Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2009, Article ID 487194, 37 pages
doi:10.1155/2009/487194

Research Article

Model and Variable Selection Procedures for Semiparametric Time Series Regression

Risa Kato and Takayuki Shiohama

Department of Management Science, Faculty of Engineering, Tokyo University of Science, Kudankita 1-14-6, Chiyoda, Tokyo 102-0073, Japan

Correspondence should be addressed to Takayuki Shiohama, shiohama@ms.kagu.tus.ac.jp

Received 13 March 2009; Accepted 26 June 2009

Recommended by Junbin Gao

Semiparametric regression models are very useful for time series analysis. They facilitate the detection of features resulting from external interventions. The complexity of semiparametric models poses new challenges for issues of nonparametric and parametric inference and model selection that frequently arise from time series data analysis. In this paper, we propose penalized least squares estimators which can simultaneously select significant variables and estimate unknown parameters. An innovative class of variable selection procedure is proposed to select significant variables and basis functions in a semiparametric model. The asymptotic normality of the resulting estimators is established. Information criteria for model selection are also proposed. We illustrate the effectiveness of the proposed procedures with numerical simulations.

1. Introduction

Non- and semiparametric regression has become a rapidly developing field of statistics in recent years. Various types of nonlinear model such as neural networks, kernel methods, as well as spline method, series estimation, local linear estimation have been applied in many fields. Non- and semiparametric methods, unlike parametric methods, make no or only mild assumptions about the trend or seasonal components and are, therefore, attractive when the data on hand does not meet the criteria for classical time series models. However, the price of this flexibility can be high; when multiple predictor variables are included in the regression equation, nonparametric regression faces the so-called curse of dimensionality.

A major problem associated with non- and semiparametric trend estimation involves the selection of a smoothing parameter and the number of basis functions. Most literature on nonparametric regression with dependent errors focuses on the kernel estimator of the trend function (see, e.g., Altman [1], Hart [2] and Herrmann et al. [3]). These results have been extended to the case with long-memory errors by Hall and Hart [4], Ray and Tsay [5], and Beran and Feng [6]. Kernel methods are affected by the so-called boundary effect. A well-known estimator with automatic boundary correction is the local polynomial approach which is asymptotically equivalent to some kernel estimates. For detailed discussions on local polynomial fitting see, for example, Fan and Gijbels [7] and Fan and Yao [8].

For semiparametric models with serially correlated errors, Gao [9] proposed the semiparametric least-square estimators (SLSEs) for the parametric component and studied its asymptotic properties. You and Chen [10] constructed a semiparametric generalized least-square estimator (SGLSE) with autoregressive errors. Aneiros-Pérez and Vilar-Fernández [11] constructed SLSE with correlated errors.

Like parametric regression models, variable selection of the smoothing parameter for the basis functions is important problem in non- and semiparametric models. It is common practice to include only important variables in the model to enhance predictability. The general approach to finding sensible parameters is to choose an optimal subset determined according to the model selection criterion. Several information criteria for evaluating models constructed by various estimation procedures have been proposed, see, for example, Konishi and Kitagawa [12]. The commonly used criteria are generalized cross-validation, the Akaike information criterion (AIC), and the Bayesian information criterion (BIC). Although best subset selection is practically useful, these selection procedures ignore stochastic errors inherited between the stages of variable selection. Furthermore, best subset selection lacks stability, see, for example, Breiman [13]. Nonconcave penalized likelihood approaches for selecting significant variables for parametric regression models have been proposed by Fan and Li [14]. This methodology can be extended to semiparametric generalized regression models with dependent errors. One of the advantages of this procedure is the simultaneous selection of variables and the estimation of unknown parameters.

The rest of this paper is organized as follows. In Section 2.1 we introduce our semiparametric regression models and explain classical partial ridge regression estimation. Rather than focus on the kernel estimator of the trend function, we use the basis functions to fit the trend component of time series. In Section 2.2, we propose a penalized weighted least-square approach with information criteria for estimation and variable selection. The estimation algorithms are explained in Section 2.3. In Section 2.4, the GIC proposed by Konishi and Kitagawa [15], the BICm proposed by Hastie and Tibshirani [16], and the BICp proposed by Konishi et al. [17] are applied to the evaluation of models estimated by penalized weighted least-square. Section 2.5 contains the asymptotic results of proposed estimators. In Section 3 the performance of these information criteria is evaluated by simulation studies. Section 4 contains the real data analysis. Section 5 concludes our results, and proofs of the theorems are given in the appendix.

2. Estimation Procedures

In this section, we present our semiparametric regression model and estimation procedures.

2.1. The Model and Penalized Estimation

We consider the semiparametric regression model:

$$y_i = \alpha(t_i) + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where y_i is the response variable and $\mathbf{x}_i$ is the $d \times 1$ covariate vector at time i, $\alpha(t_i)$ is an unspecified baseline function of t_i with $t_i = i/n$, $\boldsymbol{\beta}$ is a vector of unknown regression coefficients, and ε_i is a Gaussian and zero mean covariance stationary process.

We assume the following properties for the error terms ε_i and vectors of explanatory variables $\mathbf{x}_i$.

(A.1) It holds that $\{\varepsilon_i\}$ is a linear process given by

$$\varepsilon_i = \sum_{j=0}^{\infty} b_j e_{i-j}, \tag{2.2}$$

where $b_0 = 1$ and $\{e_i\}$ is an i.i.d. Gaussian random variable with $E\{e_i\} = 0$ and $E\{e_i{}^2\} = \sigma_e^2$.

