_t, \tag{2.2}$$

where $\mathbf{e}_t = (\widehat{\varepsilon}_{t-1}, \cdots, \widehat{\varepsilon}_{t-d})^T$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)^T$. In Section 2.1, we propose a new estimation procedure for $m(\cdot)$ and $\boldsymbol{\beta}$ based on model (2.2). We further propose an order selection procedure for the AR series by using penalized profile least squares method in section 2.2.

2.1 Profile least squares estimate

As to the partially linear model (2.2), there exist various estimation procedures, including partially spline estimate (Wahba, 1984, Heckman, 1986, Engle et al., 1986), partial residual method (Speckman, 1998) and profile least squares or likelihood method (Severini and Staniswalis, 1994). Here we will employ the profile least squares techniques to estimate $\boldsymbol{\beta}$ and $m(\cdot)$.

For given $\boldsymbol{\beta}$, denote $y_t^* = y_t - \mathbf{e}_t^T\boldsymbol{\beta}$ for $t = d+1, \cdots, n$. Then

$$y_t^* = m(x_t) + \eta_t \tag{2.3}$$

which is one-dimensional nonparametric model. We may employ existing linear smoothers, such as local polynomial regression and smoothing splines (Gu, 2002), to estimate $m(\cdot)$. Here we will employ the local linear regression. For a given x_0, we locally approximate the regression function

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) \widehat{=} a + b(x - x_0)$$

for x in the local neighborhood of x_0. Thus, the local linear estimate of $m(\cdot)$ is the minimizer of the following weighted least squares function

$$(\widehat{a}, \widehat{b})^T = \text{argmin}_{(a,b)} \sum_{t=d+1}^{n} \{y_t^* - a - b(x_t - x_0)\}^2 K_h(x_t - x_0),$$

where $K_h(u) = h^{-1}K(u/h)$ is a scaled kernel function of kernel $K(\cdot)$ with bandwidth h. It is clear that the local linear estimate is linear in terms of $\mathbf{y}^* = (y_{d+1}^*, \cdots, y_n^*)^T$. Let $\widehat{\mathbf{m}} = (\widehat{m}(x_{d+1}), \cdots, \widehat{m}(x_n))^T$. Then $\widehat{\mathbf{m}}$ can be represented by

$$\widehat{\mathbf{m}} = S_h \mathbf{y}^*, \tag{2.4}$$

where S_h is a $(n-d)\times(n-d)$ smoothing matrix depending on x_t's and the bandwidth only.

Substituting $m(x_t)$ in (2.3) by $\widehat{m}(x_t)$, we obtain a synthetic linear regression model

$$(I - S_h)\mathbf{y} = (I - S_h)\mathbf{E}\boldsymbol{\beta} + \boldsymbol{\eta},$$

where I is the identity matrix, $\mathbf{E} = (\mathbf{e}_{d+1}, \cdots, \mathbf{e}_n)^T$ and $\boldsymbol{\eta} = (\eta_{d+1}, \cdots, \eta_n)^T$. Thus, the profile least squares estimator for $\boldsymbol{\beta}$ and $\mathbf{m}$ are

$$\widehat{\boldsymbol{\beta}} = \{\mathbf{E}^T(I - S_h)^T(I - S_h)\mathbf{E}\}^{-1}\mathbf{E}^T(I - S_h)^T(I - S_h)\mathbf{y}, \tag{2.5}$$

and

$$\widehat{\mathbf{m}} = S_h(\mathbf{y} - \mathbf{E}\widehat{\boldsymbol{\beta}}), \tag{2.6}$$

respectively.

The asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ and the asymptotic bias and variance of $\widehat{m}(x_0)$ are given in the following theorem. Denote $\mu_i = \int x^i K(x)\,\mathrm{d}x$ and $\nu_i = \int x^i K^2(x)\,\mathrm{d}x$.

Theorem 1. *Suppose that Conditions A—G listed in Section 4 hold. Then*

(A) The asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ is given

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N(0, \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}^{-1})$$

where $\mathbf{f}_t = (\epsilon_{t-1}, \ldots, \epsilon_{t-d})^T$ and $\sigma^2 = var(\eta_t)$.

(B) The asymptotic distribution of $\widehat{m}(x_0, \widehat{\boldsymbol{\beta}})$, conditioning on $x_1, \cdots, x_n$, is given below

$$\sqrt{nh}\{\widehat{m}(x_0, \widehat{\boldsymbol{\beta}}) - m(x_0) - \frac{1}{2}\mu_2 m''(x_0)h^2\} \to N(0, \frac{\nu_0\sigma^2}{f(x_0)}),$$

where $f(x)$ is the density function of x.

Note that the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ is the same as that of Yule-Walker estimator for the AR model:

$$\varepsilon_t = \beta_1\varepsilon_{t-1} + \cdots + \beta_d\varepsilon_{t-d} + \eta_t.$$

(see Theorem 8.1.1 of Brockwell and Davis, 1991). In other words, Theorem 1 implies that $\widehat{\boldsymbol{\beta}}$ is as efficient as if one knew the true regression function $m(\cdot)$ in advance. The asymptotic bias and variance of $\widehat{m}(\cdot, \widehat{\boldsymbol{\beta}})$ are the same as those of the local linear regression for independent and identically distributed observations, respectively. This implies that the proposed profile least square estimate is very efficient.

2.2 SCAD procedure for the AR process

To implement the profile least squares estimation procedure, we have to determine the order of AR process. In practice, we may start with a large order AR process, and then apply variable selection procedure to select its order. The penalized likelihood procedures with the smoothly clipped absolute deviation (SCAD) penalty was proposed for variable selection in parametric models in Fan and Li (2001). The SCAD procedure is distinguished from the traditional variable selection procedures, such as the stepwise regression and the best subset selection with the AIC and BIC, in that it selects significant variables and estimates their coefficients simultaneously. Thus, it can be directly applied for high-dimensional data analysis. The SCAD procedure was further developed for partially linear model with longitudinal data in Fan and Li (2004). In this section, we apply the SCAD procedure to determine the complexity of AR process.

The SCAD penalized least squares function is defined to be

$$\frac{1}{2}\sum_{t=d+1}^{n}\{y_t - m(x_t) - \mathbf{e}_t^T\boldsymbol{\beta}\}^2 + n\sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|)$$

where $p_\lambda(|\beta|)$ is the SCAD penalty with a tuning parameter λ, defined by

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta|, & \text{if } 0 \le |\beta| < \lambda; \\ \frac{(a^2-1)\lambda^2 - (|\beta|-a\lambda)^2}{2(a-1)}, & \text{if } \lambda \le |\beta| < a\lambda; \\ \frac{(a+1)\lambda^2}{2}, & \text{if } |\beta| \ge a\lambda. \end{cases}$$

Fan & Li (2001) suggested fixing $a = 3.7$ from a Bayesian argument. Figure 1 depicts the SCAD penalty with $\lambda = 1$.

Applying the profile techniques for the penalized least squares, we can derive the penalized profile least squares estimate, the minimizer of the following penalized least squares

$$\frac{1}{2}\|(I - S_h)\mathbf{y} - (I - S_h)\mathbf{E}\boldsymbol{\beta}\|^2 + n\sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|). \tag{2.7}$$

As demonstrated in Fan and Li (2004), with proper choice of tuning parameter, the resulting estimate contains some exact zero coefficients. This is equivalent to excluding the

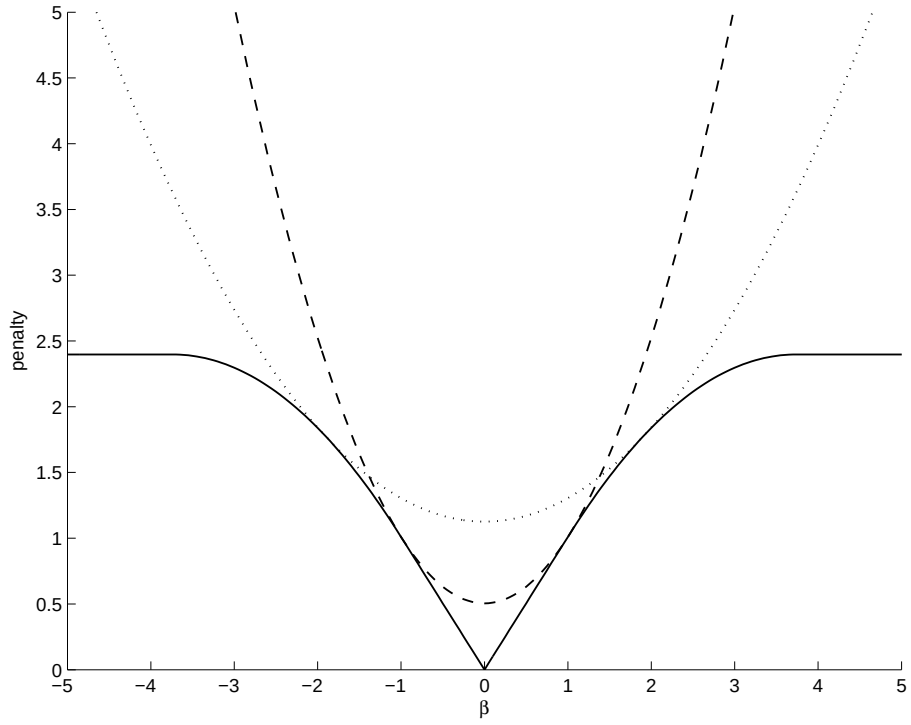

Figure 1: Scad Penalty Function and its local quadratic approximation

corresponding terms from the selected model and reducing model complexity. Since the SCAD penalty function is a nonconvex function over $[0, \infty)$, it is challenging in minimizing the SCAD penalized profile least squares function. Following Fan and Li (2004), we employ the local quadratic approximation (LQA) for the SCAD penalty function. Suppose we can get an estimate $\beta_j^{(k)}$ in the k^{th} step that is close to the true β_j. If $|\beta_j^{(k)}|$ is close to 0, then set $\widehat{\beta_j} = 0$. Otherwise, the SCAD penalty can be locally approximated by a quadratic function as

