Significant Variable Selection and Autoregressive Order Determination for Time Series Partially Linear Models*

Degao Li[1,2], Guodong Li[3] and Jinhong You[1,4]

[1]School of Statistics and Management, Shanghai University of Finance and Economics

[2]College of Mathematics Physics and Information Engineering, Jiaxing University

[3]Department of Statistics and Actuarial Science, University of Hong Kong

[4]Key Laboratory of Mathematical Economics (SUFE), Ministry of Education of China

Abstract: This paper is concerned with the regression coefficient and autoregressive order shrinkage and selection via the smoothly clipped absolute deviation (SCAD) penalty for a partially linear model with time series errors. By combining the profile semiparametric least squares method and SCAD penalty technique, a new penalized estimation for the regression and autoregressive parameters in the model is proposed. We show the asymptotic property of the resultant estimator is the same as if the order of autoregressive error structure and nonzero regression coefficients be known in advance, thus achieving the oracle property in the sense of Fan and Li (2001). In addition, based on a pre-whitening technique, we construct a two-stage local linear estimator (TSLLE) for the nonparametric component. It is shown that the TSLLE is more asymtotically efficient than the one which ignores the autoregressive time series error structure. Some simulation studies are conducted to illustrate the finite sample performance of the proposed procedure. An example of application on electricity usage data is also illustrated.

AMS 2000 classifications: primary 62H12; secondary 62A01.

Key words and phrases: Partially linear; Autoregressive error; Consistency; Asymptotic normality.

1 Introduction

It is well known that parametric models are useful tools to analyze the practical data. However, in order to avoid to suffer from large modeling biases one needs to specify the forms of parametric models correctly. With the increasing of complication of studied problems and dimensionality of data it is not easy to specify the forms of parametric models correctly. As an alternative, nonparametric smoothing can ease the concerns on modeling biases. However, the "curse of dimensionality" of the nonparametric

*Corresponding author: Jinhong You. Jinhong You's research was supported by grants from the National Natural Science Foundation of China (NSFC) (No. 11071154). The work was also partially supported by the Leading Academic Discipline Program and 211 Project for Shanghai University of Finance and Economics, and the Program for Changjiang Scholars and Innovative Research Team in University (IRT13077). Degao Li's research was supported by both grants from the Graduate Innovation Fund of Shanghai University of Finance and Economics(CXJJ-2011-426) and foundation of Zhejiang education department (Y201330003).

method hampers its wide use. The readers could refer Härdle (1990) and Fan and Gijbels (1996) for the details. One of methods for attenuating these disadvantages is to model covariate effects via a partially linear structure, a combination of linear and nonparametric parts in which the relationship between the response and some explanatory variables are linear and the relationship between the response and some explanatory variables are unspecified. This results in the famous partially linear models proposed by Engle *et al.* (1986). Generally, a partially linear model can be written as

$$y_t = \sum_{i=1}^{p} x_{ti}\beta_i + g(z_t) + \varepsilon_t, \quad t = 1, \cdots, n \tag{1.1}$$

where y_t is the response, both $\mathbf{X}_t = (x_{t1}, \ldots, x_{tp})^T$ and z_t are possibly vector-valued explanatory variables, ε_t is a random error independent of $(\mathbf{X}_t^T, z_t^T)^T$ with $E(\varepsilon_t) = 0$ and $\text{Var}(\varepsilon_t) = \sigma_\varepsilon^2$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is an unknown parameter vector having the same dimension as $\mathbf{X}_t$, $g(\cdot)$ is an unknown smooth functions, and the superscript $(_T)$ denotes the transpose of a vector or matrix.

After the model (1.1) was introduced, much effort has been made to study its corresponding estimation and statistical inference. The work in this line include, for example, Engle *et al.* (1986), Speckman (1988), Robinson (1988), Chen (1988), Chen and Shiau (1991, 1994), Gao (1995), You and Chen (2007), Aneiros-Pérez and Vilar-Fernández (2008). To mention just a few. More references could be found in the monographs such as Härdle, Liang and Gao (2000), Carroll, Ruppert and Wand (2003), Yatchew (2003), Gao (2007), Horowitz (2009) and so on.

In addition, in many problems, when each observation is recorded at specified time, strong serial correlation might arise. An example is the electricity usage data which consist of quarterly observations in Ontario for the period 1971 to 1994 (Yatchew 2003). According to Brockwell and Davis (1991) both MA and ARMA processes can be well approximated by an AR process. Therefore, an AR process is usually implemented to fit the serial correlation error structure. Until now several authors have allowed ε_t to follow an AR process

$$\varepsilon_t = \gamma_1\varepsilon_{t-1} + \cdots + \gamma_t\varepsilon_{t-q} + \eta_t, \quad t = q+1, \cdots, n \tag{1.2}$$

with $\{\eta_t\}_{t=q+1}^n$ being independent and identically distributed (i.i.d.) random errors with $E\eta_t = 0$ and $Var(\eta_t) = \sigma_\eta^2$. They include Engle *et al.* (1986), Gao (2007), You and Chen (2007), Aneiros-Pérez and Vilar-Fernández (2008), Risa and Takayuki (2009). All these authors make an assumption that the lagged order of autoregressive error structure is known. This is not true in practical problems and misspecification of the lagged order will result in the reduction of estimation efficient and prediction precision. Therefore, how to correctly determine the lagged order of autoregressive error structure deserves our investigation. Same as other modeling methods, in the absence of prior knowledge, a large number of variables may be included in model (1.1) to reduce possible model bias. This may lead to a complicated model which includes many insignificant explanatory variables, resulting in less predictive powers and difficulty in interpretation. As a result, variable selection is essential to statistical modeling. In last decade, alternative variable selection procedures have been proposed, which can simultaneously shrink regression coefficients and set minor coefficients to zero, such as Bridge regression (Frank and Friedman, 1993), LASSO (Tibshirani,1996, Fu,1998), SCAD (Fan and Li, 2001) and MCP (Zhang, 2010) and others.

The major contribution of this paper is that we utilize a unified shrinkage technique to select the significant explanatory variables in the parametric part and determine the order of lagged terms. The asymptotic property of the resulting estimator is the same as if the order of lagged terms and nonzero regression coefficients and functions be known in advance, thus achieving the oracle property in the sense of Fan and Li (2001). Our method is an extension of Wang, Li, and Tsai (2007) who focused on the parametric models. Although such an extension is straightforward, the development of statistical properties for the resulting estimators is not trivial, because when we transfer the model (1.1) to a linear model using the profile principle, such a profile procedure leads that the synthetic response and explanatory variables. Therefore, new theoretical tools are required to derive the large sample properties for the penalized profile semiparametric least squares estimators.