(A.2) The coefficients b_j satisfy the conditions that for all $|z| < 1$, $\sum_{j=0}^{\infty} b_j z^j \neq 0$ and $\sum_{j=0}^{\infty} j^2 |b_j| < \infty$.

We define $\gamma(k) = \mathrm{cov}(\varepsilon_t, \varepsilon_{t+k}) = E\{\varepsilon_t \varepsilon_{t+k}\}$.

The assumptions on covariate variables are as follows.

(B.1) Also $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})' \in \mathbb{R}^d$ and $\{x_{ij}\}, j = 1, \ldots d$, have mean zero and variance 1.

The trend function $\alpha(t_i)$ is expressed as a linear combination of a set of m underlying basis functions:

$$\alpha(t_i) = \sum_{k=1}^{m} w_k \phi_k(t_i) = \mathbf{w}' \boldsymbol{\phi}(t), \tag{2.3}$$

where $\{\boldsymbol{\phi}(t_i) = (\phi_1(t_i), \ldots, \phi_m(t_i))'\}$ is an m-dimensional vector constructed from basis functions $\{\phi_k(t_i); k = 1, \ldots, m\}$, and $\mathbf{w} = (w_1, \ldots, w_m)'$ is an unknown parameter vector to be estimated. The examples of basis functions are B-spline, P-spline, and radial basis functions. A P-spline basis is given by

$$\boldsymbol{\phi}(t_i) = \left(t_i, \ldots, t_i^p, (t_i - \kappa_1)_+^p, \ldots, (t_i - \kappa_k)_+^p\right)', \tag{2.4}$$

where $\{\kappa_k\}_{k=1,\ldots,K}$ are spline knots. This specification uses the so-called truncated power function basis. The choice of the number of knots K and the knot locations are discussed by Yu and Ruppert [18].

Radial basis function (RBF) emerged as a variant of artificial renewal network in late 80s. Nonlinear specification of using RBF has been widely used in cognitive science, engineering, biology, linguistics, and so on. If we consider the RBF modeling, a basis function can take the form

$$\phi_k(t_i) = \exp\left(-\frac{\left\|t_i - \mu_k\right\|^2}{2s_k^2}\right), \tag{2.5}$$

where μ_k determines the location and s_k^2 determines the width of the basis function.

Selecting appropriate basis functions, then the semiparametric regression model (2.1) can be expressed as a linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{B}\mathbf{w} + \boldsymbol{\varepsilon}, \tag{2.6}$$

where $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$, $\mathbf{y} = (y_1, \dots, y_n)'$, $\mathbf{B} = (\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_n)'$ with $\boldsymbol{\phi}_i = (\phi_1(i/n), \dots, \phi_m(i/n))'$. The penalized least-square estimator is then a minimizer of the function

$$\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\mathbf{w})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\mathbf{w}) + n\xi\mathbf{w}'\mathbf{K}\mathbf{w}, \tag{2.7}$$

where ξ is the smoothing parameter controlling the tradeoff between the goodness-of-fit measured by weighted least-square and the roughness of the estimated function. Also $\mathbf{K}$ is an appropriate positive semidefinite symmetric matrix. For example, if $\mathbf{K}$ satisfies $\mathbf{w}'\mathbf{K}\mathbf{w} = \int_0^1 [\alpha''(u)]^2 du$, we have the usual quadratic integral penalty (see, e.g., Green and Silverman [19]). By simple calculus, (2.7) is minimized when $\boldsymbol{\beta}$ and $\mathbf{w}$ satisfy the block matrix equation

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{B} \\ \mathbf{B}'\mathbf{X} & \mathbf{B}'\mathbf{B} + n\xi\mathbf{K} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{w} \end{pmatrix} = \begin{pmatrix} \mathbf{X}' \\ \mathbf{B}' \end{pmatrix} \mathbf{y}. \tag{2.8}$$

This equation can be solved without any iteration (see, e.g., Green [20]). First, we find $\mathbf{B}\widetilde{\mathbf{w}} = \mathbf{S}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, where $\mathbf{S} = \mathbf{B}(\mathbf{B}'\mathbf{B} + \alpha\mathbf{K})^{-1}\mathbf{B}'$ is usually called the smoothing matrix. Substituting $\mathbf{B}\widetilde{\mathbf{w}}$ into (2.6), we obtain

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.9}$$

where $\widetilde{\mathbf{y}} = (\mathbf{I} - \mathbf{S})\mathbf{y}$, $\widetilde{\mathbf{X}} = (\mathbf{I} - \mathbf{S})\mathbf{X}$,and $\mathbf{I}$ is the identity matrix of order n. Applying least-square to the linear model (2.9), we obtain the semiparametric ordinary least-square estimator (SOLSE) result:

$$\widehat{\boldsymbol{\beta}}_{\text{SOLSE}} = \left(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\right)^{-1} \widetilde{\mathbf{X}}'\widetilde{\mathbf{y}}\ , \tag{2.10}$$

$$\widehat{\mathbf{w}}_{\text{SOLSE}} = \left(\mathbf{B}'\mathbf{B} + n\xi\mathbf{K}\right)^{-1} \mathbf{B}'\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{SOLSE}}\right). \tag{2.11}$$

