



A Seemingly Unrelated Nonparametric Additive Model with Autoregressive Errors

Alan T. K. Wan, Jinhong You & Riquan Zhang


To cite this article: Alan T. K. Wan, Jinhong You & Riquan Zhang (2016) A Seemingly Unrelated Nonparametric Additive Model with Autoregressive Errors, *Econometric Reviews*, 35:5, 894-928, DOI: [10.1080/07474938.2014.998149](https://doi.org/10.1080/07474938.2014.998149)

To link to this article: <http://dx.doi.org/10.1080/07474938.2014.998149>



Accepted author version posted online: 17 Dec 2014.
Published online: 17 Dec 2014.



[Submit your article to this journal](#) 



Article views: 60



[View related articles](#) 



[View Crossmark data](#) 

A Seemingly Unrelated Nonparametric Additive Model with Autoregressive Errors

Alan T. K. Wan¹, Jinhong You², and Riquan Zhang³

¹*Department of Management Sciences, City University of Hong Kong, Kowloon,
Hong Kong, China*

²*School of Statistics and Management, Shanghai University of Finance and Economics,
Shanghai, China, and Key Laboratory of Mathematical Economics (SUFEC),
Ministry of Education of China, Shanghai, China*

³*Department of Statistics, East China Normal University, Shanghai, China*

This article considers a nonparametric additive seemingly unrelated regression model with autoregressive errors, and develops estimation and inference procedures for this model. Our proposed method first estimates the unknown functions by combining polynomial spline series approximations with least squares, and then uses the fitted residuals together with the smoothly clipped absolute deviation (SCAD) penalty to identify the error structure and estimate the unknown autoregressive coefficients. Based on the polynomial spline series estimator and the fitted error structure, a two-stage local polynomial improved estimator for the unknown functions of the mean is further developed. Our procedure applies a prewhitening transformation of the dependent variable, and also takes into account the contemporaneous correlations across equations. We show that the resulting estimator possesses an oracle property, and is asymptotically more efficient than estimators that neglect the autocorrelation and/or contemporaneous correlations of errors. We investigate the small sample properties of the proposed procedure in a simulation study.

Keywords Additive structure; Asymptotic normality; Autoregression; Local polynomial; SCAD penalty; SUR.

JEL Classification C14; C39; C51.

1. INTRODUCTION

The seemingly unrelated regression (SUR) introduced by Zellner (1962) is an important tool in econometric modeling involving pooled data. The classical SUR model consists of a set of linear regression equations in which the errors are contemporaneously

Address correspondence to Prof. Alan Wan, Department of Management Sciences, City University of Hong Kong, Tat Chee Ave., Kowloon, Hong Kong, China; E-mail: msawan@cityu.edu.hk

correlated across the equations. Because the equations are stochastically related through the error terms, efficiency may be gained by treating the equations as a system and using generalized least squares estimation. The SUR model has found considerable use in applied work. The monograph by Srivastava and Giles (1987) provides a good coverage of the SUR literature until the late 1980s. Recent applied studies in economics that involve the SUR model include Thompson et al. (2002), Wang (2010), and Shukur and Zeebari (2011), among others.

Most of the literature on SUR model estimation assumes that the functional forms of the equations are linear, but for many practical econometric problems the functional forms are actually unknown, so a more flexible nonparametric approach may be more attractive. Smith and Kohn (2000) gave two examples to illustrate their Bayesian hierarchical nonparametric SUR model. Their first example relates three different advertisement exposure scores to the position of the advertisement in an Australian women's magazine. In their second example, two separate nonparametric regressions are formulated to explain electricity load in two adjacent states in Australia. In both cases, they identified significant nonlinear relationship between the dependent and explanatory variables. Koop et al. (2005) used a two-equation nonparametric SUR model to examine the dependence of wage levels and years of schooling on a range of economic and social variables including labor market experience, arms force qualifying test score, local unemployment rate, weeks of tenure at current job, and so on. Other recent studies in econometrics that approach SUR modeling nonparametrically include Xu et al. (2008), who considered a nonparametric SUR model under a constrained covariance structure of the disturbances, Singh and Wang (2012), who analyzed the properties of a semiparametric estimator for the coefficients in a two-equation SUR equation system, and You and Zhou (2014), who considered a fixed effects panel data SUR partially linear models. In the statistics literature, recent contributions to nonparametric SUR models include Wang et al. (2000), He and Lawless (2005), Carroll et al. (2006), Welsh and Yee (2006), Xu et al. (2011), and Zhou et al. (2011). Although these studies are mostly motivated by problems in biostatistics, many of the methods developed are in fact general methods with potential for applications in economics and other fields.

Nonparametric approaches are not without drawbacks, however. There is evidence that the small sample properties of the multivariate kernel estimator can be quite unsatisfactory (Silverman, 1986). Interpretability is another problem with nonparametric regression based on kernel and smoothing splines. More importantly, as the dimension of the model grows, the convergence rate of the kernel estimator decreases, and this is the so-called curse of dimensionality. One way to ameliorate the latter problem is to reduce dimension. Methods based on this approach include the projection pursuit (Huber, 1985), single index models (Härdle and Stoker, 1989), and sliced inverse regression (Li, 1991). However, for these methods, the curse of dimensionality remains when the underlying dimension is large. Another approach is to relax the conditions on the traditional parametric models and explore the hidden structure. One prominent method

that uses this approach is the additive model proposed by Stone (1985) and Hastie and Tibshirani (1990). This model is based on an additive approximation to the nonparametric regression function, and because each of the individual additive terms is estimated using a univariate smoother, the curse of dimensionality is avoided. It also has the advantage of interpretability as estimates of the individual terms explain how the dependent variable changes with the corresponding independent variables. Moreover, Stone (1985) showed that the unknowns in the additive model can be estimated at the optimal rate of convergence for univariate functions. For more recent theoretical work on the additive model, see Linton and Nielsen (1995), Linton (1997), Huang and Yang (2004), Fan and Jiang (2005), Xue (2009), and Xue and Liang (2010). The additive model has been applied in a number of studies in economics and finance. There are a few such examples in the fields of empirical demand analysis (Lyssiotou et al., 2002), hedonic house price modeling (Martins-Filho and Bin, 2005; Bontemps et al., 2008), and stock market analysis (Eisenbeiss et al., 2007).

In this article, we extend the idea of the additive model to a system of SUR equations. Our model takes the form

$$Y_{st} = \alpha_{s0} + \alpha_{s1}(X_{st1}) + \dots + \alpha_{sp_s}(X_{stp_s}) + \varepsilon_{st}, \quad s = 1, \dots, m, \quad t = 1, \dots, T, \quad (1.1)$$

where Y_{st} 's are responses, $(\alpha_{10}, \dots, \alpha_{m0})^T$ are unknown constants, $(\alpha_{11}(\cdot), \dots, \alpha_{1p_1}(\cdot), \dots, \alpha_{mp_m}(\cdot))$'s are unknown smooth functions, $(X_{1t1}, \dots, X_{1tp_1}, \dots, X_{mtp_m})^T$'s are the design points, and ε_{st} 's are random errors. Although the equations in (1.1) are nonlinear, they each retain the interpretable additive form of linear regression through the additive approximation to the true model by the nonparametric functions. For the error terms, it is assumed that

$$\varepsilon_{st} = \mathbf{s}_{s1}\varepsilon_{s,t-1} + \dots + \mathbf{s}_{sd_s}\varepsilon_{s,t-d_s} + e_{st}, \quad s = 1, \dots, m, \quad t = 1, \dots, T, \quad (1.2)$$

with $E(e_{st}) = 0$, $E(e_{st}^2) = \sigma_{ess}^2$, and $E(e_{s_1t_1}e_{s_2t_2}) = \sigma_{s_1s_2}^2$ for $s_1 \neq s_2$, and $E(e_{s_1t_1}e_{s_2t_2}) = 0$ when $t_1 \neq t_2$ irrespective of the values of s_1 and s_2 . Thus, the m equations are stochastically related through the contemporaneous correlations of the error terms, but each of them purports to explain a different dependent variable through a different set of covariates. The error process of (1.2) also assumes that the error term in each equation follows an autoregressive (AR) process, thus allowing for serial correlations in each of the disturbances in addition to contemporaneous correlations among the disturbances. The pure AR specification is justified given that AR processes are usually good approximations to the general autoregressive moving average (ARMA) process provided that the latent roots of the MA polynomial lie inside the unit circle (Brockwell and Davis, 1991). As is well-known, the original SUR model of Zellner (1962) allows only for contemporaneous but not serial correlation. Later, Kmenta and Gilbert (1970), Guilkey and Schmidt (1973), Doran and Griffiths (1983), Turkington (2000), Foschi and Kontoghiorghes (2003), Koebel (2004), and others modified the original SUR model to

allow for both contemporaneous and serial correlations in the errors, but these findings all focused on parametric linear SUR equations. Due to the relevance of serial correlations to economic phenomena, the large majority of this work was published in econometric journals.

It is well-known that for a parametric model with serially correlated errors, efficient estimators of the mean parameters of the regression can be obtained by weighting methods. However, the same is not generally true in the case of nonparametric models. As Lin and Carroll (2000) pointed out, standard kernel or local polynomial weighting methods can often result in estimates that are worse than ignoring the serial correlations. Several recent studies, including Wang (2003), Xiao et al. (2003), Li and Li (2009), Martins-Filho and Yao (2009), and Liu et al. (2010), considered the alternative approach of a prewhitening transformation of the dependent variable. Our analysis here is complicated by the existence of both serial and contemporaneous correlations in the errors. A central question of interest is how to take both types of correlations into account to yield an efficient estimator of the additive components. We propose a method that first estimates the unknown functions by combining polynomial spline series approximation with least squares, and then uses the SCAD penalty function together with the estimated residuals to determine the order of the AR error process and estimate the unknown autoregressive coefficients. Furthermore, based on the polynomial spline series estimator and the fitted error structure, we develop a two-stage local polynomial estimator of the unknown functions of the mean. Our method extends some of the existing prewhitening methods to account for the possibility of contemporaneous correlations in the errors. We demonstrate that first, our procedure is asymptotically more efficient than methods that neglect either or both types of correlations, and second, estimators of additive components have the same distribution that they would have if the nonparametric components were known in advance; thus, our estimators have the oracle property in the sense of Fan and Li (2001) and Fan and Peng (2004). We establish these asymptotic properties under the α -mixing condition.

To summarize, the nonparametric additive model has considerable appeal and is widely popular among statisticians. By extending the additive model to SUR, which has continued to attract significant attention from economists, this article hopes to bring further awareness of the additive model to economists and strengthen the interface between econometrics and statistics. Another important strength of our analysis is that it allows errors in the SUR model to be serially correlated in addition to being contemporaneously correlated. As mentioned before, SUR model with autocorrelation has inspired a large econometric literature although nearly all previous studies on this aspect have assumed linear functional forms of the SUR equations.

The layout of the remainder of this article is as follows. In Section 2, we present an initial estimator of the unknown mean functions. Section 3 considers the identification of the error structure, and the subsequent estimation of the autoregressive coefficients, error variance and correlation parameters. In Section 4, we develop a two-stage local

polynomial estimation method for model (1.1)–(1.2). Section 5 reports results of a simulation study designed to investigate the small sample properties of estimators. The proofs of the main results are contained in the Appendix.

2. INITIAL ESTIMATION BY POLYNOMIAL SPLINE SERIES APPROXIMATION

As $E\{\alpha_{sj}(X_{stj})\} = 0$, α_{s0} can be consistently estimated by $\hat{\alpha}_{s0} = \sum_{i=1}^T Y_{st}/T$ at the rate of $1/\sqrt{T}$ —a rate faster than rate of convergence for estimating nonparametric functions. For notational convenience and without loss of generality, we assume that $\alpha_{s0} = 0$. Assume also that the nonparametric explanatory variable X_{stj} is distributed on a compact interval $[a_{sj}, b_{sj}]$ for $s = 1, \dots, m$ and $j = 1, \dots, p_s$, and without loss of generality, we let $[a_{sj}, b_{sj}] = [0, 1]$ for $s = 1, \dots, m$ and $j = 1, \dots, p_s$.

By Definition 4.1 of Schumaker (1981, p. 108), we define the polynomial splines as follows. Let $0 = \zeta_0 < \zeta_1 < \dots < \zeta_\kappa < \zeta_{\kappa+1} = 1$ be a partition of $[0, 1]$ into $\kappa + 1$ subintervals $I_{\kappa s} = [\zeta_s, \zeta_{s+1}]$, $s = 0, \dots, \kappa - 1$ and $I_{\kappa\kappa} = [\zeta_\kappa, \zeta_{\kappa+1}]$, where $\kappa \equiv \kappa_T = T^v$, with $0 < v < 0.5$ being a positive integer such that $\max_{0 \leq s \leq \kappa} |\zeta_{s+1} - \zeta_s| = O(T^v)$. Let \mathcal{S}_T be the space of polynomial splines of degree $l \geq 1$, comprising functions $f(\cdot)$ that satisfy the following conditions: (i) the restriction of $f(\cdot)$ to $I_{\kappa s}$ is a polynomial of degree l for $0 \leq s \leq \kappa$; and (ii) for $l \geq 2$ and $0 \leq l' \leq l - 2$, $f(\cdot)$ is l' times continuously differentiable on $[0, 1]$. Denote $N_T \equiv \kappa_T + l$. Then there exists a normalized polynomial spline basis $\{B_{sjk}, 1 \leq k \leq N_T\}$ in \mathcal{S}_T such that each $\alpha_{sj}(x)$ can be approximated by

$$\alpha_{sj}(x) \approx \sum_{k=1}^{N_T} \theta_{sjk} B_{sjk}(x), \quad s = 1, \dots, m \quad \text{and} \quad j = 1, \dots, p_s,$$

where $\theta_{sj} = (\theta_{sj1}, \dots, \theta_{sjN_T})^\tau$ is an unknown N_T -vector. See Wang and Yang (2007) for details. Model (1.1) can then be approximated by

$$Y_{st} \approx \sum_{k=1}^{N_T} \theta_{s1k} B_{s1k}(X_{st1}) + \dots + \sum_{k=1}^{N_T} \theta_{sp_s k} B_{sp_s k}(X_{stp_s}) + \varepsilon_{st}, \quad s = 1, \dots, m \quad \text{and} \quad t = 1, \dots, T. \tag{2.1}$$

From (2.1), for any fixed s , we can estimate $\theta_s = (\theta_{s1}^\tau, \dots, \theta_{sp_s}^\tau)^\tau$ by least squares, resulting in the estimator

$$\hat{\theta}_s = (\hat{\theta}_{s1}^\tau, \dots, \hat{\theta}_{sp_s}^\tau)^\tau = \underset{(\hat{\theta}_{s1}^\tau, \dots, \hat{\theta}_{sp_s}^\tau)^\tau}{\operatorname{argmin}} \frac{1}{n} \left\{ \mathbf{Y}_s - \sum_{j=1}^{p_s} \mathbf{B}_{sj} \theta_{sj} \right\}^\tau \left\{ \mathbf{Y}_s - \sum_{j=1}^{p_s} \mathbf{B}_{sj} \theta_{sj} \right\},$$

where $\mathbf{Y}_s = (Y_{s1}, \dots, Y_{sT})^\tau$, and $\mathbf{B}_{sj} = (\mathbf{B}_{sj}(X_{s1j}), \dots, \mathbf{B}_{sj}(X_{sTj}))^\tau$, with $\mathbf{B}_{sj}(X_{stj}) = (B_{sj1}(X_{stj}), \dots, B_{sjN_T}(X_{stj}))^\tau$. Hence, $\alpha_{sj}(x)$ may be estimated by

$$\hat{\alpha}_{sj}(x) = \{\mathbf{B}_{sj}(x)\}^\tau \hat{\theta}_{sj},$$

with $\mathbf{B}_{sj}(x) = (B_{sj1}(x), \dots, B_{sjN_T}(x))^\tau$.

To develop asymptotic properties of these initial estimators, we first introduce some notations and technical assumptions. The following definitions and notations are adopted from Fan and Yao (2003, Ch. 2).

Definition 1. A sequence of random vectors $\{\mathbf{Z}_t, t = 0, \pm 1, \pm 2, \dots\}$ is said to be strictly stationary if $\{\mathbf{Z}_1, \dots, \mathbf{Z}_T\}$ and $\{\mathbf{Z}_{1+k}, \dots, \mathbf{Z}_{T+k}\}$ have the same joint distribution for any integer $T \geq 1$ and integer k .

Denote \mathcal{F}_i^j as the σ -algebra generated by events $\{\mathbf{Z}_i, i \leq t \leq j\}$, and let $\mathcal{L}^2(\mathcal{F}_i^j)$ consist of \mathcal{F}_i^j -measurable random variables all with finite second moment. Intuitively, \mathcal{F}_i^j assembles all information on the sequence collected between time i and j . Define

$$\alpha(T) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_T^\infty} |P(A)P(B) - P(AB)|.$$

Assumption 2.1. For any fixed s and j , the random variable X_{stj} has a bounded support on $[0, 1]$, and its density function $p_{sj}(\cdot)$ is Lipschitz continuous and bounded away from 0 on its support.

Assumption 2.2. The function $\alpha_{sj}(\cdot)$ has continuous second derivative in $[0, 1]$.

Assumption 2.3. For any fixed s , the sequence of random vectors $(X_{st1}, \dots, X_{stp_s}, \boldsymbol{\varepsilon}_{st})^T, t = 1, 2, \dots$, is strictly stationary and satisfies the following mixing condition for the α -mixing process: for some $\delta_1 > 2$ and $\delta_2 > 1 - 2/\delta_1$,

$$\sum_l l^{\delta_2} [\alpha(l)]^{1-2/\delta_1} < \infty, \quad E|\boldsymbol{\varepsilon}_{s1}|^{\delta_1} < \infty, \quad P_{(X_{st1}, \dots, X_{stp_s}, \boldsymbol{\varepsilon}_{st}) | \boldsymbol{\varepsilon}_{st}}((x_1, \dots, x_{p_s}, \boldsymbol{\varepsilon}) | \boldsymbol{\varepsilon}) \leq c_1 < \infty,$$

where $p_{(X_{st1}, \dots, X_{stp_s}, \boldsymbol{\varepsilon}_{st}) | \boldsymbol{\varepsilon}_{st}}(\cdot | \cdot)$ is a density function of $(X_{st1}, \dots, X_{stp_s}, \boldsymbol{\varepsilon}_{st})$ conditional on $\boldsymbol{\varepsilon}_{st}$, and c_1 is a constant.