$$[p_{\lambda_j}(|\beta_j|)]' = p'_{\lambda_j}(|\beta_j|) \cdot \text{sgn}(\beta_j) \approx p'_{\lambda_j}(|\beta_j^{(k)}|)/|\beta_j^{(k)}|\beta_j$$

This is equivalent to

$$p_{\lambda_j}(|\beta_j|) \approx p_{\lambda_j}(|\beta_{j0}|) + \frac{1}{2}\{p'_{\lambda_j}(|\beta_j^{(k)}|)/|\beta_j^{(k)}|\}(\beta_j^2 - \beta_j^{(k)2})$$

With the aid of LQA, we may employ the following iterative ridge regression to find the

minimizer of (2.7):

$$\boldsymbol{\beta}^{(k+1)} = \{\mathbf{E}^T(I - S_h)^T(I - S_h)\mathbf{E} + n\Sigma_\lambda(\boldsymbol{\beta}^{(k)})\}^{-1}\mathbf{E}^T(I - S_h)^T(I - S_h)\mathbf{y} \tag{2.8}$$

where $\Sigma_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(k)}) = \text{diag}\{p'_{\lambda_1}(|\beta_1^{(k)}|)/|\beta_1^{(k)}|, \ldots, p'_{\lambda_d}(|\beta_d^{(k)}|)/|\beta_d^{(k)}|\}$ for nonvanished $\boldsymbol{\beta}^{(k)}$.

2.3 Tuning parameter selection and bandwidth selection

In this section, we address how to determine $\boldsymbol{\lambda}$ in the SCAD procedure and how to select a bandwidth for the profile least squares estimation procedure, two important issues in the practical implementation of the proposed methodology.

Tuning parameter selection. In the implementation of the SCAD procedure, we need to choose the tuning parameter $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_d)$. Following the advocation of Wang, Li and Tsai (2007), we use the BIC selector to find the optimal $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_d)$. From (2.8), we define the effective number of parameters of the penalized least square estimator (2.8) to be

$$e(\boldsymbol{\lambda}) = \text{tr}[\{\tilde{D} + \Sigma_{\boldsymbol{\lambda}}(\widehat{\beta})\}^{-1}\tilde{D}]$$

where $\tilde{D} = \mathbf{E}^T(I - S_h)^T(I - S_h)\mathbf{E}$ for nozero $\widehat{\beta}$.

The BIC score is defined to be

$$BIC(\boldsymbol{\lambda}) = \log\left\{\frac{RSS(\boldsymbol{\lambda})}{n}\right\} + e(\boldsymbol{\lambda})\frac{\log n}{n}$$

where $RSS(\boldsymbol{\lambda}) = \|(I - S_h)\mathbf{y} - (I - S_h)\mathbf{E}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|^2$ is the residual sum of squares with $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$, the penalized profile least squares estimate of $\boldsymbol{\beta}$ with tuning parameter $\boldsymbol{\lambda}$.

It is challenging to minimize $BIC(\boldsymbol{\lambda})$ over a d-dimensional space of $\boldsymbol{\lambda}$. By heuristic, the magnitude of λ_j is proportional to the standard error of the profile least squares estimate of β_j. In our implementation, we set $\boldsymbol{\lambda} = \lambda\,\text{se}(\widehat{\beta}_{LS})$, where $\text{se}(\widehat{\beta}_{LS})$ is the standard error of the unpenalized profile least square estimates and λ is a scalar variable. Thus, the original d-dimensional optimization becomes a 1-dimensional problem. In section 3, we minimize $BIC(\lambda)$ over a grid of points evenly distributed in the interval $[\frac{0.1}{\sqrt{n}}, \frac{2\sqrt{\log n}}{\sqrt{n}}]$, and set $\widehat{\lambda} = \text{argmin}_\lambda BIC(\lambda)$.

Bandwidth selection. Xiao et al. (2003) pointed out it was challenging in selecting a bandwidth for their procedure, and the authors simply used the rule of thumb bandwidth, $h = 1.06\, S_X\, n^{-\frac{1}{5}}$, to prewhite AR process, where S_X is the standard error of x_t. Note that $h = 1.06\, S_X\, n^{-\frac{1}{5}}$ is the rule of thumb bandwidth for kernel density estimate (Silverman, 1986), and we doubt that it may be a careless error. From our limited simulation experience, the bandwidth is not appropriate for the regression problem. Here we propose a bandwidth selector for the profile least squares estimate.

We use local linear regression to get the initial estimate $\widehat{m}_I(\cdot)$ with the plug-in bandwidth selector (Ruppert, Sheather and Wand, 1995), pretending the data are independent. Since model (2.2) is a partially linear model, we can use the existing bandwidth selector for partially linear model in the literature. Here we suggest using the proposal of Fan and Li (2004). Specifically, we calculate the difference-based estimate for $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}_{dbe}$. Plug-in the difference-based estimate in (2.3), and further apply the plug-in bandwidth selector, we can select an appropriate bandwidth for the profile least squares procedures. The selected bandwidth is used for the SCAD procedure in (2.7).

3 Numerical comparison and application

In this section, we investigate the finite sample performance of the proposed procedures by Monte Carlo simulation, and compare the performance of proposed procedures with existing ones by the mean squares errors, defined by

$$\mathrm{MSE}\{\widehat{m}(\cdot)\} = \frac{1}{n}\sum_{t=1}^{n}\{\widehat{m}(x_t) - m(x_t)\}^2.$$

We summarize our simulation results in terms of relative MSE (RMSE), defined by the ratio of the MSE of an estimation procedure to the MSE of $\widehat{m}_I(\cdot)$, the estimate of $m(\cdot)$ pretending the error ε_t being independent. We report the percentage of accuracy gain, defined by $(1 - RMSE) * 100\%$.

Example 1. In this example, a random sample of size n, either $n = 100$ or $n = 500$, is

generated from

$$y_t = m(x_t) + \varepsilon_t.$$

In this example, we consider two scenarios for $m(x)$. The first one is

$$m(x) = 4\cos(2\pi x),$$

and the second one is

$$m(x) = \exp(2x).$$

The mean function $m(x)$ is not monotone in the first scenario, while it is monotone in the second scenario. The error process ε_t is an AR process of order $d = 10$ or $d = 20$, i.e.,

$$\varepsilon_t = \sum_{j=1}^{d} \beta_j \varepsilon_{t-j} + \eta_t,$$

where $\eta_t \sim N(0, \sigma^2)$ with $\sigma = 0.5$ or 1. In our simulation we consider two situations: the first one is $\beta_1 = 0.5$, or 0.7, and all other β_j's equal 0, the second one is $\beta_1 = 0.5$, $\beta_2 = 0.4$ or $\beta_1 = 0.7$, $\beta_2 = 0.2$ and all others equal 0. In the first situation, the error process indeed is an AR(1), while in the second situation, the error process is an AR(2). The number of replication is 500.

To understand how the sampling scheme of x_t affects the proposed procedure, we consider three sampling schemes in our simulation.

I. x_t is independent and identically distributed according to the uniform distribution over $[0, 1]$.

II. u_t is independent and identically distributed according to the standard normal distribution for $t = 1, 2, ...$ Let $x_t = \Phi\{(au_t + bu_{t-1})/\sqrt{a^2 + b^2}\}$ for $t = 2, 3, \cdots$, where $\Phi(u)$ is the cumulative distribution function of the standard normal distribution. Thus, x_t is 1-dependent process. In our simulation, we take $a = 0.9$ and $b = 0.1$.

