Statistica Sinica **25** (2015), 709-723
doi:http://dx.doi.org/10.5705/ss.2012.301

VARYING COEFFICIENT MODELS FOR DATA WITH AUTO-CORRELATED ERROR PROCESS

Zhao Chen, Runze Li and Yan Li

Princeton University, Pennsylvania State University and eBay Inc

Abstract: The varying coefficient model has been popular in the literature. In this paper, we propose a profile least squares estimation procedure for its regression coefficients when the random error is an auto-regressive (AR) process. We study the asymptotic properties of the proposed procedure, and establish asymptotic normality for the resulting estimate. We show that the resulting estimate for the regression coefficients has the same asymptotic bias and variance as the local linear estimate for varying coefficient models with independent and identically distributed observations. We apply the SCAD variable selection procedure (Fan and Li (2001)) to reduce model complexity of the AR error process. Numerical comparison and finite sample performance of the resulting estimate are examined in Monte Carlo studies. Our simulation results demonstrate the proposed procedure is more efficient than the one ignoring the error correlation. The proposed methodology is illustrated by a data example.

Key words and phrases: Auto-regressive error, profile least squares, SCAD, varying coefficient model.

1. Introduction

Suppose that a random sample $\{(u_t, x_{t1}, \ldots, x_{tp}, y_t),\ t = 1, \ldots, n\}$, is collected from the varying coefficient model

$$y_t = \alpha_0(u_t)x_{t0} + \alpha_1(u_t)x_{t1} + \cdots + \alpha_p(u_t)x_{tp} + \varepsilon_t, \tag{1.1}$$

where $\alpha_0(\cdot), \alpha_1(\cdot), \ldots, \alpha_p(\cdot)$ are unknown coefficient functions, and ε_t is the random error. We set x_{t0} to be 1 to include the intercept $\alpha_0(\cdot)$. In practice, it is common that data are collected over a period of time and are auto-correlated. This paper is concerned with the varying coefficient model for data with auto-regressive error ε_t.

The varying coefficient model was systematically introduced in Hastie and Tibshirani (1993). The statistical estimation and inference procedures for varying coefficient model with independent data or longitudinal data have been studied intensively. See Fan and Zhang (2008) and references therein for details. More recent developments can be found in Cheng, Zhang, and Chen (2009), Li and Zhang

(2011) and references therein. Varying coefficient models have been proposed for time series data. Cai, Fan, and Li (2000) proposed a local linear estimation procedure for functional-coefficient models with nonlinear time series, while Huang and Shen (2004) developed an estimation procedure for the functional coefficient models with nonlinear time series using a polynomial spline approach. Cai, Yao, and Zhang (2001) proposed a local smoothed maximum likelihood estimator for discrete time series data. Cai (2007) developed a local linear approach to estimating the time trend and coefficient functions. In the absence of x-covariates, the model reduces to a nonparametric model. Altman (1990) and Hart (1991) studied a nonparametric model for time series error with $u_t = t/n$ (i.e., the fixed design cases). When $\{u_t\}$ resumes from random design, Xiao et al. (2003) proposed a prewhiting estimation procedure for the nonparametric model to deal with correlation among the errors. Li and Li (2009) proposed an estimation for nonparametric regression with AR error by using techniques related to partial linear models.

In this paper, we study the varying coefficient model with a stationary AR error process with order d. Thus, the model may be regarded as a semiparametric regression model. We propose a profile least squares estimation procedure based on local linear regression techniques. The profile least squares estimator can estimate the functional coefficients and autoregressive coefficients effectively. We establish asymptotic normality for the resulting estimator of the coefficient functions and autoregressive coefficients. The asymptotic bias and variance of the resulting estimates for the coefficient functions are the same as those of local linear estimate for the varying coefficient model with independent and identically distributed observations. We further extend the SCAD method (Fan and Li (2001)) to reduce model complexity of the AR process. Monte Carlo simulations are conducted to compare the proposed procedure with the one ignoring the error correlation, under different sampling schemes. The simulation results demonstrate that the newly proposed procedure significantly outperforms the one ignoring the error correlation in moderate samples.

The rest of this paper is organized as follows. In Section 2, we propose a new procedure to estimate the coefficients and to reduce model complexity of the AR order. Monte Carlo simulations and applications are presented in Section 3. Concluding remarks are presented in Section 4. Regularity conditions and technical proofs are given in an online supplemental appendix.

2. Semi-Varying Coefficient Model and Estimation Procedure

Assume that ε_t is an AR series

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

where $\{\eta_t\}$ is independent and identically distributed random error with mean zero and variance σ^2. Throughout, it is assumed that the covariate processes $\{(u_t, x_{t1}, \ldots, x_{tp}) : t \geq 1\}$ are independent of both error processes $\{\varepsilon_t : t \geq 1\}$ and $\{\eta_t : t \geq 1\}$. The order d for the AR error model is fixed, but may be large, and variable selection for the AR error will be discussed in next section. Thus, the model can be written as

$$y_t = \alpha_0(u_t) + \alpha_1(u_t)x_{t1} + \cdots + \alpha_p(u_t)x_{tp} + \beta_1\varepsilon_{t-1} + \cdots + \beta_d\varepsilon_{t-d} + \eta_t. \quad (2.1)$$

If the values for ε_t were available, then the coefficient functions $\alpha_j(\cdot)$'s and AR coefficients β_j's may be obtained by directly using existing estimation procedures for semiparametric varying-coefficient partially linear models. See, for example, Fan and Huang (2005) and Fan, Huang, and Li (2007). In practice, ε_t is not available but may be estimated by $\hat{\varepsilon}_t = y_t - \tilde{\alpha}_0(u_t) - \tilde{\alpha}_1(u_t)x_{t1} - \cdots - \tilde{\alpha}_p(u_t)x_{tp}$, where $\{\tilde{\alpha}_j(\cdot), j = 0, \ldots, p\}$ is obtained pretending the errors are independent. In this paper, we employ the local linear estimator $\tilde{\alpha}_j$ to estimate $\alpha_j(\cdot)$. Detailed implementation can be found in Fan and Zhang (1999).