Assumption 2.4. As $T \rightarrow \infty, N_T \rightarrow \infty, N_T = o(T^{1/2})$ and $T^{1/2}N_T^{-4} = o(1)$.

Theorem 1 describes the asymptotic properties of $\hat{\boldsymbol{\theta}}_{sj}$ and $\hat{\alpha}_{sj}(x)$.

Theorem 1. *Let Assumptions 2.1 to 2.4 hold. Then we have as follows:*

- i) $\|\hat{\boldsymbol{\theta}}_{sj} - \boldsymbol{\theta}_{sj}\| = O_p(\sqrt{N_T/T} + N_T^{-2})$ for $s = 1, \dots, m$ and $j = 1, \dots, p_s$, where $\|\cdot\|$ is the Euclidean norm, given by $\|\mathbf{a}\| = \sqrt{\mathbf{a}^T \mathbf{a}}$ for any column vector \mathbf{a} ; and
- ii) $\int_{x \in [0,1]} [\hat{\alpha}_{sj}(x) - \alpha_{sj}(x)]^2 p_{sj}(x) = O_p(N_T/T + N_T^{-4})$ for $s = 1, \dots, m$ and $j = 1, \dots, p_s$.

Remark 1. If we let $N_T = O(T^{1/5})$, then $\int_{x \in [0,1]} [\hat{\alpha}_{sj}(x) - \alpha_{sj}(x)]^2 p_{sj}(x) = O_p(T^{-4/5})$ for $s = 1, \dots, m$ and $j = 1, \dots, p_s$. That is, $\hat{\alpha}_{sj}(x)$ converges to $\alpha_{sj}(x)$ at the optimal rate of convergence.

Notwithstanding the above theoretical development, it is difficult to derive the asymptotic distribution of $\hat{\alpha}_{sj}(\cdot)$. More importantly, this estimator does not take into account the information of the serial and contemporaneous correlations of the disturbances, and thus cannot be efficient. An improved estimation method is developed in the next section.

3. IDENTIFYING THE ERROR STRUCTURE

Traditionally, the order of autoregressive processes is identified by information criterion-based methods. These methods are based on best subset selection or its stepwise variants. However, it is well-known that best subset selection is computationally infeasible when the AR process is of high order. Likewise, these methods are sensitive to small changes in the data and thus are unstable. In this section, we will apply the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) penalty variable selection method to identify the order of the AR process. The SCAD penalty method selects the significant variables and estimates their coefficients simultaneously.

3.1. Determining the Order of the AR Process

The residuals result from the estimator $\hat{\alpha}_{sj}(x)$ are

$$\hat{\epsilon}_{st} = Y_{st} - \hat{\alpha}_{s1}(X_{st1}) - \dots - \hat{\alpha}_{sp_s}(X_{stp_s}), \quad \text{for } s = 1, \dots, m \text{ and } t = 1, \dots, T.$$

Suppose that the true order of the error process is d_s^0 . The penalized least squares approach is based on the objective function

$$\mathcal{L}(\boldsymbol{\varsigma}_s) = \frac{1}{2} \sum_{t=d_s+1}^T (\hat{\epsilon}_{st} - \boldsymbol{\varsigma}_{s1}\hat{\epsilon}_{s,t-1} - \dots - \boldsymbol{\varsigma}_{sd_s}\hat{\epsilon}_{s,t-d_s})^2 + T \sum_{j=1}^{d_s} \lambda_{jT} p_j(|\boldsymbol{\varsigma}_{sj}|), \quad (3.1)$$

where the $p_j(\cdot)$'s are penalty functions, and λ_{jT} 's are tuning parameters that control the model complexity. For purposes of simplicity, we denote $\lambda_{jT} p_j(\cdot)$ by $p_{\lambda_{jT}}(\cdot)$.

Antoniadis and Fan (2001) and Fan and Li (2001) maintained that a good penalty function should yield an estimator satisfying the properties of unbiasedness, sparsity (i.e., assigning a zero estimate to a coefficient close to zero to reduce model complexity), as well as continuity to avoid the unnecessary variations in model prediction. Fan and Li (2001) proposed the SCAD penalty $p_\lambda(\cdot)$, which differs from the L_q penalties in that it can produce estimators that possess all of the above three properties simultaneously. The SCAD penalty has the first derivative

$$p'_\lambda(\chi) = \lambda \{ I(\chi \leq \lambda) + \frac{(a\lambda - \chi)_+}{(a - 1)\lambda} I(\chi > \lambda) \}, \quad \text{for some } a > 2 \text{ and } \chi > 0,$$

with $p_\lambda(0) = 0$. The SCAD penalty $p_\lambda(\cdot)$ involves two unknown parameters, λ and a . Fan and Li (2001) suggested setting $a = 3.7$ from a Bayesian point of view.

For our purpose, we assume that $p_{\lambda_{jT}}(\cdot)$'s are negative and nondecreasing, and $p_{\lambda_{jT}}(0) = 0$. Denote

$$a_T = \max_j \left\{ |p'_{\lambda_{jT}}(|\mathbf{s}_{sj}|)| : \mathbf{s}_{sj} \neq 0 \right\}, \quad \text{and} \quad b_T = \max_j \left\{ |p''_{\lambda_{jT}}(|\mathbf{s}_{sj}|)| : \mathbf{s}_{sj} \neq 0 \right\}.$$

We have the following theorem.

Theorem 2. *Let Assumptions 2.1 to 2.4 hold. If a_T and b_T tend to zero as $T \rightarrow \infty$, then with probability tending to one, there exists a local minimizer $\hat{\boldsymbol{\zeta}}_s$ of $\mathcal{L}(\boldsymbol{\zeta}_s)$ such that $\|\hat{\boldsymbol{\zeta}}_s - \boldsymbol{\zeta}_s\| = O_p(T^{-1/2} + a_T)$, where $\boldsymbol{\zeta}_s = (\boldsymbol{\zeta}_{s1}, \dots, \boldsymbol{\zeta}_{sd_s})^\tau$ and $\hat{\boldsymbol{\zeta}}_s = (\hat{\boldsymbol{\zeta}}_{s1}, \dots, \hat{\boldsymbol{\zeta}}_{sd_s})^\tau$.*

The minimizer $\hat{\boldsymbol{\zeta}}_s$ is called the penalized residual-based least squares estimator of $\boldsymbol{\zeta}_s$. Theorem 2 exhibits the manner in which the rate of convergence of $\hat{\boldsymbol{\zeta}}_s$ depends on λ_{jT} . Specifically, to achieve the \sqrt{T} convergence rate, λ_{jT} must be sufficiently small in order for $a_T = O(T^{-1/2})$. To further study the properties of $\hat{\boldsymbol{\zeta}}_s$, we assume, without the loss of generality, that the first d_s^0 components of the true $\boldsymbol{\zeta}_s$ are nonzero, and all other components are 0. Also, define

$$\mathbf{D} = \text{diag} \left\{ p''_{\lambda_{1T}}(|\mathbf{s}_{s1}|), \dots, p''_{\lambda_{d_s^0 T}}(|\mathbf{s}_{sd_s^0}|) \right\} \quad \text{and} \\ \mathbf{b} = \left(p'_{\lambda_{1T}}(|\mathbf{s}_{s1}|) \text{sgn}(\boldsymbol{\zeta}_{s1}), \dots, p'_{\lambda_{d_s^0 T}}(|\mathbf{s}_{sd_s^0}|) \text{sgn}(\boldsymbol{\zeta}_{sd_s^0}) \right),$$

and let $\boldsymbol{\zeta}_s^{(1)}$ and $\boldsymbol{\zeta}_s^{(2)}$ contain the first d_s^0 and last $d_s - d_s^0$ components of $\boldsymbol{\zeta}_s$, respectively, and $\hat{\boldsymbol{\zeta}}_s^{(1)}$ and $\hat{\boldsymbol{\zeta}}_s^{(2)}$ contain the first d_s^0 and last $d_s - d_s^0$ components of $\hat{\boldsymbol{\zeta}}_s$, respectively.

Theorem 3 (Oracle property). *For $j = 1, \dots, d_s$, assume that as $T \rightarrow \infty$, $\lambda_{jT} \rightarrow 0$, and $\sqrt{T}\lambda_{jT} \rightarrow \infty$, and that $p_{\lambda_{jT}}(\cdot)$ satisfies*

$$\liminf_{T \rightarrow \infty} \liminf_{\mathbf{s}_{sj} \rightarrow 0^+} p_{\lambda_{jT}}(\mathbf{s}_{sj}) / \lambda_{jT} > 0.$$

Suppose that Assumptions 2.1 to 2.4 hold, and $b_T \rightarrow 0$. If $a_T = O(T^{-1/2})$, then with probability tending to 1, the local minimizer $\hat{\boldsymbol{\zeta}}_s = (\hat{\boldsymbol{\zeta}}_s^{(1)\tau}, \hat{\boldsymbol{\zeta}}_s^{(2)\tau})^\tau$ satisfies the following properties:

- (i) **Sparsity:** $\hat{\boldsymbol{\zeta}}_s^{(2)} = \mathbf{0}_{d_s - d_s^0}$;
- (ii) **Asymptotic normality:**

$$\sqrt{T} \{ \boldsymbol{\Sigma}_{s11} + \mathbf{D} \} \left[\hat{\boldsymbol{\zeta}}_s^{(1)} - \boldsymbol{\zeta}_s^{(1)} + \{ \boldsymbol{\Sigma}_{s11} + \mathbf{D} \}^{-1} \mathbf{b} \right] \xrightarrow{D} N(0, \boldsymbol{\Sigma}_{s11}),$$

where Σ_{s11} consists of the first d_s^0 rows and columns of Σ_s , with

$$\Sigma_s = \begin{bmatrix} \gamma_s(0) & \gamma_s(1) & \dots & \gamma_s(d_s - 1) \\ \gamma_s(1) & \gamma_s(0) & \dots & \gamma_s(d_s - 2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_s(d_s - 1) & \gamma_s(d_s - 2) & \dots & \gamma_s(0) \end{bmatrix} \quad \text{and} \quad \gamma_s(l) = E(\varepsilon_{s1} \varepsilon_{s,1+l}).$$

In other words, provided that the assumptions stated under Theorem 3 are satisfied, the penalized least squares procedure would select the correct covariates and estimates the unknown coefficients as efficiently as if the true model were known in advance. If the SCAD penalty function is applied to all cases, then $a_T = 0$ when T is sufficiently large, and $a_T = O(T^{-1/2})$ is satisfied by default.

On the other hand, it is challenging to find a minimizer to the penalized weighted partial least squares objective function of (3.1) because the SCAD penalty is irregular at the origin, and its second derivative may not exist at certain points. To reconcile this difficulty, we locally approximate the penalty function by quadratic functions as in Fan and Li (2001). The procedure is as follows. Let ς_s^0 be an initial value close to the minimizer of (3.1). If $|\varsigma_{sj}^0| \geq c_0$ (a predetermined value), then we approximate $p_{\lambda_{jT}}(\cdot)$ using the relationship

$$[p_{\lambda_{jT}}(|\varsigma_{sj}|)]' = p'_{\lambda_{jT}}(|\varsigma_{sj}|) \text{sgn}(\varsigma_{sj}) \approx \left\{ p'_{\lambda_{jT}}(|\varsigma_{sj}^0|) / |\varsigma_{sj}^0| \right\} \varsigma_{sj}.$$

The Newton–Raphson algorithm is then implemented on the approximated penalty function to minimize $\mathcal{L}(\varsigma_s)$.

The tuning parameters $(\lambda_{1T}, \dots, \lambda_{d_s T})$ control the model complexity and can be selected by data-driven methods such as the Bayesian Information Criterion (BIC) (Wang et al., 2007), or the Generalized Information Criterion (GIC) (Zhang et al., 2010).

3.2. Estimation of the Error Variance and Contemporaneous Correlation Coefficients

By our assumption, $E(e_{s_1 t} e_{s_2 t}) = \sigma_{e_{s_1 s_2}}^2$ and $E(e_{s_1 t_1} e_{s_2 t_2}) = 0$ for $s_1, s_2 = 1, \dots, m$; $t, t_1, t_2 = 1, \dots, T$ and $t_1 \neq t_2$. Thus, by using

$$\hat{e}_{st} = \hat{\varepsilon}_{st} - \hat{\varsigma}_{s1} \hat{\varepsilon}_{s,t-1} - \dots - \hat{\varsigma}_{s\hat{d}_s} \hat{\varepsilon}_{s,t-\hat{d}_s}, \quad s = 1, \dots, m, \quad t = \hat{d}_s + 1, \dots, T,$$

we can estimate $\sigma_{e_{s_1 s_2}}^2$ by

$$\hat{\sigma}_{e_{s_1 s_2}}^2 = \frac{1}{T - \max(\hat{d}_{s_1} + 1, \hat{d}_{s_2} + 1)} \sum_{t=\max(\hat{d}_{s_1} + 1, \hat{d}_{s_2} + 1)}^T \hat{e}_{s_1 t} \hat{e}_{s_2 t},$$

where $\hat{d}_s = \max_{1 \leq j \leq d_s} \{\hat{s}_{sj} \neq 0\}$. For $\widehat{\Sigma}_e = (\hat{\sigma}_{e_{s_1 s_2}}^2)_{s_1, s_2=1}^m$, we have the following asymptotic result.

Theorem 4. *Suppose that all of the assumptions stated in Theorem 3 hold. Then*

$$\sqrt{T}(\text{Vech}(\widehat{\Sigma}_e) - \text{Vech}(\Sigma_e)) \rightarrow_D N(0, \mathbf{L}_m \text{Cov}(\mathbf{e}_t \otimes \mathbf{e}_t) \mathbf{L}_m^\tau), \quad \text{as } T \rightarrow \infty,$$

where *Vech* is a column stacking operator that stacks only the elements on and below the main diagonal of a matrix, \mathbf{L}_m is the $\frac{1}{2}m(m+1) \times m^2$ elimination matrix, and $\mathbf{e}_t = (e_{1t}, \dots, e_{mt})^\tau$.

3.3. Estimation of Autoregressive Coefficients

The estimator $\hat{\zeta}_s$ does not take the into account the contemporaneous correlations across equations. Here, we propose an improved estimator of the autoregressive coefficients $\zeta = (\zeta_1^\tau, \dots, \zeta_m^\tau)^\tau$ along the lines of weighted least squares estimation (Zellner, 1962). Let $\hat{d} = \max\{\hat{d}_s, 1 \leq s \leq m\}$,

$$\hat{\boldsymbol{\epsilon}}_s^* = \begin{bmatrix} \hat{\boldsymbol{\epsilon}}_{s, \hat{d}} & \hat{\boldsymbol{\epsilon}}_{s, \hat{d}-1} & \cdots & \hat{\boldsymbol{\epsilon}}_{s, (\hat{d}-\hat{d}_s)+1} \\ \hat{\boldsymbol{\epsilon}}_{s, \hat{d}+1} & \hat{\boldsymbol{\epsilon}}_{s, \hat{d}} & \cdots & \hat{\boldsymbol{\epsilon}}_{s, (\hat{d}-\hat{d}_s)+2} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\boldsymbol{\epsilon}}_{s, T-1} & \hat{\boldsymbol{\epsilon}}_{s, T-2} & \cdots & \hat{\boldsymbol{\epsilon}}_{s, T-\hat{d}_s} \end{bmatrix},$$

and $\hat{\boldsymbol{\epsilon}}_s^{**} = (\hat{\boldsymbol{\epsilon}}_{s, \hat{d}+1}, \dots, \hat{\boldsymbol{\epsilon}}_{s, T})^\tau$. The improved estimator of $\zeta = (\zeta_1^\tau, \dots, \zeta_m^\tau)^\tau$ has the form

$$\hat{\zeta}^w = (\hat{\zeta}_1^{w\tau}, \dots, \hat{\zeta}_m^{w\tau})^\tau = \{\text{diag}(\hat{\boldsymbol{\epsilon}}_1^*, \dots, \hat{\boldsymbol{\epsilon}}_m^*)^\tau (\widehat{\Sigma}_e \otimes \mathbf{I}_{T-\hat{d}}) [\text{diag}(\hat{\boldsymbol{\epsilon}}_1^*, \dots, \hat{\boldsymbol{\epsilon}}_m^*)]\}^{-1} \cdot \text{diag}(\hat{\boldsymbol{\epsilon}}_1^*, \dots, \hat{\boldsymbol{\epsilon}}_m^*) (\widehat{\Sigma}_e \otimes \mathbf{I}_{T-\hat{d}}) \begin{pmatrix} \hat{\boldsymbol{\epsilon}}_1^{**} \\ \vdots \\ \hat{\boldsymbol{\epsilon}}_m^{**} \end{pmatrix}$$

with $\widehat{\Sigma}_e = (\hat{\sigma}_{e_{s_1 s_2}}^2)_{s_1, s_2=1}^m$.

Denote $\gamma_{s_1 s_2}(l) = E(\boldsymbol{\epsilon}_{s_1 t} \boldsymbol{\epsilon}_{s_2, t+l})$ for $s_1, s_2 = 1, \dots, m$, and let

$$\Sigma_{s_1 s_2} = \begin{bmatrix} \gamma_{s_1 s_2}(0) & \gamma_{s_1 s_2}(1) & \cdots & \gamma_{s_1 s_2}(d_{s_2}^0 - 1) \\ \gamma_{s_1 s_2}(1) & \gamma_{s_1 s_2}(0) & \cdots & \gamma_{s_1 s_2}(d_{s_2}^0 - 2) \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{s_1 s_2}(d_{s_1}^0 - 1) & \gamma_{s_1 s_2}(d_{s_1}^0 - 2) & \cdots & \gamma_{s_1 s_2}(d_{s_1}^0 - d_{s_2}^0) \end{bmatrix}$$

and

$$\Sigma^w = (\sigma_e^{s_1 s_2} \Sigma_{s_1 s_2})_{s_1, s_2=1}^m,$$

with $\Sigma_e^{-1} = [(\hat{\sigma}_{e_{s_1 s_2}}^2)_{s_1, s_2=1}^m]^{-1} = (\hat{\sigma}_{e_{s_1 s_2}}^2)_{s_1, s_2=1}^m$. Then we have the following theorem relating to some of the asymptotic properties of the estimator $\hat{\xi}^w = (\hat{\xi}_1^{w\tau}, \dots, \hat{\xi}_m^{w\tau})^\tau$.