III. x_t is a fixed design point evenly distributed over $[0, 1]$, i.e., $x_t = (t - 0.5)/n$.

Table 1: Simulation Results for Sampling Scheme I when $d = 10$

(β_1, β_2)	σ	XLCM	Profile	SCAD	Oracle	XLCM	Profile	SCAD	Oracle
		$m(x) = 4\cos(2\pi x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	1.24	1.74	6.22	8.71	10.04	11.00	13.62	14.06
(0.7,0)	0.5	11.35	8.88	17.69	19.1	25.14	25.24	27.01	27.14
(0.5,0.4)	0.5	15.03	14.45	18.24	19.4	27.34	27.42	28.07	28.07
(0.7,0.2)	0.5	17.44	17.86	20.74	21.94	30.17	30.25	30.79	30.84
(0.5,0)	1	4.03	4.98	6.38	7.66	9.53	9.44	12.45	12.70
(0.7,0)	1	9.29	9.03	15.63	16.18	23.24	23.12	25.08	25.13
(0.5,0.4)	1	13.27	12.11	16.19	17.18	27.39	27.46	28.12	28.13
(0.7,0.2)	1	16.25	15.17	18.92	19.80	30.18	30.25	30.80	30.85
		$m(x) = \exp(2x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	0.19	0.98	2.87	3.12	5.05	4.44	5.81	6.10
(0.7,0)	0.5	4.38	4.60	7.02	8.88	15.42	14.48	16.54	16.45
(0.5,0.4)	0.5	11.12	8.83	13.26	14.34	21.94	21.73	22.32	22.37
(0.7,0.2)	0.5	12.28	14.39	15.67	17.09	24.48	24.25	24.78	24.86
(0.5,0)	1	5.06	3.83	6.25	6.71	5.72	4.75	7.76	7.97
(0.7,0)	1	3.54	4.20	7.32	7.90	15.31	14.24	16.52	16.42
(0.5,0.4)	1	11.00	8.53	13.33	14.29	22.05	21.85	22.45	22.50
(0.7,0.2)	1	14.11	13.86	15.67	16.88	24.51	24.27	24.79	24.89

For each sampling scheme, three methods, Xiao, Linton, Carroll and Mammen (2003) method (XLCM), profile least squares method (Profile) and penalized profile least square method with SCAD penalty function (SCAD) are compared with regard to the efficiency improvement. In addition, oracle procedure by substituting the true autoregressive coefficient and order is listed as a benchmark.

For sample scheme I, the covariate x_t's are independent and identically distributed, only the random error is correlated. Tables 1 and 2 summarize the simulation for sampling scheme

I for $d = 10$ and 20, respectively. The overall pattern for $d = 10$ and $d = 20$ is the same, although the gain in term of RMSE with $d = 20$ is slightly more than that with $d = 10$. For both $d = 10$ and 20, the SCAD procedures performs better than the Xiao's et al (2003)'e method and the profile least squares estimate, and its performance is very close to the oracle procedure. The performance of Xiao et al's method and the profile least squares procedure is very close to each other, and no one dominates the other one.

When the sample size is large, such as $n = 500$, the performance of Xiao et al's method, the profile least squares method and the SCAD procedure are very closely to each other, although the SCAD procedure is slightly better than the other two. The gain for these three methods in terms of RMSE with large sample is more than the one with the smaller sample size ($n = 100$). This is expected because with the large sample size, all three methods can estimate $\boldsymbol{\beta}$ more accurate. This leads the decorrelation method works better.

Simulation results for sampling scheme II are summarized in Tables 3 and 4. From Tables 3 and 4, we can see that results for $d = 10$ and $d = 20$ are almost the same in terms of RMSE. The overall pattern of Tables 3 and 4 is similar to that in Tables 1 and 2. Although the sampling scheme II is different from the sampling scheme I in that the covariate x_t's is dependent in the sample scheme II, while they are independent in the sample scheme I. For the sample scheme II, the SCAD procedures performs best among the three methods in the comparison, and its performance is very close to the oracle procedure. The performances of Xiao et al's method and the profile least squares procedure are similar, and no one dominates the other one.

As a summary, the performance of the proposed profile least squares procedure and the SCAD procedures seems not to rely on the sampling scheme of covariate x_t.

Although the sampling scheme III does not satisfy the regularity conditions and is not the focus of this paper, we include this sample scheme to demonstrate that the proposed method may work well for this kind sampling scheme.

For the sampling scheme III, $\{x_t\}$ is a fixed design. As demonstrated in Altman (1990)

Table 2: Simulation Results for Sampling Scheme I when $d = 20$

(β_1, β_2)	σ	XLCM	Profile	SCAD	Oracle	XLCM	Profile	SCAD	Oracle
		$m(x) = 4\cos(2\pi x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	3.04	1.08	7.05	8.71	12.27	12.35	13.72	14.07
(0.7,0)	0.5	14.84	13.63	17.72	18.46	26.03	26.08	27.00	27.18
(0.5,0.4)	0.5	17.18	17.05	18.59	19.12	30.32	30.41	30.73	30.70
(0.7,0.2)	0.5	19.96	19.92	20.99	21.67	33.08	33.16	33.47	33.50
(0.5,0)	1	1.16	1.07	5.65	6.47	10.91	10.87	12.48	12.72
(0.7,0)	1	12.87	11.21	15.62	15.78	24.08	24.01	25.10	25.18
(0.5,0.4)	1	15.54	15.23	16.73	17.09	27.68	27.71	28.07	28.09
(0.7,0.2)	1	18.39	18.09	19.42	19.66	30.44	30.47	30.75	30.82
		$m(x) = \exp(2x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	3.79	4.70	5.00	4.21	4.39	3.57	6.53	5.91
(0.7,0)	0.5	12.26	11.29	13.45	14.80	16.00	15.79	16.40	16.47
(0.5,0.4)	0.5	13.48	12.31	13.79	14.33	22.50	22.38	22.70	22.75
(0.7,0.2)	0.5	16.07	14.99	16.06	16.63	25.00	24.87	25.18	25.23
(0.5,0)	1	5.53	3.85	6.90	6.71	6.40	6.02	8.13	8.18
(0.7,0)	1	14.24	13.27	14.39	15.25	15.66	15.69	16.75	16.64
(0.5,0.4)	1	13.60	12.36	14.06	14.51	22.20	22.09	22.42	22.47
(0.7,0.2)	1	16.41	15.29	16.56	16.59	24.64	24.52	24.78	24.90

and Hart (1991) for kernel regression estimator, the ordinary bandwidth selector will tend to undersmooth the true regression function when the error is positively correlated. From our simulation experience, the local linear regression estimator also suffer from this difficulty: the plug-in bandwidth proposed by Ruppert, Sheather and Wand (1995) always picks a small bandwidth to undersmooth the fitting. Many authors have proposed various adjustment methods on bandwidth selection to overcome the difficulty in fixed design. For example, Altman (1990) suggested revising CV and GCV criteria by incorporating the estimation

Table 3: Simulation Results for Sampling Scheme II when $d = 10$

(β_1, β_2)	σ	XLCM	Profile	SCAD	Oracle	XLCM	Profile	SCAD	Oracle
		$m(x) = 4\cos(2\pi x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	4.82	3.06	8.20	10.30	12.94	13.09	13.85	14.06
(0.7,0)	0.5	16.33	14.92	18.42	20.28	25.95	26.05	26.58	26.71
(0.5,0.4)	0.5	19.39	19.47	20.93	21.46	30.43	30.54	30.65	30.71
(0.7,0.2)	0.5	20.15	20.22	21.24	22.04	33.12	33.25	33.35	33.41
(0.5,0)	1	2.21	3.17	5.57	7.72	11.86	11.90	12.81	12.86
(0.7,0)	1	17.32	16.33	19.93	20.45	24.01	24.02	24.67	24.65
(0.5,0.4)	1	18.10	17.81	18.72	19.44	27.66	27.70	27.86	27.88
(0.7,0.2)	1	23.29	24.23	25.38	25.41	30.29	30.35	30.44	30.52
		$m(x) = \exp(2x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	2.11	4.15	6.05	7.85	4.14	3.79	4.67	4.34
(0.7,0)	0.5	13.63	13.21	14.46	15.10	14.91	14.65	15.69	15.55
(0.5,0.4)	0.5	16.65	15.99	16.64	16.90	20.84	20.68	20.86	20.87
(0.7,0.2)	0.5	19.28	18.77	19.2	19.67	23.46	23.32	23.42	23.50
(0.5,0)	1	6.29	5.80	7.80	10.34	5.76	5.68	6.74	6.51
(0.7,0)	1	14.17	14.69	15.63	15.99	16.99	17.12	17.77	17.64
(0.5,0.4)	1	17.07	16.41	17.11	17.46	20.51	20.35	20.62	20.60
(0.7,0.2)	1	19.93	19.50	20.17	20.44	22.87	22.72	22.91	22.98

of covariance structure. Hart (1991) proposed a risk estimation procedure. For simplicity, a ratio based on the pilot study is multiplied on the the plug-in bandwidth to adjust the undersmoothness in our simulation study,

Because $\{x_t\}$ is the fixed design in scheme III, the results of Table 3 are quite different from Table 1 and 2, although all XLCM, Profile, SCAD methods improve the estimation efficiency as expected. The magnitude of the gain at the same correlation level is much more significant than that in sampling schemes I and II, especially when $\beta_1 = 0.7$ in $AR(1)$ model.