In addition, based on a pre-whitening technique and the estimated parametric component, we construct a two-stage local linear estimator (TSLLE) for the nonparametric component. It is shown that the TSLLE is more asymptotically efficient than the one which ignores the autoregressive time series error structure. The pre-whitening technique is used by Xiao, *et al.* (2003) and Liu, Chen and Yao (2009). However, they just investigate the estimating problem of purely nonparametric models.

The rest of this paper is organized in the following way. In Section 2 we describe the profile semiparametric least squares estimation with the SCAD penalty. In Section 3, we establish the selection consistency and the oracle property of the resultant estimator. In Section 4, we develop a pre-whitening local linear estimator for the nonparametric component. An algorithm is presented in Section 5. We demonstrate the performance of the proposed methodology through comprehensive simulations and an application to the electricity sales data in Section 6 and 7, respectively. The paper is concluded by Section 8. All proofs of main results are relegated to the supplementary material.

2 Profile Least Squares and SCAD Penalty Estimation

For notational simplicity, throughout this paper we assume that z_t is scalar. According to Fan and Gijbels (1996), compared with the traditional kernel methods the local polynomial approach can automatically correct boundary effect. Threfore, we apply the local linear smoothing to $g(z_t)$. On the other hand, the profile least squares technique is a useful approach in semiparametric models (Speckman, 1988, Carroll *et al.*, 1997, Murphy *et al.*, 2000, Fan and Huang, 2005, Lam and Fan, 2008). Same as the authors mentioned above, we also implement the profile least squares technique to estimate regression coefficients in the model (1.1).

For given z, we can locally approximate $g(z_t)$ with a linear function

$$g(z_t) \approx g(z) + g'(z)(z_t - z) \equiv a + b\frac{z_t - z}{h}. \tag{2.1}$$

Then the estimator of $a = g(z)$ and $b = hg'(z)$ can be obtained by minimizing the following least square type objective function

$$\sum_{t=1}^{n}\left[y_t - \sum_{i=1}^{p} x_{ti}\beta_i - (a + b\frac{z_t - z}{h})\right]^2 K_h(z_t - z), \tag{2.2}$$

where $K(\cdot)$ is a kernel function, $K_h(z_t - z) = K((z_t - z)/h)/h$ and h is a bandwidth.

Denote $\mathbf{Y} = (y_1, \cdots, y_n)^T$, $\mathbf{Z} = (z_1, \cdots, z_n)^T$, $\mathbf{X}_t = (x_{t1}, \cdots, x_{tp})^T$, $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)^T$, $\mathbf{G} = (g(z_1), \cdots, g(z_n))^T$, $\boldsymbol{\varepsilon}_t = (\varepsilon_1, \cdots, \varepsilon_n)^T$. Then, (1.1) can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G} + \boldsymbol{\varepsilon}. \tag{2.3}$$

As a result, an estimator of a and b from (2.2) can be expressed as

$$(\widehat{a}, \widehat{b})^T = (\mathbf{D}_z^T \mathbf{W}_z \mathbf{D}_z)^{-1}\mathbf{D}_z^T \mathbf{W}_z(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}),$$

where $\mathbf{D}_z = \begin{pmatrix} 1 & \cdots & 1 \\ \frac{z_1 - z}{h} & \cdots & \frac{z_n - z}{h} \end{pmatrix}^T$ and $\mathbf{W}_z = \mathrm{diag}(K_h(z_1 - z), \cdots, K_h(z_n - z))$. For given $\boldsymbol{\beta}$, it holds that

$$\widehat{g}(z) = (1, 0)(\mathbf{D}_z^T \mathbf{W}_z \mathbf{D}_z)^{-1}\mathbf{D}_z^T \mathbf{W}_z(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Let

$$\mathbf{S} = \begin{pmatrix} (1,0)(\mathbf{D}_{z_1}^T \mathbf{W}_{z_1}\mathbf{D}_{z_1})^{-1}\mathbf{D}_{z_1}^T \mathbf{W}_{z_1} \\ \vdots \\ (1,0)(\mathbf{D}_{z_n}^T \mathbf{W}_{z_n}\mathbf{D}_{z_n})^{-1}\mathbf{D}_{z_n}^T \mathbf{W}_{z_n} \end{pmatrix} = \begin{pmatrix} \mathbf{S}_1 \\ \vdots \\ \mathbf{S}_n \end{pmatrix}.$$

Substituting $\widehat{g}(z)$ into (2.3), we obtain

$$\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \approx \mathbf{S}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\varepsilon}. \tag{2.4}$$

Combining the error structure model (1.2), it holds that

$$y_t - \mathbf{S}_t\mathbf{Y} = (\mathbf{X}_t^T - \mathbf{S}_t\mathbf{X})\boldsymbol{\beta} + \varepsilon_t = (\mathbf{X}_t^T - \mathbf{S}_t\mathbf{X})\boldsymbol{\beta} + \sum_{j=1}^{q}\gamma_j\varepsilon_{t-j} + \eta_t, \ t = q+1, \cdots, n.$$

Let $y_t^* = y_t - \mathbf{S}_t\mathbf{Y}$, $\mathbf{X}_t^* = \mathbf{X}_t - \mathbf{S}_t\mathbf{X}$, then the estimator for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_q)^T$ can be obtained by minimizing the following least squares type objective function

$$L_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{t=q+1}^{n}\left[y_t^* - \mathbf{X}_t^{*T}\boldsymbol{\beta} - \sum_{j=1}^{q}\gamma_j(y_{t-j}^* - \mathbf{X}_{t-j}^{*T}\boldsymbol{\beta})\right]^2. \tag{2.5}$$

In order to shrink insignificant coefficients of $(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ to zero, same as Fan and Li (2001), we obtain the estimator of $(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ by minimizing the following SCAD penalty criterion

$$V_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) = L_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) + n\sum_{i=1}^{p} p_\lambda(|\beta_i|) + n\sum_{j=1}^{q} p_\mu(|\gamma_j|), \tag{2.6}$$

where $p_\lambda(\cdot)$ and $p_\mu(\cdot)$ are defined same as Fan and Li (2001).

3 Selection Consistency and Asymptotic Normality of the Resultant Estimators

In order to present the selection consistency and asymptotic normality of the resultant estimators in last section, we first introduce some notations and technical assumptions.