Theorem 5. *Suppose that all of the assumptions in Theorem 3 hold.*

- (i) $\sqrt{T} [(\hat{\xi}_1^{w\tau}, \dots, \hat{\xi}_m^{w\tau})^\tau - (\xi_1^\tau, \dots, \xi_m^\tau)^\tau] \xrightarrow{D} N(0, (\Sigma^w)^{-1})$ as $T \rightarrow \infty$.
- (ii) $\hat{\xi}_s^w$ is asymptotically more efficient than $\hat{\xi}_s$, for $s = 1, \dots, m$.

The implementation of Theorem 5 requires a consistent estimator of Σ^w . Let this estimator be

$$\hat{\Sigma}^w = (\hat{\sigma}_{e_{s_1 s_2}}^{s_1 s_2} \hat{\Sigma}_{s_1 s_2})_{s_1, s_2=1}^m,$$

where

$$\hat{\Sigma}_{s_1 s_2} = \begin{bmatrix} \hat{\gamma}_{s_1 s_2}(0) & \hat{\gamma}_{s_1 s_2}(1) & \dots & \hat{\gamma}_{s_1 s_2}(\hat{d}_{s_2} - 1) \\ \hat{\gamma}_{s_1 s_2}(1) & \hat{\gamma}_{s_1 s_2}(0) & \dots & \hat{\gamma}_{s_1 s_2}(\hat{d}_{s_2} - 2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\gamma}_{s_1 s_2}(\hat{d}_{s_1} - 1) & \hat{\gamma}_{s_1 s_2}(\hat{d}_{s_1} - 2) & \dots & \hat{\gamma}_{s_1 s_2}(\hat{d}_{s_1} - \hat{d}_{s_2}) \end{bmatrix},$$

with $\hat{\Sigma}_e^{-1} = [(\hat{\sigma}_{e_{s_1 s_2}}^2)_{s_1, s_2=1}^m]^{-1} = (\hat{\sigma}_{e_{s_1 s_2}}^2)_{s_1, s_2=1}^m$ and $\hat{\gamma}_{s_1 s_2}(l) = \frac{1}{T-\hat{d}} \sum_{t=1}^{T-\hat{d}} \hat{\varepsilon}_{s_1 t} \hat{\varepsilon}_{s_2 t+l}$.

The following theorem shows that $\hat{\Sigma}^w$ is a consistent estimator of Σ^w .

Theorem 6. *Suppose that all of the assumptions in Theorem 3 hold. Then $\hat{\Sigma}^w \rightarrow_p \Sigma^w$ as $T \rightarrow \infty$.*

The estimators of $(\xi_1^\tau, \dots, \xi_m^\tau)^\tau$ and Σ_e facilitate the construction of efficient estimators of the unknown functions.

4. TWO-STAGE LOCAL POLYNOMIAL ESTIMATION

With the results developed in Sections 2 and 3, we are now ready to construct efficient estimators of the unknown functions by a two-stage local polynomial estimation procedure that takes full account of the contemporaneous and serial correlations of the residuals. Our two-stage procedure combines prewhitening transformation (Xiao et al., 2003; Liu et al., 2010) with the seemingly local linear estimation method developed by You et al. (2007). We show that the estimator obtained from this two-stage method is asymptotically normal, as well as being more efficient than the estimator that neglects the contemporaneous and/or serial correlations.

To describe our procedure, assume, temporarily, that d_s^0 , $(\zeta_1^\tau, \dots, \zeta_m^\tau)^\tau$ and $(\alpha_{s1}(\cdot), \dots, \alpha_{sp_s}(\cdot))^\tau$ are known, $s = 1, \dots, m$. We can then construct a pseudo response with a mean of $\alpha_{sj}(X_{stj})$ and an error which has a smaller variance than that of e_{st} . Let

$$Y_{st}^* = Y_{st} - \mathbf{s}_{s1} \left(Y_{s,t-1} - \sum_{j=1}^{p_s} \alpha_{sj}(X_{s,t-1,j}) \right) - \dots - \mathbf{s}_{sd_s^0} \left(Y_{s,t-d_s^0} - \sum_{j=1}^{p_s} \alpha_{sj}(X_{s,t-d_s^0,j}) \right),$$

$s = 1, \dots, m$; $t = d_s^0 + 1, \dots, T$. Denote $\mathbf{Y}_{\cdot t}^* = (Y_{1t}^*, \dots, Y_{mt}^*)^\tau$, $\mathbf{e}_t = (e_{1t}, \dots, e_{mt})^\tau$, $\Sigma_e = E(\mathbf{e}_t \mathbf{e}_t^\tau) = (\sigma_{e_{s_1 s_2}}^2)_{s_1, s_2=1}^m$, and $\Sigma_e^{-1} = (\sigma_e^{s_1 s_2})_{s_1, s_2=1}^m$. The s th element of $\Sigma_e^{-1} \mathbf{Y}_{\cdot t}^*$ has the form

$$\sum_{s_1=1}^m \sigma_e^{s s_1} Y_{s_1 t}^* = \sigma_e^{ss} Y_{st}^* + \sum_{s_1 \neq s} \sigma_e^{s s_1} Y_{s_1 t}^*.$$

It is easy to see that

$$Y_{stj}^{**} = (\sigma_e^{ss})^{-1} \left(\sigma_e^{ss} Y_{st}^* + \sum_{s_1 \neq s} \sigma_e^{s s_1} Y_{s_1 t}^* - \sigma_e^{ss} \sum_{j_1 \neq j} \alpha_{s j_1}(X_{stj_1}) - \sum_{s_1 \neq s} \sigma_e^{s s_1} \sum_{j_2=1}^{p_{s_1}} \alpha_{s_1 j_2}(X_{s_1 t j_2}) \right)$$

has mean $\alpha_{sj}(X_{stj})$ and variance $(\sigma_e^{ss})^{-1}$, which is smaller than $\sigma_{e_{ss}}^2$. Now, we can apply the local polynomial estimation method to Y_{stj}^{**} to construct estimators of the unknown functions of the mean in model (1.1). Fan and Gijbels (1996) showed that the local polynomial smoother has more desirable properties than the kernel estimator. For example, it has smaller bias than the Nadaraya–Watson estimator and smaller variance than the Gasser–Müller estimator; it also adapts automatically to the boundary of design points and requires no boundary modification, as well as being design adaptive.

Let $d^0 = \max_{1 \leq s \leq m} \{d_s^0\}$. For any X_{stj} in a close neighborhood of x , $\alpha_{sj}(X_{stj})$ can be approximated by

$$\alpha_{sj}(X_{stj}) \approx \alpha_{sj}(x) + \mathcal{D}\alpha_{sj}(x)(X_{stj} - x) \equiv a_{sj} + b_{sj}(X_{stj} - x),$$

where $\mathcal{D}\alpha_{sj}(x) = \partial\alpha_{sj}(x)/\partial x$. This approximation results in the following local least squares problem (see, for example, Fan and Gijbels, 1996).

Find $\{(a_{sj}, b_{sj})\}$ to minimize

$$\sum_{t=d^0+1}^T [Y_{stj}^{**} - \{a_{sj} + b_{sj}(X_{stj} - x)\}]^2 K_{h_T}(X_{stj} - x), \tag{4.1}$$

where $K(\cdot)$ is a kernel function, h_T is a bandwidth, and $K_{h_T}(\cdot) = h_T^{-1}K(\cdot/h_T)$. Straightforward algebra yields

$$(\hat{a}_{sj}^{*TS}, \hat{b}_{sj}^{*TS})^\tau = (\mathbf{D}_{sjx}^\tau \mathbf{W}_{sjx} \mathbf{D}_{sjx})^{-1} \mathbf{D}_{sjx}^\tau \mathbf{W}_{sjx} \mathbf{Y}_{sj}^{**}$$

as the solution to (4.1), where $\mathbf{Y}_{sj}^{**} = (Y_{s,d^0+1,j}^{**}, \dots, Y_{sTj}^{**})^\tau$, $\mathbf{D}_{sjx} = \begin{pmatrix} 1 & (X_{s,d^0+1,j}-x) \\ \vdots & \vdots \\ 1 & (X_{sTj}-x) \end{pmatrix}$, and

$$\mathbf{W}_{sjx} = \text{diag}(K_{h_T}(X_{s,d^0+1,j} - x), \dots, K_{h_T}(X_{sTj} - x)).$$

The unknown function $\alpha_{sj}(x)$ can be estimated by

$$\hat{\alpha}_{sj}^{*TS}(x) = \frac{1}{(T - d^0)h_T} \sum_{t=1}^{T-d^0} K_{sj}^* \left(\frac{X_{stj} - x}{h_T}, x \right) Y_{stj}^{**},$$

where

$$K_{sj}^*(x_1, x_2) = (1, 0)(\mathbf{S}_{sj}(x_2))^{-1}(1, x_1)^\tau K_{h_T}(x_1),$$

$\mathbf{S}_{sj}(x)$ is a 2×2 matrix with its (i_1, i_2) th element being $s_{sj,i_1+i_2-2}(x)$, and

$$s_{sj,i_1+i_2-2}(x) = \frac{1}{(T - d^0)} \sum_{t=1}^{T-d^0} \left(\frac{X_{stj} - x}{h_T} \right)^{i_1+i_2-2} K_{h_T}(X_{stj} - x), \quad i_1, i_2 = 1, 2.$$

However, this estimator is of no practical utility because $\hat{\alpha}_{sj}^{*TS}(x)$ depends on d_s^0 and Y_{st}^{**} depends on $\alpha_{sj}(\cdot)$, $(\xi_1^\tau, \dots, \xi_m^\tau)^\tau$ and $\widehat{\Sigma}_e$. To overcome this difficulty, we estimate d^0 by $\hat{d}^0 = \max_{1 \leq s \leq m} \{\hat{d}_s^0\}$, $\alpha_{sj}(\cdot)$ by $\hat{\alpha}_{sj}(\cdot)$, $(\xi_1^\tau, \dots, \xi_m^\tau)^\tau$ by $(\hat{\xi}_1^{w\tau}, \dots, \hat{\xi}_m^{w\tau})^\tau$, and Σ_e by $\widehat{\Sigma}_e$. Then, $(\alpha_{sj}(x), \alpha'_{sj}(x))$ may be estimated by

$$(\hat{a}_{sj}^{TS}, \hat{b}_{sj}^{TS})^\tau = (\mathbf{D}_{sjx}^\tau \mathbf{W}_{sjx} \mathbf{D}_{sjx})^{-1} \mathbf{D}_{sjx}^\tau \mathbf{W}_{sjx} \widehat{\mathbf{Y}}_{sj}^{**},$$

and our two-stage estimator of $\alpha_j(x_0)$ is

$$\hat{\alpha}_{sj}^{TS}(x) = \frac{1}{(T - \hat{d}^0)h_T} \sum_{t=1}^{T-\hat{d}^0} K_s^* \left(\frac{X_{stj} - x}{h_T}, x \right) \widehat{Y}_{stj}^{**}, \quad s = 1, \dots, m, \quad \text{and } j = 1, \dots, p_s,$$

where

$$\begin{aligned} \widehat{\mathbf{Y}}_{stj}^{**} &= (\hat{\sigma}_e^{ss})^{-1} \left(\hat{\sigma}_e^{ss} \widehat{Y}_{st}^* + \sum_{s_1 \neq s}^m \hat{\sigma}_e^{ss_1} \widehat{Y}_{s_1 t}^* - \hat{\sigma}_e^{ss} \sum_{j_1 \neq j}^{p_s} \hat{\alpha}_{s j_1}(X_{stj_1}) - \sum_{s_1 \neq s}^m \hat{\sigma}_e^{ss_1} \sum_{j_2=1}^{p_{s_1}} \hat{\alpha}_{s_1 j_2}(X_{s_1 t j_2}) \right), \\ \widehat{Y}_{st}^* &= Y_{st} - \hat{\varsigma}_{s1}^w \left(Y_{s,t-1} - \sum_{j=1}^{p_s} \hat{\alpha}_{sj}(X_{s,t-1,j}) \right) - \dots - \hat{\varsigma}_{sd_s}^w \left(Y_{s,t-\hat{d}_s} - \sum_{j=1}^{p_s} \hat{\alpha}_{sj}(X_{s,t-\hat{d}_s,j}) \right), \\ t &= \hat{d}^0 + 1, \dots, T \end{aligned}$$

and $\widehat{\mathbf{Y}}_{sj}^{**} = (\widehat{Y}_{s,d^0+1,j}^{**}, \dots, \widehat{Y}_{sTj}^{**})^\tau$.

In order to develop asymptotic properties of $\hat{\alpha}_{sj}^{TS}(x)$, we require the following technical assumptions.

Assumption 4.1. $\kappa = c_1 T^{1/5} / (\log T)$ for some constant c_1 .

Assumption 4.2. The kernel function $K(\cdot)$ is a density function with a compact support.

Assumption 4.3. The bandwidth $h_T = c_2 T^{-1/5}$ for some constant c_2 .

Now, define

$$\mu_k = \int_{-\infty}^{\infty} x^k K(x) dx, \quad \text{and} \quad \nu_k = \int_{-\infty}^{\infty} x^k K^2(x) dx, \quad k = 0, 1, 2, 3.$$

Then for $(\hat{\alpha}_{sj}^{TS}(x), \hat{\alpha}'_{sj}{}^{TS}(x))^\tau = (\hat{a}_{sj}^{TS}, \hat{b}_{sj}^{TS})^\tau$, we have the following theorem.

Theorem 7. *Let Assumptions 4.1 to 4.3 and all of the assumptions in Theorem 3 hold. Then*

$$\sqrt{Th_T} \left[\mathbf{H}_T^{-1} \left\{ \begin{pmatrix} \hat{\alpha}_{sj}^{TS}(x) \\ \hat{\alpha}'_{sj}{}^{TS}(x) \end{pmatrix} - \begin{pmatrix} \alpha_{sj}(x) \\ \alpha'_{sj}(x) \end{pmatrix} \right\} - \frac{h_T^2}{2} \begin{pmatrix} \mathfrak{S}_1 \alpha''_{sj}(x) \\ \mathfrak{S}_2 \alpha''_{sj}(x) \end{pmatrix} + o(h_T^2) \right] \xrightarrow{D} N(0, \boldsymbol{\Sigma}_{(\alpha_{sj}, \alpha'_{sj})}^{TS}),$$

as $T \rightarrow \infty$, where $\mathbf{H}_T = \text{diag}(1, h_T)$, $\alpha''_{sj}(x) = \partial^2 \alpha_{sj}(x) / \partial x^2$,

$$\begin{aligned} \boldsymbol{\Sigma}_{(\alpha_{sj}, \alpha'_{sj})}^{TS} &= (\sigma_e^{ss})^{-1} \{p_{sj}(x)\}^{-1} \begin{pmatrix} \mathfrak{S}_{11} \mathfrak{S}_{12} \\ \mathfrak{S}_{21} \mathfrak{S}_{22} \end{pmatrix}, \quad \mathfrak{S}_1 = \frac{\mu_2 - \mu_1 \mu_3}{\mu_2 - \mu_1^2}, \quad \mathfrak{S}_2 = \frac{\mu_3 - \mu_1 \mu_2}{\mu_2 - \mu_1^2}, \\ \mathfrak{S}_{11} &= \mu_2^2 \nu_0 - 2\mu_1 \mu_2 \nu_1 + \mu_1^2 \nu_2, \quad \mathfrak{S}_{12} = (\mu_1^2 + \mu_2) \nu_1 - \mu_1 \mu_2 \nu_0 - \mu_1 \nu_2, \\ \mathfrak{S}_{21} &= (\mu_1^2 + \mu_2) \nu_1 - \mu_1 \mu_2 \nu_0 - \mu_1 \nu_2, \quad \text{and} \quad \mathfrak{S}_{22} = \nu_2 - \mu_1 (2\nu_1 + \mu_1 \nu_0). \end{aligned}$$

Theorem 7 leads to the following corollary relating to the properties of $\hat{\alpha}_{sj}^{TS}(x)$.

Corollary 1. *Let Assumptions 4.1 to 4.3 and all of the assumptions in Theorem 3 hold. Then*

$$\sqrt{Th_T} \left\{ \hat{\alpha}_{sj}^{TS}(x) - \alpha_{sj}(x) - \frac{h_T^2}{2} \frac{\mu_2^2 - \mu_1 \mu_3}{\mu_2 - \mu_1^2} \alpha''_{sj}(x) + o(h_T^2) \right\} \xrightarrow{D} N(0, \sigma_{\alpha_{sj}^{TS}}) \quad \text{as } T \rightarrow \infty,$$

where

$$\sigma_{\alpha_{sj}}^{TS} = (\sigma_e^{ss})^{-1} \{p_{sj}(x)\}^{-1} (\mathfrak{S}_3^2 v_0 + 2\mathfrak{S}_3 \mathfrak{S}_4 v_1 + \mathfrak{S}_4^2 v_2),$$

with $\mathfrak{S}_3 = \mu_2/(\mu_2 - \mu_1^2)$ and $\mathfrak{S}_4 = -\mu_1/(\mu_2 - \mu_1^2)$.

Remark 2. If we consider only serial correlation but ignore contemporaneous correlation in the disturbances, then the two-stage estimator is denoted by $(\alpha_{sj}(\cdot), \alpha'_{sj}(\cdot))^\tau$, which has the form

$$(\check{\alpha}_{sj}^{TS}(x), \check{\alpha}'_{sj}{}^{\tau}(x))^\tau = (\mathbf{D}_{sjx}^\tau \mathbf{W}_{sjx} \mathbf{D}_{sjx})^{-1} \mathbf{D}_{sjx}^\tau \mathbf{W}_{sjx} \widehat{\mathbf{Y}}_s^*,$$

where $\widehat{\mathbf{Y}}_s^* = (\widehat{Y}_{s,d^0+1}^*, \dots, \widehat{Y}_{sT}^*)^\tau$. Applying Theorem 7, for $(\check{\alpha}_{sj}^{TS}(x), \check{\alpha}'_{sj}{}^{\tau}(x))^\tau$, we have the following asymptotic result:

$$\sqrt{Th_T} \left[\mathbf{H}_T^{-1} \left\{ \begin{pmatrix} \check{\alpha}_{sj}^{TS}(x) \\ \check{\alpha}'_{sj}{}^{\tau}(x) \end{pmatrix} - \begin{pmatrix} \alpha_{sj}(x) \\ \alpha'_{sj}(x) \end{pmatrix} \right\} - \frac{h_T^2}{2} \begin{pmatrix} \mathfrak{S}_1 \alpha''_{sj}(x) \\ \mathfrak{S}_2 \alpha_{sj}(x) \end{pmatrix} + o(h_T^2) \right] \xrightarrow{D} N(0, \check{\Sigma}_{(\alpha_{sj}, \alpha'_{sj})}^{TS})$$

as $T \rightarrow \infty$, where

$$\check{\Sigma}_{(\alpha_{sj}, \alpha'_{sj})}^{TS} = \sigma_{ess}^2 \{p_{sj}(x)\}^{-1} \begin{pmatrix} \mathfrak{S}_{11} & \mathfrak{S}_{12} \\ \mathfrak{S}_{21} & \mathfrak{S}_{22} \end{pmatrix}.$$

Since $\sigma_{ess}^2 \geq (\sigma_e^{ss})^{-1}$, we have $\check{\Sigma}_{(\alpha_{sj}, \alpha'_{sj})}^{TS} \geq \Sigma_{(\alpha_{sj}, \alpha'_{sj})}^{TS}$; that is, $(\hat{\alpha}_{sj}^{TS}(x), \hat{\alpha}'_{sj}{}^{\tau}(x))^\tau$ is asymptotically more efficient than $(\check{\alpha}_{sj}^{TS}(x), \check{\alpha}'_{sj}{}^{\tau}(x))^\tau$.