Table 4: Simulation Results for Sampling Scheme II when $d = 20$

(β_1, β_2)	σ	XLCM	Profile	SCAD	Oracle	XLCM	Profile	SCAD	Oracle
		$m(x) = 4\cos(2\pi x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	0.70	2.10	6.33	10.05	11.7	12.07	13.85	14.00
(0.7,0)	0.5	13.23	10.62	19.41	20.97	25.07	25.38	26.52	26.57
(0.5,0.4)	0.5	17.22	17.54	20.52	21.87	27.19	27.36	27.79	27.76
(0.7,0.2)	0.5	19.55	20.48	23.52	24.42	29.89	30.09	30.46	30.47
(0.5,0)	1	1.45	1.80	6.41	8.35	10.74	10.94	12.85	12.89
(0.7,0)	1	13.19	10.44	19.45	20.06	23.38	23.56	24.81	24.76
(0.5,0.4)	1	16.53	16.54	19.51	20.41	27.26	27.43	27.87	27.83
(0.7,0.2)	1	19.11	19.35	22.27	22.93	29.97	30.17	30.54	30.55
		$m(x) = \exp(2x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	0.23	0.96	1.17	3.82	1.62	0.95	2.54	2.59
(0.7,0)	0.5	9.30	8.83	12.66	13.69	13.37	12.75	13.83	13.74
(0.5,0.4)	0.5	14.74	13.65	15.91	16.57	20.16	19.96	20.54	20.42
(0.7,0.2)	0.5	17.48	16.62	18.93	19.48	22.60	22.41	22.91	22.88
(0.5,0)	1	3.39	3.38	7.23	8.19	1.27	2.40	5.97	5.97
(0.7,0)	1	9.72	7.03	14.44	14.74	12.54	11.76	13.13	13.03
(0.5,0.4)	1	15.31	14.24	16.73	17.23	20.21	20.02	20.60	20.48
(0.7,0.2)	1	17.79	16.88	19.49	19.89	22.70	22.53	22.99	23.03

The profile least squares procedure is similar to Xiao et al's method. More interestingly, all three methods have more gain for monotone regression function $m(x) = \exp(2x)$ than non-monotone function $m(x) = 4\cos(2\pi x)$ only in this scheme.

Example 2. In this example, we illustrate the proposed methodology by analysis of a data set about U.S. macroeconomics, collected from January 1980 to December 2006 in a monthly basis. Our interest here is to investigate the relationship between the unemployment rate and house price index change.

Table 5: Simulation Results for Sampling Scheme III when $d = 10$

(β_1, β_2)	σ	XLCM	Profile	SCAD	Oracle	XLCM	Profile	SCAD	Oracle
		$m(x) = 4\cos(2\pi x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	9.32	9.74	18.58	23.12	16.87	18.56	27.06	27.56
(0.7,0)	0.5	21.36	21.47	29.87	32.48	37.68	39.90	43.27	43.66
(0.5,0.4)	0.5	18.80	17.51	19.62	20.67	39.56	43.33	44.95	45.03
(0.7,0.2)	0.5	21.03	23.33	24.72	26.09	42.17	45.80	47.33	47.33
(0.5,0)	1	8.64	9.24	22.00	27.44	20.63	22.74	28.96	29.53
(0.7,0)	1	21.30	24.60	32.78	36.28	40.03	42.54	46.96	47.74
(0.5,0.4)	1	21.32	24.68	25.29	28.71	39.80	43.72	45.29	45.59
(0.7,0.2)	1	21.55	25.14	27.08	30.24	42.39	36.15	47.64	47.78
		$m(x) = \exp(2x)$							
		$n = 100$				$n = 500$			
(0.5,0)	1	8.10	11.40	23.54	30.50	24.75	27.09	36.27	37.19
(0.7,0)	1	21.16	25.27	33.64	37.63	40.97	43.53	48.58	49.38
(0.5,0.4)	1	21.74	25.58	28.08	32.05	39.80	43.74	45.52	45.66
(0.7,0.2)	1	21.71	25.77	28.67	32.11	42.41	46.18	47.83	47.85
(0.5,0)	1	7.63	11.18	24.39	31.24	25.50	27.85	37.29	38.51
(0.7,0)	1	21.09	25.26	33.83	37.84	41.03	43.60	48.57	49.50
(0.5,0.4)	1	21.95	25.98	27.51	32.59	39.83	43.79	45.40	45.71
(0.7,0.2)	1	21.66	25.80	28.41	32.28	42.40	46.16	47.68	47.84

In the past few years, house price in U.S. has shown a strong upward trend, although the bubble warning always exists. Many home buyers who do not have sound credit history nor sufficient financial capability become home owners with the help of sub-prime mortgage. They bear with high level of interest payments but believe the property will keep appreciating. In the meantime, the mortgage agent packages the debt and sell it to other institutional investors. This long chain prospers and works well when the housing market is booming. However, when the house price began to plummet in spring 2007, borrowers had to default

Table 6: Simulation Results for Sampling Scheme III when $d = 20$

(β_1, β_2)	σ	XLCM	Profile	SCAD	Oracle	XLCM	Profile	SCAD	Oracle
		$m(x) = 4\cos(2\pi x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	8.55	9.26	9.58	22.40	14.80	17.44	22.01	22.13
(0.7,0)	0.5	22.23	22.36	25.90	32.25	36.07	39.24	43.11	43.56
(0.5,0.4)	0.5	18.92	14.06	19.61	20.73	38.78	43.27	44.93	45.04
(0.7,0.2)	0.5	21.62	19.53	24.69	26.13	41.41	45.84	47.32	47.35
(0.5,0)	1	11.98	12.18	19.63	27.62	15.40	19.18	27.92	28.69
(0.7,0)	1	23.75	22.30	30.95	35.91	37.31	41.14	46.50	47.43
(0.5,0.4)	1	21.86	21.41	25.30	28.64	38.94	43.77	45.39	45.67
(0.7,0.2)	1	22.51	20.72	27.79	30.47	41.56	45.64	47.09	47.51
		$m(x) = \exp(2x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	13.59	10.84	22.80	31.60	16.81	21.09	33.93	35.17
(0.7,0)	0.5	24.11	18.28	31.17	37.78	37.81	41.74	47.79	48.81
(0.5,0.4)	0.5	21.93	21.32	27.18	32.04	38.85	43.73	45.51	45.71
(0.7,0.2)	0.5	22.07	20.98	28.39	32.21	41.47	46.11	47.77	47.85
(0.5,0)	1	13.02	10.31	25.29	32.52	16.94	21.33	34.45	48.94
(0.7,0)	1	23.90	18.61	33.02	38.10	37.85	41.84	47.82	48.94
(0.5,0.4)	1	21.67	21.47	27.31	32.43	38.87	43.76	43.43	45.73
(0.7,0.2)	1	21.84	21.18	28.77	32.35	41.54	46.20	47.71	47.92

and many houses went to foreclosure. Consequently, a number of big financial institutions that have heavy investment in sub-prime mortgage market claimed billions dollars write-off due to the crisis.

In this example, we are interested in the effect of unemployment rate on the house price. By classical economics theory, unemployment rate is an important indicator of the overall economics. If many people claim unemployment, the purchase power is definitely be hurt. However, to our best knowledge, there are no many literatures to study the relationship

between the unemployment rate and the housing market in a quantitative manner. Motivated by the sub-prime mortgage turmoil and recent suspicion of recession, it is believed that the historical data might shed some interesting insights on how these two indexes are related. Thus, we take the unemployment rate as the covariate x and House Price Index Change as the response variable y, and consider the following model

$$y_t = m(x_t) + \epsilon_t. \tag{3.1}$$

Initial estimate and residual analysis. We ignore the correlation of the random errors temporarily and estimate (3.1) by the conventional local linear model as Fan and Gijbels (1996). The Ruppert, Sheather and Wand (1995)'s direct plug-in bandwidth is 0.2969.

When the initial estimate $\tilde{m}(x_t)$ is obtained, we can estimate the residual ϵ_t as $y_t - \tilde{m}(x_t)$. There is an obvious correlation pattern present in the autocorrelation plot of $\widehat{\epsilon}_t$ (Figure 2 (a)). The partial-autocorrelation plot (Figure 2 (b)) indicates an autoregressive model and the first lag effect is most outstanding. Furthermore, the Ljung-Box-Pierce test used to check the autocorrelation pattern for white noise has the P-value less than 0.0001. It also verifies that autocorrelation exists in $\widehat{\epsilon}_t$.