Let $\boldsymbol{\theta}^0 = (\boldsymbol{\theta}_1^{0T}, \boldsymbol{\theta}_2^{0T})^T$ being the true coefficient vectors, where $\boldsymbol{\theta}_1^0 = (\boldsymbol{\beta}_{\mathcal{C}_1}^{0T}, \boldsymbol{\gamma}_{\mathcal{C}_2}^{0T})^T$, $\boldsymbol{\theta}_2^0 = (\boldsymbol{\beta}_{\mathcal{C}_1^c}^{0T}, \boldsymbol{\gamma}_{\mathcal{C}_2^c}^{0T})^T$ with $\mathcal{C}_1 = \{1 \le i \le p, \beta_i^0 \neq 0\}, \mathcal{C}_2 = \{1 \le j \le q, \gamma_j^0 \neq 0\}$, and $\mathcal{C}_1^c = \{1 \le i \le p, \beta_i^0 = 0\}, \mathcal{C}_2^c = \{1 \le j \le q, \gamma_j^0 = 0\}$. Correspondingly, denote estimators of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ by $\widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_2$, respectively.

The following assumptions are needed, and they are not the weakest possible conditions.

Assumption 1. The sequence $\mathbf{X}_t$ is independent ε_t for all $t = 1, \cdots, n$.

Assumption 2. $\gamma(\zeta) = 1 - \gamma_1^0\zeta - \cdots - \gamma_q^0\zeta^q \neq 0$ for all ζ such that $|\zeta| \le 1$ on the complex plane.

Assumption 3. The errors $\{\eta_t\}_{t=1}^n$ are i.i.d random variables with mean zero and variance σ_η^2. In addition, η_t has the finite fourth order moment.

Assumption 4. $g(z)$ and $M_i(z) = E(x_{ti}|z_t = z)$ have the continuous second derivative in $[0, 1]$, where $i = 1, \ldots, p$, and we let $\mathbf{M}(z) = (M_1(z), \cdots, M_p(z))^T$.

Assumption 5. The random variable z_t has a bounded support $[0, 1]$, and its density function $f_Z(z)$ is Lipschitz continuous and bounded away from 0 on its support.

Assumption 6. The sequence of $\{(z_t, \varepsilon_t)\}$ is a strictly stationary sequence satisfying the mixing condition $\alpha(l) \le c_1 l^{-\gamma}$, $p_{z_1,z_t|\varepsilon_1,\varepsilon_t}\,(u_1, u_t\,|\varepsilon_1, \varepsilon_t) \le c_2 < \infty$, for $c_1, c_2 > 0, \gamma > 2.5$, and $\forall\ l \ge 1$, where p denotes the joint density of (z_t, ε_t). Further, for some $s > 2$, $\lambda > 1 - 2/s$ and interval $z \in [0, 1]$, then

$$\sum_l l^\lambda [\alpha(l)]^{1-2/s} < \infty, E\,|\varepsilon_t|^s < \infty,\ \sup_{z\in[0,1]} \int |\varepsilon|^s\, p(z, \varepsilon) d\varepsilon < \infty.$$

Assumption 7. Both$\mathbf{X}_t$ and z_t have finite second order moment. Furthermore, let $\mathbf{L}_t = \mathbf{X}_t - \mathbf{M}(z_t) - \sum\limits_{j=1}^q \gamma_j^0\{\mathbf{X}_{t-j} - \mathbf{M}(z_{t-j})\}$, then $\mathbf{B} = E(\mathbf{L}_t\mathbf{L}_t^T)$ is positive definite.

Assumption 8. The function $K(\cdot)$ is a bounded symmetric density function with second order, $h = o(n^{-1/5})$, and $nh^4 \to \infty$, as $n \to \infty$.

Remark 1 *These assumptions are commonly adopted in the literature of time series regression and nonparametric technique. Assumption 1, 2, 3 and 6 are reasonable and verifiable and cover many linear and nonlinear time series models. See, for example, Fan and Yao (2003), Wang, Li, and Tsai (2007) and so on. Assumption 4, 5 are about some commonly-used conditions for nonparametric smoothing. Assumption 7 is same as condition (D) in Wang, Li, and Tsai (2007), we just extend it to the partially linear model accordingly. For Assumption 8, the smoothing parameter h for the initial estimators $\widehat{g}(\cdot)$ should be of a smaller order than the standard one in order to control the bias in the preliminary step of the estimation. As a result, undersmoothing is needed. In addition, Liu, Chen and Yao (2009) suggested $h = O(n^{-1/4})$ can be selected.*

Let $\widehat{\boldsymbol{\theta}} = \left(\widehat{\boldsymbol{\theta}}_1^T, \widehat{\boldsymbol{\theta}}_2^T\right)^T = \text{argmin} V_n(\boldsymbol{\theta})$, and $a_n = \max\{p'_{\lambda_n}(|\beta_i^0|), p'_{\mu_n}(|\gamma_j^0|) : i \in \mathcal{C}_1, j \in \mathcal{C}_2\}$, $b_n = \max\left\{p''_{\lambda_n}(|\beta_i^0|), p''_{\mu_n}(|\gamma_j^0|) : i \in \mathcal{C}_1, j \in \mathcal{C}_2\right\}$. The following theorem shows the $\sqrt{n}$-consistency of the resulting estimators.

Theorem 1 *Suppose that Assumption 1 to Assumption 8 hold. If $b_n = o(1)$, then there is a local minimizer $\widehat{\boldsymbol{\theta}}$ of $V_n(\boldsymbol{\theta})$ such that*

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 = O_p(n^{-1/2} + a_n).$$

The next theorem shows that the resulting estimators shrink to 0 with probability 1 for insignificant regression coefficients.

Theorem 2 *Suppose that Assumption 1 to 8 hold. If $\liminf_{n\to\infty}\liminf_{\beta_i\to 0+} p'_{\lambda_n}(\beta_i)/\lambda_n > 0, \lambda_n \to 0, \sqrt{n}\lambda_n \to \infty$, and $\liminf_{n\to\infty}\liminf_{\gamma_j\to 0+} p'_{\mu_n}(\gamma_j)/\mu_n > 0, \mu_n \to 0, \sqrt{n}\mu_n \to \infty, a_n = O(n^{-1/2})$, it holds that*

$$Pr(\widehat{\beta}_i = 0) \to 1, and\ P(\widehat{\gamma}_j = 0) \to 1, \quad as \ \ n \to \infty,$$

where $i \in \mathcal{C}_1^c, j \in \mathcal{C}_2^c$.

Theorem 2 implies that SCAD penalty estimators possess the sparsity property for insignificant regression and autoregressive coefficients.