On the other hand, if we incorporate only serial correlation but ignore contemporaneous correlation, the two-stage estimator is denoted by $(\alpha_{sj}(\cdot), \alpha'_{sj}(\cdot))^\tau$, which has the form

$$(\tilde{\alpha}_{sj}^{TS}(x), \tilde{\alpha}'_{sj}{}^{\tau}(x))^\tau = (\mathbf{D}_{sjx}^\tau \mathbf{W}_{sjx} \mathbf{D}_{sjx})^{-1} \mathbf{D}_{sjx}^\tau \mathbf{W}_{sjx} \widehat{\mathbf{Y}}_s^*,$$

where $\widehat{\mathbf{Y}}_s^* = (\widehat{Y}_{s,d^0+1}^*, \dots, \widehat{Y}_{sT}^*)^\tau$ and

$$\widehat{Y}_{st}^* = (\hat{\sigma}_\varepsilon^{ss})^{-1} \left(\hat{\sigma}_\varepsilon^{ss} Y_{st} + \sum_{s_1 \neq s}^m \hat{\sigma}_\varepsilon^{ss_1} \widehat{Y}_{s_1 t} - \hat{\sigma}_\varepsilon^{ss} \sum_{j_1 \neq j}^{p_s} \hat{\alpha}_{sj_1}(X_{stj_1}) - \sum_{s_1 \neq s}^m \hat{\sigma}_\varepsilon^{ss_1} \sum_{j_2=1}^{p_{s_1}} \hat{\alpha}_{s_1 j_2}(X_{s_1 t j_2}) \right).$$

Applying Theorem 7, for $(\tilde{\alpha}_{sj}^{TS}(x), \tilde{\alpha}'_{sj}{}^{\tau}(x))^\tau$, we have the following asymptotic result:

$$\sqrt{Th_T} \left[\mathbf{H}_T^{-1} \left\{ \begin{pmatrix} \tilde{\alpha}_{sj}^{TS}(x) \\ \tilde{\alpha}'_{sj}{}^{\tau}(x) \end{pmatrix} - \begin{pmatrix} \alpha_{sj}(x) \\ \alpha'_{sj}(x) \end{pmatrix} \right\} - \frac{h_T^2}{2} \begin{pmatrix} \mathfrak{S}_1 \alpha''_{sj}(x) \\ \mathfrak{S}_2 \alpha_{sj}(x) \end{pmatrix} + o(h_T^2) \right] \xrightarrow{D} N(0, \check{\Sigma}_{(\alpha_{sj}, \alpha'_{sj})}^{TS})$$

as $T \rightarrow \infty$, where

$$\tilde{\Sigma}_{(\alpha_{sj}, \alpha'_{sj})}^{TS} = (\sigma_e^{ss})^{-1} \{p_{sj}(x)\}^{-1} \begin{pmatrix} \mathfrak{S}_{11} & \mathfrak{S}_{12} \\ \mathfrak{S}_{21} & \mathfrak{S}_{22} \end{pmatrix}.$$

Because $(\sigma_e^{ss})^{-1} \geq (\sigma_e^{ss})^{-1}$, we have $\tilde{\Sigma}_{(\alpha_{sj}, \alpha'_{sj})}^{TS} \geq \Sigma_{(\alpha_{sj}, \alpha'_{sj})}^{TS}$, that is, $(\hat{\alpha}_{sj}^{TS}(x), \hat{\alpha}'_{sj}{}^T(x))^\tau$ is asymptotically more efficient than $(\tilde{\alpha}_{sj}^{TS}(x), \tilde{\alpha}'_{sj}{}^T(x))^\tau$.

We require a consistent estimator of $\sigma_{\alpha_{sj}^{TS}}$ or $\Sigma_{(\alpha_{sj}, \alpha'_{sj})}^{TS}$ in order to apply Corollary 1 or Theorem 7 to conduct statistical inference for $\alpha_{sj}(\cdot)$ or $(\alpha_{sj}(\cdot), \alpha'_{sj}(\cdot))^\tau$. Since $\mu_1, \mu_2, \mu_3, v_0, v_1$, and v_2 are known constants, we only need to estimate σ_e^{ss} and $p_{sj}(\cdot)$ for $s = 1, \dots, m$ and $j = 1, \dots, p_s$. According to Theorem 4, $\hat{\sigma}_e^{ss}$ is a consistent estimator of σ_e^{ss} . As well, we can use the usual kernel density method to estimate $p_{sj}(\cdot)$, namely,

$$\hat{p}_{sj}(x) = \frac{1}{h_T} \sum_{t=1}^T K_{h_T}(X_{stj} - x).$$

It should be noted that $\hat{\alpha}_{sj}^{TS}(\cdot)$ involves the smoothing parameters h_T and N_T . The asymptotic result of Theorem 7 shows that the smoothing parameter h_T is of the standard order. However, the smoothing parameter N_T of the initial estimators $\hat{\alpha}_{sj}(\cdot)$ should be of order larger than the standard one. That is, undersmoothing is required in the preliminary stage of the estimation. In practice, standard smoothing parameter selection can be used in the second stage. Our simulation results show that the findings are relatively insensitive to the choice of the smoothing parameter N_T , and the usual optimal smoothing parameters multiplied by a constant, say, 1.5 or 2, can be used. Undersmoothing is widely applied in two-stage estimation. See, for instance, Wang and Yang (2007), Horowitz and Mammen (2004), and Liu et al. (2010).

5. A SIMULATION STUDY

In this section, we conduct a simulation study to evaluate the finite sample performance of the proposed estimators. The data are generated from the following nonparametric additive SUR model

$$Y_{st} = \alpha_{s1}(X_{st1}) + \alpha_{s2}(X_{st2}) + \varepsilon_{st}, \quad s = 1, 2 \quad \text{and} \quad t = 1, \dots, T, \quad (5.1)$$

where $X_{st1} = X_{st1}^*/2$ and $X_{st2} = X_{st2}^*/2$ with $X_{st1}^* = 0.2X_{s,t-1,1}^* + \mu_{st1} + v_{t1}$, $X_{st2}^* = -0.1X_{s,t-1,2}^* + \mu_{st2} + v_{t2}$, $\mu_{st1} \sim i.i.d. U(-0.5, 0.5)$, $\mu_{st2} \sim i.i.d. U(-0.5, 0.5)$, $v_{t1} \sim i.i.d. U(-0.5, 0.5)$, $v_{t2} \sim i.i.d. U(-0.5, 0.5)$,

$$\alpha_{11}(X_{1t1}) = 2 \sin(2\pi(X_{1t1} + 0.5)) - 2E \{ \sin(2\pi(X_{1t1} + 0.5)) \},$$

$$\begin{aligned} \alpha_{12}(X_{1t2}) &= 2(0.5X_{1t2} - 1)^2 - 2E\{(0.5X_{1t2} - 1)^2\}, \\ \alpha_{21}(X_{2t1}) &= 2 \cos(2\pi(X_{2t1} + 1.5)) - 2E\{\cos(2\pi(X_{2t1} + 1.5))\}, \\ \alpha_{22}(X_{2t2}) &= g(X_{2t2}) - E\{g(X_{2t2})\} \quad \text{and} \\ g(X_{2t2}) &= 3 [0.1 \sin(0.75\pi X_{2t2}) + 0.2 \cos(0.75\pi X_{2t2}) + 0.3\{\sin(0.75\pi X_{2t2})\}^2 \\ &\quad + 0.4\{\cos(0.75\pi X_{2t2})\}^4 + 0.5\{\sin(0.5\pi X_{2t2})\}^3]. \end{aligned}$$

Additionally, we let $\varepsilon_{st} = \varsigma_s \varepsilon_{s,t-1} + e_{st}$ for $s = 1, 2$ and $t = 1, \dots, T$, $T = 200, 300, 500$, $(\varsigma_1, \varsigma_2) = (0.3, 0.3), (0.5, 0.5)$, and $(0.75, 0.75)$, and $(\sigma_{e11}^2, \sigma_{e12}^2, \sigma_{e22}^2) = \sqrt{0.75}(1, 0, 1)$ and $\sqrt{0.75}(1, 0.5, 1)$.

In each case, we set the number of simulated samples to 1000. We use univariate quadratic polynomial splines with uniform knots to approximate each function $\alpha_{sj}(\cdot)$. As in Wang and Yang (2007), N_T is determined by the sample size T , and the tuning constant is taken to be 0.5. The tuning parameter λ_T is selected by the the Bayesian Information Criterion (e.g., (Wang et al., 2007)). In Table 1, we compare the variable selection results for different T , σ_{e12}^2 and $(\varsigma_1, \varsigma_2)$ over the replicated samples. We see from Table 1 that for each σ_{e12}^2 and $(\varsigma_1, \varsigma_2)$, the percentage of the method selecting the correct model increases as T increases, and approaches 100% very quickly.

TABLE 1
Autoregressive Error Order Selection Results

σ_{e12}^2	$(\varsigma_1, \varsigma_2)$	T	ς_1			ς_2		
			Number of correct-fitting	Number of under-fitting	Number of over-fitting	Number of correct-fitting	Number of under-fitting	Number of over-fitting
0	(0.3,0.3)	200	828	51	121	830	50	120
		300	936	13	51	943	7	50
		500	978	0	22	975	0	25
	(0.5,0.5)	200	866	0	134	870	0	130
		300	919	0	81	913	0	87
		500	980	0	20	973	0	27
	(0.75,0.75)	200	777	0	223	758	0	242
		300	872	0	128	853	0	147
		500	918	0	82	922	0	78
0.5	(0.3,0.3)	200	853	55	92	850	52	98
		300	929	14	57	929	18	53
		500	982	0	18	975	2	23
	(0.5,0.5)	200	857	0	143	855	0	145
		300	917	0	83	919	0	81
		500	978	0	22	974	0	26
	(0.75,0.75)	200	756	0	244	757	0	243
		300	884	0	116	874	0	126
		500	914	0	86	916	0	84

TABLE 2
Finite Sample Performance of Estimators of the Autoregressive Coefficients

$\sigma_{\epsilon_{12}}^2$	(s_1, s_2)	T	\hat{s}_1		\hat{s}_2		\hat{s}_1^w		\hat{s}_2^w	
			<i>sm</i>	<i>std</i>	<i>sm</i>	<i>std</i>	<i>sm</i>	<i>std</i>	<i>sm</i>	<i>std</i>
0	(0.3,0.3)	200	0.2682	0.0670	0.2760	0.0684	0.2680	0.0667	0.2759	0.0682
		300	0.2811	0.0567	0.2826	0.0549	0.2812	0.0565	0.2824	0.0548
		500	0.2825	0.0444	0.2810	0.0426	0.2825	0.0444	0.2810	0.0425
	(0.5,0.5)	200	0.4578	0.0651	0.4522	0.0629	0.4580	0.0650	0.4521	0.0627
		300	0.4603	0.0523	0.4717	0.0522	0.4603	0.0521	0.4718	0.0522
		500	0.4728	0.0408	0.4738	0.0394	0.4729	0.0407	0.4738	0.0394
	(0.75,0.75)	200	0.6837	0.0529	0.6870	0.0535	0.6837	0.0529	0.6871	0.0534
		300	0.7088	0.0417	0.7145	0.0408	0.7089	0.0416	0.7146	0.0408
		500	0.7208	0.0314	0.7243	0.0317	0.7207	0.0314	0.7243	0.0316
0.5	(0.3,0.3)	200	0.2729	0.0677	0.2682	0.0665	0.2679	0.0612	0.2653	0.0604
		300	0.2786	0.0564	0.2794	0.0534	0.2746	0.0512	0.2756	0.0487
		500	0.2857	0.0435	0.2843	0.0426	0.2821	0.0397	0.2811	0.0377
	(0.5,0.5)	200	0.4543	0.0647	0.4480	0.0646	0.4475	0.0590	0.4426	0.0594
		300	0.4685	0.0534	0.4676	0.0537	0.4633	0.0484	0.4617	0.0481
		500	0.4800	0.0382	0.4844	0.0395	0.4761	0.0339	0.4803	0.0349
	(0.75, 0.75)	200	0.6964	0.0538	0.6955	0.0509	0.6884	0.0492	0.6875	0.0464
		300	0.7125	0.0413	0.7132	0.0404	0.7065	0.0388	0.7079	0.0371
		500	0.7204	0.0315	0.7238	0.0308	0.7160	0.0294	0.7183	0.0287

We also calculate the sample means (sms) and standard deviations (stds) of (\hat{s}_1, \hat{s}_2) and $(\hat{s}_1^w, \hat{s}_2^w)$. The results are summarized in Table 2. The results clearly show that taking the contemporaneous correlation into account improves the performance of the estimator of the autoregressive coefficients.

Furthermore, we compute the root average squared error (RASE)

$$RASE(\check{\alpha}_{sj}) = \left[T^{-1} \sum_{t=1}^T \{ \check{\alpha}_{sj}(X_{stj}) - \alpha_{sj}(X_{stj}) \}^2 \right]^{1/2},$$

where $\check{\alpha}_{sj}$ is one of $\hat{\alpha}_{sj}(x)$, which ignores both types of correlations, $\tilde{\alpha}_{sj}^{TS}(x)$, which takes into account only the serial correlation, and $\hat{\alpha}_{sj}^{TS}(x)$ which takes both serial and contemporaneous correlations into account. The performance of these three estimators of the unknown additive function $\alpha_{sj}(x)$ in (5.1) is assessed through the sample means (sms) and standard deviations (stds) of the RASE values. The results are reported in Tables 3. The results show that the estimator that takes both the contemporaneous and serial correlations into account invariably outperforms the estimators that neglect either one or both of the correlations. The improvement is especially significant when (s_1, s_2) and $\sigma_{\epsilon_{12}}^2$ are large.

TABLE 3
Finite Sample Performance of the Estimators of the Unknown Additive Functions