For simplicity, write $\boldsymbol{\alpha}(u_t) = (\alpha_0(u_t), \cdots, \alpha_p(u_t))^T, \mathbf{X}_t = (1, x_{t1}, \ldots, x_{tp})^T$, and $\mathbf{e}_t = (\hat{\varepsilon}_{t-1}, \cdots, \hat{\varepsilon}_{t-d})^T$. Replacing ε_t's with $\hat{\varepsilon}_t$'s, model (2.1) becomes

$$y_t \approx \boldsymbol{\alpha}(u_t)^T \mathbf{X}_t + \mathbf{e}_t^T \boldsymbol{\beta} + \eta_t, \quad (2.2)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$. In Section 2.1, we propose an estimation procedure for $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}$ based on (2.2). We further propose a variable selection procedure for the AR series by using the penalized profile least squares method in Section 2.2.

2.1. An estimation procedure

There exist various estimation methods for (2.2). We use the profile least squares estimation procedure proposed by Fan and Huang (2005) to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}(\cdot)$.

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

$$y_t^* = \sum_{j=0}^{p} \alpha_j(u_t)x_{tj} + \eta_t \quad (2.3)$$

which is a varying coefficient model. We can employ local linear regression (Fan and Gijbels (1996)) to estimate $\{\alpha_j(\cdot), j = 0, \ldots, p\}$. Specifically for a given u_0, we locally approximate the coefficient function as

$$\alpha_j(u) \approx \alpha_j(u_0) + \alpha_j'(u_0)(u - u_0) \hat{=} a_j + b_j(u - u_0).$$

Let $K_h(u) = h^{-1}K(u/h)$ be a scaled kernel function of kernel $K(\cdot)$ with bandwidth h. Local linear regression is used to estimate the local parameter $\{(a_j, b_j), j = 0, \ldots, p\}$ via minimizing the weighted least squares function

$$\sum_{t=d+1}^{n}[y_t^* - \sum_{j=0}^{p}\{a_j - b_j(u_t - u_0)\}x_{tj}]^2 K_h(u_t - u_0),$$

with respect to $\{(a_j, b_j), j = 0, \ldots, p\}$. Denote the resulting estimate by $\{(\hat{a}_j, \hat{b}_j), j = 0, \ldots, p\}$. Then, $\hat{\alpha}_j(u_0) = \hat{a}_j$, and $\hat{\alpha}'_j(u_0) = \hat{b}_j$, for $j = 0, \ldots, p$.

It is clear that the local linear estimate of $\mathbf{M} = (\boldsymbol{\alpha}(u_{d+1})^T\mathbf{X}_{d+1}, \cdots, \boldsymbol{\alpha}(u_n)^T\mathbf{X}_n)$ is linear in terms of $\mathbf{y}^* = (y_{d+1}^*, \cdots, y_n^*)^T$. Let $\hat{\mathbf{M}}$ be the estimator of $\mathbf{M}$. Then it can be represented as

$$\hat{\mathbf{M}} = \mathbf{S}_h\mathbf{y}^*, \tag{2.4}$$

where $\mathbf{S}_h$ is a $(n-d)\times(n-d)$ smoothing matrix depending on $\{u_t, \mathbf{X}_t\}$ and the bandwidth only.

Let $\mathbf{E} = (\mathbf{e}_{d+1}, \ldots, \mathbf{e}_n)^T$ and $\boldsymbol{\eta} = (\eta_{d+1}, \ldots, \eta_n)^T$. Then (2.2) can be written in a matrix form

$$\mathbf{y} = \mathbf{M} + \mathbf{E}\boldsymbol{\beta} + \boldsymbol{\eta}.$$

Substituting for $\mathbf{M}(u_t, \mathbf{X}_t)$ by $\hat{\mathbf{M}}(u_t, \mathbf{X}_t)$, we obtain a synthetic linear regression model

$$(I - \mathbf{S}_h)\mathbf{y} \approx (I - \mathbf{S}_h)\mathbf{E}\boldsymbol{\beta} + \boldsymbol{\eta},$$

where I is the identity matrix. Thus, the profile least squares estimators for $\boldsymbol{\beta}$ and $\mathbf{M}$ are

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

$$\hat{\mathbf{M}} = \mathbf{S}_h(\mathbf{y} - \mathbf{E}\hat{\boldsymbol{\beta}}). \tag{2.6}$$

Let $\mu_l = \int u^l K(u)\,\mathrm{d}u$ and $\nu_l = \int u^l K^2(u)\,\mathrm{d}u$. Theorem 1 states the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ and the asymptotic bias and variance of $\hat{\alpha}_j(u_0), j = 0, \ldots, p$.