Speckman [21] studied similar solutions for partial linear models with independent observations. Since the errors are serially correlated in model (2.1), $\widehat{\boldsymbol{\beta}}_{\text{SOLSE}}$ is not asymptotically efficient. To obtain an asymptotically efficient estimator for $\boldsymbol{\beta}$, we use the prewhitening transformation. Note that the errors $\{\varepsilon_i\}$ in (2.6) are invertible. Let $b(L) = \sum_{j=1}^{\infty} b_j e_{i-j}$, where L is the lag operator and $a(L) = b(L)^{-1} = a_0 - \sum_{j=1}^{\infty} a_j L^j$ with $a_0 = 1$. Applying $a(L)$ to the model (2.6) and rewriting the corresponding equation, we obtain the new model:

$$\underline{\mathbf{y}} = \underline{\mathbf{X}}\boldsymbol{\beta} + \underline{\mathbf{B}}\mathbf{w} + \mathbf{e}, \tag{2.12}$$

where $\underline{\mathbf{y}} = (\underline{y}_1, \ldots, \underline{y}_n)'$, $\underline{\mathbf{X}} = (\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_n)'$, $\underline{\mathbf{B}} = (\underline{\boldsymbol{\phi}}_1, \ldots, \underline{\boldsymbol{\phi}}_n)'$ and $\mathbf{e} = (e_1, \ldots, e_n)'$. Here

$$\underline{y}_i = y_i - \sum_{j=1}^{\infty} a_j y_{i-j}, \qquad \underline{\boldsymbol{\phi}}_i = \boldsymbol{\phi}_i - \sum_{j=1}^{\infty} a_j \boldsymbol{\phi}_{i-j},$$
$$\underline{\mathbf{x}}_i = \mathbf{x}_i - \sum_{j=1}^{\infty} a_j \mathbf{x}_{i-j}. \tag{2.13}$$

The regression errors in (2.12) are i.i.d. Because, in practice, the response variable y_i is unknown, we use a reasonable approximation by $\underline{y}_i$ based on the work by Xiao et al. [22] and Aneiros-Pérez and Vilar-Fernández [11].

Under the usual regularity conditions the coefficients a_j decrease geometrically so, letting $\tau = \tau(n)$ denote a truncation parameter, we may consider the truncated autoregression on ε_i:

$$e_i = \varepsilon_i - \sum_{j=1}^{\infty} a_j \varepsilon_{i-j}, \tag{2.14}$$

where e_i are i.i.d. random variables with $E(e_i) = 0$. We make the following assumption about the truncation parameter.

(C.1) The truncation parameter τ satisfies $\tau(n) = c \log n$ for some $c > 0$.

The expansion rate of the truncation parameter given in (C.1) is also for convenience. Let $\mathbf{T}_\tau$ be the $n \times n$ transformation matrix such that $\mathbf{e}_\tau = \mathbf{T}_\tau \boldsymbol{\varepsilon}$. Then the model (2.12) can be expressed as

$$\mathbf{T}_\tau \mathbf{y} = \mathbf{T}_\tau \mathbf{X} \boldsymbol{\beta} + \mathbf{T}_\tau \mathbf{B} \mathbf{w} + \mathbf{T}_\tau \boldsymbol{\varepsilon}, \tag{2.15}$$

where

$$\mathbf{T}_\tau = \begin{pmatrix} \delta_{11} & 0 & \cdots & & & & 0 \\ \delta_{21} & -\delta_{22} & 0 & \cdots & & & 0 \\ \vdots & & & & & & \\ \delta_{\tau 1} & \cdots & -\delta_{\tau\tau} & & & & \\ -a_\tau & \cdots & -a_1 & 1 & & & \\ 0 & -a_\tau & \cdots & -a_1 & 1 & & \\ \vdots & & & & & & \\ 0 & \cdots & 0 & -a_\tau & \cdots & -a_1 & 1 \end{pmatrix} \tag{2.16}$$

with $\delta_{11} = \sigma_e / \sqrt{\gamma(0)}$, $\delta_{22} = \sigma_e / \sqrt{(1 - \rho^2(1))\gamma(0)}$, $\delta_{21} = \rho(1)(\sigma_e / \sqrt{(1 - \rho^2(1))\gamma(0)}), \ldots$. Here $\rho(h) = \gamma(h) / \rho(0)$ denotes the lag h autocorrelation function of $\{\varepsilon_i\}$.

Now our estimation problem for the semiparametric time series regression model can be expressed as the minimization of the function

$$\mathcal{L}(\boldsymbol{\beta}, \mathbf{w}) = \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\mathbf{w})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\mathbf{w}) + \alpha\mathbf{w}'\mathbf{K}\mathbf{w}, \tag{2.17}$$

where $\mathbf{V}^{-1} = {\sigma_e}^{-2}\mathbf{T}_\tau'\mathbf{T}_\tau$ and ${\sigma_e}^2 = n^{-1}\|\mathbf{T}_\tau\boldsymbol{\varepsilon}\|^2$. Based on the work by Aneiros-Pérez and Vilar-Fernández [11], an estimator for $\mathbf{T}_\tau$ is constructed as follows. We use the residuals $\widehat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{SOLSE}} - \mathbf{B}\widehat{\mathbf{w}}_{\text{SOLSE}}$ to construct an estimate of $\mathbf{T}_\tau$ using the ordinary least square method applied to the model

$$\widehat{\varepsilon}_i = a_1\widehat{\varepsilon}_{i-1} + \cdots + a_\tau\widehat{\varepsilon}_{i-\tau} + \text{residual}_i. \tag{2.18}$$