$\sigma_{\epsilon_{12}}^2$	(s_1, s_2)	T	$\alpha_{11}(\cdot)$		$\alpha_{12}(\cdot)$		$\alpha_{21}(\cdot)$		$\alpha_{22}(\cdot)$		
			<i>sm</i> (RASE)	<i>std</i> (RASE)	<i>sm</i> (RASE)	<i>std</i> (RASE)	<i>sm</i> (RASE)	<i>std</i> (RASE)	<i>sm</i> (RASE)	<i>std</i> (RASE)	
0	(0.3,0.3)	200	$\hat{\alpha}$	0.1771	0.0431	0.1556	0.0498	0.1950	0.0445	0.1687	0.0520
			$\hat{\alpha}^{TS}$	0.1681	0.0455	0.1368	0.0432	0.1622	0.0439	0.1377	0.0423
			$\hat{\alpha}^{TS}$	0.1684	0.0458	0.1370	0.0432	0.1625	0.0439	0.1378	0.0427
		300	$\hat{\alpha}$	0.1684	0.0331	0.1426	0.0409	0.1543	0.0388	0.1483	0.0411
			$\hat{\alpha}^{TS}$	0.1554	0.0391	0.1237	0.0351	0.1461	0.0367	0.1225	0.0357
			$\hat{\alpha}^{TS}$	0.1554	0.0393	0.1239	0.0351	0.1460	0.0367	0.1225	0.0357
		500	$\hat{\alpha}$	0.1391	0.0251	0.1141	0.0296	0.1403	0.0287	0.1301	0.0314
			$\hat{\alpha}^{TS}$	0.1218	0.0307	0.0990	0.0273	0.1132	0.0259	0.0969	0.0261
			$\hat{\alpha}^{TS}$	0.1219	0.0308	0.0990	0.0274	0.1132	0.0259	0.0968	0.0261
	(0.5,0.5)	200	$\hat{\alpha}$	0.2180	0.0615	0.1923	0.0628	0.2382	0.0475	0.1902	0.0591
			$\hat{\alpha}^{TS}$	0.1957	0.0583	0.1494	0.0503	0.1787	0.0458	0.1496	0.0451
			$\hat{\alpha}^{TS}$	0.1958	0.0585	0.1495	0.0505	0.1788	0.0459	0.1497	0.0456
		300	$\hat{\alpha}$	0.1777	0.0399	0.1556	0.0441	0.1677	0.0404	0.1532	0.0450
			$\hat{\alpha}^{TS}$	0.1478	0.0391	0.1308	0.0394	0.1478	0.0373	0.1226	0.0343
			$\hat{\alpha}^{TS}$	0.1478	0.0393	0.1307	0.0395	0.1477	0.0372	0.1227	0.0342
		500	$\hat{\alpha}$	0.1397	0.0317	0.1178	0.0363	0.1519	0.0304	0.1270	0.0366
			$\hat{\alpha}^{TS}$	0.1227	0.0345	0.1017	0.0304	0.1170	0.0269	0.0974	0.0288
			$\hat{\alpha}^{TS}$	0.1227	0.0345	0.1017	0.0303	0.1171	0.0269	0.0974	0.0287
	(0.75,0.75)	200	$\hat{\alpha}$	0.2309	0.0683	0.2303	0.0832	0.2338	0.0729	0.2187	0.0684
			$\hat{\alpha}^{TS}$	0.1799	0.0550	0.1853	0.0631	0.1678	0.0486	0.1522	0.0510
			$\hat{\alpha}^{TS}$	0.1802	0.0552	0.1854	0.0634	0.1681	0.0486	0.1523	0.0512
		300	$\hat{\alpha}$	0.2046	0.0561	0.1871	0.0600	0.1933	0.0494	0.1848	0.0569
			$\hat{\alpha}^{TS}$	0.1604	0.0454	0.1337	0.0452	0.1428	0.0367	0.1255	0.0390
			$\hat{\alpha}^{TS}$	0.1605	0.0454	0.1337	0.0453	0.1429	0.0368	0.1256	0.0390
500		$\hat{\alpha}$	0.1764	0.0415	0.1503	0.0468	0.1586	0.0412	0.1513	0.0438	
		$\hat{\alpha}^{TS}$	0.1354	0.0371	0.1025	0.0305	0.1175	0.0266	0.0975	0.0268	
		$\hat{\alpha}^{TS}$	0.1355	0.0372	0.1026	0.0304	0.1175	0.0266	0.0976	0.0268	
0.5	(0.3,0.3)	200	$\hat{\alpha}$	0.1781	0.0473	0.1686	0.0543	0.1764	0.0476	0.1714	0.0506
			$\hat{\alpha}^{TS}$	0.1714	0.0474	0.1557	0.0502	0.1664	0.0435	0.1358	0.0431
			$\hat{\alpha}^{TS}$	0.1568	0.0434	0.1396	0.0461	0.1507	0.0401	0.1204	0.0383
		300	$\hat{\alpha}$	0.1594	0.0316	0.1291	0.0412	0.1476	0.0377	0.1403	0.0404
			$\hat{\alpha}^{TS}$	0.1529	0.0364	0.1193	0.0350	0.1412	0.0340	0.1172	0.0337
			$\hat{\alpha}^{TS}$	0.1362	0.0337	0.1054	0.0309	0.1249	0.0300	0.1043	0.0297
		500	$\hat{\alpha}$	0.1385	0.0239	0.1161	0.0300	0.1212	0.0302	0.1204	0.0325
			$\hat{\alpha}^{TS}$	0.1201	0.0290	0.0952	0.0286	0.1138	0.0269	0.1037	0.0274
			$\hat{\alpha}^{TS}$	0.1083	0.0270	0.0842	0.0248	0.1005	0.0247	0.0918	0.0242
	(0.5,0.5)	200	$\hat{\alpha}$	0.1979	0.0483	0.1758	0.0536	0.1879	0.0542	0.1957	0.0536
			$\hat{\alpha}^{TS}$	0.1764	0.0461	0.1434	0.0437	0.1684	0.0465	0.1467	0.0452
			$\hat{\alpha}^{TS}$	0.1588	0.0421	0.1284	0.0400	0.1555	0.0426	0.1313	0.0403
		300	$\hat{\alpha}$	0.1661	0.0392	0.1479	0.0431	0.1789	0.0384	0.1559	0.0467
			$\hat{\alpha}^{TS}$	0.1491	0.0404	0.1268	0.0354	0.1350	0.0346	0.1233	0.0352
			$\hat{\alpha}^{TS}$	0.1344	0.0359	0.1122	0.0322	0.1223	0.0322	0.1105	0.0317
		500	$\hat{\alpha}$	0.1363	0.0301	0.1158	0.0350	0.1184	0.0315	0.1263	0.0309
			$\hat{\alpha}^{TS}$	0.1207	0.0302	0.1000	0.0278	0.1136	0.0271	0.0992	0.0262
			$\hat{\alpha}^{TS}$	0.1098	0.0289	0.0888	0.0247	0.1022	0.0251	0.0883	0.0247
	(0.75,0.75)	200	$\hat{\alpha}$	0.2314	0.0675	0.1969	0.0651	0.2178	0.0652	0.2224	0.0719
			$\hat{\alpha}^{TS}$	0.1913	0.0557	0.1508	0.0445	0.1684	0.0454	0.1524	0.0498
			$\hat{\alpha}^{TS}$	0.1741	0.0516	0.1363	0.0407	0.1554	0.0422	0.1400	0.0461
		300	$\hat{\alpha}$	0.1994	0.0544	0.1841	0.0575	0.1891	0.0540	0.1980	0.0604
			$\hat{\alpha}^{TS}$	0.1529	0.0394	0.1326	0.0411	0.1440	0.0379	0.1275	0.0398
			$\hat{\alpha}^{TS}$	0.1380	0.0358	0.1203	0.0382	0.1341	0.0371	0.1162	0.0372
500		$\hat{\alpha}$	0.1936	0.0366	0.1595	0.0478	0.1496	0.0438	0.1452	0.0415	
		$\hat{\alpha}^{TS}$	0.1355	0.0377	0.1100	0.0333	0.1167	0.0284	0.1033	0.0296	
		$\hat{\alpha}^{TS}$	0.1222	0.0332	0.0999	0.0300	0.1063	0.0266	0.0942	0.0270	

6. CONCLUDING REMARKS

In this article, we have developed a nonparametric additive SUR model with autoregressive errors, and an inference procedure for the model. Our procedure first estimates the unknown functions by combining the polynomial spline series approximation and least squares, then uses the estimated residuals and the SCAD penalty to fit the error structure. Based on the polynomial spline series estimator and the fitted error structure, a two-stage local polynomial estimator for the unknown functions of the mean is further proposed to improve efficiency. Our procedure applies a prewhitening transformation of the dependent variable and the additive structure, and also takes into account the contemporaneous correlations. The resulting estimator has several advantages including the possession an oracle property, and being asymptotically more efficient than the estimators that neglect the autocorrelation and/or contemporaneous correlation of errors.

Compared with nonparametric regressions, parametric or semiparametric regression models can often provide a more parsimonious description of the relationship between the response variable and its covariates. For this reason, one may be interested in checking whether all or some of $\alpha_{sj}(\cdot)$ can be described by a parametric structure. This amounts to testing if all or some of $\alpha_{sj}(\cdot)$'s are in a certain parametric form. Huang et al. (2012) proposed a semiparametric model pursuit method for identifying the covariates with a linear effect. It remains an interesting avenue for further research to extend Huang et al. (2012)'s method to model (1.1)–(1.2) considered here.

APPENDIX: PROOFS OF TECHNICAL RESULTS

To prove the technical results we first introduce some lemmas. The following Lemma 1 is adopted from Chapter 2 of Fan and Yao (2003).

Lemma 1. *Let $(U_1, \varepsilon_1), \dots, (U_T, \varepsilon_T)$ 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 $[0, 1]$,*

$$E|\varepsilon_t|^s < \infty \quad \text{and} \quad \sup_{u \in [0,1]} \int |\varepsilon_t|^s p(u, \varepsilon) d\varepsilon < \infty,$$

where $p(\cdot)$ is the joint density of (U_t, ε_t) . In addition, the conditional density $p_{U_1, U_t | \varepsilon_1, \varepsilon_t}$ satisfies $(u_1, u_t | \varepsilon_1, \varepsilon_t) \leq c_2 < \infty$ for all $t \geq 1$, and $K(\cdot)$ satisfies Assumption 4.2. Then

$$\sup_{u \in [0,1]} \left| \frac{1}{T} \sum_{t=1}^T \{K_{h_T}(U_t - u)\varepsilon_t - E[K_{h_T}(U_t - u)\varepsilon_t]\} \right| = O_p \left(\left\{ \frac{\log T}{Th_T} \right\}^{1/2} \right)$$

Downloaded by [City University of Hong Kong Library] at 17:55 28 April 2016

provided that $h_T \rightarrow 0$, $T^{1-2s^{-1}-2\zeta} h_T \rightarrow \infty$ and $T^{(\tau+1.5)(s^{-1}+\zeta)-\tau/2+5/4} h_T^{-\tau/2-5/4} \rightarrow 0$ as $T \rightarrow \infty$ for some $\zeta > 0$.

Let

$$\mathbf{Z}_t = \begin{pmatrix} K_{h_T}(X_{stj} - x) \\ \frac{X_{stj}-x}{h_T} K_{h_T}(X_{stj} - x) \end{pmatrix} \sum_{s_1=1}^m (\sigma_e^{ss})^{-1} \sigma_e^{ss_1} e_{st}, \quad \mathbf{Q}_T = \frac{1}{T} \sum_{t=1}^T \mathbf{Z}_t \quad \text{and} \quad \boldsymbol{\omega} = \begin{pmatrix} v_0 & v_1 \\ v_1 & v_2 \end{pmatrix}.$$

Lemma 2. Suppose that Assumptions 4.1 to 4.3 hold. Then, as $T \rightarrow \infty$, we have as follows

- (a) $h_T \text{Cov}(\mathbf{Z}_t) \rightarrow (\sigma_e^{ss})^{-1} p_{sj}(x) \boldsymbol{\omega}$;
- (b) $h_T \sum_{t=1}^{T-1} \|\text{Cov}(\mathbf{Z}_t, \mathbf{Z}_{t+1})\| = o(1)$; and
- (c) $h_T \text{Cov}(\mathbf{Q}_T) \rightarrow (\sigma_e^{ss})^{-1} p_{sj}(x) \boldsymbol{\omega}$, where $\|\cdot\|$ denotes the Euclidean norm.

Proof. It is easy to see that \mathbf{Z}_t can be written as

$$\mathbf{Z}_t = \begin{pmatrix} K_{h_T}(X_{stj} - x) \\ \frac{X_{stj}-x}{h_T} K_{h_T}(X_{stj} - x) \end{pmatrix} (0, \dots, (\sigma_e^{ss})^{-1}, \dots, 0) \boldsymbol{\Sigma}_e^{-1} (e_{1t}, \dots, e_{mt})^\tau.$$

Thus, we have

$$h^* \text{Cov}(\mathbf{Z}_t) = (\sigma_e^{ss})^{-1} h_T E \left\{ \begin{pmatrix} K_{h_T}(X_{stj} - x) \\ \frac{X_{stj}-x}{h_T} K_{h_T}(X_{stj} - x) \end{pmatrix} \begin{pmatrix} K_{h_T}(X_{stj} - x) \\ \frac{X_{stj}-x}{h_T} K_{h_T}(X_{stj} - x) \end{pmatrix}^\tau \right\}.$$

Using Theorem 1 of Sun (1984), we can show that

$$\begin{aligned} h_T E \{K_{h_T}(X_{stj} - x) K_{h_T}(X_{stj} - x)\} &= h_T \int p_{sj}(x) \{K_{h_T}(x_{sj} - x)\}^2 dx_{sj} \\ &= \int p_{sj}(x_{sj}^* h_T + x) \{K(x_{sj}^*)\}^2 dx_{sj}^* = p_{sj}(x) v_0 + o(1). \end{aligned}$$

Along the same lines of argument, we can show that

$$h_T E \left\{ \frac{X_{stj} - x}{h_T} (K_{h_T}(X_{stj} - x))^2 \right\} = p_{sj}(x) v_1 + o(1),$$

and

$$h_T E \left\{ \left(\frac{X_{stj} - x}{h_T} \right)^2 (K_{h_T}(X_{stj} - x))^2 \right\} = p_{sj}(x) v_2 + o(1).$$

This proves result (a). The proof of result (b) is same as that of Lemma 1 (b) of Cai et al. (1999). The proof of result (c) follows straightforwardly by using (a) and (b) together with

$$\text{Cov}(\mathbf{Q}_T) = \frac{1}{T} \text{Cov}(\mathbf{Z}_1) + \frac{2}{T} \sum_{t=1}^{T-1} \left(1 - \frac{t}{T}\right) \text{Cov}(\mathbf{Z}_1, \mathbf{Z}_{t+1}).$$

Proof of Theorem 1. Let $\mathbf{B}_s = (\mathbf{B}_{s1}, \dots, \mathbf{B}_{sp_s})$. Based on the definition of $\hat{\boldsymbol{\theta}}_s$, we have

$$\hat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_s = (\mathbf{B}_s^\tau \mathbf{B}_s)^{-1} \mathbf{B}_s^\tau \mathbf{Y}_s - \boldsymbol{\theta}_s = (\mathbf{B}_s^\tau \mathbf{B}_s)^{-1} \mathbf{B}_s^\tau \boldsymbol{\varepsilon}_s + \left\{ (\mathbf{B}_s^\tau \mathbf{B}_s)^{-1} \mathbf{B}_s^\tau \sum_{j=1}^{p_s} \boldsymbol{\alpha}_{sj} - \boldsymbol{\theta}_s \right\} = J_1 + J_2,$$

where $\boldsymbol{\alpha}_{sj} = (\alpha_{sj}(X_{s1j}), \dots, \alpha_{sj}(X_{sTj}))^\tau$. Now,

$$\begin{aligned} \|J_1\|^2 &= \boldsymbol{\varepsilon}_s^\tau \mathbf{B}_s (\mathbf{B}_s^\tau \mathbf{B}_s)^{-2} \mathbf{B}_s^\tau \boldsymbol{\varepsilon}_s \leq O_p(T^{-2}) \cdot \max_{1 \leq j \leq p_s} \left[\lambda_{\min} \left(\frac{1}{T} \mathbf{B}_s^\tau \mathbf{B}_{sj} \right) \right]^{-2} (\boldsymbol{\varepsilon}_s^\tau \mathbf{B}_s \mathbf{B}_s^\tau \boldsymbol{\varepsilon}_s) \\ &= O_p(T^{-2}) \cdot O_p(N_T T) = O_p(T^{-1} N_T), \end{aligned}$$

which implies $\|J_1\| = O_p(\sqrt{N_T/T})$. Additionally,

$$\|J_2\|^2 = \left\| (\mathbf{B}_s^\tau \mathbf{B}_s)^{-1} \mathbf{B}_s^\tau \left(\sum_{j=1}^{p_s} \boldsymbol{\alpha}_{sj}^\tau - \mathbf{B}_s \boldsymbol{\theta}_s \right) \right\|^2 = O_p(T^{-1}) \cdot \left\| \sum_{j=1}^{p_s} \boldsymbol{\alpha}_{sj}^\tau - \mathbf{B}_s \boldsymbol{\theta}_s \right\|^2 = O_p(N_T^{-2}),$$

implying $\|J_2\| = O_p(N_T^{-1})$. Hence, result (i) holds.

By the definition of $\hat{\alpha}_{sj}(x)$,

$$\begin{aligned} &\int_{x \in [0,1]} \{\hat{\alpha}_{sj}(x) - \alpha_{sj}(x)\}^2 p_{sj}(x) dx \\ &= \int_{x \in [0,1]} \{(\hat{\boldsymbol{\alpha}}_{sj}^\tau(x) - (\mathbf{B}_{sj}(u)^\tau \boldsymbol{\theta}_{sj}) - (\alpha_{sj}(u) - (\mathbf{B}_{sj}(u)^\tau \boldsymbol{\theta}_{sj}))\}^2 p_{sj}(x) dx \\ &\leq 2\lambda \max \left(\int_{x \in [0,1]} (\mathbf{B}_{sj}(x)^\tau \mathbf{B}_{sj}(x) p_{sj}(x) dx \right) \cdot \|\hat{\boldsymbol{\theta}}_{sj} - \boldsymbol{\theta}_{sj}\|^2 \\ &\quad + 2 \int_{x \in [0,1]} (\alpha_{sj}(x) - (\mathbf{B}_{sj}(u)^\tau \boldsymbol{\theta}_{sj}))^2 p_{sj}(x) dx = O_p(N_T^{-2} + N_T/T) + O_p(N_T^{-2}). \end{aligned}$$

This shows that result (ii) is also true.

Proof of Theorem 2. For convenience purposes, write

$$\hat{\boldsymbol{\epsilon}}_s^o = \begin{bmatrix} \hat{\boldsymbol{\epsilon}}_{s,d_s} & \hat{\boldsymbol{\epsilon}}_{s,d_s-1} & \cdots & \hat{\boldsymbol{\epsilon}}_{s1} \\ \hat{\boldsymbol{\epsilon}}_{s,d_s+1} & \hat{\boldsymbol{\epsilon}}_{sd_s} & \cdots & \hat{\boldsymbol{\epsilon}}_{s2} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\boldsymbol{\epsilon}}_{s,T-1} & \hat{\boldsymbol{\epsilon}}_{s,T-2} & \cdots & \hat{\boldsymbol{\epsilon}}_{s,T-d_s} \end{bmatrix}, \quad \hat{\boldsymbol{\epsilon}}_s^{oo} = (\hat{\boldsymbol{\epsilon}}_{s,d_s+1}, \dots, \hat{\boldsymbol{\epsilon}}_{sT})^\tau,$$

$$\hat{\Lambda}_s^o = \begin{bmatrix} \sum_{j=1}^{p_s} \hat{\alpha}_{sj}(X_{sd_s,j}) & \sum_{j=1}^{p_s} \hat{\alpha}_{sj}(X_{s,d_s-1,j}) & \cdots & \sum_{j=1}^{p_s} \hat{\alpha}_{sj}(X_{s1j}) \\ \sum_{j=1}^{p_s} \hat{\alpha}_{sj}(X_{s,d_s+1,j}) & \sum_{j=1}^{p_s} \hat{\alpha}_{sj}(X_{sd_sj}) & \cdots & \sum_{j=1}^{p_s} \hat{\alpha}_{sj}(X_{s2j}) \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{j=1}^{p_s} \hat{\alpha}_{sj}(X_{s,T-1,j}) & \sum_{j=1}^{p_s} \hat{\alpha}_{sj}(X_{s,T-2,j}) & \cdots & \sum_{j=1}^{p_s} \hat{\alpha}_{sj}(X_{s,T-d_s,j}) \end{bmatrix},$$

and

$$\hat{\Lambda}_s^{oo} = \left(\sum_{j=1}^{p_s} \hat{\alpha}_{sj}(X_{s,d_s+1,j}), \dots, \sum_{j=1}^{p_s} \hat{\alpha}_{sj}(X_{sTj}) \right)^\tau.$$

Also, define $\boldsymbol{\epsilon}_s^o$ and $\boldsymbol{\epsilon}_s^{oo}$ as the matrices that result when $\hat{\boldsymbol{\epsilon}}_{st}$ in $\hat{\boldsymbol{\epsilon}}_s^o$ and $\hat{\boldsymbol{\epsilon}}_s^{oo}$ are replaced by $\boldsymbol{\epsilon}_{st}$. Similarly, by replacing $\sum_{j=1}^{p_s} \hat{\alpha}_{sj}(X_{stj})$ by $\sum_{j=1}^{p_s} \alpha_{sj}(X_{stj})$ in $\hat{\Lambda}_s^o$ and $\hat{\Lambda}_s^{oo}$, we obtain Λ_s^o and Λ_s^{oo} .