Theorem 1. *If Conditions A—H in the online supplemental appendix hold, the following statements are valid.*

(A) *If* $\mathbf{f}$ *has the same distribution as that of* $\mathbf{f}_t = (\varepsilon_{t-1}, \ldots, \varepsilon_{t-d})^T$ *and* $\sigma^2 = var(\eta_t)$,

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

(B) *If* $\hat{\alpha}_j(u_0, \hat{\boldsymbol{\beta}})$ *stands for* $\hat{\alpha}_j(u_0)$ *and* $g(u)$ *is the density function of* u,

$$\sqrt{nh}\{\hat{\alpha}_j(u_0, \hat{\boldsymbol{\beta}}) - \alpha_j(u_0) - \frac{1}{2}\mu_2\alpha''_j(u_0)h^2\} \to N(0, \frac{\nu_0\sigma^2}{g(u_0)}),\ j = 0, \ldots, p.$$

According to Fan and Huang (2005), $\sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}^{-1}$ is the semiparametric efficiency bound for general varying coefficient partially linear model. In addition, the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ 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)). Thus, Theorem 1 (A) implies that $\hat{\boldsymbol{\beta}}$ is as efficient as if the one knew the true functional coefficients $\alpha_j(\cdot)$'s in advance. Theorem 1 (B) indicates that $\hat{\alpha}_j(\cdot, \hat{\boldsymbol{\beta}})$ shares the same asymptotic bias and variance as those of the local linear regression for independent and identically distributed observations.

2.2. Issues related to practical implementation

One needs to specify two bandwidths, one for initial estimate and one for the profile least squares method, and has to specify the order of the AR error model. We discuss these issues in this subsection.

Bandwidths selection. Following Cai, Fan, and Li (2000), we employ multifold cross-validation to choose the bandwidths. We randomly partition the entire data set into J subsets d_j, $j = 1, \ldots, J$, and use $J-1$ data subsets to estimate the functional coefficients. Denote the resulting estimate without data set d_j by $\tilde{\boldsymbol{\alpha}}_{(-j)}(\cdot)$. The cross-validation score for bandwidth selection in the initial estimate is defined to be

$$CV_0(h) = \sum_{j=1}^{J}\sum_{t\in d_j}\{y_t - \mathbf{X}_t^T\tilde{\boldsymbol{\alpha}}_{(-j)}(u_t)\}^2.$$

We minimize $CV_0(h)$ over a set of grid points to choose the optimal bandwidth for the initial estimate. Similarly, we can define the cross-validation scores for the profile least squares method. Denote the resulting estimate based on data excluding d_j-data subset by $\hat{\boldsymbol{\alpha}}_{(-j)}(\cdot)$ and $\hat{\boldsymbol{\beta}}_{(-j)}$. The cross-validation score of bandwidth selection for the profile least squares estimate is defined to be

$$CV_1(h) = \sum_{j=1}^{J}\sum_{t\in d_j}\{y_t - \mathbf{X}_t^T\hat{\boldsymbol{\alpha}}_{(-j)}(u_t) - \mathbf{e}_t^T\hat{\boldsymbol{\beta}}_{(-j)}\}^2.$$

We minimize this cross-validation score over a set of grid points to choose the optimal bandwidth for the profile least squares estimate. It is typical to set $J = 5$ or 10 in practice.

Variable selection for the AR error model. Regarding model (1.1), we can start from a large order AR model and establish an algorithm to reduce the model complexity. Motivated by the variable selection mechanism in linear regression, we add a penalty term to the squared loss function as:

$$\frac{1}{2}\sum_{t=d+1}^{n}\{y_t-\alpha_0(u_t)-\alpha_1(u_t)x_{t1}-\cdots-\alpha_p(u_t)x_{tp}-\mathbf{e}_t^T\boldsymbol{\beta}\}^2+n\sum_{j=1}^{d}p_{\lambda_j}(|\beta_j|), \quad (2.7)$$

where $p_{\lambda_j}(\cdot)$ is a penalty function with tuning parameter λ_j controlling the model complexity. The tuning parameter can be selected by a data driven method. The choice of λ_j will be discussed later.

With a proper choice of penalty function and λ_j, we expect to get some exact zero estimates by minimizing (2.7) with respect to $\boldsymbol{\beta}$. This is equivalent to removing the corresponding terms from the original model. However, it is challenging to minimize (2.7) directly because the functional coefficient $\alpha_l(\cdot)$ has not been parameterized. Using the profile technique as introduced in the previous section, we can substitute the functional part by a linear form of $\boldsymbol{\beta}$ and get the profile least squared loss function

$$\frac{1}{2}(\mathbf{y}-\mathbf{E}^T\boldsymbol{\beta})^T(I-\mathbf{S}_h)^T(I-\mathbf{S}_h)(\mathbf{y}-\mathbf{E}^T\boldsymbol{\beta})+n\sum_{j=1}^{d}p_{\lambda_j}(|\beta_j|). \quad (2.8)$$

There are various choices for the penalty function $p_{\lambda_j}(\cdot)$. Fan and Li (2001) provided insights into this choice and advocated a penalty which can (a) automatically force the estimators of nonsignificant β_j to zero, (b) keep the estimators of large β_j unbiased, and (c) make the resulting estimate of regression coefficients continuous in some sense. Such commonly used penalty functions as the family of L_q penalties ($q \geq 0$) do not have these properties. Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty that does, and we use it here. The derivative of the SCAD penalty is

$$p'_\lambda(\beta)=\lambda\{I(\beta\leq\lambda)+\frac{(a\lambda-\beta)_+}{(a-1)\lambda}I(\beta>\lambda)\}$$

for $\beta > 0$, with $a = 3.7$ as suggested by Fan and Li (2001). We refer the penalized profile least squares with the SCAD penalty as the SCAD procedure for simplicity.