Define the estimate $\widehat{\mathbf{a}}_\tau = (\widehat{a}_1, \widehat{a}_2, \ldots, \widehat{a}_\tau)'$ of $\mathbf{a}_\tau = (a_1, a_2, \ldots, a_\tau)'$, where

$$\widehat{\mathbf{a}}_\tau = \left(\widehat{\mathbf{E}}_\tau'\widehat{\mathbf{E}}_\tau\right)^{-1}\widehat{\mathbf{E}}_\tau'\widehat{\boldsymbol{\varepsilon}}, \tag{2.19}$$

where $\widehat{\boldsymbol{\varepsilon}} = (\widehat{\varepsilon}_{\tau+1}, \ldots, \widehat{\varepsilon}_n)$ and $\widehat{\mathbf{E}}_\tau$ is the $(n-\tau)\times\tau$ matrix of regressors with the typical element $\widehat{\varepsilon}_{i-j}$. Then $\widehat{\mathbf{T}}_\tau$ is obtained from $\mathbf{T}_\tau$ by replacing a_j with $\widehat{a}_j$, σ_e^2 with $\widehat{\sigma}_e^2$, and so forth. Applying least-square to the linear model, we obtain

$$\widehat{\mathbf{T}}_\tau\mathbf{y} = \widehat{\mathbf{T}}_\tau\mathbf{X}\boldsymbol{\beta} + \widehat{\mathbf{T}}_\tau\mathbf{B}\mathbf{w} + \widehat{\mathbf{T}}_\tau\boldsymbol{\varepsilon}. \tag{2.20}$$

Then

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\text{SGLSE}} &= \left(\widetilde{\mathbf{X}}_{\widehat{\tau}}'\widetilde{\mathbf{X}}_{\widehat{\tau}}\right)^{-1}\widetilde{\mathbf{X}}_{\widehat{\tau}}'\widetilde{\mathbf{y}}_{\widehat{\tau}}, \\ \widehat{\mathbf{w}}_{\text{SGLSE}} &= \left(\mathbf{B}_{\widehat{\tau}}'\mathbf{B}_{\widehat{\tau}} + n\xi\mathbf{K}\right)^{-1}\mathbf{B}_{\widehat{\tau}}'\left(\mathbf{y}_{\widehat{\tau}} - \mathbf{X}_{\widehat{\tau}}\widehat{\boldsymbol{\beta}}_{\text{SGLSE}}\right), \end{aligned} \tag{2.21}$$

where $\widetilde{\mathbf{X}}_{\widehat{\tau}} = (\mathbf{I} - \mathbf{S})\mathbf{X}_{\widehat{\tau}}$ and $\widetilde{\mathbf{y}}_{\widehat{\tau}} = (\mathbf{I} - \mathbf{S})\mathbf{y}_{\widehat{\tau}}$, with $\mathbf{y}_{\widehat{\tau}} = \widehat{\mathbf{T}}_\tau\mathbf{y}$ and $\mathbf{X}_{\widehat{\tau}} = \widehat{\mathbf{T}}_\tau\mathbf{X}$. The following theorem shows that the loss in efficiency associated with the estimation of the autocorrelation structure is modest in large samples.

Theorem 2.1. *Let the conditions of (A.1), (A.2), (B.1), and (C.1) hold, and assume that* $\boldsymbol{\Sigma}_1 = \lim_{n\to\infty} n^{-1}\widetilde{\mathbf{X}}'\mathbf{V}^{-1}\widetilde{\mathbf{X}}$ *is nonsingular. Let* $\boldsymbol{\beta}_0$ *denote the true value of* $\boldsymbol{\beta}$*, then*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) = \sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{SGLSE}} - \boldsymbol{\beta}_0\right) + O_p\left(\left(\frac{\tau}{n}\right)^{1/2}\right), \tag{2.22}$$

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{SGLSE}} - \boldsymbol{\beta}_0\right) \underset{D}{\rightarrow} N\left(0, \boldsymbol{\Sigma}_1^{-1}\right), \tag{2.23}$$

where $\underset{D}{\rightarrow}$ *denotes convergence in distribution and* $\widehat{\boldsymbol{\beta}} = (\mathbf{X}_\tau'\mathbf{X}_\tau)^{-1}\mathbf{X}_\tau'\mathbf{y}_\tau$. *Assume that* $\boldsymbol{\Sigma}_2 = \lim_{n\to\infty} n^{-1}\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}$ *is nonsingular and let* $\mathbf{w_0}$ *denote the true value of* $\mathbf{w}$, *then one has*

$$\sqrt{n}(\widehat{\mathbf{w}} - \mathbf{w}_0) = \sqrt{n}(\widehat{\mathbf{w}}_{\mathrm{SGLSE}} - \mathbf{w}_0) + O_p\left(\left(\frac{\tau}{n}\right)^{1/2}\right), \tag{2.24}$$

$$\sqrt{n}(\widehat{\mathbf{w}}_{\mathrm{SGLSE}} - \mathbf{w}_0) \underset{D}{\rightarrow} N\left(0, \boldsymbol{\Sigma}_2{}^{-1}\right), \tag{2.25}$$

where $\widehat{\mathbf{w}} = (\mathbf{B}_\tau'\mathbf{B}_\tau + n\xi\mathbf{K})^{-1}\mathbf{B}_\tau'(\mathbf{y}_\tau - \mathbf{X}_\tau\widehat{\boldsymbol{\beta}})$.