From the definition of $\mathcal{L}(\boldsymbol{\varsigma}_s)$,

$$\mathcal{L}(\boldsymbol{\varsigma}_s) = \frac{1}{2} (\hat{\boldsymbol{\epsilon}}_s^{oo} - \hat{\boldsymbol{\epsilon}}_s^o \boldsymbol{\varsigma}_s)^\tau (\hat{\boldsymbol{\epsilon}}_s^{oo} - \hat{\boldsymbol{\epsilon}}_s^o \boldsymbol{\varsigma}_s) + T \sum_{j=1}^{d_s} p_{\lambda_{jT}}(|s_{sj}|).$$

Denote $\alpha_T = T^{-1/2} + a_T$. It suffices to show that for any given $\zeta > 0$, there exists a large constant c such that

$$P \left\{ \inf_{\|\mathbf{u}\|=c} \mathcal{L}(\boldsymbol{\varsigma}_s + \alpha_T \mathbf{u}) \geq \mathcal{L}(\boldsymbol{\varsigma}_s) \right\} \geq 1 - \zeta.$$

This implies, with probability no smaller than $1 - \zeta$, that there exists a local minimizer in $\{\boldsymbol{\varsigma}_s + \alpha_n \mathbf{u} : \|\mathbf{u}\| \leq c\}$. Define

$$D_T(\mathbf{u}) = \mathcal{L}(\boldsymbol{\varsigma}_s + \alpha_T \mathbf{u}) - \mathcal{L}(\boldsymbol{\varsigma}_s).$$

Note that $p_{\lambda_{jT}}(0) = 0$, and $p_{\lambda_{jT}}(|s_{sj}|)$ is nonnegative. Therefore, it holds that

$$T^{-1} D_T(\mathbf{u}) \geq \frac{1}{2T} \{ (\hat{\boldsymbol{\epsilon}}_s^{oo} - \hat{\boldsymbol{\epsilon}}_s^o (\boldsymbol{\varsigma}_s + \alpha_T \mathbf{u}))^\tau (\hat{\boldsymbol{\epsilon}}_s^{oo} - \hat{\boldsymbol{\epsilon}}_s^o (\boldsymbol{\varsigma}_s + \alpha_T \mathbf{u})) - (\hat{\boldsymbol{\epsilon}}_s^{oo} - \hat{\boldsymbol{\epsilon}}_s^o \boldsymbol{\varsigma}_s)^\tau (\hat{\boldsymbol{\epsilon}}_s^{oo} - \hat{\boldsymbol{\epsilon}}_s^o \boldsymbol{\varsigma}_s) \} \\ + \sum_{j=1}^{d_s} \{ p_{\lambda_{jT}}(|s_{sj} + \alpha_T u_j|) - p_{\lambda_{jT}}(|s_{sj}|) \}.$$

Obviously,

$$\begin{aligned} & \frac{1}{2T} \{(\hat{\boldsymbol{\epsilon}}_s^{oo} - \hat{\boldsymbol{\epsilon}}_s^o(\boldsymbol{\varsigma}_s + \alpha_T \mathbf{u}))^\tau (\hat{\boldsymbol{\epsilon}}_s^{oo} - \hat{\boldsymbol{\epsilon}}_s^o(\boldsymbol{\varsigma}_s + \alpha_T \mathbf{u})) - (\hat{\boldsymbol{\epsilon}}_s^{oo} - \hat{\boldsymbol{\epsilon}}_s^o \boldsymbol{\varsigma}_s)^\tau (\hat{\boldsymbol{\epsilon}}_s^{oo} - \hat{\boldsymbol{\epsilon}}_s^o \boldsymbol{\varsigma}_s)\} \\ &= \frac{\alpha_T^2}{2T} \mathbf{u}^\tau \left\{ \boldsymbol{\epsilon}_s^o + (\boldsymbol{\Lambda}_s^o - \hat{\boldsymbol{\Lambda}}_s^o) \right\}^\tau \left\{ \boldsymbol{\epsilon}_s^o + (\boldsymbol{\Lambda}_s^o - \hat{\boldsymbol{\Lambda}}_s^o) \right\} \mathbf{u} \\ & \quad - \frac{\alpha_T}{T} \mathbf{u}^\tau \left\{ \boldsymbol{\epsilon}_s^o + (\boldsymbol{\Lambda}_s^o - \hat{\boldsymbol{\Lambda}}_s^o) \right\} \left\{ \mathbf{e}_s - ((\boldsymbol{\Lambda}_s^{oo} - \hat{\boldsymbol{\Lambda}}_s^{oo}) - (\boldsymbol{\Lambda}_s^o - \hat{\boldsymbol{\Lambda}}_s^o) \boldsymbol{\varsigma}_s) \right\} = J_1 + J_2 \text{ (say)}. \end{aligned}$$

From Theorem 1, we have

$$J_1 = \frac{\alpha_T^2}{2T} \mathbf{u}^\tau \boldsymbol{\epsilon}_s^{o\tau} \boldsymbol{\epsilon}_s^o \mathbf{u} + \alpha_T^2 \|\mathbf{u}\|^2 \cdot \{O_p(N_T^{-4} + N_T/T) + O_p(T^{-1/2})\}$$

and

$$J_2 = O_p(\alpha_T \|\mathbf{u}\|) \cdot O_p(T^{-\frac{1}{2}} \alpha_T \|\mathbf{u}\|).$$

Note that when T is sufficiently large,

$$\frac{1}{T} \boldsymbol{\epsilon}_s^{o\tau} \boldsymbol{\epsilon}_s^o = \boldsymbol{\Sigma}_s + O_p(T^{-\frac{1}{2}}) > 0,$$

so J_1 is of the order $c^2 \alpha_T^2$. Also note that $T^{-1/2} \alpha_T = O_p(\alpha_T^2)$. By choosing a sufficiently large c , J_1 will dominate the second term uniformly in $\|\mathbf{u}\| = c$. Furthermore, by the Taylor series expansion and Cauchy–Schwarz inequality,

$$\sum_{j=1}^{d_s} \{p_{\lambda_{jT}}(|s_{sj} + \alpha_T u_j|) - p_{\lambda_{jT}}(|s_{sj}|)\}$$

is bounded by

$$\sqrt{d_s} \alpha_T a_T \|\mathbf{u}\| + \alpha_T^2 b_T \|\mathbf{u}\|^2 = c \alpha_T^2 (\sqrt{d_s^0} + b_T c).$$

By taking a sufficiently large c , $c \alpha_T^2 (\sqrt{d_s^0} + b_T c)$ is dominated by J_1 as $b_T \rightarrow 0$. This completes the proof of the theorem.

Proof of Theorem 3. (i) The proof is same as that of Lemma A.1 in Fan and Li (2004). We will show that with probability tending to 1, as $T \rightarrow \infty$, for any d_s^0 dimensional $\boldsymbol{\varsigma}_s^{(1)*}$ satisfying $\|\boldsymbol{\varsigma}_s^{(1)*} - \boldsymbol{\varsigma}_s^{(1)}\| = O_p(T^{-1/2})$ and $d_s - d_s^0$ dimensional $\boldsymbol{\varsigma}_s^{(2)*}$ satisfying $\|\boldsymbol{\varsigma}_s^{(2)*}\| \leq cT^{-1/2}$, $\partial \mathcal{L}(\boldsymbol{\varsigma}_s^*) / \partial \boldsymbol{\varsigma}_{sj}$ and $\boldsymbol{\varsigma}_{sj}^*$ have the same sign, $\boldsymbol{\varsigma}_{sj}^* \in (-cT^{-1/2}, cT^{-1/2})$ for $j = d_s^0 + 1, \dots, d_s$, where $\boldsymbol{\varsigma}_s^* = (\boldsymbol{\varsigma}_s^{(1)*\tau}, \boldsymbol{\varsigma}_s^{(2)*\tau})^\tau$ and $\boldsymbol{\varsigma}_{sj}^*$ is the j th element of $\boldsymbol{\varsigma}_s^*$. Thus, a minimum is attained at $\boldsymbol{\varsigma}_s^{(2)} = \mathbf{0}$.

For $\boldsymbol{\varsigma}_{sj}^* \neq \mathbf{0}$ and $j = d_s^0 + 1, \dots, d_s$,

$$\frac{\partial \mathcal{L}(\boldsymbol{\varsigma}_s^*)}{\partial \boldsymbol{\varsigma}_{sj}} = \mathcal{L}'_j(\boldsymbol{\varsigma}_s^*) + T p'_{\lambda_{jT}}(|\boldsymbol{\varsigma}_{sj}^*|) \text{sgn}(\boldsymbol{\varsigma}_{sj}^*),$$

where $\mathcal{L}'_j(\boldsymbol{\varsigma}_s^*) = \partial \mathcal{L}(\boldsymbol{\varsigma}_s^*) / \partial \boldsymbol{\varsigma}_{sj}^*$. It is easy to see that

$$\begin{aligned} \mathcal{L}'_j(\boldsymbol{\varsigma}_s^*) &= - \sum_{t=1}^{T-d_s} \left\{ \boldsymbol{\epsilon}_{stj}^o + (\boldsymbol{\Lambda}_{stj}^o - \hat{\boldsymbol{\Lambda}}_{stj}^o) \right\} \left[\left\{ \boldsymbol{\epsilon}_{st}^{oo} + (\boldsymbol{\Lambda}_{st}^{oo} - \hat{\boldsymbol{\Lambda}}_{st}^{oo}) \right\} - \left\{ \boldsymbol{\epsilon}_{st}^o + (\boldsymbol{\Lambda}_{st}^o - \hat{\boldsymbol{\Lambda}}_{st}^o) \right\}^\tau \boldsymbol{\varsigma}_s \right] \\ &\quad - \sum_{t=1}^{T-d_s} \left\{ \boldsymbol{\epsilon}_{stj}^o + (\boldsymbol{\Lambda}_{stj}^o - \hat{\boldsymbol{\Lambda}}_{stj}^o) \right\} \left\{ \boldsymbol{\epsilon}_{st}^o + (\boldsymbol{\Lambda}_{st}^o - \hat{\boldsymbol{\Lambda}}_{st}^o) \right\}^\tau (\boldsymbol{\varsigma}_s^* - \boldsymbol{\varsigma}_s). \end{aligned}$$

Note that $\|\boldsymbol{\varsigma}_s^* - \boldsymbol{\varsigma}_s\| = O_p(T^{-1/2})$ by assumption. Then, using Theorem 1, we can show that $T^{-1} \mathcal{L}'_j(\boldsymbol{\varsigma}_s)$ is of the order $O_p(T^{-1/2})$. Therefore,

$$\frac{\partial \mathcal{L}(\boldsymbol{\varsigma}_s^*)}{\partial \boldsymbol{\varsigma}_{sj}} = T \lambda_{jT} \left\{ \lambda_{jT}^{-1} p'_{\lambda_{jT}}(|\boldsymbol{\varsigma}_{sj}^*|) \text{sgn}(\boldsymbol{\varsigma}_{sj}^*) + O_p(T^{-\frac{1}{2}}) \right\}.$$

As

$$\liminf_{T \rightarrow \infty} \liminf_{\boldsymbol{\varsigma}_{sj}^* \rightarrow \mathbf{0}^+} \lambda_{jT}^{-1} p'_{\lambda_{jT}}(|\boldsymbol{\varsigma}_{sj}^*|) > 0 \quad \text{and} \quad T^{-\frac{1}{2}} \lambda_{jT} \rightarrow 0,$$

the sign of the derivative is completely determined by that of $\boldsymbol{\varsigma}_{sj}^*$. This completes the proof of part (i) of the theorem.

To prove part (ii), using an argument similar to the proof of Theorem 2, it can be shown that there exists a $\hat{\boldsymbol{\zeta}}^{(1)}$, a \sqrt{T} consistent minimizer of $\mathcal{L}\{(\boldsymbol{\zeta}^{(1)\tau}, \mathbf{0}^\tau)^\tau\}$, that satisfies the penalized least squares equations

$$\partial \mathcal{L} \{(\hat{\boldsymbol{\zeta}}^{(1)\tau}, \mathbf{0}^\tau)^\tau\} / \partial \boldsymbol{\zeta}^{(1)} = 0.$$

Furthermore,

$$\frac{\partial \mathcal{L} \{(\hat{\boldsymbol{\zeta}}^{(1)\tau}, \mathbf{0}^\tau)^\tau\}}{\partial \boldsymbol{\zeta}^{(1)}} = -\hat{\boldsymbol{\epsilon}}_s^{o(1)} (\hat{\boldsymbol{\epsilon}}_s^{oo} - \hat{\boldsymbol{\epsilon}}_s^{o(1)} \boldsymbol{\zeta}^{(1)}) - T \{ \mathbf{b} + \{ \mathbf{D} + o_p(1) \} (\hat{\boldsymbol{\zeta}}^{(1)} - \boldsymbol{\zeta}^{(1)}) \},$$

with $\hat{\boldsymbol{\epsilon}}_s^{o(1)}$ containing the first d_s^0 columns of $\hat{\boldsymbol{\epsilon}}_s^o$. We can show that

$$-\frac{1}{\sqrt{T}} \hat{\boldsymbol{\epsilon}}_s^{o(1)} (\hat{\boldsymbol{\epsilon}}_s^{oo} - \hat{\boldsymbol{\epsilon}}_s^{o(1)} \boldsymbol{\zeta}^{(1)}) \rightarrow_D N(\mathbf{0}, \boldsymbol{\Sigma}_{s11}) \quad \text{as } T \rightarrow \infty,$$

with Σ_{s11} containing the first d_s^0 rows and columns of Σ_s . Thus, by Slutsky's Theorem, it follows that

$$\sqrt{T} \{ \Sigma_{s11} + \mathbf{D} \} \{ \hat{\zeta}^{(1)} - \zeta^{(1)} + (\Sigma_{s11} + \mathbf{D})^{-1} \mathbf{b} \} \xrightarrow{D} N(0, \Sigma_{s11}) \quad \text{as } T \rightarrow \infty.$$

This completes the proof of (ii).

Proof of Theorem 4. It is easy to see that with probability tending to 1,

$$\begin{aligned} \hat{e}_{st} &= \hat{\varepsilon}_{st} - \hat{s}_{s1} \hat{\varepsilon}_{s,t-1} - \dots - \hat{s}_{s\hat{d}_s} \hat{\varepsilon}_{s,t-\hat{d}_s} \\ &= e_{st} + \{ (s_{s1} - \hat{s}_{s1}) \varepsilon_{s,t-1} - \dots - (s_{s\hat{d}_s} - \hat{s}_{s\hat{d}_s}) \varepsilon_{s,t-\hat{d}_s} \} \\ &\quad + \sum_{j=1}^{p_s} (\alpha_{sj}(X_{stj}) - \hat{\alpha}_{sj}(X_{stj})) - \hat{s}_{s1} \sum_{j=1}^{p_s} (\alpha_{sj}(X_{s,t-1,j}) - \hat{\alpha}_{sj}(X_{s,t-1,j})) - \dots \\ &\quad - \hat{s}_{s\hat{d}_s} \sum_{j=1}^{p_s} (\alpha_{sj}(X_{s,t-\hat{d}_s,j}) - \hat{\alpha}_{sj}(X_{s,t-\hat{d}_s,j})) = e_{st} + J_{1st} + J_{2st} \quad (\text{say}). \end{aligned}$$

Thus, $\hat{\sigma}_{e_{s_1 s_2}}^2$ can be decomposed as

$$\begin{aligned} \hat{\sigma}_{e_{s_1 s_2}}^2 &= \frac{1}{T - \max(\hat{d}_{s_1} + 1, \hat{d}_{s_2} + 1)} \sum_{t=\max(\hat{d}_{s_1}+1, \hat{d}_{s_2}+1)}^T (e_{s_1 t} e_{s_2 t} + J_{1s_1 t} J_{1s_2 t} + J_{2s_1 t} J_{2s_2 t} + J_{1s_1 t} J_{2s_2 t} \\ &\quad + J_{2s_1 t} J_{1s_2 t} + e_{s_1 t} J_{1s_2 t} + e_{s_1 t} J_{2s_2 t} + J_{1s_1 t} e_{s_2 t} + J_{2s_1 t} e_{s_2 t}) \\ &= J_1^0 + \dots + J_9^0, \quad (\text{say}) \end{aligned}$$

with probability tends to 1. Combining the \sqrt{T} consistency property of $\hat{\zeta}_s$, we have

$$|J_2^0| = O_p(T^{-2}) \cdot \sum_{t=\max(\hat{d}_{s_1}+1, \hat{d}_{s_2}+1)}^T (\varepsilon_{s_1 t}^2 + \varepsilon_{s_2 t}^2) = O_p(T^{-1}).$$

Using this same property together with Theorem 1, we obtain

$$|J_3^0| = O_p(\sqrt{N_T/T} + N_T^{-2}) \cdot O_p(\sqrt{N_T/T} + N_T^{-2}) = O_p(N_T/T + N_T^{-4}) = o_p(T^{-\frac{1}{2}}).$$

Denote $T - \max(\hat{d}_{s_1} + 1, \hat{d}_{s_2} + 1) = T_{s_1 s_2}$, and $\max(\hat{d}_{s_1} + 1, \hat{d}_{s_2} + 1) = d_{s_1 s_2}$. By the definition of $\hat{\alpha}_{sj}(x)$, we have

$$\frac{1}{T_{s_1 s_2}} \sum_{t=d_{s_1 s_2}}^T (\alpha_{s_1 j}(X_{s_1 t j}) - \hat{\alpha}_{s_1 j}(X_{s_1 t j})) \varepsilon_{s_2 t}$$

$$\begin{aligned}
 &= \frac{1}{T_{s_1 s_2}} \sum_{t=d_{s_1 s_2}}^T \left\{ \alpha_{s_1 j}(X_{s_1 t j}) - (\mathbf{0}_{N_T}^\tau, \dots, (\mathbf{B}_{s_1 j}(X_{s_1 t j}))^\tau, \dots, \mathbf{0}_{N_T}^\tau)(\mathbf{B}_{s_1}^\tau \mathbf{B}_{s_1})^{-1} \mathbf{B}_{s_1}^\tau \right. \\
 &\quad \left. \left(\sum_{j=1}^{p_{s_1}} \alpha_{s_1 j}(X_{s_1 1 j}), \dots, \sum_{j=1}^{p_{s_1}} \alpha_{s_1 j}(X_{s_1 T j}) \right)^\tau \right\} \boldsymbol{\varepsilon}_{s_2 t} \\
 &\quad - \frac{1}{T_{s_1 s_2}} \sum_{t=d_{s_1 s_2}}^T (\mathbf{0}_{N_T}^\tau, \dots, (\mathbf{B}_{s_1 j}(X_{s_1 t j}))^\tau, \dots, \mathbf{0}_{N_T}^\tau)(\mathbf{B}_{s_1}^\tau \mathbf{B}_{s_1})^{-1} \mathbf{B}_{s_1}^\tau (\boldsymbol{\varepsilon}_{s_1 1}, \dots, \boldsymbol{\varepsilon}_{s_1 T})^\tau \boldsymbol{\varepsilon}_{s_2 t} \\
 &= J_{10}^0 - J_{11}^0 \quad (\text{say}).
 \end{aligned}$$