Algorithm. The minimization of the SCAD penalized profile least squares is not easy because the objective function is irregular at the origin and does not have a second derivative at some points. To proceed, we take the local quadratic approximation to the SCAD penalty function suggested by Fan and Li (2001). Suppose we can get an estimate $\beta_j^{(k)}$ in the k^{th} step iteration that is close to the true β_j. If $|\beta_j^{(k)}|$ is close to 0, then we set $\hat{\beta}_j = 0$. Otherwise, the SCAD penalty is locally approximated by a quadratic function as

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

We can employ Newton-Raphson algorithm to minimize (2.8). In practice, we use an iterative ridge regression to find the minimizer of (2.8),

$$\boldsymbol{\beta}^{(k+1)} = \{\mathbf{E}^T(I-\mathbf{S}_h)^T(I-\mathbf{S}_h)\mathbf{E} + n\Sigma_\lambda(\boldsymbol{\beta}^{(k)})\}^{-1}\mathbf{E}^T(I-\mathbf{S}_h)^T(I-\mathbf{S}_h)\mathbf{y}, \quad (2.9)$$

where $\Sigma_\lambda(\boldsymbol{\beta}^{(k)}) = \mathrm{diag}\{p'_{\lambda_1}(|\beta_1^{(k)}|)/|\beta_1^{(k)}|, \ldots, p'_{\lambda_d}(|\beta_d^{(k)}|)/|\beta_d^{(k)}|\}$ for nonvanished $\boldsymbol{\beta}^{(k)}$. The unpenalized profile least squares estimator is taken as the initial value to update $\boldsymbol{\beta}^{(1)}$.

Tuning parameter selection. The other important issue in implementation is to select the tuning parameter λ_j. The minimization of (2.8) with respect to $(\lambda_1, \ldots, \lambda_d)$ is a challenging high dimensional optimization problem, but the magnitude of λ_j is believed to be proportional to the standard error of the estimate of β_j. Following Fan and Li (2004), we set $\lambda_j = \lambda\,\mathrm{se}(\hat{\beta}_j)$ where $\mathrm{se}(\hat{\beta}_j)$ is the standard error of the unpenalized least squares estimates. To this end, the original d-dimensional optimization reduces to a 1-dimensional problem. We can minimize the BIC or GCV score to find the optimal λ. Here we use the BIC selector.

Define the effective number of parameters of the penalized least squares estimator (2.9) to be

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

where $\tilde{D} = \mathbf{E}^T(I-\mathbf{S}_h)^T(I-\mathbf{S}_h)\mathbf{E}$ for nozero $\hat{\boldsymbol{\beta}}$.

Wang, Li, and Tsai (2007) advocated using the BIC tuning parameter selector for linear regression, and showed that it yields an oracle estimate in an asymptotic sense. Thus, we use the BIC tuning parameter selector to select λ. The BIC score can be written as

$$BIC(\lambda) = \log\Big(\frac{RSS}{n}\Big) + e(\lambda)\frac{\log n}{n},$$

where $RSS = \|(I-\mathbf{S}_h)\mathbf{y} - (I-\mathbf{S}_h)\mathbf{E}\hat{\boldsymbol{\beta}}\|^2$ is the residual sum of squares given λ. The primary goal of the SCAD procedure is to reduce the model complexity of the AR error with large order d, and is slightly different from the order selection of the AR error process. It is of interest to extend the SCAD procedure for the order selection of the AR error model while taking additional prior information for the AR model into account. The SCAD procedure also involves a bandwidth in the profile least squares method. It is typical to take the bandwidth for the profile least squares without a penalty for the SCAD procedure (Fan and Li (2004)).

3. Simulation 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 the local linear estimator without considering the error structure. We also illustrate the proposed procedure by an empirical analysis of a data example. All numerical studies were conducted by Matlab code.

3.1. Simulation studies

A random sample of size n, $n = 250$ or 500, was generated from

$$y_t = \alpha_0(u_t) + \alpha_1(u_t)x_{t1} + \alpha_2(u_t)x_{t2} + \varepsilon_t,$$

where $\alpha_0(u) = 3u^2 - 2u + 1, \alpha_1(u) = \cos(2\pi u), \alpha_2(u) = 2\sin(2\pi u)$ were the functional coefficients with the same degree of smoothness, and $\{u_t, x_{t1}, x_{t2}\}$ were covariate with distribution that will be spelled out in two ways. The error process ε_t is an AR process of order $d = 10$, 20,

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

with $\eta_t \sim N(0, \sigma^2)$ and $\sigma = 0.5$ or 1. We considered two situations: first an AR model with $\beta_1 = 0.5$ or 0.7, and all other β_j's zero; the second is another AR model with $\beta_1 = 0.5$, $\beta_2 = 0.4$ or $\beta_1 = 0.7$, $\beta_2 = 0.2$, and all others zero. Although the true model for the error process is either AR(1) or AR(2), the profile least squares procedure and the SCAD procedure used AR(d) with $d = 10$ or 20 in our simulations. The number of replications for each case was 500.

To understand how the sampling scheme of covariates affects the proposed procedure, we considered two sampling schemes in our simulation.

I. $\{u_t\}$ is i.i.d. uniform on $[0, 1]$. $\{x_{t1}, x_{t2}\}$ is a multivariate normal distribution with mean vector $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and covariance matrix $\begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$.