Denote $\xi_k = E(\varepsilon_t\varepsilon_{t+k}), \mathbf{F} = (\xi_{|i-j|}), \boldsymbol{\Sigma} = diag\{\mathbf{B}, \mathbf{F}\}$, then the asymptotic distribution of SCAD penalty estimator is given below .

Theorem 3 *Suppose that Assumption 1 to 8 hold. If $\liminf_{n\to\infty}\liminf_{\beta_i\to 0+} p'_{\lambda_n}(\beta_i)/\lambda_n > 0, \lambda_n \to 0, \sqrt{n}\lambda_n \to \infty$, and $\liminf_{n\to\infty}\liminf_{\gamma_j\to 0+} p'_{\mu_n}(\gamma_j)/\mu_n > 0, \mu_n \to 0, \sqrt{n}\mu_n \to \infty, a_n = o(n^{-1/2}), b_n = o(1)$, we have that*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^0) \xrightarrow{\mathcal{D}} N(0, \sigma_\eta^2\boldsymbol{\Sigma}_0^{-1}), \quad as \ \ n \to \infty,$$

where $\boldsymbol{\Sigma}_0$ is the submatrix of $\boldsymbol{\Sigma}$ corresponding to $\boldsymbol{\theta}_1^0$.

Theorem 3 shows that the resulting estimators can be as efficient as the oracle estimator in an asymptotic sense.

In order to apply Theorem 3 to make statistical inference, the consistent estimators of σ_η^2 and $\boldsymbol{\Sigma}$ are needed. Let $\widehat{\sigma}_\eta^2 = \frac{1}{n}\sum\limits_t \left\{ y_t^* - \mathbf{X}_t^{*T}\widehat{\boldsymbol{\beta}} - \sum\limits_{j=1}^q \widehat{\gamma}_j\widehat{\varepsilon}_{t-j} \right\}^2$, $\widehat{\mathbf{B}} = \frac{1}{n}\sum\limits_t \left\{ \mathbf{X}_t^* - \sum\limits_{j=1}^q \widehat{\gamma}_j\mathbf{X}_{t-j}^* \right\} \left\{ \mathbf{X}_t^* - \sum\limits_{j=1}^q \widehat{\gamma}_j\boldsymbol{X}_{t-j}^* \right\}^T$, $\widehat{\mathbf{F}} = (\widehat{\xi}_{|i-j|})$ with $\widehat{\xi}_k = \frac{1}{n}\sum\limits_t \widehat{\varepsilon}_t\widehat{\varepsilon}_{t-k}$, $\widehat{\varepsilon}_t = y_t^* - \mathbf{X}_t^{*T}\widehat{\boldsymbol{\beta}}$. We here omit the proof of the property for estimators.

4 Pre-whitening Estimation for Nonparametric Component

If we ignore the autoregressive error structure, it is easy to obtain an estimator of $(g(z), hg'(z))^T$ which has the form

$$(\widehat{g}(z), h\widehat{g}'(z))^T = (\mathbf{D}_z^T\mathbf{W}_z\mathbf{D}_z)^{-1}\mathbf{D}_z^T\mathbf{W}_z(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}).$$

Obviously, $(\widehat{g}(z), h\widehat{g}'(z))^T$ is not asymptotically efficient since it does not take the error structure into account. Same as Xiao, *et al.* (2003) and Liu, Chen and Yao (2009), we will implement the pre-whitening technique to construct a two-stage local linear estimator for $(g(z), hg'(z))^T$.

Denote

$$\check{y}_t = y_t - \mathbf{X}_t^T\boldsymbol{\beta} - \sum_{j=1}^{q}\gamma_j\varepsilon_{t-j}.$$

Then, $E(\check{y}_t|z_t) = g(z_t)$ and $\mathrm{Var}(\check{y}_t|z_t) = \sigma_\eta^2$. Due to the fact that $\sigma_\eta^2 < \sigma_\varepsilon^2$, we can use $\check{y}_t$ to construct a more asymptotically efficient estimator for $g(z_t)$. However, $\check{y}_t$ involves the unknown parameters $\boldsymbol{\beta}$, γ_j and ε_{t-j}. Therefore, we replace $\boldsymbol{\beta}$, γ_j and ε_{t-j} by $\widehat{\boldsymbol{\beta}}$, $\widehat{\gamma}_j$ and $\widehat{\varepsilon}_{t-j}$ in $\check{y}_t$. As a result, we construct a two-stage local linear estimator of $(g(z), h^*g'(z))^T$ which has the form

$$(\widehat{g}^{TS}(z), h^*\widehat{g}'^{TS}(z))^T = (\mathbf{D}_z^{*T}\mathbf{W}_z^*\mathbf{D}_z^*)^{-1}\mathbf{D}_z^{*T}\mathbf{W}_z^*\widehat{\check{\mathbf{Y}}},$$

where $\mathbf{D}_z^* = \begin{pmatrix} 1 & \cdots & 1 \\ \frac{z_{q+1}-z}{h^*} & \cdots & \frac{z_n-z}{h^*}\end{pmatrix}^T$ and $\mathbf{W}_z^* = \mathrm{diag}(K_{h^*}(z_{q+1}-z), \cdots, K_{h^*}(z_n-z))$, $\widehat{\check{\mathbf{Y}}} = (\widehat{\check{y}}_{q+1}, \cdots, \widehat{\check{y}}_n)^T$ with $\widehat{\check{y}}_t = y_t - \mathbf{X}_t^T\widehat{\boldsymbol{\beta}} - \sum_{j=1}^{q}\widehat{\gamma}_j\widehat{\varepsilon}_{t-j}$.

Then, the theorem below shows that the asymptotic property of $(\widehat{g}^{TS}(z), h^*\widehat{g}'^{TS}(z))^T$.

Theorem 4 *Suppose that Assumption 1 to Assumption 8 hold, denote* $\zeta_j = \int z^jK^2(z)dz$, $\rho_j = \int z^jK(z)dz, h^* = O(n^{-1/5})$. *Then we have that*

$$\sqrt{nh^*}\left\{\begin{pmatrix}\widehat{g}^{TS}(z)-g(z)\\ h^*\widehat{g}'^{TS}(z)-h^*g'(z)\end{pmatrix} - \frac{h^{*2}}{2}\begin{pmatrix}\rho_2 g''(z)\\ 0\end{pmatrix}\right\} \xrightarrow{\mathcal{D}} N\left(0, \boldsymbol{\Sigma}_1^*\right), \quad as \quad n\to\infty,$$

where $\boldsymbol{\Sigma}_1{}^* = \frac{\sigma_\eta^2}{f(z)\rho_2^2}\begin{pmatrix}\rho_2^2\zeta_0 & \rho_2\zeta_1\\ \rho_2\zeta_1 & \zeta_2\end{pmatrix}$.