Since

$$\begin{aligned}
 &\alpha_{s_1 j}(X_{s_1 t j}) - (\mathbf{0}_{N_T}^\tau, \dots, (\mathbf{B}_{s_1 j}(X_{s_1 t j}))^\tau, \dots, \mathbf{0}_{N_T}^\tau)(\mathbf{B}_{s_1}^\tau \mathbf{B}_{s_1})^{-1} \mathbf{B}_{s_1}^\tau \\
 &\quad \cdot \left(\sum_{j=1}^{p_{s_1}} \alpha_{s_1 j}(X_{s_1 1 j}), \dots, \sum_{j=1}^{p_{s_1}} \alpha_{s_1 j}(X_{s_1 T j}) \right)^\tau \\
 &= \alpha_{s_1 j}(X_{s_1 t j}) - (\mathbf{B}_{s_1 j}(X_{s_1 t j}))^\tau \boldsymbol{\theta}_{s_1} - (\mathbf{0}_{N_T}^\tau, \dots, (\mathbf{B}_{s_1 j}(X_{s_1 t j}))^\tau, \dots, \mathbf{0}_{N_T}^\tau)(\mathbf{B}_{s_1}^\tau \mathbf{B}_{s_1})^{-1} \mathbf{B}_{s_1}^\tau \\
 &\quad \cdot \left\{ \left(\sum_{j=1}^{p_{s_1}} \alpha_{s_1 j}(X_{s_1 1 j}), \dots, \sum_{j=1}^{p_{s_1}} \alpha_{s_1 j}(X_{s_1 T j}) \right)^\tau - \mathbf{B}_{s_1} \boldsymbol{\theta}_{s_1} \right\},
 \end{aligned}$$

we can show that $J_{10}^0 = O_p(T^{-1/2}) \cdot O_p(N_T/T + N_T^{-2}) = o_p(T^{-1/2})$. Also,

$$\begin{aligned}
 J_{11}^0 &= \frac{1}{T_{s_1 s_2}} \boldsymbol{\varepsilon}_{s_2}^\tau (\mathbf{0}_{N_T}^\tau, \dots, (\mathbf{B}_{s_1 j}(X_{s_1 t j}))^\tau, \dots, \mathbf{0}_{N_T}^\tau)(\mathbf{B}_{s_1}^\tau \mathbf{B}_{s_1})^{-1} \mathbf{B}_{s_1}^\tau \boldsymbol{\varepsilon}_{s_1} \\
 &\leq \frac{1}{T_{s_1 s_2}} \sqrt{\boldsymbol{\varepsilon}_{s_1}^\tau \mathbf{B}_{s_1} (\mathbf{B}_{s_1}^\tau \mathbf{B}_{s_1})^{-1} \mathbf{B}_{s_1}^\tau \boldsymbol{\varepsilon}_{s_1}} \\
 &\quad \cdot \sqrt{\boldsymbol{\varepsilon}_{s_2}^\tau (\mathbf{0}_{N_T}^\tau, \dots, (\mathbf{B}_{s_1 j}(X_{s_1 t j}))^\tau, \dots, \mathbf{0}_{N_T}^\tau)(\mathbf{B}_{s_1}^\tau \mathbf{B}_{s_1})^{-1} (\mathbf{0}_{N_T}^\tau, \dots, (\mathbf{B}_{s_1 j}(X_{s_1 t j}))^\tau, \dots, \mathbf{0}_{N_T}^\tau) \boldsymbol{\varepsilon}_{s_2}} \\
 &= O_p(T^{-1} N_T) = o_p(T^{-\frac{1}{2}}).
 \end{aligned}$$

Therefore,

$$\frac{1}{T_{s_1 s_2}} \sum_{t=d_{s_1 s_2}}^T (\alpha_{s_1 j}(X_{s_1 t j}) - \hat{\alpha}_{s_1 j}(X_{s_1 t j})) \boldsymbol{\varepsilon}_{s_2 t} = o_p(T^{-\frac{1}{2}}).$$

Along the same lines, we can show that $J_s^0 = o_p(T^{-1/2})$, with $s = 4, \dots, 9$. Thus,

$$\hat{\sigma}_{e_{j_1 j_2}}^2 = \frac{1}{T - \max(\hat{d}_{s_1} + 1, \hat{d}_{s_2} + 1)} \sum_{t=\max(d_{s_1}^0+1, d_{s_2}^0+1)}^T e_{s_1 t} e_{s_2 t} + o_p(T^{-\frac{1}{2}}).$$

The proof is completed by combining the above with the Central Limit Theorem.

Proof of Theorem 5. To prove part (i), along the lines of proving Theorem 4, we can show that

$$\begin{aligned} \sqrt{T}(\hat{\zeta}^w - \zeta) &= \sqrt{T} \left\{ \text{diag}(\hat{\boldsymbol{\epsilon}}_1^*, \dots, \hat{\boldsymbol{\epsilon}}_m^*)^\tau (\hat{\boldsymbol{\Sigma}}_e \otimes \mathbf{I}_{T-\hat{d}}) [\text{diag}(\hat{\boldsymbol{\epsilon}}_1^*, \dots, \hat{\boldsymbol{\epsilon}}_m^*)]^{-1} \right. \\ &\quad \cdot \text{diag}(\hat{\boldsymbol{\epsilon}}_1^*, \dots, \hat{\boldsymbol{\epsilon}}_m^*) (\hat{\boldsymbol{\Sigma}}_e \otimes \mathbf{I}_{T-\hat{d}}) \begin{pmatrix} \hat{\boldsymbol{\epsilon}}_1^{**} \\ \vdots \\ \hat{\boldsymbol{\epsilon}}_m^{**} \end{pmatrix} - \zeta \left. \right\} \\ &= \sqrt{T} \{ \text{diag}(\boldsymbol{\epsilon}_1^*, \dots, \boldsymbol{\epsilon}_m^*)^\tau (\boldsymbol{\Sigma}_e \otimes \mathbf{I}_{T-\hat{d}}) [\text{diag}(\boldsymbol{\epsilon}_1^*, \dots, \boldsymbol{\epsilon}_m^*)]^{-1} \\ &\quad \cdot \text{diag}(\boldsymbol{\epsilon}_1^*, \dots, \boldsymbol{\epsilon}_m^*) (\boldsymbol{\Sigma}_e \otimes \mathbf{I}_{T-\hat{d}}) \begin{pmatrix} \mathbf{e}_1^{**} \\ \vdots \\ \mathbf{e}_m^{**} \end{pmatrix} + o_p(1). \end{aligned}$$

Also, with probability that tends to 1,

$$\begin{aligned} &\sqrt{T} \{ \text{diag}(\boldsymbol{\epsilon}_1^*, \dots, \boldsymbol{\epsilon}_m^*)^\tau (\boldsymbol{\Sigma}_e \otimes \mathbf{I}_{T-\hat{d}}) [\text{diag}(\boldsymbol{\epsilon}_1^*, \dots, \boldsymbol{\epsilon}_m^*)]^{-1} \text{diag}(\boldsymbol{\epsilon}_1^*, \dots, \boldsymbol{\epsilon}_m^*) (\boldsymbol{\Sigma}_e \otimes \mathbf{I}_{T-\hat{d}}) \begin{pmatrix} \mathbf{e}_1^{**} \\ \vdots \\ \mathbf{e}_m^{**} \end{pmatrix} \\ &= \sqrt{T} \{ \text{diag}(\boldsymbol{\epsilon}_1^*, \dots, \boldsymbol{\epsilon}_m^*)^\tau (\boldsymbol{\Sigma}_e \otimes \mathbf{I}_{T-d}) [\text{diag}(\boldsymbol{\epsilon}_1^*, \dots, \boldsymbol{\epsilon}_m^*)]^{-1} \text{diag}(\boldsymbol{\epsilon}_1^*, \dots, \boldsymbol{\epsilon}_m^*) \\ &\quad \times (\boldsymbol{\Sigma}_e \otimes \mathbf{I}_{T-d}) \begin{pmatrix} \mathbf{e}_1^{**} \\ \vdots \\ \mathbf{e}_m^{**} \end{pmatrix} \}. \end{aligned}$$

Using the above results together with the Central Limit Theorem, the proof of part (i) is completed. To prove part (ii), for purposes of convenience, and without loss of generality, we focus on the case of $m = 2$. The asymptotic covariance matrix of $\hat{\zeta}^w$ for this case is

$$\{ (\sigma_e^{11})^{-1} \boldsymbol{\Sigma}_{11} - (\sigma_e^{12})^{-2} (\sigma_e^{22})^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \}^{-1}.$$

Since $\Sigma_{11} \geq \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, we have

$$\begin{aligned} \{(\sigma_e^{11})^{-1}\Sigma_{11} - (\sigma_e^{12})^{-2}(\sigma_e^{22})^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\}^{-1} &\leq \{(\sigma_e^{11})^{-1} - (\sigma_e^{12})^{-2}(\sigma_e^{22})^{-1}\}^{-1}\Sigma_{11}^{-1} \\ &= \{\sigma_{e11}^2\sigma_{e22}^2 - (\sigma_{e12}^2)^2\} \{\sigma_{e22}^2 - (\sigma_{e12}^2)^2(\sigma_{e11}^2)^{-1}\}^{-1}\Sigma_{11}^{-1} = \sigma_{e11}^2\Sigma_{11}^{-1}. \end{aligned}$$

This implies that $\hat{\xi}_1^w$ has a smaller asymptotic covariance than $\hat{\xi}_1$. By the same argument we can show that $\hat{\xi}_2^w$ has a smaller asymptotic covariance than $\hat{\xi}_2$.

Proof of Theorem 6. Theorem 6 can be easily proved by applying Theorems 1 and 4. We omit the details for brevity.

Proof of Theorem 7. Denote

$$(\tilde{\alpha}_{sj}^{TS}(x), \tilde{\alpha}'_{sj}{}^T(x))^\tau = (\mathbf{D}_{sjx}^\tau \mathbf{W}_{sjx} \mathbf{D}_{sjx})^{-1} \mathbf{D}_{sjx}^\tau \mathbf{W}_{sjx} \tilde{\mathbf{Y}}_{sj}^{**},$$

where

$$\begin{aligned} \tilde{\mathbf{Y}}_{stj}^{**} &= (\sigma_e^{ss})^{-1} \left(\sigma_e^{ss} \hat{\mathbf{Y}}_{st}^* + \sum_{s_1 \neq s}^m \sigma_e^{ss_1} \hat{\mathbf{Y}}_{s_1 t}^* - \sigma_e^{ss} \sum_{j_1 \neq j}^{p_s} \hat{\alpha}_{sj_1}(X_{stj_1}) - \sum_{s_1 \neq s}^m \sigma_e^{ss_1} \sum_{j_2=1}^{p_{s_1}} \hat{\alpha}_{s_1 j_2}(X_{s_1 t j_2}) \right), \\ \hat{\mathbf{Y}}_{st}^* &= Y_{st} - \hat{s}_{s1}^w \left(Y_{s,t-1} - \sum_{j=1}^{p_s} \hat{\alpha}_{sj}^w(X_{s,t-1,j}) \right) - \dots - \hat{s}_{sd_s} \left(Y_{s,t-\hat{d}_s} - \sum_{j=1}^{p_s} \hat{\alpha}_{sj}(X_{s,t-\hat{d}_s,j}) \right), \end{aligned}$$

and \tilde{Y}_{stj}^{**} is the t th element of $\tilde{\mathbf{Y}}_{sj}^{**}$.

By the definition of $(\tilde{\alpha}_{sj}^{TS}(x), \tilde{\alpha}'_{sj}{}^T(x))^\tau$, $(\tilde{\alpha}_{sj}^{TS}(x), \tilde{\alpha}'_{sj}{}^T(x))^\tau - (\alpha_{sj}(x), \alpha'_{sj}(x))^\tau$ can be decomposed as

$$\begin{aligned} &(\tilde{\alpha}_{sj}^{TS}(x), \tilde{\alpha}'_{sj}{}^T(x))^\tau - (\alpha_{sj}(x), \alpha'_{sj}(x))^\tau \\ &= J_1 + J_2 + (\mathbf{D}_{sjx}^\tau \mathbf{W}_{sjx} \mathbf{D}_{sjx})^{-1} \sum_{t=1}^{T-\hat{d}} \begin{pmatrix} 1 \\ X_{stj} \end{pmatrix} \frac{1}{h_T} K\left(\frac{X_{stj} - x}{h_T}\right) (\sigma_e^{ss})^{-1} \\ &\quad \cdot \left\{ \sigma_e^{ss} \left(\sum_{j_1 \neq j}^{p_s} \alpha_{sj_1}(X_{stj_1}) - \sum_{j_1 \neq j}^{p_s} \hat{\alpha}_{sj_1}(X_{stj_1}) \right) + \sum_{s_1 \neq s}^m \sigma_e^{ss_1} \left(\sum_{j_1=1}^{p_{s_1}} \alpha_{s_1 j_1}(X_{s_1 t j_1}) - \sum_{j_1=1}^{p_{s_1}} \hat{\alpha}_{s_1 j_1}(X_{s_1 t j_1}) \right) \right\} \\ &\quad - (\mathbf{D}_{sjx}^\tau \mathbf{W}_{sjx} \mathbf{D}_{sjx})^{-1} \sum_{t=1}^{T-\hat{d}} \begin{pmatrix} 1 \\ X_{stj} \end{pmatrix} \frac{1}{h_T} K\left(\frac{X_{stj} - x}{h_T}\right) \sum_{s_1=1}^m (\sigma_e^{ss})^{-1} \sigma_e^{ss_1} \\ &\quad \cdot \left\{ \sum_{j_1=1}^{p_{s_1}} \alpha_{s_1 j_1}(X_{s_1 t j_1}) - \sum_{j_1=1}^{p_{s_1}} \hat{\alpha}_{s_1 j_1}(X_{s_1 t j_1}) \right\} \end{aligned}$$

$$\begin{aligned}
 & - (\mathbf{D}_{s_{jx}}^\tau \mathbf{W}_{s_{jx}} \mathbf{D}_{s_{jx}})^{-1} \sum_{t=1}^{T-\hat{d}} \binom{1}{X_{stj}} \frac{1}{h_T} K\left(\frac{X_{stj}-x}{h_T}\right) \\
 & \cdot \left\{ \hat{\mathbf{s}}_{s1}^w \left(\sum_{j=1}^{p_s} \alpha_{sj}(X_{s,t-1,j}) - \sum_{j=1}^{p_s} \hat{\alpha}_{sj}(X_{s,t-1,j}) \right) + \cdots + \hat{\mathbf{s}}_{sd_s} \right. \\
 & \left. \left(\sum_{j=1}^{p_s} \alpha_{sj}(X_{s,t-\hat{d}_s,j}) - \sum_{j=1}^{p_s} \hat{\alpha}_{sj}(X_{s,t-\hat{d}_s,j}) \right) \right\} \\
 & - (\mathbf{D}_{s_{jx}}^\tau \mathbf{W}_{s_{jx}} \mathbf{D}_{s_{jx}})^{-1} \sum_{t=1}^{T-\hat{d}} \binom{1}{X_{stj}} \frac{1}{h_T} K\left(\frac{X_{stj}-x}{h_T}\right) \\
 & \cdot \left\{ (\hat{\mathbf{s}}_{s1}^w \boldsymbol{\varepsilon}_{s,t-1} + \cdots + \hat{\mathbf{s}}_{sd_s}^w \boldsymbol{\varepsilon}_{s,t-\hat{d}_s}) - (\mathbf{s}_{s1} \boldsymbol{\varepsilon}_{s,t-1} + \cdots + \mathbf{s}_{sd_s}^0 \boldsymbol{\varepsilon}_{s,t-\hat{d}_s}^0) \right\} \\
 & = J_1 + J_2 + J_3 - J_4 - J_5 - J_6, \text{ (say),}
 \end{aligned}$$

with

$$J_1 = (\mathbf{D}_{s_{jx}}^\tau \mathbf{W}_{s_{jx}} \mathbf{D}_{s_{jx}})^{-1} \sum_{t=1}^{T-\hat{d}} \binom{1}{X_{stj}} \frac{1}{h_T} K\left(\frac{X_{stj}-x}{h_T}\right) \alpha_{sj}(X_{stj}) - (\alpha_{sj}(x), \alpha'_{sj}(x))^\tau$$

and

$$J_2 = (\mathbf{D}_{s_{jx}}^\tau \mathbf{W}_{s_{jx}} \mathbf{D}_{s_{jx}})^{-1} \sum_{t=1}^{T-\hat{d}} \binom{1}{X_{stj}} \frac{1}{h_T} K\left(\frac{X_{stj}-x}{h_T}\right) \sum_{s1=1}^m (\sigma_e^{ss})^{-1} \sigma_e^{ss1} e_{s1t}.$$

Note that

$$\mathbf{D}_{s_{jx}}^\tau \mathbf{W}_{s_{jx}} \mathbf{D}_{s_{jx}} = \begin{pmatrix} \sum_{t=1}^{T-\hat{d}} K_{h_T}(X_{stj}-x) & \sum_{t=1}^{T-\hat{d}} \left(\frac{X_{stj}-x}{h_T}\right) K_{h_T}(X_{stj}-x) \\ \sum_{t=1}^{T-\hat{d}} \left(\frac{X_{stj}-x}{h_T}\right) K_{h_T}(X_{stj}-x) & \sum_{t=1}^{T-\hat{d}} \left(\frac{X_{stj}-x}{h_T}\right)^2 K_{h_T}(X_{stj}-x) \end{pmatrix},$$

and each element of the above matrix is in the form of kernel regression. By Lemma 1, it holds that

$$\frac{1}{T} \mathbf{D}_{s_{jx}}^\tau \mathbf{W}_{s_{jx}} \mathbf{D}_{s_{jx}} = p_{sj}(x) \otimes \mathbf{H}_T \begin{pmatrix} 1 & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix} \mathbf{H}_T \cdot O_p \left[1 + \left\{ \frac{\log T}{Th_T} \right\}^{\frac{1}{2}} \right]$$

with probability tending to 1. Therefore, by the conventional properties of nonparametric regression, we have

$$\sqrt{Th_T} \left[\mathbf{H}_T^{-1} J_1 - \frac{h_T^2}{2} \begin{pmatrix} \mathfrak{S}_1 \alpha''_{sj}(x) \\ \mathfrak{S}_2 \alpha''_{sj}(x) \end{pmatrix} + o(h_T^2) \right] = o_p(1).$$

Our next task is to show

$$\sqrt{Th_T} \mathbf{H}_n^{-1} J_2 \xrightarrow{D} N(0, \boldsymbol{\Sigma}_{(\alpha_{sj}, \alpha'_{sj})}^{TS}) \text{ as } T \rightarrow \infty. \tag{7.1}$$

Let

$$\mathbf{Z}_t = \left(\frac{K_{h_T}(X_{stj} - x)}{X_{stj} - x} K_{h_T}(X_{stj} - x) \right) \sum_{s_1=1}^m (\sigma_e^{ss})^{-1} \sigma_e^{ss_1} e_{s_1 t} \quad \text{and} \quad \mathbf{Q}_T = \frac{1}{T} \sum_{t=1}^T \mathbf{Z}_t.$$

By Lemma 2, we can employ the Doob’s small-block and large technique (as in Cai et al., 2000) to prove that

$$\frac{\sqrt{Th_T}}{T} \mathbf{Q}_T \xrightarrow{D} N(0, \boldsymbol{\Sigma}) \text{ as } T \rightarrow \infty, \text{ with } \boldsymbol{\Sigma} = (\sigma_e^{ss})^{-1} p_{sj}(x) \begin{pmatrix} v_0 & v_1 \\ v_1 & v_2 \end{pmatrix}.$$

Therefore, (7.1) holds.