From a conservative point of view, we suspect that the house price might have a year lag. So we assume AR(12) model on errors and employ the penalized profile least square method to select the AR order and estimate $m(\cdot)$ simultaneously. The plug-in bandwidth in the profile least square estimation is 0.2140. By the BIC criterion, the optimal tuning parameter used in the order selection procedure is 0.000019.

As an result, AR(1) model with a strong autocorrelation coefficient 0.9438 is most appropriate. It means that the error has only one month lag, which agrees with the partial-autocorrelation plot in Figure 2 (b). After accounting for the autocorrelation, the correlogram of $\widehat{\eta}_t$ does not have any significant pattern. (See Figure 2 (c) and (d)) In addition, the P-value in the Ljung-Box-Pierce test at the first 24 lags, 0.9134, also shows that the autocorrelation has been successfully removed.

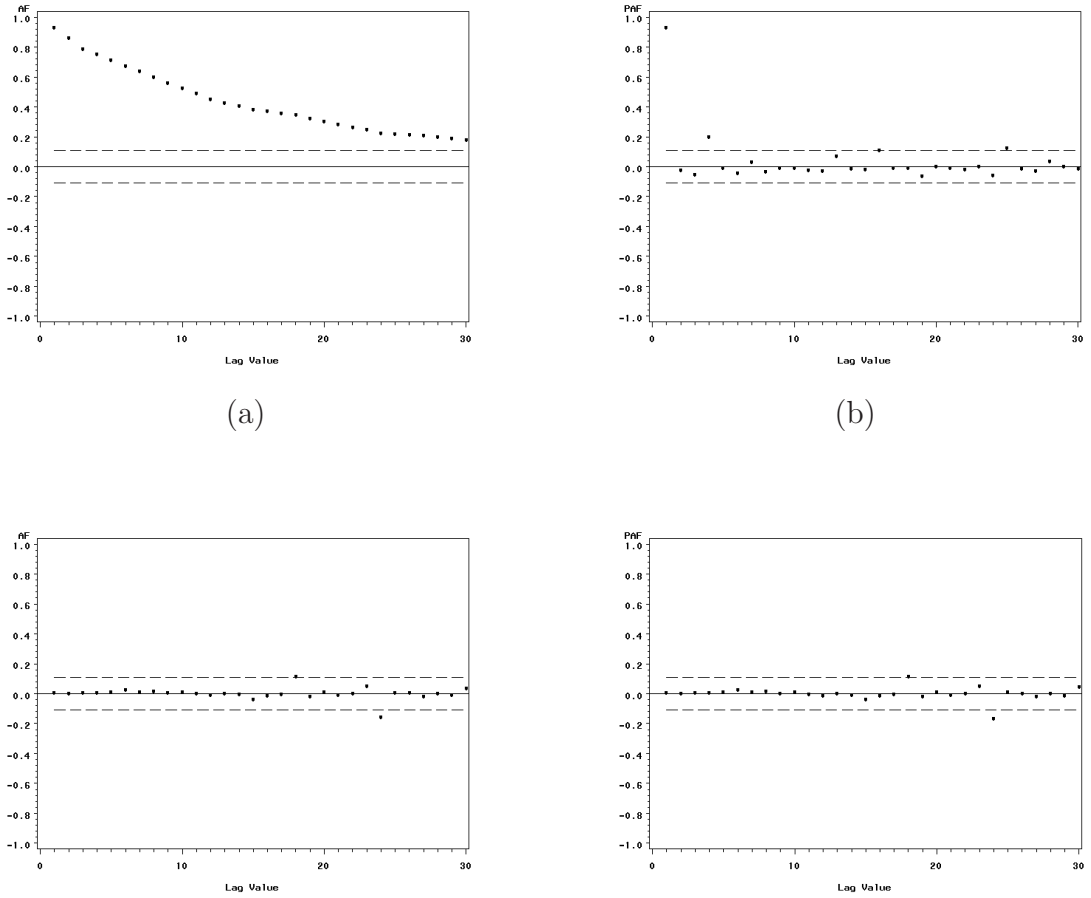

Figure 2: Correlogram of residual $\widehat{\epsilon}_t$ and $\widehat{\eta}_t$. Plot (a) and (b) are the autocorrelation and partial autocorrelation for $\widehat{\epsilon}_t$. Plot (c) and (d) are the autocorrelation and partial autocorrelation for $\widehat{\eta}_t$. In each plot, the upper and lower dashed lines represent 95% confidence interval.

Final model. By applying the penalized profile least squares estimation method, the relationship between the House Price Index Change and the unemployment rate turns out to be

$$\widehat{y}_t = \widehat{m}(x_t) + 0.9438\widehat{\epsilon}_{t-1} \tag{3.2}$$

where $\widehat{m}(\cdot)$ is displayed in Figure 3. The penalized profile least square approach yields a smoother estimate than the conventional local linear regression because it takes the corre-

lation into account. As expected, the unemployment rate has a negative correlation with house price index change. But this effect is most significant when the unemployment varies between 4% and 5% or between 8% and 10%.

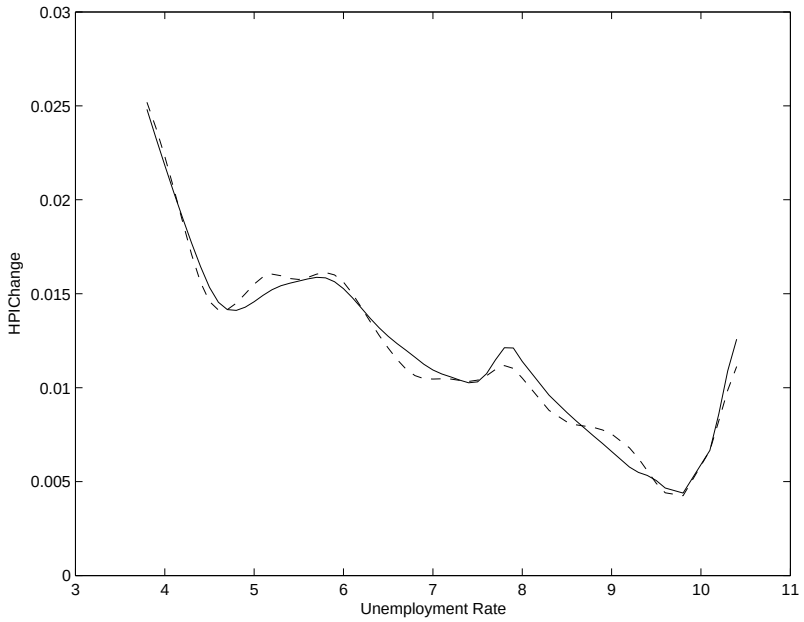

Figure 3: Estimation of $m(\cdot)$. Dashed curves are the initial estimates; Solid curves are the penalized profile least squares estimate.

4 Proofs

4.1 Preliminaries

To present the regularity conditions, we need the following definitions for a sequence of random vectors $\{\mathbf{z}_t, t = 0, \pm 1, \pm 2, \cdots\}$. The following notation and definitions are adopted from Chapter 2 of Fan and Yao (2003).

Definition 1. *A sequence of random vectors* $\{\mathbf{z}_t, t = 0, \pm 1, \pm 2, \cdots\}$ *is said to be strictly stationary if* $\{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$ *and* $\{\mathbf{z}_{1+k}, \cdots, \mathbf{z}_{1+n}\}$ *have the same joint distributions for any integer* $n \geq 1$ *and any integer* k.

Denote $\mathcal{F}_i^j$ to be the σ-algebra generated by events $\{\mathbf{z}_t, i \leq t \leq j\}$, and $\mathcal{L}^2(\mathcal{F}_i^j)$ consists of $\mathcal{F}_i^j$-measurable random variables with finite second moment. Intuitively, $\mathcal{F}_i^j$ assembles all

information on the sequence collected between time i and j. Define

$$\alpha(n) = \sup_{A \in \mathcal{F}_{-\infty}^{0}, B \in \mathcal{F}_{n}^{\infty}} |P(A)P(B) - P(AB)| \tag{4.1}$$

Definition 2. *A sequence of random vectors* $\{\mathbf{z}_t, t = 0, \pm 1, \pm 2, \cdots\}$ *is said to be* α*-mixing if it is strictly stationary and* $\alpha(n) \to 0$ *as* $n \to \infty$.

4.2 Regularity conditions and proofs

To make the argument concise, denote $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)^T$ with $\mathbf{f}_t = (\epsilon_{t-1}, \ldots, \epsilon_{t-d})^T$, and $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_n)^T$ with $\mathbf{e}_t = (\widehat{\epsilon}_{t-1}, \ldots, \widehat{\epsilon}_{t-d})^T$. Define $\mathbf{\Delta} = \mathbf{E} - \mathbf{F}$. Our proof follows the same strategy as that in Fan and Huang (2005). The following conditions are imposed to facilitate the proof and are adopted from Fan and Huang (2005). They are not the weakest possible conditions.