By Lemma 4 in the supplementary material, noted that $\sigma_\eta^2 < \sigma_\varepsilon^2$, this implies that the two-stage local linear estimator $(\widehat{g}^{TS}(z), \widehat{g}'^{TS}(z))^T$ is more asymptotically efficient than $(\widehat{g}(z), \widehat{g}'(z))^T$.

5 Computational Algorithm

Owing to objective function (2.6) is neither differentiable at the origin nor concave, we adapt the local quadratic approximation (Fan and Li, 2001, and Fan and Peng, 2004) procedure to this issue.

Since (2.6) contains the order q as well as regression and autoregressive parameters, it is reasonable to minimize the two following objective functions iteratively:

$$L_n(\boldsymbol{\beta},\boldsymbol{\gamma})+n\sum_{i=1}^{p}p_\lambda(|\beta_i|)\ \text{ with a fixed }\ \boldsymbol{\gamma}, \qquad \text{and} \qquad L_n(\boldsymbol{\beta},\boldsymbol{\gamma})+n\sum_{j=1}^{q}p_\mu(|\gamma_j|)\ \text{ with a fixed }\ \boldsymbol{\beta}.$$

When γ is fixed, we use the local quadratic approximation for the first formula at the true value for $\boldsymbol{\beta}^0$, and it reduces to a quadratic minimization problem:

$$L_n(\boldsymbol{\beta},\boldsymbol{\gamma})+n\sum_{i=1}^{p}\frac{p'_\lambda(|\beta_i^{(0)}|)}{2|\beta_i^{(0)}|}\beta_i^2. \tag{5.1}$$

In order to complete the whole iterative process, we minimize the following two objective functions

$$\widehat{\boldsymbol{\beta}}^{(l)}=\arg\min_{\boldsymbol{\beta}}\left\{L_n(\boldsymbol{\beta},\widehat{\boldsymbol{\gamma}}^{(l-1)})+n\sum_{i=1}^{p}\frac{p'_\lambda(|\widehat{\beta}_i^{(l-1)}|)}{2|\widehat{\beta}_i^{(l-1)}|}\beta_i^2\right\}, \tag{5.2}$$

with a fixed $\widehat{\boldsymbol{\gamma}}^{(l-1)}$, and

$$\widehat{\boldsymbol{\gamma}}^{(l)}=\arg\min_{\boldsymbol{\gamma}}\left\{L_n(\widehat{\boldsymbol{\beta}}^{(l)},\gamma)\right\}, \tag{5.3}$$

with a fixed $\widehat{\boldsymbol{\beta}}^{(l)}$. When $\max\left\{|\widehat{\beta}_i^{(l)}-\widehat{\beta}_i^{(l-1)}|,|\widehat{\gamma}_j^{(l)}-\widehat{\gamma}_j^{(l-1)}|,i=1,\cdots,p;j=1,\cdots,q\right\}<10^{-6}$, we stop the algorithm process. At the same time, we use the same approach for the estimator of $\boldsymbol{\gamma}$.

We use the least squares estimator without considering of autocorrelation as an initial estimator for the regression coefficients. Then, computing the ordinary residual, we combine the least squares and the estimated residual to obtain the estimator of autoregression coefficients.

In order to complete iterative process, we need to select the tuning parameters (λ,μ) for the SCAD. Same as Wang, Li and Tsai (2007), we use the BIC criterion:

$$BIC(\lambda,\mu)=\log\left\{\frac{RSS(\lambda,\mu)}{n}\right\}+m(\lambda,\mu)\frac{\log(n)}{n}, \tag{5.4}$$

where $RSS(\lambda,\mu)=L_n(\widehat{\boldsymbol{\beta}},\widehat{\gamma})$ and $m(\lambda,\mu)$ is the number of nonzero coefficients of (λ,μ). We minimize $BIC(\lambda,\mu)$ over a grid of points distributed as $(0,\frac{5\sqrt{log(n)}}{\sqrt{n}})\times(0,\frac{5\sqrt{log(n)}}{\sqrt{n}})$. After the (λ,μ) are selected, we can obtain the tuning parameters and the order q.

As Xiao *et al* (2003), we use the rule of thumb bandwidth, $h=1.06S_xn^{-1/4}$, and $h^*=1.06S_xn^{-1/5}$, where S_x is the standard error of $\mathbf{X}_t$. From our limited simulation experience, the bandwidths are appropriate for the model proposed in this article.

6 Simulation Studies

In this section, we conduct several Monte Carlo simulation studies to investigate the finite sample performance of the SCAD penalty estimators proposed in previous sections. The data are generated from the following partially linear model with autoregressive error

$$y_t = x_{t1} + 1.5x_{t2} + 2x_{t5} + 3\cos(2\pi z_t) + \varepsilon_t,$$

where $\varepsilon_t = 0.6\varepsilon_{t-1} - 0.3\varepsilon_{t-3} + \eta_t$. The explanatory variables $\mathbf{X}_t = (x_{t1}, \cdots, x_{t6})^T$ are independently generated from the multivariate normal distribution with mean $\mathbf{0}_{6\times 1}$, and the pairwise correlation between x_{tk_1} and x_{tk_2} is $0.3^{|k_1-k_2|}$. In addition, we let $z_t = x_{t3}$, $\eta_t \overset{i.i.d.}{\sim} N(0, \sigma_\eta^2)$. The regression and autocorrelated coefficients are $\boldsymbol{\beta}^0 = (1, 1.5, 0, 0, 2, 0)^T$ and $\boldsymbol{\gamma}^0 = (0.6, 0, -0.3, 0, 0)^T$, respectively.

In our simulation, we take n = 100, 200 and 400, and the number of replication is 1000. We consider the following scenarios: $\sigma_\eta = 1$ or 2. Define signal-to-noise ratio(SNR) as follows:

$$SNR = \frac{Var(x_{t1} + 1.5x_{t2} + 2x_{t5} + 3\cos(2\pi z_t))}{Var(\varepsilon_t)}.$$

Then, $SNR = 7.83$ with $\sigma_\eta = 1$ and $SNR = 1.96$ with $\sigma_\eta = 2$, respectively. Both of these settings may reflect the real data encountered in practice.