Now, we will show sequentially that $J_s = o_p(n^{-\frac{2}{5}})$, $s = 3, 4, 5$, and 6. First, let us show that

$$\left\| (\mathbf{D}_{sjx}^\tau \mathbf{W}_{sjx} \mathbf{D}_{sjx})^{-1} \sum_{t=1}^{T-\hat{d}} \begin{pmatrix} 1 \\ X_{stj} \end{pmatrix} \frac{1}{h_T} K \left(\frac{X_{stj} - x}{h_T} \right) (\alpha_{sj_1}(X_{stj_1}) - \hat{\alpha}_{sj_1}(X_{stj_1})) \right\| = o_p(T^{-\frac{2}{5}}). \tag{7.2}$$

Let $\mathbf{B}_{sj}(x) = (B_{sj1}(x), \dots, B_{sjN_T}(x))^\tau$, $\boldsymbol{\alpha}_{sj} = (\alpha_{sj}(X_{stj}), \dots, \alpha_{sj}(X_{sTj}))^\tau$, $\boldsymbol{\varepsilon}_s = (\varepsilon_{s1}, \dots, \varepsilon_{sT})^\tau$, and $\boldsymbol{\Delta} = (\mathbf{0}_{N_T \times N_T(j-1)}, \mathbf{I}_{N_T}, \mathbf{0}_{N_T \times N_T(p_s-j-1)}) \{\mathbf{B}_s^\tau \mathbf{B}_s\}^{-1} \mathbf{B}_s^\tau$, with $\mathbf{B}_s = (\mathbf{B}_{s1}, \dots, \mathbf{B}_{sp_s})$. Then we have

$$\begin{aligned} \alpha_{sj_1}(X_{stj_1}) - \hat{\alpha}_{sj_1}(X_{stj_1}) &= \alpha_{sj_1}(X_{stj_1}) - \mathbf{B}_{sj_1}^\tau(X_{stj_1}) \boldsymbol{\Delta} \mathbf{Y}_s \\ &= \{ \alpha_{sj_1}(X_{stj_1}) - \mathbf{B}_{sj_1}^\tau(X_{stj_1}) \boldsymbol{\theta}_{sj_1} \} - \mathbf{B}_{sj_1}^\tau(X_{stj_1}) \boldsymbol{\Delta} \left(\sum_{j=1}^{p_s} \boldsymbol{\alpha}_{sj} - \mathbf{B}_s \boldsymbol{\theta}_s \right) - \mathbf{B}_{sj_1}^\tau(X_{stj_1}) \boldsymbol{\Delta} \boldsymbol{\varepsilon}. \end{aligned}$$

By the polynomial spline property,

$$\frac{1}{T} \left(\sum_{j=1}^{p_s} \boldsymbol{\alpha}_{sj} - \mathbf{B}_s \boldsymbol{\theta}_s \right)^\tau \left(\sum_{j=1}^{p_s} \boldsymbol{\alpha}_{sj} - \mathbf{B}_s \boldsymbol{\theta}_s \right) = O_p(N_T^{-2}).$$

It is easy to see that

$$\left\| (\mathbf{D}_{sjx}^\tau \mathbf{W}_{sjx} \mathbf{D}_{sjx})^{-1} \sum_{t=1}^{T-\hat{d}} \begin{pmatrix} 1 \\ X_{stj} \end{pmatrix} \frac{1}{h_T} K \left(\frac{X_{stj} - x}{h_T} \right) \{ \alpha_{sj_1}(X_{stj_1}) - \mathbf{B}_{sj_1}^\tau(X_{stj_1}) \boldsymbol{\theta}_{sj_1} \} \right\|$$

$$\leq \left\| (\mathbf{D}_{s_{jx}}^\tau \mathbf{W}_{s_{jx}} \mathbf{D}_{s_{jx}})^{-1} \sum_{t=1}^{T-\hat{d}} \begin{pmatrix} 1 \\ X_{stj} \end{pmatrix} \frac{1}{h_T} K\left(\frac{X_{stj} - x}{h_T}\right) \right\|$$

$$\cdot \max_{1 \leq t \leq T} |\alpha_{s_{j1}}(X_{stj_1}) - \mathbf{B}_{s_{j1}}^\tau(X_{stj_1}) \boldsymbol{\theta}_{s_{j1}}| = O_p(N_T^{-2}) = O_p\left\{T^{-\frac{2}{3}}(\log T)^{-2}\right\} = o_p(T^{-\frac{2}{3}}),$$

and

$$\left\| (\mathbf{D}_{s_{jx}}^\tau \mathbf{W}_{s_{jx}} \mathbf{D}_{s_{jx}})^{-1} \sum_{t=1}^{T-\hat{d}} \begin{pmatrix} 1 \\ X_{stj} \end{pmatrix} \frac{1}{h_T} K\left(\frac{X_{stj} - x}{h_T}\right) \mathbf{B}_{s_{j1}}^\tau(X_{stj_1}) \Delta \left(\sum_{j=1}^{p_s} \boldsymbol{\alpha}_{sj} - \mathbf{B}_s \boldsymbol{\theta}_s \right) \right\|$$

$$\leq \left\| (\mathbf{D}_{s_{jx}}^\tau \mathbf{W}_{s_{jx}} \mathbf{D}_{s_{jx}})^{-1} \sum_{t=1}^{T-\hat{d}} \begin{pmatrix} 1 \\ X_{stj} \end{pmatrix} \frac{1}{h_T} K\left(\frac{X_{stj} - x}{h_T}\right) \right\|$$

$$\cdot \max_{1 \leq t \leq T} \left| \mathbf{B}_{s_{j1}}^\tau(X_{stj_1}) \Delta \left(\sum_{j=1}^{p_s} \boldsymbol{\alpha}_{sj} - \mathbf{B}_s \boldsymbol{\theta}_s \right) \right|$$

$$\leq O_p(1) \cdot \max_{1 \leq t \leq T} \sqrt{\left(\sum_{j=1}^{p_s} \boldsymbol{\alpha}_{sj} - \mathbf{B}_s \boldsymbol{\theta}_s \right)^\tau \Delta^\tau \mathbf{B}_{s_{j1}}(X_{stj_1}) \mathbf{B}_{s_{j1}}^\tau(X_{stj_1}) \Delta \left(\sum_{j=1}^{p_s} \boldsymbol{\alpha}_{sj} - \mathbf{B}_s \boldsymbol{\theta}_s \right)}$$

$$= O_p(N_T^{-2}) = O_p\left\{T^{-\frac{2}{3}}(\log T)^{-2}\right\} = o_p(T^{-\frac{2}{3}}).$$

Along the lines of the proof of Lemma 5.1 in Wang and Yang (2007), we obtain

$$\left\| (\mathbf{D}_{s_{jx}}^\tau \mathbf{W}_{s_{jx}} \mathbf{D}_{s_{jx}})^{-1} \sum_{t=1}^T \begin{pmatrix} 1 \\ X_{stj} \end{pmatrix} \frac{1}{h_T} K\left(\frac{X_{stj} - x}{h_T}\right) \mathbf{B}_{s_{j1}}(X_{stj_1}) \Delta \boldsymbol{\varepsilon} \right\|$$

$$= O_p\left\{N_T(\log T)^2 T^{-1}\right\} = o_p(T^{-\frac{2}{3}}).$$

Combining these results, we obtain (7.2). Based on the latter, we can show that $J_3 = o_p(T^{-2/5})$, $J_4 = o_p(T^{-2/5})$, and $J_5 = o_p(T^{-2/5})$. Additionally, combining Theorems 4 and 5, we can show that

$$J_6 = (\mathbf{D}_{s_{jx}}^\tau \mathbf{W}_{s_{jx}} \mathbf{D}_{s_{jx}})^{-1} \sum_{t=1}^{T-\hat{d}} \begin{pmatrix} 1 \\ X_{stj} \end{pmatrix} \frac{1}{h_T} K\left(\frac{X_{stj} - x}{h_T}\right)$$

$$\cdot \left\{ (\hat{\mathbf{S}}_{s_1}^w \boldsymbol{\varepsilon}_{s,t-1} + \cdots + \hat{\mathbf{S}}_{s_{\hat{d}_s}}^w \boldsymbol{\varepsilon}_{s,t-\hat{d}_s}) - (\mathbf{s}_{s_1} \boldsymbol{\varepsilon}_{s,t-1} + \cdots + \mathbf{s}_{s_{\hat{d}_s}^0} \boldsymbol{\varepsilon}_{s,t-\hat{d}_s^0}) \right\} = O_p(T^{-\frac{1}{2}}).$$

Combining these results with the \sqrt{T} consistency property of $\hat{\sigma}_e^{SS1}$, the proof of Theorem 7 is completed.

ACKNOWLEDGMENTS

The authors thank the editor Esfandiar Maasoumi and three anonymous referees for helpful comments. The order of authorship carries only alphabetical significance.

FUNDING

Alan Wan's research was supported by a strategic grant from the City University of Hong Kong (No.7008134). Jinhong You's research was supported by grants from the following bodies: National Natural Science Foundation of China (NSFC) (No. 11471203), Program for New Century Excellent Talents in University (NCET), Program for Changjiang Scholars and Innovative Research Team in University (IRT13077), Shanghai University of Finance and Economics through Project 211 (Phase IV), and Shanghai Leading Academic Discipline Project (B803).

REFERENCES

- Antoniadis, A., Fan, J. (2001). Regularization of wavelet approximations (with discussion). *Journal of the American Statistical Association* 96:939–967.
- Bontemps, C., Simioni, M., Surry, Y. (2008). Semiparametric hedonic price models: Assessing the effects of agricultural nonpoint source pollution. *Journal of Applied Econometrics* 23:825–842.
- Brockwell, P. J., Davis, R. A. (1991). *Time Series Theory and Methods*. New York: Springer.
- Cai, Z., Fan, J., Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association* 95:941–956.
- Carroll, R. J., Midthune, D., Freedman, L. S., Kipnis, V. (2006). Seemingly unrelated measurement error models, with application to nutritional epidemiology. *Biometrics* 62:75–84.
- Doran, H. E., Griffiths, W. E. (1983). On the relative efficiency of estimators which include the initial observations in the estimation of seemingly unrelated regressions with first order autoregressive disturbances. *Journal of Econometrics* 23:165–191.
- Eisenbeiss, M., Kauermann, G., Semmler, W. (2007). Estimating beta-coefficients of German stock data: a non-parametric approach. *European Journal of Finance* 13:503–522.
- Fan, J., Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. London: Chapman and Hall.
- Fan, J., Jiang, J. (2005). Nonparametric inferences for additive models. *Journal of American Statistical Association* 100:890–907.
- Fan, J., Li, R. (2001). Variable selection via penalized likelihood. *Journal of the American Statistical Association* 95:1348–1360.
- Fan, J., Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association* 99:710–723.
- Fan, J., Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters. *Annals of Statistics* 32:928–961.
- Fan, J., Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer.
- Foschi, P., Kontoghiorghes, E. J. (2003). Estimating seemingly unrelated regression models with vector autoregressive disturbances. *Journal of Economic Dynamics and Control* 28:27–44.
- Guilkey, D. K., Schmidt, P. (1973). Estimation of seemingly unrelated regressions with vector autoregressive errors. *Journal of the American Statistical Association* 68:642–647.
- Härdle, W., Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association* 84:986–995.

- Hastie, T. J., Tibshirani, R. J. (1990). *Generalized Additive Models*. London: Chapman and Hall.
- He, W., Lawless, J. F. (2005). Bivariate location-scale models for regression analysis with applications to lifetime data. *Journal of the Royal Statistical Society (Series B)* 67:63–78.
- Horowitz, J. L., Mammen, E. (2004). Nonparametric estimation of an additive model with a link function. *Annals of Statistics* 32:2412–2443.
- Huang, J., Wei, R. R., Ma, S. G. (2012). Semiparametric regression pursuit. *Statistica Sinica* 22:1403–1426.
- Huang, J., Yang, L. (2004). Identification of nonlinear additive autoregressive models. *Journal of Royal Statistical Society (Series B)* 66:463–477.
- Huber, P. J. (1985). Projection pursuit. *Annals of Statistics* 13:435–475.
- Kmenta, J., Gilbert, R. F. (1970). Estimation of seemingly unrelated regressions with autoregressive disturbances. *Journal of the American Statistical Association* 65:186–197.
- Koebel, B. M. (2004). First-order serial correlation in seemingly unrelated regressions. *Economics Letters* 82:1–7.
- Koop, G., Poirier, D. J., Tobias, J. (2005). Semiparametric Bayesian inference in multiple equation models. *Journal of Applied Econometrics* 20:723–748.
- Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association* 86:316–342.
- Li, R., Li, Y. (2009). Local linear regression for data with AR errors. *Acta Mathematicae Applicatae Sinica (English Series)* 25:427–444.
- Lin, X., Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with errors. *Journal of the American Statistical Association* 96:520–534.
- Linton, O. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika* 84:469–473.
- Linton, O., Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* 82:93–100.
- Liu, J., Chen, R., Yao, Q. (2010). Nonparametric transfer function models. *Journal of Econometrics* 157:151–164.
- Lyssiotou, P., Pashardes, P., Stengos, T. (2002). Age effects on consumer demand: An additive partially linear regression model. *Canadian Journal of Economics* 35:153–165.
- Martins-Filho, C., Bin, O. (2005). Estimation of hedonic price functions via additive nonparametric regression. *Empirical Economics* 30:93–114.
- Martins-Filho, C., Yao, F. (2009). Nonparametric regression estimation with general parametric error covariance. *Journal of Multivariate Analysis* 100:309–333.
- Schumaker, L. L. (1981). *Spline Functions*. New York: Wiley.
- Shukur, G., Zeebari, Z. (2011). Developing median regression for SUR models with application to 3-generation immigrants' data in Sweden. *Economic Modelling* 28:2566–2578.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
- Singh, R., Wang, L. (2012). A note on estimation in seemingly unrelated semi-parametric regression models. *Journal of Quantitative Economics* 10:56–69.
- Smith, M., Kohn, R. (2000). Nonparametric seemingly unrelated regression. *Journal of Econometrics* 98:257–281.
- Srivastava, V. K., Giles, D. E. A. (1987). *Seemingly Unrelated Regression Equations Models: Estimation and Inference*. New York: Marcel Dekker.
- Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics* 13:689–705.
- Sun, Z. (1984). Asymptotic unbiased and strong consistency for density function estimator. *Acta Mathematica Sinica* 27:769–782.
- Thompson, S. R., Sul, D., Bohl, M. T. (2002). Spatial market efficiency and policy regime change: seemingly unrelated error correction model estimation. *American Journal of Agricultural Economics* 84:1042–1053.
- Turkington, D. (2000). Generalised vec operators and the seemingly unrelated regression equations model with vector correlated disturbances. *Journal of Econometrics* 99:225–253.
- Wang, H. (2010). Sparse seemingly unrelated regression modelling: Applications in finance and econometrics. *Computational Statistics and Data Analysis* 54:2866–2877.
- Wang, H., Li, R., Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 94:553–568.

- Wang, L., Yang, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *Annals of Statistics* 35:2474–2503.
- Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* 90:43–52.
- Wang, Y. D., Guo, W.S., Brown, B. (2000). Spline smoothing for bivariate data with application between hormones. *Statistica Sinica* 10:377–397.
- Welsh, A. H., Yee, T. W. (2006). Local regression for vector responses. *Journal of Statistical Planning and Inference* 135:307–331.
- Xiao, Z., Linton, O. B., Carroll, R. J., Mammen, E. (2003). More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *Journal of the American Statistical Association* 98:980–992.
- Xu, Q., You, J., Li, X. (2011). Statistical inference on seemingly unrelated non-parametric regression models with serially correlated errors. *Statistica Neerlandica* 65:297–318.
- Xu, Q., You, J., Zhou, B. (2008). Seemingly unrelated nonparametric models with positive correlation and constrained error variances. *Economics Letters* 99:223–227.
- Xue, L. (2009). Variable selection in additive models. *Statistica Sinica* 19:1281–1296.
- Xue, L., Liang, H. (2010). Polynomial spline estimation for the generalized additive coefficient model. *Scandinavian Journal of Statistics* 37:26–46.
- You, J., Xie, S., Zhou, Y. (2007). Two-stage estimation for seemingly unrelated nonparametric regression models. *Journal of Systems Science and Complexity* 20:509–520.
- You, J., Zhou, X. (2014). Asymptotic theory in fixed effects panel data seemingly unrelated partially linear regression models. *Econometric Theory* 30:407–435.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* 57:348–368.
- Zhang, Y., Li, R., Tsai C. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* 105:312–323.
- Zhou, B., Xu, Q., You, J. (2011). Efficient estimation for error component seemingly unrelated nonparametric regression models. *Metrika* 73:121–138.