A. The random variable x_t has a bounded support Ω. Its density function $f(.)$ is Lipschitz continuous and bounded away from 0 on its support.

B. There is an $s > 2$ such that $E||\mathbf{f}_t||^s < \infty$ and for some $\xi > 0$ such that $n^{1-2s^{-1}-2\xi}h \to \infty$.

C. $m(\cdot)$ has the continuous second derivative in $x \in \Omega$.

D. The function $K(\cdot)$ is a bounded symmetric density function with bounded support $[-M, M]$, satisfying the Lipschitz condition.

E. $nh^8 \to 0$ and $nh^2/(\log n)^2 \to \infty$.

F. $\sup_{x \in \Omega} |\widehat{m}_I(x) - m(x)| = o_p(n^{-\frac{1}{4}})$ where $\widehat{m}_I(x_t)$ is obtained by local linear regression pretending that data are i.i.d.

G. The sequence of random vector (x_t, ε_t), $t = 1, 2, \cdots$, is a strictly stationary and satisfies the mixing condition for α-mixing processes: assume that for some $\delta > 2$ and $a > 1 - 2/\delta$,

$$\sum_l l^a [\alpha(l)]^{1-2/\delta} < \infty, \quad E|\varepsilon_1|^\delta < \infty, \quad g_{x_1|\varepsilon_1}(x|\varepsilon) \leq A_1 < \infty$$

Lemma 4.1 is taken from Lemma 6.1 of Fan and Yao (2003) and will be used in our proof repeatedly.

Lemma 4.1. *Let $(x_1, \varepsilon_1), \ldots, (x_n, \varepsilon_n)$ be a strictly stationary sequence satisfying the mixing condition $\alpha(l) \leq cl^{-\tau}$ for some $c > 0$ and $\tau > 5/2$. Assume further that for some $s > 2$ and interval $[a, b]$,*

$$E|\varepsilon_t|^s < \infty \quad \textit{and} \quad \sup_{\forall x \in [a,b]} \int |\varepsilon_t|^s g(x, \epsilon) \mathrm{d}\epsilon < \infty,$$

where g denote the joint density of (x_t, ε_t). In addition, Condition G holds, and the conditional density $g_{x_1, x_l|\varepsilon_1, \varepsilon_l}(x_1, x_l|\varepsilon_1, \varepsilon_l) \leq A_2 < \infty, \forall l \geq 1$. Let K satisfy Condition D. Then

$$\sup_{x \in [a,b]} |\frac{1}{n} \sum_{i=1}^{n} \{K_h(x_i - x)\varepsilon_i - E[K_h(x_i - x)\varepsilon_i]\}| = O_p(\left\{ \frac{\log n}{nh} \right\}^{1/2})$$

provided that $h \to 0$, for some $\xi > 0, n^{1-2s^{-1}-2\xi} h \to \infty$ and $n^{(\tau+1.5)(s^{-1}+\xi)-\tau/2+5/4} h^{-\tau/2-5/4} \to 0$.

Lemma 4.2. *Under Conditions A—G, it follows that*

$$\frac{1}{n} \mathbf{F}^T (I - S)^T (I - S) \mathbf{F} \xrightarrow{P} E(\mathbf{f}\mathbf{f}^T).$$

Proof Denote W_x be a $n \times n$ diagonal matrix with j-th diagonal element $K_h(x_j - x)$ and

$$D_x = \begin{pmatrix} 1 & \frac{x_1 - x}{h} \\ \vdots & \vdots \\ 1 & \frac{x_n - x}{h} \end{pmatrix}$$

Then the smoothing matrix $\mathbf{S}$ for the local linear regression can be expressed as

$$\mathbf{S} = \begin{pmatrix} [1, 0]\{D_{x_1}^T W_{x_1} D_{x_1}\}^{-1} D_{x_1}^T W_{x_1} \\ \vdots \\ [1, 0]\{D_{x_n}^T W_{x_n} D_{x_n}\}^{-1} D_{x_n}^T W_{x_n} \end{pmatrix}$$

where

$$D_x^T W_x D_x = \begin{pmatrix} \sum_{i=1}^n K_h(x_i - x) & \sum_{i=1}^n (x_i - x) K_h(x_i - x)/h \\ \sum_{i=1}^n (x_i - x) K_h(x_i - x)/h & \sum_{i=1}^n (x_i - x)^2 K_h(x_i - x)/h^2 \end{pmatrix}$$

The generic element of matrix $D_x^T W_x D_x$ is in the form of $\sum_{i=1}^n (\frac{x_i - x}{h})^j K_h(x_i - x)$, $j = 0, 1, 2$. Denote $S_{n,j} = \sum_{i=1}^n (\frac{x_i - x}{h})^j K_h(x_i - x)$. By using the formula $S_{n,j} = E(S_{n,j}) + O_p(\sqrt{\mathrm{Var}(S_{n,j})})$, it is easy to show that if j is even,

$$\begin{aligned} S_{n,j} &= n \int v^j K(v) f(x + hv) \mathrm{d}v + O_p(\sqrt{nE\{(x_1 - x)^{2j} K_h^2(x_1 - x)\}}) \\ &= nf(x)\mu_j + O_p(h^2 + 1/\sqrt{nh}) \end{aligned}$$

Because of the symmetry of kernel function, for any odd numbered j, $\mu_j = 0$ and then $S_{n,j} = O_p(h + 1/\sqrt{nh})$. Indeed, with Lemma 4.1, it can be further shown that for even j,

$$S_{n,j} = nf(x)\mu_j + O_p(h^2 + \sqrt{\log(n)/nh}),$$

and for odd j,

$$S_{n,j} = O_p(h + \sqrt{\log(n)/nh})$$

holds uniformly in x. Therefore,

$$\frac{1}{n}D_x^T W_x D_x = \begin{pmatrix} f(x)(1 + O_p(h^2 + \sqrt{\log(n)/nh})) & O_p(h + \sqrt{\log(n)/nh}) \\ O_p(h + \sqrt{\log(n)/nh}) & f(x)\mu_2(1 + O_p(h^2 + \sqrt{\log(n)/nh})) \end{pmatrix}$$

holds uniformly in x.

Since $h + \sqrt{\log(n)/nh} = o_p(1)$, we can regard the above matrix as being approximately diagonal. Then its inverse is

$$\{\frac{1}{n}D_x^T W_x D_x\}^{-1} = \begin{pmatrix} \{f(x)\}^{-1}(1 + O_p(h^2 + \sqrt{\log(n)/nh}) & O_p(h + \sqrt{\log(n)/nh}) \\ O_p(h + \sqrt{\log(n)/nh}) & \{f(x)\mu_2\}^{-1}(1 + O_p(h^2 + \sqrt{\log(n)/nh})) \end{pmatrix}$$

holds uniformly in x.

Similarly, by Lemma 4.1 and the assumption of independence between the process ε_t and the process x_t, it follows that

$$\frac{1}{n}D_x^T W_x \mathbf{F} = \begin{pmatrix} O_p(h^2 + \sqrt{\frac{\log n}{nh}}) \\ O_p(h + \sqrt{\frac{\log n}{nh}}) \end{pmatrix}$$

holds uniformly in x.

Consequently,

$$\begin{aligned} &[1,0]\{\frac{1}{n}D_x^T W_x D_x\}^{-1}\{\frac{1}{n}D_x^T W_x \mathbf{F}\} \\ &= [1,0] \begin{pmatrix} \{f(x)\}^{-1}(1 + O_p(h^2 + \sqrt{\log(n)/nh})) & O_p(h + \sqrt{\log(n)/nh}) \\ O_p(h + \sqrt{\log(n)/nh}) & \{f(x)\mu_2\}^{-1}(1 + O_p(h^2 + \sqrt{\log(n)/nh})) \end{pmatrix} \begin{pmatrix} O_p(h^2 + \sqrt{\frac{\log n}{nh}}) \\ O_p(h + \sqrt{\frac{\log n}{nh}}) \end{pmatrix} \\ &= \{f(x)\}^{-1} O_p(h^2 + \sqrt{\frac{\log n}{nh}})(1 + o_p(1)) = o_p(1) \end{aligned}$$

Substituting this result into the smoothing matrix S, we have

$$S\mathbf{F} = \begin{pmatrix} [1,0]\{D_{x_1}^T W_{x_1} D_{x_1}\}^{-1} D_{x_1}^T W_{x_1}\mathbf{F} \\ \vdots \\ [1,0]\{D_{x_n}^T W_{x_n} D_{x_n}\}^{-1} D_{x_n}^T W_{x_n}\mathbf{F} \end{pmatrix} = \begin{pmatrix} o_p(1) \\ \vdots \\ o_p(1) \end{pmatrix}.$$