II. $\{v_t\}$ is i.i.d. standard normal for $t = 1, \ldots, n$. We took $u_t = \Phi\{(av_t + bv_{t-1})/\sqrt{a^2 + b^2}\}$ for $t = 2, 3, \ldots, n + 1$, with $\Phi(v)$ the distribution function of the standard normal. Thus, $\{u_t\}$ was a 1-dependent process. $\{x_{t1}, x_{t2}\}$ was generated as a multidimensional 1-dependent process as well. We took $\{z_{t1}, z_{t2}\}$ to be multivariate normal with mean vector $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and covariance matrix $\begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$. $\{x_{t1}, x_{t2}\} = c\{z_{t1}, z_{t2}\} + d\{z_{t-1,1}, z_{t-1,2}\}$ with $c^2 + d^2 = 1$ for $t = 2, 3, \ldots, n + 1$. In our simulation, we took $a = 0.9, b = 0.1, c = 0.8$, and $d = 0.6$.

The scheme to generate $\{u_t\}$ has been used in Li and Li (2009). For each sampling scheme, profile least squares and penalized profile least squares with SCAD penalty were compared with the oracle estimator obtained by substituting

Table 1. Simulation results under sampling scheme I. $(1 - \text{RMSE}) * 100\%$ is summarized for comparison. 'Profile' stands for profile least squares method, 'SCAD' for penalized profile least squares method with SCAD penalty, and 'Oracle' for the oracle estimator that uses the true order of the AR error, and h_1 and h_2 are the bandwidths in the initial estimation and profile least squares estimation, respectively.

(β_1, β_2)	σ	h_1	h_2	Profile	SCAD	Oracle	h_1	h_2	Profile	SCAD	Oracle
$d = 10$				$n = 250$					$n = 500$		
(0.5,0)	0.5	0.104	0.104	10.922	11.976	14.175	0.104	0.086	23.802	24.333	26.054
(0.7,0)	0.5	0.124	0.124	23.720	24.111	26.501	0.104	0.104	30.132	30.149	30.717
(0.5,0.4)	0.5	0.104	0.104	12.507	13.664	16.500	0.104	0.086	23.976	23.562	22.870
(0.7,0.2)	0.5	0.124	0.124	19.328	20.660	21.993	0.086	0.086	37.554	37.655	37.970
(0.5,0)	1	0.215	0.149	18.618	20.744	23.518	0.124	0.149	20.594	21.239	22.115
(0.7,0)	1	0.310	0.215	50.444	52.022	52.524	0.149	0.149	37.105	37.352	38.354
(0.5,0.4)	1	0.215	0.179	21.684	24.762	26.534	0.149	0.149	20.030	21.263	20.897
(0.7,0.2)	1	0.215	0.179	32.005	32.895	36.098	0.179	0.124	39.004	39.975	41.305
$d = 20$				$n = 250$					$n = 500$		
(0.5,0)	0.5	0.104	0.104	5.285	11.093	16.503	0.086	0.086	10.748	12.785	16.601
(0.7,0)	0.5	0.124	0.124	16.584	19.664	24.934	0.086	0.104	24.338	25.641	37.084
(0.5,0.4)	0.5	0.104	0.104	8.055	11.497	17.961	0.104	0.072	14.379	16.796	20.822
(0.7,0.2)	0.5	0.124	0.104	25.212	26.430	32.535	0.086	0.086	33.411	33.698	36.879
(0.5,0)	1	0.179	0.179	6.059	10.444	17.485	0.124	0.149	14.212	16.033	19.289
(0.7,0)	1	0.179	0.149	25.267	29.270	36.821	0.149	0.149	37.071	38.894	40.808
(0.5,0.4)	1	0.215	0.179	17.228	22.528	26.835	0.124	0.149	22.897	24.895	25.496
(0.7,0.2)	1	0.179	0.179	28.700	31.454	34.479	0.149	0.149	35.839	37.454	38.123

the true autoregressive order. The oracle estimator was included as a benchmark. We compared their performances according to

$$\text{MSE}\{\boldsymbol{\alpha}(\cdot)\} = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=0}^{p}\{\hat{\alpha}_j(u_t) - \alpha_j(u_t)\}^2.$$

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 $\tilde{\boldsymbol{\alpha}}(\cdot)$, the estimate of $\boldsymbol{\alpha}(\cdot)$, pretending the error ε_t independent. We report the percentage of accuracy gain, defined by $(1 - \text{RMSE}) * 100\%$.

The multi-fold cross-validation method for bandwidth selection was time-consuming in simulation studies as we had to repeat each case 500 times. To reduce computational burden, we determined the bandwidths for each case as follows. For a given bandwidth, take

$$\text{MSE}(h) = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=0}^{p}\{\hat{\alpha}_j(u_t) - \alpha_j(u_t)\}^2,$$

where $\hat{\alpha}_j(\cdot)$ is the local linear estimator or the profile least squares estimator respectively. In our simulation, we set the bandwidth that minimized $\text{MSE}(h)$ in a pilot study. Tables 1 and 2 depict the median of percentages of accuracy gain,

Table 2. Simulation results under sampling scheme II. $(1 - \text{RMSE}) * 100\%$ is summarized for comparison. Caption of this table is the same as that for Table 1.