We compare four type estimators of $\boldsymbol{\beta}_C^0 = (\beta_1, \beta_2, \beta_5)$. They are, (1) : the profile local linear estimator which ignores the autoregressive error structure ($\widehat{\boldsymbol{\beta}}_C^I$). (2): the profile local linear estimator which takes the autoregressive error structure into account, but the order of autoregressive error structure is misspecified ($\widehat{\boldsymbol{\beta}}_C^E$). (3): the profile local linear estimator which takes the autoregressive error structure into account and the autoregressive error structure is fitted by the SCAD penalty ($\widehat{\boldsymbol{\beta}}_C^S$). (4): the profile local linear estimator which takes the autoregressive error structure into account and the autoregressive error structure is totally known ($\widehat{\boldsymbol{\beta}}_C^O$).

The same as Fan and Li (2004), the performance of $\widehat{g}(z)$ is assessed by the square root of average squared errors(RASE):

$$RASE\{\widehat{g}(z)\} = \left\{ \frac{1}{n} \sum_{j=1}^{n} (\widehat{g}(z_j) - g(z_j))^2 \right\}^{\frac{1}{2}}.$$

We use the Gaussian kernel $K(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})$, and the Rule of thumb is applied to select bandwidth. The parameters estimators and the corresponding standard deviation under different sample size are given in the following Table 1.

From Table 1 we make the following observations:

1. Under the model studied, $\widehat{\boldsymbol{\beta}}_C^S$ is more efficient than both of $\widehat{\boldsymbol{\beta}}_C^I$ and $\widehat{\boldsymbol{\beta}}_C^E$. In addition, $\widehat{\boldsymbol{\beta}}_C^E$ is more efficient than $\widehat{\boldsymbol{\beta}}_C^I$. Therefore, it is necessary to take the autoregressive error structure into account.
2. The finite sample performances of $\widehat{\boldsymbol{\beta}}_C^S$ are very close to those of $\widehat{\boldsymbol{\beta}}_C^O$. This is consistent with the theoretical results.
3. With the increase of sample size, the finite sample performances all types of estimator for parameters have been improved.

At the same time, both the percentage of correctly(under, over) estimated numbers of regression variables and the percentage of correctly(under, over) estimated numbers of autoregressive orders are

Table 1: Finite sample performances of estimators for the regression coefficient.

	$\widehat{\beta}^I_{\mathcal{C}}$	$\widehat{\beta}^E_{\mathcal{C}}$	$\widehat{\beta}^S_{\mathcal{C}}$	$\widehat{\beta}^O_{\mathcal{C}}$
$n = 100, \sigma_\eta = 1, SNR = 7.83$				
β_1	0.996(0.166)	1.008(0.111)	0.994(0.104)	0.994(0.102)
β_2	1.497(0.126)	1.488(0.123)	1.505(0.115)	1.504(0.112)
β_5	2.002(0.096)	1.997(0.093)	2.003(0.087)	2.002(0.087)
$n = 200, \sigma_\eta = 1, SNR = 7.83$				
β_1	0.997(0.065)	1.004(0.062)	0.998(0.059)	0.998(0.058)
β_2	1.502(0.075)	1.490(0.072)	1.505(0.069)	1.505(0.067)
β_5	1.995(0.053)	1.994(0.051)	2.001(0.048)	2.001(0.047)
$n = 400, \sigma_\eta = 1, SNR = 7.83$				
β_1	0.997(0.041)	1.001(0.034)	0.999(0.028)	0.999(0.028)
β_2	1.503(0.044)	1.497(0.038)	1.500(0.026)	1.500(0.026)
β_5	1.998(0.034)	1.997(0.033)	1.999(0.025)	1.999(0.025)
$n = 100, \sigma_\eta = 2, SNR = 1.96$				
β_1	1.019(0.177)	1.014(0.142)	1.009(0.128)	1.009(0.127)
β_2	1.488(0.186)	1.484(0.162)	1.495(0.142)	1.495(0.141)
β_5	2.003(0.139)	1.988(0.118)	1.998(0.097)	1.998(0.095)
$n = 200, \sigma_\eta = 2, SNR = 1.96$				
β_1	1.007(0.109)	1.006(0.087)	1.001(0.082)	1.001(0.081)
β_2	1.491(0.129)	1.496(0.097)	1.502(0.091)	1.502(0.090)
β_5	1.994(0.099)	1.991(0.072)	1.994(0.067)	1.995(0.067)
$n = 400, \sigma_\eta = 2, SNR = 1.96$				
β_1	1.005(0.074)	1.005(0.059)	1.003(0.057)	1.003(0.057)
β_2	1.503(0.083)	1.498(0.065)	1.501(0.061)	1.500(0.061)
β_5	1.994(0.058)	1.997(0.048)	1.997(0.042)	1.997(0.042)

given in Tables 2, where “correct” of regression variables denotes $\widehat{\boldsymbol{\beta}}^S_{\mathcal{C}_1} \neq \mathbf{0}$ and $\widehat{\boldsymbol{\beta}}^{Sc}_{\mathcal{C}_1} = \mathbf{0}$, “under” of regression variables denotes $\widehat{\boldsymbol{\beta}}^{Sc}_{\mathcal{C}_1} = \mathbf{0}$ and at least one of $\widehat{\boldsymbol{\beta}}^S_{\mathcal{C}_1}$ is zero. Similarly, “correct” of autoregressive orders denotes $\widehat{\boldsymbol{\gamma}}^S_{\mathcal{C}_1} \neq \mathbf{0}$ and $\widehat{\boldsymbol{\gamma}}^{Sc}_{\mathcal{C}_1} = \mathbf{0}$.

Table 2: Finite sample percent of SCAD-estimators for the regression coefficient.

	n=100			n=200			n=400		
	Under	Correct	Over	Under	Correct	Over	Under	Correct	Over
σ_η=1									
AR order	0.162	0.713	0.125	0.024	0.923	0.053	0.024	0.969	0.007
Reg.coef	0.008	0.951	0.041	0.005	0.967	0.028	0	0.982	0.018
σ_η=2									
AR order	0.151	0.701	0.148	0.011	0.908	0.081	0.013	0.955	0.032
Reg.coef	0.015	0.932	0.053	0.012	0.956	0.032	0	0.970	0.030

From Table 2, we make see that the finite sample performances of determining the regression