Thus,

$$\mathbf{F} - S\mathbf{F} = \mathbf{F}\{1 + o_P(1)\}.$$

Finally, by the WLLN,

$$\frac{1}{n}\mathbf{F}^T(I-S)^T(I-S)\mathbf{F} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{f}_i\mathbf{f}_i^T\right)\{1+o_p(1)\}^2 \xrightarrow{P} E(\mathbf{f}\mathbf{f}^T)$$

Lemma 4.3. *Under Conditions A—G, we have*

$$\frac{1}{n}\mathbf{E}^T(I-S)^T(I-S)\mathbf{E} \xrightarrow{P} E(\mathbf{f}\mathbf{f}^T).$$

Proof Since $\mathbf{\Delta} = \mathbf{E} - \mathbf{F}$, the generic element of $\mathbf{\Delta}$ is of the form $m(x_t) - \widehat{m}(x_t)$, which is of order $o_P(n^{-1/4})$ uniformly in x by Condition F. Thus, $\mathbf{\Delta} = o_P(n^{-1/4})$. Therefore

$$\frac{1}{n}\mathbf{E}^T(I-S)^T(I-S)\mathbf{E} = \frac{1}{n}(\mathbf{F}+\mathbf{\Delta})^T(I-S)^T(I-S)(\mathbf{F}+\mathbf{\Delta})$$

By using similar argument in the proof of Lemma 4.2, it can be shown that

$$\frac{1}{n}\mathbf{E}^T(I-S)^T(I-S)\mathbf{E} = \frac{1}{n}\mathbf{F}^T(I-S)^T(I-S)\mathbf{F} + o_P(1)$$

Thus, Lemma 4.3 follows by Lemma 4.2.

Lemma 4.4. *Suppose Conditions A—G hold. It follows*

$$\frac{1}{\sqrt{n}}\mathbf{F}^T(I-\mathbf{S})^T(I-\mathbf{S})\mathbf{m} = o_p(1)$$

Proof It is noted that

$$\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\mathbf{m} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathbf{f}_i - (S\mathbf{f})_i][m(x_i) - [1,0]\{D_{x_i}^T W_{x_i} D_{x_i}\}^{-1} D_{x_i}^T W_{x_i}\mathbf{m}] \quad (4.2)$$

Similar to the argument in the proof of Lemma 4.2, we can show that

$$[1,0]\{\frac{1}{n}D_x^T W_x D_x\}^{-1}\{\frac{1}{n}D_x^T W_x\mathbf{m}\} = m(x)(1 + O_p(h^2 + \sqrt{\log(n)/nh}))$$

holds uniformly in $x \in \Omega$. Plugging this in (4.2), it follows that

$$\begin{aligned}\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\mathbf{m} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathbf{f}_i-(S\mathbf{f})_i][m(x_i)-m(x_i)(1+O_p(h^2+\sqrt{\log(n)/nh})] \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{f}_i m(x_i)[1+o_p(1)]O_p(h^2+\sqrt{\log(n)/nh})\end{aligned}$$

Note that $E\{\mathbf{f}_i m(x_i)\} = 0$, and covariance matrix for $\{\mathbf{f}_i m(x_i)\}$ is finite. Thus, using $R = E(R) + O_p(\sqrt{\mathrm{Var}(R)})$, it follows that $\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\mathbf{m} = o_p(1)$.

Lemma 4.5. *Under Conditions A—G, we have*

$$\frac{1}{\sqrt{n}}\mathbf{E}^T(I-S)^T(I-S)\mathbf{m} = o_p(1)$$

Proof Since $\mathbf{E} = \mathbf{F} + \mathbf{\Delta}$, we can break $\frac{1}{\sqrt{n}}\mathbf{E}^T(I-S)^T(I-S)\mathbf{m}$ into two terms: $\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\mathbf{m}$, which is $o_P(1)$ by Lemma 4.4, and $\frac{1}{\sqrt{n}}\mathbf{\Delta}^T(I-S)^T(I-S)\mathbf{m}$, which is also $o_P(1)$ as $\mathbf{\Delta} = o_P(n^{-1/4})$.

Lemma 4.6. *Suppose that Conditions A—G hold. We have*

$$\frac{1}{\sqrt{n}}\mathbf{E}^T(I-S)^T(I-S)\mathbf{\Delta}\boldsymbol{\beta} = o_p(1)$$

Proof This is a direct result from the proof of Lemma 4.3.

Lemma 4.7. *Under Conditions A—G, let $\eta = (\eta_1, \cdots, \eta_n)^T$. Then*

$$\sqrt{n}[\mathbf{F}^T(I-S)^T(I-S)\mathbf{F}]^{-1}\mathbf{F}^T(I-S)^T(I-S)\eta \to N(0, \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}^{-1})$$

Proof We observe that

$$\mathbf{F}^T(I-S)^T(I-S)\eta = \sum_{i=1}^{n}\mathbf{f}_i[\eta_i - [1,0]\{D_{x_i}^T W_{x_i} D_{x_i}\}^{-1}D_{x_i}^T W_{x_i}\eta][1+o_p(1)] \tag{4.3}$$

By using Lemma 4.1 on $\{x_i, \eta_i\}$, we can show that

$$\begin{aligned}&[1,0]\{\frac{1}{n}D_x^T W_x D_x\}^{-1}\{\frac{1}{n}D_x^T W_x \eta\} \\ &= [1,0]\begin{pmatrix}\{f(x)\}^{-1}(1+O_p(h^2+\sqrt{\log(n)/nh}) & O_p(h+\sqrt{\log(n)/nh}) \\ O_p(h+\sqrt{\log(n)/nh}) & \{f(x)\mu_2\}^{-1}(1+O_p(h^2+\sqrt{\log(n)/nh})\end{pmatrix}\begin{pmatrix}O_p(h^2+\sqrt{\frac{\log n}{nh}}) \\ O_p(h+\sqrt{\frac{\log n}{nh}})\end{pmatrix} \\ &= o_p(1)\end{aligned}$$

Then $\eta_i - [1,0]\{D_{x_i}^T W_{x_i} D_{x_i}\}^{-1} D_{x_i}^T W_{x_i}\eta = \eta_i\{1+o_p(1)\}$. Plugging this in (4.3), we obtain that

$$\mathbf{F}^T(I-S)^T(I-S)\eta = \sum_{i=1}^{n} \mathbf{f}_i\eta_i\{1+o_p(1)\}$$

Since $E(\mathbf{f}_i\eta_i) = 0$, $\mathrm{Var}(\mathbf{f}_i\eta_i) = \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\} < \infty$, and $E(\mathbf{f}_i\eta_i\mathbf{f}_j\eta_j) = 0$ for $i \neq j$ since η_i is independent of $\mathbf{f}_i$. By Central Limit Theorem for strictly stationary sequence (see Theorem 2.21 of Fan and Yao, 2003),

$$\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\eta \xrightarrow{L} N(0, \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}).$$

By Lemma 4.2, $\frac{1}{n}\mathbf{F}^T(I-S)^T(I-S)\mathbf{F} \xrightarrow{P} E(\mathbf{f}\mathbf{f}^T)$. Apply the Slutsky theorem, it follows that

$$\sqrt{n}[\mathbf{F}^T(I-S)^T(I-S)\mathbf{F}])^{-1}\mathbf{F}^T(I-S)^T(I-S)\eta \xrightarrow{L} N(0, \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}^{-1}).$$

Lemma 4.8. *Under Conditions A—G, we have*

$$\sqrt{n}[\mathbf{E}^T(I-S)^T(I-S)\mathbf{E}])^{-1}\mathbf{E}^T(I-S)^T(I-S)\eta \xrightarrow{P} N(0, \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}^{-1})$$

Proof Since $\mathbf{E} = \mathbf{F} + \mathbf{\Delta}$, we may write $\mathbf{E}^T(I-S)^T(I-S)\eta = \mathbf{F}^T(I-S)^T(I-S)\eta + \mathbf{\Delta}^T(I-S)^T(I-S)\eta$. Note that $\mathbf{\Delta} = o_P(n^{-1/4})$ by Condition F, it can be shown that

$$\frac{1}{\sqrt{n}}\Delta^T(I-S)^T(I-S)\eta = o_p(1).$$

Furthermore, we have shown in the last lemma that $\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\eta \to N(0, \sigma^2 E(\mathbf{f}\mathbf{f}^T))$. So $\frac{1}{\sqrt{n}}\mathbf{E}^T(I-S)^T(I-S)\eta \to N(0, \sigma^2 E(\mathbf{f}\mathbf{f}^T))$ as well. The proof is completed by the Slutsky theorem and Lemma 4.3.