2.2. Variable Selection and Penalized Least Squares

Variable and model selection are an indispensable tool for statistical data analysis. However, it has rarely been studied in the semiparametric context. Fan and Li [23] studied penalized weighted least-square estimation with variable selection in semiparametric models for longitudinal data. In this section, we introduce the penalized weighted least-square approach. We propose an algorithm for calculating the penalized weighted least-square estimator of $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathbf{w}')'$ in Section 2.3. In Section 2.4 we present the information criteria for the model selection.

From now on, we assume that the matrices $\mathbf{X}_\tau$ and $\mathbf{B}_\tau$ are standardized so that each column has mean 0 and variance 1. The first term in (2.7) can be regarded as a loss function of $\boldsymbol{\beta}$ and $\mathbf{w}$, which we will denote by $l(\boldsymbol{\beta}, \mathbf{w})$. Then expression (2.7) can be written as

$$\mathcal{L}(\boldsymbol{\beta}, \mathbf{w}) = l(\boldsymbol{\beta}, \mathbf{w}) + n\xi\mathbf{w}'\mathbf{K}\mathbf{w}. \tag{2.26}$$

The methodology in the previous section can be applied to the variable selection via penalized least-square. A form of penalized weighted least-square is

$$\mathcal{S}(\boldsymbol{\beta}, \mathbf{w}) = l(\boldsymbol{\beta}, \mathbf{w}) + n\left(\sum_{i=1}^{d} p_{\lambda_1}(|\beta_i|) + \sum_{j=1}^{m} p_{\lambda_2}(|w_j|)\right) + n\xi\mathbf{w}'\mathbf{K}\mathbf{w}. \tag{2.27}$$

where $p_{\lambda_i}(\cdot)$ are penalty functions and λ_i are regularization parameters, which control the model complexity. By minimizing (2.27) with a special construction of the penalty function given in what following some coefficients are estimated as 0, which deletes the corresponding variables, whereas others are not. Thus, the procedure selects variables and estimates coefficients simultaneously. The resulting estimate is called a penalized weighted least-square estimate.

Many penalty functions have been used for penalized least-square and penalized likelihood in various non- and semiparametric models. There are strong connections between the penalized weighted least-square and the variable selection. Denote by $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathbf{w}')'$ and $\mathbf{z} = (z_1, \ldots, z_{d+m})'$ the true parameters and the estimates, respectively. By taking the hard thresholding penalty function

$$p_\lambda(|\theta|) = \lambda^2 + (|\theta| - \lambda)^2 I(|\theta| < \lambda), \tag{2.28}$$

we obtain the hard thresholding rule

$$\widehat{\theta} = zI(|z| > \lambda). \tag{2.29}$$

The L_2 penalty $P_\lambda(|\theta|) = \lambda|\theta|^2$ results in a ridge regression and the L_1 penalty $P_\lambda(|\theta|) = \lambda|\theta|$ yields a soft thresholding rule

$$\widehat{\theta} = \operatorname{sgn}(z)I(|z| > \lambda)_+. \tag{2.30}$$

This solution gives the best subset selection via stepwise deletion and addition. Tibshirani [24, 25] has proposed LASSO, which is the penalized least-square estimate with the L_1 penalty, in the general least-square and likelihood settings.

2.3. An Estimation Algorithm

In this section we describe an algorithm for calculating the penalized least-square estimator of $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathbf{w}')'$. The estimate of $\boldsymbol{\theta}$ minimizes the penalized sum of squares $\mathcal{L}(\boldsymbol{\theta})$ given by (2.17). First we obtain $\widehat{\boldsymbol{\theta}}_{\text{SOLSE}}$ in *Step* 1. In *Step* 2, we estimate $\mathbf{T}_\tau$ by using ε obtained in *Step* 1. Then $\widehat{\boldsymbol{\theta}}^{\text{HT}}_{\text{SGLSE}}$ is obtained using $\widehat{\mathbf{T}}_\tau$(*Step* 3). Here the penalty parameters λ, and ξ, and the number of basis functions m are chosen using information criteria that will be discussed in Section 2.4.

Step 1. First we obtain $\widehat{\boldsymbol{\beta}}_{\text{SOLSE}}$ and $\widehat{\mathbf{w}}_{\text{SOLSE}}$ by (2.10) and (2.11), respectively. Then we have the model

$$\widehat{\mathbf{y}} = \mathbf{B}\widehat{\mathbf{w}}_{\text{SOLSE}} + \mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{SOLSE}} + \boldsymbol{\varepsilon}. \tag{2.31}$$

Step 2. An estimator for $\mathbf{T}_\tau$ is constructed followings the work of Aneiros-Pérez and Vilar-Fernández [4]. We use the residuals $\widehat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{B}\widehat{\mathbf{w}}_{\text{SOLSE}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{SOLSE}}$ to construct an estimate of $\mathbf{T}_\tau$ using the ordinary least square method applied to the model

$$\widehat{\varepsilon}_i = a_1\widehat{\varepsilon}_{i-1} + \cdots + a_\tau\widehat{\varepsilon}_{i-\tau} + \text{residual}_i. \tag{2.32}$$

The estimator $\widehat{\mathbf{T}}_\tau$ is obtained from $\mathbf{T}_\tau$ by replacing parameters with their estimates.