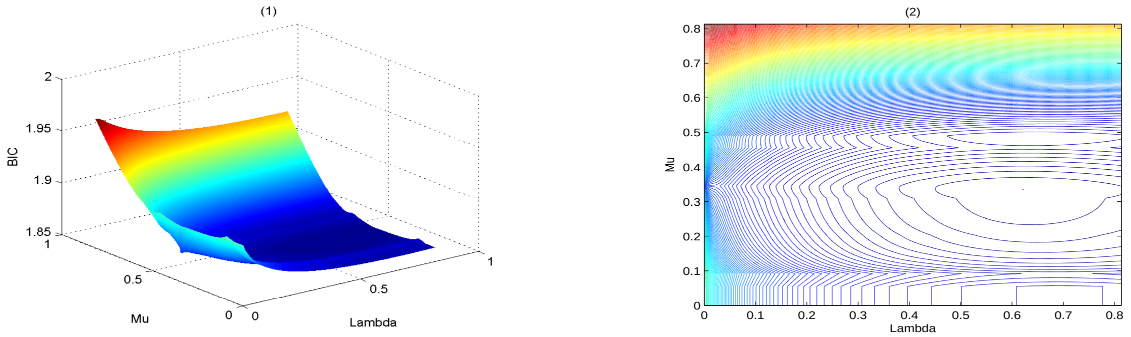

Figure 1: Plots for parameters selection process. Where $n = 200, \sigma_\eta = 2$, and (1) denotes surface of BIC subject to the tuning parameters (λ, μ); (2) denotes contour plot of tuning parameters (λ, μ). $BIC_{min} = 1.8586$ with the optimal parameter values of (λ, μ) are $(0.622, 0.335)$.

variable and the autoregressive error order by SCAD penalty are satisfied, especially when the sample size is moderate or large.

In addition to the estimation of nonparametric function, two types estimation proposed in this article for nonparametric component are presented in Table 3.

Table 3: Estimators for the nonparametric function.

	n=100		n=200		n=400	
	RASE		RASE		RASE	
$\sigma_\eta = 1$						
$(\widehat{g}, \widehat{g}^{TS})$	0.0353	0.0221	0.0237	0.0151	0.0205	0.0112
$\sigma_\eta = 2$						
$(\widehat{g}, \widehat{g}^{TS})$	0.0368	0.0341	0.0311	0.0259	0.0232	0.0204

From Tables 3, we can see that the finite sample performances of estimators for the nonparametric function is also satisfied. Its finite sample performances is improved with the increasing of the sample size.

Furthermore, An anonymous referee suggested presenting in plots for the tuning parameter selection via BIC, we give the corresponding selection process in the Figure 1.

7 An Application Example

Engle *et al.* (1986) proposed the partially linear model, they analyzed the relationship between temperature and electricity usage. Yatchew (2003) estimated a similar model. As both high and low temperatures lead to the increasing of electricity usage, it is natural to assume the impact of temperature on electricity consumption to be nonlinear, and a linear relationship may be assumed for other regressors. Here, SCAD-penalty is applied to select significant explanatory variables and autoregressive errors order.

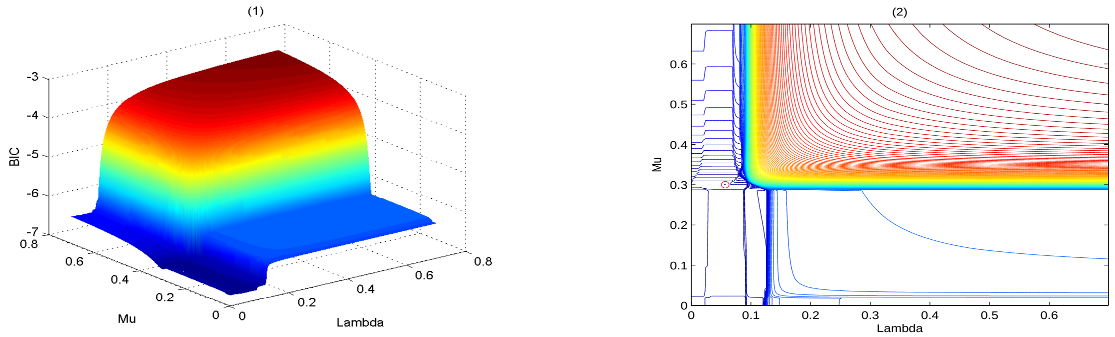

Figure 2: Plots for parameters selection process. Where (1) denotes surface of BIC subject to the tuning parameters (λ, μ); (2) denotes contour plot of tuning parameters (λ, μ). $BIC_{min} = -6.7524$ with the optimal parameter values of (λ, μ) are $(0.057, 0.3)$.

The data consist of quarterly observations in Ontario for the period 1971 to 1994 (Yatchew, 2003). The specification is

$$Elec_t = f(Temp_t) + \beta_1 Relprice_t + \beta_2 Gdp_t + \beta_3 Cdd_t + \beta_4 Hdd_t + \varepsilon_t, \tag{7.1}$$

where $\varepsilon_t = \gamma_1 \varepsilon_{t-1} + \cdots + \gamma_q \varepsilon_{t-q} + \eta_t$, where q is unknown.

$Elec$ is the log of electricity sales, $Temp$ is heating and cooling degree days measured relative to 68 $^\circ F$, $Relprice$ is the log of the ratio of the price of electricity ($pelec$) to the price of natural gas ($Pgas$), Cdd is the cooling degree days per quarter, Hdd is the heating degree days per quarter, and Gdp is the log of gross provincial.

We first fit the data with classical partially linear model, and initial resulting estimated coefficients are $\widehat{\beta}_1 = -0.0664, \widehat{\beta}_2 = 1.0616, \widehat{\beta}_3 = 0.0001, \widehat{\beta}_4 = -0.0001$.

Considering the correlation of data, we naturally recommend the partially linear model with autoregressive errors. Engle *et al.* (1986) just let $q = 1$. In general, we choose maximum autoregressive order 4 for the data. That is to say, we can assume:

$$\varepsilon_t = \gamma_1 \varepsilon_{t-1} + \cdots + \gamma_4 \varepsilon_{t-4} + \eta_t. \tag{7.2}$$

In the same way, we fit the autoregressive errors, and the resulting estimators are $\widehat{\gamma}_1 = 0.4286, \widehat{\gamma}_2 = 0.0285, \widehat{\gamma}_3 = 0.1811, \widehat{\gamma}_4 = 0.1165$.

By the formula of (5.4), we can obtain that $(\lambda, \mu) \in (0, 0.7011) \times (0, 0.7011)$. The surfaces plot for parameters selection process is illustrated as figure 2.