Proof of Theorem 1

Let us first show the asymptotic normality of $\widehat{\boldsymbol{\beta}}$. According to the expression of $\widehat{\boldsymbol{\beta}}$ in (2.5), we can break $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ into the sum of the following three terms (a), (b) and (c)

$$(a)\widehat{=}\sqrt{n}[\{\mathbf{E}^T(I-S)^T(I-S)\mathbf{E}\}^{-1}\mathbf{E}^T(I-S)^T(I-S)\mathbf{m}]$$

$$(b)\widehat{=}\sqrt{n}[\{\mathbf{E}^T(I-S)^T(I-S)\mathbf{E}\}^{-1}\mathbf{E}^T(I-S)^T(I-S)\mathbf{\Delta}\beta]$$

$$(c)\widehat{=}\sqrt{n}[\{\mathbf{E}^T(I-S)^T(I-S)\mathbf{E}\}^{-1}\mathbf{E}^T(I-S)^T(I-S)\eta]$$

For term (a), it is a product of two terms $[\frac{\{\mathbf{E}^T(I-S)^T(I-S)\mathbf{E}\}}{n}]^{-1}$ and $[\frac{\mathbf{E}^T(I-S)^T(I-S)\mathbf{m}}{\sqrt{n}}]$. From Lemmas 4.3 and 4.5, the asymptotic properties of these two terms lead to the conclusion that $(a) = o_p(1)$. Similarly, applying Lemmas 4.3 and 4.6 on two product components of term (b) results in $(b) = o_p(1)$ as well. In addition, Lemma 4.8 states that term (c) converges to $N(0, \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}^{-1})$. Put three terms together and we get the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$.

Next we derive the asymptotic bias and variance of $\widehat{m}(\cdot)$. From Lemmas 4.1— 4.8, we have

$$\widehat{m}(x_0, \widehat{\boldsymbol{\beta}}) = [1, 0]\{D_{x_0}^T W_{x_0} D_{x_0}\}^{-1} D_{x_0}^T W_{x_0}(\mathbf{m} + \eta)\{1 + o_P(1)\}$$

Note that $E(\eta|\mathcal{X}) = 0$, where $\mathcal{X} = (x_1, \cdots, x_n)$. Thus, So

$$E\{\widehat{m}(x_0, \widehat{\boldsymbol{\beta}})|\mathcal{X}\} = [1, 0]\{D_{x_0}^T W_{x_0} D_{x_0}\}^{-1} D_{x_0}^T W_{x_0}\mathbf{m}\{1 + o_p(1)\}$$

which is same as the conditional expected mean for local linear regression derived in Fan and Gijbels(1992). So the asymptotic bias is $\frac{1}{2}m''(x_0)h^2 \int x^2 K(x)$.

Regarding to the asymptotic variance of $\widehat{m}(\cdot)$, conditioning on $x_1, \cdots, x_n$,

$$\mathrm{Var}[\widehat{m}(x_0, \widehat{\boldsymbol{\beta}})|\mathcal{X}] = [1, 0]\{D_{x_0}^T W_{x_0} D_{x_0}\}^{-1} D_{x_0}^T W_{x_0} \mathrm{Var}\{\eta\} W_{x_0} D_{x_0}\{D_{x_0}^T W_{x_0} D_{x_0}\}^{-1}[1, 0]^T$$

Using the same argument as that in the proof of Lemma 4.2, we have

$$\mathrm{Var}[\widehat{m}(x_0, \widehat{\boldsymbol{\beta}})|\mathcal{X}] = \frac{\sigma^2}{nhf(x_0)} \int K^2(x)\,\mathrm{d}x$$

As to the asymptotic normality,

$$\widehat{m}(x_0, \widehat{\boldsymbol{\beta}}) - E\{\widehat{m}(x_0, \widehat{\boldsymbol{\beta}})|\mathcal{X}\} = [1, 0]\{D_{x_0}^T W_{x_0} D_{x_0}\}^{-1} D_{x_0}^T W_{x_0}\eta\{1 + o_P(1)\}$$

Thus, conditioning on $\mathcal{X}$, the asymptotic normality can be established using the CLT since η_i is independent and identically distributed with mean zero and variance σ^2.

5 Discussions

In this paper, we proposed a new estimation procedure for the nonparametric regression model with AR error by using profile least squares techniques. We further propose to determine the order of the AR error process using the penalized profile least squares with the SCAD penalty. We studied the asymptotic properties of the proposed estimators, and established their asymptotic normality. We conducted extensive Monte Carlo simulation studies to examine the finite sample performance of the proposed procedure and compare the proposed procedure with the proposal of Xiao et al. (2003).

It is of interest to theoretic compare the asymptotic mean squares errors of the proposed procedures and the local linear estimator without taking into account the error correlation. It is also of interest to investigate the effect of misspecification of error model. This needs more research in future.

Acknowledgements

Runze Li's research was supported by National Institute on Drug Abuse grant R21 DA024260, and Yan Li is supported by National Science Foundation grant DMS 0348869 as a graduate research assistant.

References

[1] Altman, N.S. (1990). Kernel Smoothing of Data with Correlated Errors. *Journal of the American Statistical Association*, **85**, 749-759.

[2] Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theorey and Methods*, Springer, New York.

[3] Engle, R.F., Granger,C.W.J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Stat. Assoc.,* , **81**, 310-320.

[4] Fan, J. and Gijbels, I. (1996). Local Polynomial Modeling and Its Applications, *Chapman and Hall*, London.

[5] Fan, J. and Huang, T. (2005) Profile Likelihood Inferences on Semiparametric Varying-coefficient Partially Linear Models. *Bernoulli*. **11**, 1031-1059.

[6] Fan, J., Huang, T. and Li, R. (2007) Analysis of Longitudinal Data with Semiparametric Estimation of Covariance Function *Journal of American Statistical Association*, **102**, 632-641.

[7] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of American Statistical Association*, **96**. 1348-1360.

[8] Fan, J. and Li, R. (2004). New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis. *Journal of American Statistical Association*, **99**, 710-723.

[9] Fan, J. and Yao, Q. (2003) *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag, New York.

[10] Gu, C. (2002). *Smoothing Spline ANOVA Models*, Springer-Verlag. New York.

[11] Hart, J.D. (1991). Kernel Regression Estimation with Time Series Errors. *Journal of the Royal Statistical Society*, Series B, **53**, 173-187.

[12] Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrics,* **59**, 1029-1054.

[13] Heckman, N. (1986). Spline smoothing in partly linear models, *J. Royal Stat. Soc., Ser. B*, **48**, 244-248.

[14] Li, R. and Li, Y. (2008). Local linear regression for data with AR errors. Technical Report 08-88, The Methodology Center, The Pennsylvania State University. Available at http://www.stat.psu.edu/˜rli/research/Report1v6.pdf.

[15] Liang, K. and Zeger, S. (1986) Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika*, **73**, 13-22.

[16] Lin, X. and Carroll, R. (2000) Nonparametric Function Estimation for Clustered Data When the Predictor is Measured without/with Error. *Journal of The American Statistical Association*, **95**, 520-534.

[17] Opsomer, J. (1995) Estimating a Function by Local Linear Regression when the Errors are Correlated. *preprint 95-42*. Department of Statistics, Iowa State University.

[18] Opsomer, J., Wang, Y. and Yang, Y. (2001) Nonparametric Regression with Correlated Errors. *Statistical Science*, **16**, 134-153.

[19] Qu, A., Lindsay, B. and Li, B. (2000) Improving Generalized Estimating Equations Using Quadratic Inference Functions. *Biometrika*, **87**, 823-836.

[20] Ruppert, D., Sheather, S.J. and Wand, M.P. (1995). An Effective Bandwidth Selector for Local Least Squares Regression. *Journal of American Statistical Association*, **90**, 1257-1270.

[21] Severini, T.A. and Staniswalis, J.G. (1994). Quasi-likelihood estimation in semiparametric models. *J. Amer. Stat. Assoc.,* , **89**, 501–511.

[22] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

[23] Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal. Royal Statist. Soc. B*, **50**, 413–436.

[24] Wahba, G. (1984). Partial spline models for semiparametric estimation of functions of several variables, in *Statist. Analysis of Time Ser.*, Proceedings of the Japan U.S. Joint Seminar, Tokyo, 319–329.

[25] Wang, H., Li, R. and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*. **94**, 553-568.

[26] Wang, N. (2003). Marginal nonparametric kernel regression accounting within-subject correlation. *Biometrika*, **90**, 29-42.

[27] Xiao, Z., Linton, O., Carroll, R.J. and Mammen, E. (2003) More Efficient Local Polynomial Estimation in Nonparametric Regression with Autocorrelated Errors. *Journal of the AMerican Statistical Association*, **98**, 980-992.

[28] Zeger, S. and Liang, K. (1986) Longitudinal Data-Analysis for Discrete and Continuous Outcomes. *Biometrika*, **42**, 121-130.