Step 3. Our SGLSE of $\boldsymbol{\theta}$ is obtained by using the model

$$\mathbf{y}_{\widehat{\tau}} = \mathbf{B}_{\widehat{\tau}}\mathbf{w} + \mathbf{X}_{\widehat{\tau}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{\widehat{\tau}}, \tag{2.33}$$

where $\mathbf{y}_{\widehat{\tau}} = \widehat{\mathbf{T}}_\tau\mathbf{y}$, $\mathbf{B}_{\widehat{\tau}} = \widehat{\mathbf{T}}_\tau\mathbf{B}$, $\mathbf{X}_{\widehat{\tau}} = \widehat{\mathbf{T}}_\tau\mathbf{X}$, and $\boldsymbol{\varepsilon}_\tau = \widehat{\mathbf{T}}_\tau\boldsymbol{\varepsilon}$. Finding the solution of the penalized least-square of (2.27) needs the local quadratic approximation, because the L_1 and hard thresholding penalty are irregular at the origin and may not have second derivatives at some points. We follow the methodology of Fan and Li [14]. Suppose that we are given an initial value $\boldsymbol{\theta}^{(0)}$ that is close to the minimizer of (2.27). If $\theta_j^{(0)}$ is very close to 0, then set $\widehat{\theta}_j^{(0)} = 0$. Otherwise they can be locally approximated by a quadratic function as

$$\left[p_{\lambda_j}(\theta_j)\right]' = p'_{\lambda_j}(|\theta_j|)\operatorname{sgn}(\theta_j) \approx \left\{\frac{p'_{\lambda_j}\left(\left|\theta_j^{(0)}\right|\right)}{\left|\theta_j^{(0)}\right|}\right\}\theta_j, \tag{2.34}$$

when $\widehat{\theta}_j^{(0)} \neq 0$. Therefore, the minimization problem (2.27) can be reduced to a quadratic minimization problem and the Newton-Raphson algorithm can be used. The right-hand side of equation (2.27) can be locally approximated by

$$\begin{aligned} &l(\boldsymbol{\beta}_0, \mathbf{w}_0) + \nabla l_\beta(\boldsymbol{\beta}, \mathbf{w})'(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \nabla l_w(\boldsymbol{\beta}_0, \mathbf{w}_0)'(\mathbf{w} - \mathbf{w}_0) \\ &\quad + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\nabla^2_{\beta\beta}(\boldsymbol{\beta}_0, \mathbf{w}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)'\nabla^2_{ww}(\boldsymbol{\beta}_0, \mathbf{w}_0)(\mathbf{w} - \mathbf{w}_0), \\ &\quad \frac{1}{2}(\boldsymbol{\beta}_0 - \boldsymbol{\beta})'\nabla^2_{\beta w}(\mathbf{w} - \mathbf{w}_0) + n\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\lambda_1}(\boldsymbol{\beta}_0)\boldsymbol{\beta} + n\mathbf{w}'\boldsymbol{\Sigma}_{\lambda_2}(\mathbf{w_0})\mathbf{w}, \end{aligned} \tag{2.35}$$

where

$$\begin{gathered} \nabla l_\beta(\boldsymbol{\beta}_0, \mathbf{w}_0) = \frac{\partial l(\boldsymbol{\beta}_0, \mathbf{w}_0)}{\partial \boldsymbol{\beta}}, \qquad \nabla l_w(\boldsymbol{\beta}_0, \mathbf{w}_0) = \frac{\partial l(\boldsymbol{\beta}_0, \mathbf{w}_0)}{\partial \mathbf{w}}, \\ \nabla^2 l_{\beta\beta}(\boldsymbol{\beta}_0, \mathbf{w}_0) = \frac{\partial^2 l(\boldsymbol{\beta}_0, \mathbf{w}_0)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}, \qquad \nabla^2 l_{ww}(\boldsymbol{\beta}_0, \mathbf{w}_0) = \frac{\partial^2 l(\boldsymbol{\beta}_0, \mathbf{w}_0)}{\partial \mathbf{w}\partial \mathbf{w}'}, \\ \nabla^2 l_{\beta,w}(\boldsymbol{\beta}_0, \mathbf{w}_0) = \frac{\partial^2 l(\boldsymbol{\beta}_0, \mathbf{w}_0)}{\partial \boldsymbol{\beta}\partial \mathbf{w}}, \\ \boldsymbol{\Sigma}_{\lambda_1}(\boldsymbol{\beta}_0) = \operatorname{diag}\left\{\frac{p'_{\lambda_1}\left(\left|\beta_1^{(0)}\right|\right)}{\left|\beta_1^{(0)}\right|}, \ldots, \frac{p'_{\lambda_1}\left(\left|\beta_d^{(0)}\right|\right)}{\left|\beta_d^{(0)}\right|}\right\}, \\ \boldsymbol{\Sigma}_{\lambda_2}(\mathbf{w}_0) = \operatorname{diag}\left\{\frac{p'_{\lambda_2}\left(\left|w_1^{(0)}\right|\right)}{\left|w_1^{(0)}\right|}, \ldots, \frac{p'_{\lambda_2}\left(\left|w_m^{(0)}\right|\right)}{\left|w_m^{(0)}\right|}\right\}. \end{gathered} \tag{2.36}$$