(β_1, β_2)	σ	h_1	h_2	Profile	SCAD	Oracle	h_1	h_2	Profile	SCAD	Oracle
$d = 10$				$n = 250$					$n = 500$		
(0.5,0)	0.5	0.100	0.100	8.409	11.442	16.385	0.075	0.075	16.096	16.412	17.925
(0.7,0)	0.5	0.100	0.100	28.352	30.144	34.264	0.100	0.100	28.735	28.315	29.105
(0.5,0.4)	0.5	0.100	0.100	14.861	16.234	19.186	0.100	0.075	23.019	24.162	24.501
(0.7,0.2)	0.5	0.100	0.100	33.225	34.287	36.634	0.100	0.100	28.995	29.493	36.981
(0.5,0)	1	0.175	0.175	10.061	13.215	17.335	0.150	0.125	15.697	17.197	18.660
(0.7,0)	1	0.175	0.175	30.141	32.737	35.722	0.150	0.125	35.101	37.152	37.399
(0.5,0.4)	1	0.175	0.150	17.940	18.952	21.047	0.150	0.150	19.560	19.657	20.903
(0.7,0.2)	1	0.200	0.175	35.434	37.394	39.523	0.150	0.125	36.598	37.240	38.974
$d = 20$				$n = 250$					$n = 500$		
(0.5,0)	0.5	0.100	0.100	3.126	7.624	16.385	0.075	0.100	5.160	6.745	8.315
(0.7,0)	0.5	0.100	0.100	22.804	27.181	34.264	0.100	0.075	32.334	32.230	29.105
(0.5,0.4)	0.5	0.100	0.100	8.200	11.421	19.186	0.100	0.075	21.513	22.645	24.501
(0.7,0.2)	0.5	0.100	0.100	27.940	29.979	36.634	0.100	0.075	33.900	35.655	36.981
(0.5,0)	1	0.175	0.175	3.324	9.963	17.335	0.150	0.125	13.030	16.125	18.660
(0.7,0)	1	0.175	0.175	25.994	30.771	35.722	0.150	0.125	32.003	35.008	37.399
(0.5,0.4)	1	0.175	0.150	12.325	15.752	21.047	0.150	0.150	18.423	19.379	20.903
(0.7,0.2)	1	0.200	0.175	30.464	33.339	39.523	0.150	0.150	34.745	36.106	38.974

defined by $(1-\text{RMSE})*100\%$, over the 500 simulations. In these tables, 'Profile' stands for profile least squares method, 'SCAD' for penalized profile least squares method with SCAD penalty, and 'Oracle' for the oracle estimator that uses the true order of the AR error process, and h_1 and h_2 are the bandwidths in the initial estimation and profile least squares estimation, respectively.

Simulation results for sampling scheme I are summarized in Table 1. Under this sampling scheme, the covariates $\{u_t\}$ and $\{x_{t1}, x_{t2}\}$ were independent. Our proposed methods improve the estimation accuracy, especially when the correlation is strong. The SCAD procedure always outperforms the profile least squares method. This superiority is more significant when the sample size is small. On the other hand, when $n = 250$ and $\sigma = 1$, the profile least squares approach at $d = 10$ performs better than that at $d = 20$. This implies that variable selection for the AR error may be a necessary step when d is large.

When the sample size is large, such as $n = 500$, both the profile least squares method and the SCAD method have larger gain over the working independence procedure. In addition, their performances are close to each other and also close to the oracle estimator. This is expected because the estimate should be more accurate as the sample size increases. The difference between $d = 10$ and $d = 20$ is not remarkable. This implies that the proposed procedure is not very sensitive to the assumption of the AR order provided a variable selection for the AR error is conducted.

For sampling II, both $\{u_t\}$ and $\{x_{t1}, x_{t2}\}$ are 1-dependent processes. However, the overall pattern of Table 2 is very similar to that in Table 1. The SCAD

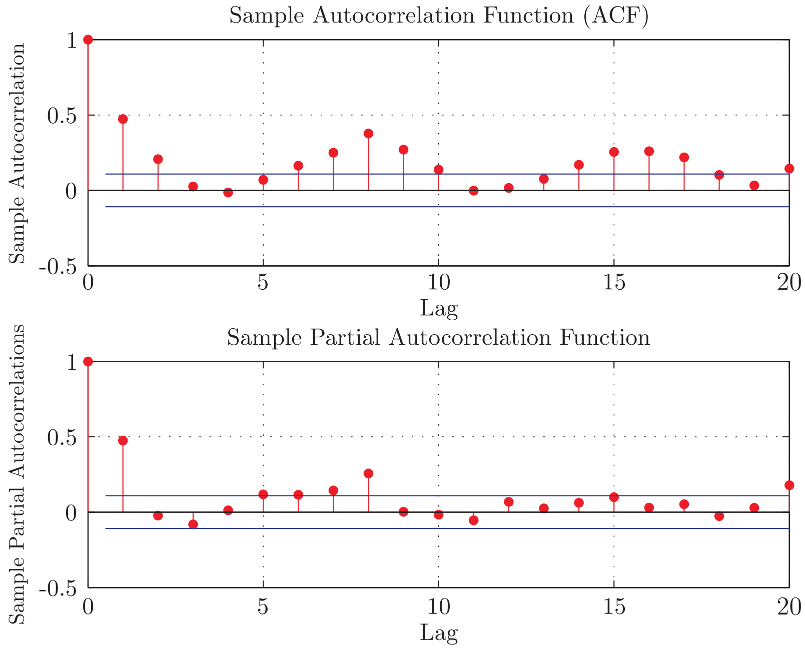


Figure 1. Correlogram of residual $\hat{\varepsilon}_t$. Plots are the autocorrelation and partial autocorrelation for $\hat{\varepsilon}_t$. In each plot, the upper and lower dashed lines represent 95% confidence intervals.

method has a better performance than the profile least squares method for all cases, and is close to the oracle estimator, especially when the sample size is large. We can conclude that our proposed estimation methods work well with either independent or 1-dependent errors.