Then, under the SCAD-penalty, $\widehat{\beta}_{3S} = \widehat{\beta}_{4S} = 0$. Combined with initial coefficients, it shows that neither Cdd nor Hdd is significant. Furthermore, $\widehat{\gamma}_{2S} = \widehat{\gamma}_{3S} = \widehat{\gamma}_{4S} = 0$, which show that the estimated residuals satisfy AR(1) process, and we obtain the following function estimator:

$$Elec_t = f(Temp_t) - \underset{(0.0577)}{0.0699} Relprice_t + \underset{(0.0657)}{1.0531} Gdp_t + \varepsilon_t, \quad \varepsilon_t = 0.5655 \varepsilon_{t-1} + \eta_t, \tag{7.3}$$

where values in $(\cdot)$ denote the standard error of estimator. To improve the efficiency of regression coefficients estimators, we apply pre-whitening technique to nonparametric function estimation and obtain the following function estimator:

$$Elec_t = f(Temp_t) - \underset{(0.0571)}{0.0673} Relprice_t + \underset{(0.0648)}{1.0523} Gdp_t + \varepsilon_t, \quad \varepsilon_t = 0.5655\varepsilon_{t-1} + \eta_t. \tag{7.4}$$

The estimator for unknown function $f(Temp_t)$ is given as figure 3. Both (a) and (b) indicate the relationship between $Elec$ and $Temp$ is nonlinear, $Elec$ is get much bigger when $Temp$ is low as well as it is high.

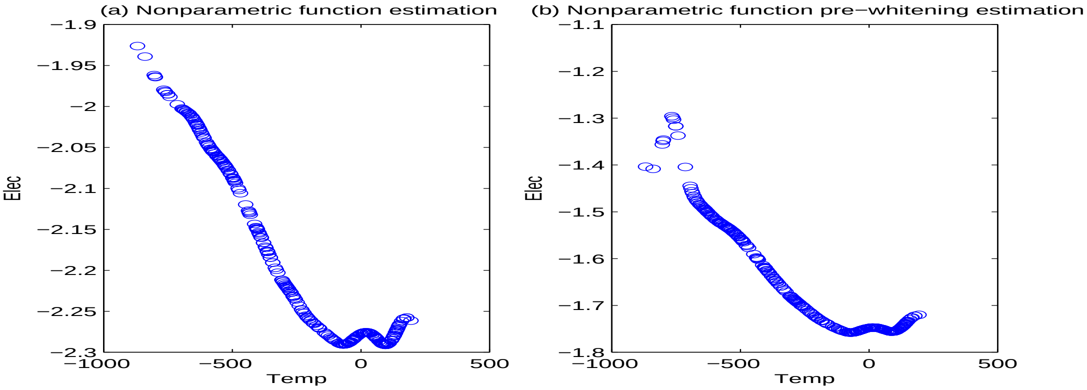

Figure 3: Estimators of the unknown function

8 Concluding Remarks

In this paper, we have investigated the estimating and selecting problem of a partially linear model with correlated errors. By combining the linear local approximation and semiparametric least squares, the estimator and selector for both regression coefficients and autoregressive parameters have been constructed. Their consistency, asymptotic normality and oracle property were established. Based on the estimated time series error structure and pre-whitening technique, we further constructed a two-stage local linear estimator for the nonparametric component of the model. The two-stage local linear estimator was shown to be asymptotically normal and more efficient than the one ignoring the estimated time series error structure.

Important and interesting further studies can be conducted on inference problems such as how to identify which explanatory variables are linear and which explanatory variables are nonlinear relevant. One may also be interested in check whether the nonparametric component has a parametric form. In addition, it may be more reasonable to allow p and q to increase with the increasing of sample size. These issues are not easy and yet deserve our further investigations in the future.

Acknowledgements

The authors would like to appreciate the comments and suggedtions of the associate editor and referees, which significantly improved the appearance of this paper.

References

[1] Aneiros-pérez, G. and Vilar-Fernández, J. M. (2008). Local polynomial estimation in partial linear regression models under dependence. *J. Statist. Plann. Infer*, **52**, 2757-2777.

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

[3] Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc*, **92**, 477-489.

[4] Carroll, R. J., D. Ruppert and M. P. Wand (2003). *Semiparametric Regression.* Cambridge University Press.

[5] Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist*, **16**, 136-146.

[6] Chen, H. and Shiau, J. H. (1991). A two-stage spline smoothing method for partially linear models. *J. Statist. Plann. Infer*, **27**, 187-202.

[7] Chen, H. and Shiau, J. H. (1994). Data-driven efficient estimators for a partially linear model. *Ann. Statist*, **22**, 211-237.

[8] 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. Statist. Assoc*, **81**, 310-320.

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

[10] Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, **11**, 1031-1057.

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

[12] 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.

[13] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist*, **32**, 928-961.

[14] Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer, New York, NY, USA.

[15] Frank, I. E., and Friedman, J. H. (1993). A satistical view of some chemometrics regression tools. *Technometrics*, **35**, 109-148.

[16] Fu, W. J. (1998). Penalized regression: the bridge versus the lasso. *J. Comput. and Graph. Statist*, **7**, 397-416.

[17] Gao, J. T. (1995). Asympototic theory for paratially linear models. *Commun stat-theor m*, **22**, 3327-3354.

[18] Gao, J. T. (2007). *Nonlinear Time Series: Semiparametric and Nonparametric Methods.* Chapman & Hall/CRC, London.

[19] Härdle, W. (1990). *Applied Nonparametric Regression.* New York, Cambridge University Press.

[20] Härdle, W., Liang, H. and Gao, J. T. (2000). *Partially Linear Models*, Physica-Verlag, Heidelberg.

[21] Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics.* Springer-Verlag.

[22] Lam, C. and Fan, J. (2008). Profile-kernel liklelihood inference with diverging number of parameters. *Ann. Statist*, **36**, 2232-2260.

[23] Liu, J., Chen, R. and Yao, Q. (2009). Nonparametric transfer function models. *Journal of Econometrics*, **157**, 151-164.

[24] Murphy, S. A. and Van Der Vaart, A. W. (2000). On profile likelihood (with discussion). *J. Amer. Statist. Assoc*, **95**, 449-485.

[25] Risa, K. and Takayuki, S. (2009). Model and variable selection procedures for semiparametric time series regression. *J. Probab Statist*, Article ID 487194,doi:10.1155/2009/487194.

[26] Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, **56**, 931-954.

[27] Speckman, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc., Ser. B*, **50**, 413-436.

[28] Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc., Ser. B*, **58**, 267-288.

[29] Wang, H., Li, G. and Tsai, C.-L. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *J. Roy. Statist. Soc., Ser. B*, **69**, 63-78.

[30] Xiao, Z., Linton, O. B., 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.

[31] Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician.* Campbridge University Press, New York.

[32] You, J. H. and Chen, G. (2007). Semiparametric generalized least squares estimation in partially linear regression models with correlated errors. *J. Statist. Plann. Infer*, **137**, 117-132.

[33] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist*, **38**, 894-942.