The solution can be found by iteratively computing the block matrix equation:

$$\begin{pmatrix} \mathbf{X}'_{\widehat{\tau}}\mathbf{X}_{\widehat{\tau}} + n\boldsymbol{\Sigma}_{\lambda_1}\left(\boldsymbol{\beta}^{(0)}\right) & \mathbf{X}'_{\widehat{\tau}}\mathbf{B}_{\widehat{\tau}} \\ \mathbf{B}'_{\widehat{\tau}}\mathbf{X}_{\widehat{\tau}} & \mathbf{B}'_{\widehat{\tau}}\mathbf{B}_{\widehat{\tau}} + \alpha\mathbf{K} + n\boldsymbol{\Sigma}_{\lambda_2}\left(\mathbf{w}^{(0)}\right) \end{pmatrix}\begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{w} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'_{\widehat{\tau}} \\ \mathbf{B}'_{\widehat{\tau}} \end{pmatrix}\mathbf{y}. \tag{2.37}$$

This gives the estimators

$$\widehat{\boldsymbol{\beta}}_{\mathrm{SGLSE}}^{\mathrm{HT}} = \left(\widetilde{\mathbf{X}}_{\hat{\tau}}'\widetilde{\mathbf{X}}_{\hat{\tau}} + n\boldsymbol{\Sigma}_{\lambda_1}(\boldsymbol{\beta}^{(0)})\right)^{-1}\widetilde{\mathbf{X}}_{\hat{\tau}}'\widetilde{\mathbf{y}}_{\hat{\tau}},$$
$$\widehat{\mathbf{w}}_{\mathrm{SGLSE}}^{\mathrm{HT}} = \left(\mathbf{B}_{\hat{\tau}}'\mathbf{B}_{\hat{\tau}} + n\xi\mathbf{K} + n\boldsymbol{\Sigma}_{\lambda_2}(\mathbf{w}^{(0)})\right)^{(-1)}\mathbf{B}_{\hat{\tau}}'\left(\mathbf{y}_{\hat{\tau}} - \mathbf{X}_{\hat{\tau}}\widehat{\boldsymbol{\beta}}_{\mathrm{SGLSE}}^{\mathrm{HT}}\right), \tag{2.38}$$

where $\widetilde{\mathbf{y}}_{\hat{\tau}} = (\mathbf{I} - \mathbf{S}_{\hat{\tau}})\mathbf{y}_{\hat{\tau}}$, $\widetilde{\mathbf{X}}_{\hat{\tau}} = (\mathbf{I} - \mathbf{S}_{\hat{\tau}})\mathbf{X}_{\hat{\tau}}$, and $\mathbf{S}_{\hat{\tau}} = \mathbf{B}_{\hat{\tau}}(\mathbf{B}_{\hat{\tau}}'\mathbf{B}_{\hat{\tau}} + n\xi\mathbf{K} + n\boldsymbol{\Sigma}_{\lambda_2}(\mathbf{w}^{(0)}))^{-1}\mathbf{B}_{\hat{\tau}}'$.

2.4. Information Criteria

Selecting suitable values for the penalty parameters and number of basis functions is crucial to obtaining good curve fitting and variable selection. The estimate of $\boldsymbol{\theta}$ minimizes the penalized sum of squares $\mathcal{L}(\boldsymbol{\theta})$ given by (2.17). In this section, we express the model (2.15) as

$$\mathbf{y}_{\tau} = \mathbf{A}_{\tau}\boldsymbol{\theta} + \mathbf{e}, \tag{2.39}$$

where $\mathbf{A}_{\tau} = (\mathbf{X}_{\tau}, \mathbf{B}_{\tau})$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathbf{w}')'$. In many applications, the number of basis functions m needs to be large to adequately capture the trend. To determine the number of basis functions, all models with $m \leq m_{\max}$ are fitted and the preferred model minimizes some model selection criteria.

The Schwarz BIC is given by

$$\mathrm{BIC} = n\log\left(2\pi\widehat{\sigma}_e^2\right) + \log n\left(\text{the number of parameters}\right), \tag{2.40}$$

where $\widehat{\sigma}_e^2$ is the least-square estimate of σ_e^2 without a degree of freedom correction. Hastie and Tibshirani [16] used the trace of the smoother matrix as an approximation to the effective number of parameters. By replacing the number of parameters in BIC by $\mathrm{tr}\mathbf{S}_{\beta}$, we formally obtain information criteria for the basis function Gaussian regression model in the form

$$\mathrm{BIC}m = n\log\left(2\pi\widehat{\sigma}_e^2\right) + (\mathrm{tr}\mathbf{S}_{\boldsymbol{\theta}})\log n, \tag{2.41}$$

where $\widehat{\sigma}_e^2 = n^{-1}\|y - \mathbf{S}_{\boldsymbol{\theta}}y\|^2$ and

$$\mathrm{tr}\mathbf{S}_{\boldsymbol{\theta}} = \mathbf{A}_{\tau}\left(\mathbf{A}_{\tau}'\mathbf{A}_{\tau} + n\xi\widetilde{\mathbf{K}} + n\boldsymbol{\Sigma}_{\lambda}(\boldsymbol{\theta})\right)^{-1}\mathbf{A}_{\tau}'. \tag{2.42}$$

Here $\boldsymbol{\Sigma}_{\lambda}(\boldsymbol{\theta})$ is defined by (2.44) in what follows.