3.2. An application

We illustrate the proposed methodology with an empirical analysis of a data set collected from the website of Pennsylvania-New Jersey-Maryland Interconnections (PJM), the largest regional transmission organization (RTO) in the U.S. electricity market. The data set includes 340 daily observations of electricity price, electricity load, and prices of oil, natural gas and coal in the Pennsylvania electric (PENELEC) district. It is of interest to study the relationship between the electricity price and electricity load, and prices of oil, natural gas, and coal. In this illustration, we take electricity price as the response variable y_t, the normalized electricity load as u_t, and prices of oil, natural gas, and coal as covariates x_{1t}, x_{2t}, and x_{3t}, respectively. We fit the data using the model

$$y_t = \alpha_0(u_t) + \alpha_1(u_t)x_{1t} + \alpha_2(u_t)x_{2t} + \alpha_3(u_t)x_{3t} + \varepsilon_t.$$

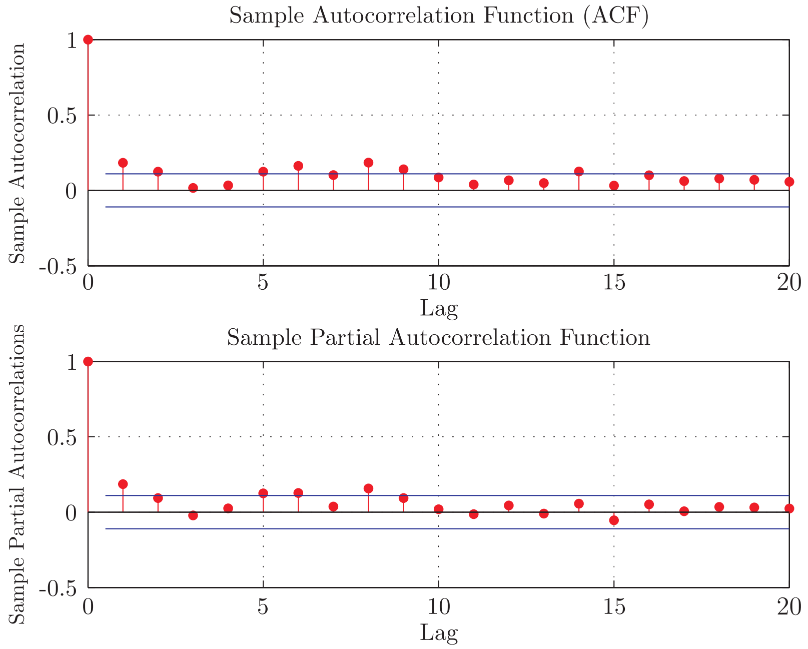


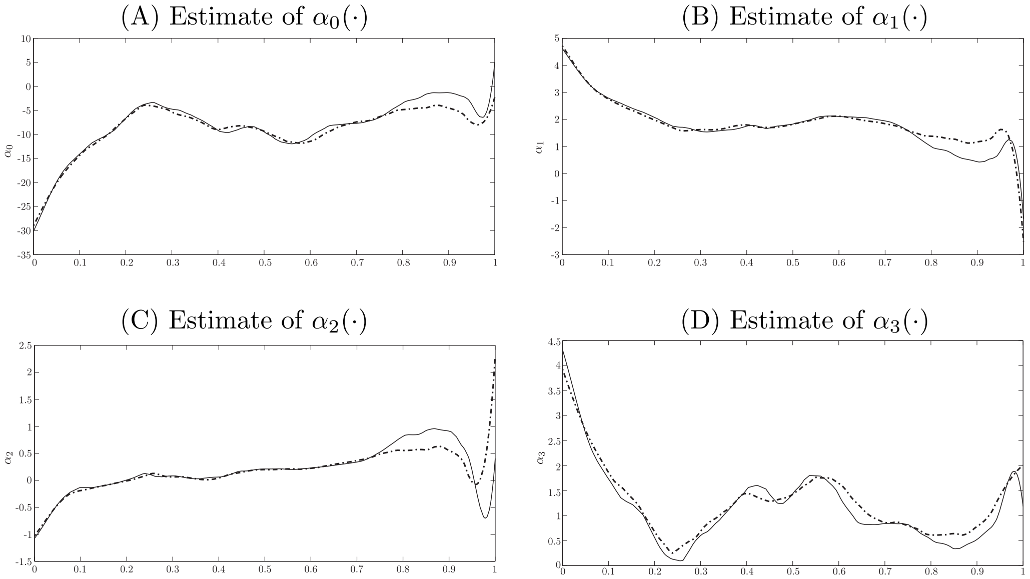
Figure 2. Correlogram of residual $\hat{\eta}_t$. Plots are the autocorrelation and partial autocorrelation for $\hat{\eta}_t$. In each plot, the upper and lower dashed lines represent 95% confidence intervals.

Initial estimate $\tilde{\alpha}_j(\cdot)$. Local linear regression was used to obtain the initial estimate of $\alpha_j(\cdot)$. The correlation structure of errors was ignored in the initial estimate. The 10-fold cross-validation method was used to select a bandwidth. The selected bandwidth for the initial estimate was 0.2089, which minimized $CV_0(h)$, defined in Section 2.2.

Residual analysis. Based on the initial estimate, we further calculated its residuals. The autocorrelation plot of the residuals is depicted in Figure 1, which shows that there exists a periodic structure in $\{\hat{\varepsilon}_t\}$. The partial-autocorrelation plot is displayed in Figure 1, which suggests that an AR(d) with $d \leq 10$ may fit the errors well.

We now applied the proposed estimation procedure for the data. The bandwidth was selected by another 10-fold cross-validation, and the selected bandwidth $h = 0.1899$ at which the $CV_1(h)$ defined in Section 2.2 reached its minimum.

With the selected bandwidth, we applied penalized profile least squares with SCAD penalty to select the AR order and estimated $\alpha_j(\cdot)$ and $\boldsymbol{\beta}$. The BIC tuning parameter selector for the SCAD had $\lambda = 0.0290$. AR coefficients at lag 1 and 9 were significant, the AR(9) model was selected. After taking the autocorrelation


(A) Estimate of $\alpha_0(\cdot)$

(B) Estimate of $\alpha_1(\cdot)$

(C) Estimate of $\alpha_2(\cdot)$

(D) Estimate of $\alpha_3(\cdot)$

Figure 3. Plot of $\hat{\alpha}_j(\cdot)$'s in model (3.1). Dashed curves are the initial estimates; Solid curves are the penalized profile least squares estimate.

into account, the autocorrelation and partial-autocorrelation plots, displayed in Figure 2, confirm that the residuals $\hat{\eta}_t$ look like a white noise process.

Final model. We came to the final model

$$\hat{y}_t = \hat{\alpha}_0(u_t) + \hat{\alpha}_1(u_t)x_{t1} + \hat{\alpha}_2(u_t)x_{t2} + \hat{\alpha}_3(u_t)x_{t3} + 0.0851\hat{\varepsilon}_{t-1} + 0.0139\hat{\varepsilon}_{t-9}, \quad (3.1)$$

where $\hat{\alpha}_j$, $j = 0, 1, 2, 3$, as depicted in Figure 3. Compared with the initial estimate, the estimate $\hat{\alpha}_j(\cdot)$ is smoother when the correlation of errors is accounted.

4. Discussions

In this paper, we proposed a new estimation procedure for the varying coefficient model with AR error by using profile least squares techniques. We further proposed to select the order of the AR 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. Monte Carlo simulation studies showed that our proposed method can effectively improve estimation accuracy under different sampling schemes. Finally, we applied the penalized profile least squares method in a data example.

As a topic for future research, one can consider a more general AR structured error in the varying-coefficient setting. That is to assume the autoregressive coefficients are smoothing functions depending on another covariate rather than constant coefficients. The estimation and inference of such an extended model needs more systematic studies.

Acknowledgements

The authors thank the Editor, the AE and reviewers for their constructive comments, which have led to a dramatic improvement of the earlier version of this paper. All authors equally contribute to this paper, and the authors are listed in the alphabetic order. Runze Li is the corresponding author. This research was supported by NIDA grants P50-DA10075 and P50 DA036107, a NCI grant R01 CA168676, and National Natural Science Foundation of China grants 11028103 and 10911120395. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA, NCI, or the NIH.

References

Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc.* **85**, 749-759.

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

Cai, Z. (2007). Trending time-varying coefficient time series models with serially correlated errors. *J. Econom.* **136**, 163-188.

Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.* **95**, 888-902.

Cai, Z., Yao, Q. and Zhang, W. (2001). Smoothing for discrete-values time series. *J. Roy. Statist. Soc. Ser. B* **63**, 357-375.

Cheng, M. Y., Zhang, W. and Chen, L. (2009). Statistical estimation in generalized multiparameter likelihood models. *J. Amer. Statist. Assoc.* **104**, 1179-1191.

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

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11**, 1031-1059.

Fan, J. Huang, T. and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *J. Amer. Statist. Assoc.* **102**, 632-641.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Amer. Statist. Assoc.* **99**, 710-723.

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

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27**, 1491-1518.

Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statist. and Its Interface* **1**, 179-195.

Hart, J. D. (1991). Kernel regression estimation with time series errors. *J. Roy. Statist. Soc. Ser. B* **53**, 173-187.

Hastie, T. and Tibshirani, R (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55**, 757-796.

Huang, J. and Shen, H. (2004). Functional coefficient regression models for non-linear time series: A polynomial spline approach. *Scand. J. Statist.* **31**, 515-534.

Li, J. and Zhang, W. (2011). A semiparametric threshold model for censored longitudinal data analysis. *J. Amer. Statist. Assoc.* **106**, 685-696.

Li, R. and Li, Y. (2009). Local linear regression for data with AR errors. Special Issue of *Acta Mathematicae Applicatae Sinica* (English Series) **25**, 427-444.

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

Xiao, Z., Linton, O., Carroll, R. J. and Mammen, E. (2003) More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *J. Amer. Statist. Assoc.* **98**, 980-992.

Department of Operation Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA.

E-mail: chenzhao1985@gmail.com

Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, Pennsylvania, 16802-2111, USA

E-mail: rzli@psu.edu

eBay Inc, 618 East Yan'an Road, Shanghai, 200001, China.

E-mail: Maggie.Li.Stat@gmail.com

(Received October 2012; accepted April 2014)