We also consider the use of the BICp criterion to choose appropriate values for these unknown parameters. Denote

$$\boldsymbol{\Sigma}_{\lambda_1}(\boldsymbol{\beta}) = \operatorname{diag}\left\{p''_{\lambda_1}(|\beta_{10}|), \ldots, p''_{\lambda_1}(|\beta_{d0}|)\right\},$$
$$\boldsymbol{\Sigma}_{\lambda_2}(\mathbf{w}) = \operatorname{diag}\left\{p''_{\lambda_2}(|w_{10}|), \ldots, p''_{\lambda_2}(|w_{m0}|)\right\}, \tag{2.43}$$

$$\boldsymbol{\Sigma}_{\lambda}(\boldsymbol{\theta}) = (\boldsymbol{\Sigma}_{\lambda_1}(\boldsymbol{\beta}), \boldsymbol{\Sigma}_{\lambda_2}(\mathbf{w})). \tag{2.44}$$

Let N_1 and N_2 be the number of zero components in $\boldsymbol{\beta}_0$ and $\mathbf{w}_0$, respectively. Then the BICp criterion is

$$\begin{aligned}\mathrm{BIC}p = {} & n\log\left(2\pi\widehat{\sigma}_e^2\right) + n\widehat{\boldsymbol{\theta}}'\boldsymbol{\Sigma}_{\lambda}(\boldsymbol{\theta})\widehat{\boldsymbol{\theta}} + n\xi\widehat{\boldsymbol{\theta}}\widetilde{\mathbf{K}}\widehat{\boldsymbol{\theta}} + \log\left|\mathbf{J}_G\left(\widehat{\boldsymbol{\theta}}\right)\right| \\ & - \log\left|\widetilde{\mathbf{K}}\right|_+ - \log|\Sigma_{\lambda}(\boldsymbol{\theta})|_+ - (m - N_2)\log\xi + \mathrm{Const},\end{aligned} \tag{2.45}$$

where $\mathbf{J}_G(\widehat{\boldsymbol{\theta}})$ is the $(d+m+1)\times(d+m+1)$ matrix of second derivatives of the penalized likelihood defined by

$$\mathbf{J}_G = \frac{1}{n\widehat{\sigma}_e^2}\begin{pmatrix} \mathbf{A}'_{\tau}\mathbf{A}_{\tau} + n\Sigma_{\lambda}(\boldsymbol{\theta}) + n\xi\widetilde{\mathbf{K}} & \dfrac{\mathbf{A}'_{\tau}\Lambda\mathbf{1}_n}{\sigma_e^2} \\ \dfrac{\mathbf{1}'_n\Lambda\mathbf{A}_{\tau}}{\widehat{\sigma}_e^2} & \dfrac{n}{2\widehat{\sigma}_e^2}\end{pmatrix}. \tag{2.46}$$

Here Λ is a diagonal matrix with ith element $\boldsymbol{\Lambda}_i = \operatorname{diag}[e_1, \ldots, e_n]$ and $\mathbf{1}_n = (1, \ldots, 1)'$. The n-dimensional vector $\mathbf{q}$ has ith element $(\mathbf{T}_{ij}\mathbf{y}_j - \mathbf{A}_{\tau,ij}\boldsymbol{\theta}_j)^2/2\widehat{\sigma}_e^4 - 1/2\widehat{\sigma}_e^2$ where $\mathbf{T}_{ij}$ is the element in the ith row and jth column of $\mathbf{T}_{\tau}$. Also $\widetilde{\mathbf{K}}$ is the $(d+m)\times(d+m)$ matrix defined by

$$\widetilde{\mathbf{K}} = \begin{pmatrix} \mathbf{K} & \mathbf{O}_{d,m} \\ \mathbf{O}_{m,d} & \mathbf{O}_{m,m}\end{pmatrix}, \tag{2.47}$$

and $|\widetilde{\mathbf{K}}|_+$ and $|\Sigma_{\lambda}(\boldsymbol{\theta})|_+$ are the product of the $(m-N_1)$ and $(d+m-N_1-N_2)$ nonzero eigenvalues of $\widetilde{\mathbf{K}}$ and $\Sigma_{\lambda}(\boldsymbol{\theta})$, respectively.

Konishi and Kitagawa [15] proposed a framework of Generalized Information Criteria (GIC) to the case where the models are not estimated by maximum likelihood. Hence, we also consider the use of GIC for the model evaluations. The GIC for the hard thresholding penalty function is given by

$$\mathrm{GIC} = n\log\left(2\pi\widehat{\sigma}_e^2\right) + n + 2\operatorname{tr}\left\{\mathbf{I}_G\mathbf{J}_G^{-1}\right\}, \tag{2.48}$$

where $\mathbf{I}_G$ is a $(m+d+1)\times(m+d+1)$ matrix. Also $\mathbf{I}_G$ is basically the product of the empirical influence function and the score function. It is defined by

$$\mathbf{I}_G = \frac{1}{n\widehat{\sigma}_e^2}\begin{pmatrix} \dfrac{\mathbf{A}_\tau'\Lambda}{\sigma_e^2} - \boldsymbol{\Sigma}_\lambda(\boldsymbol{\theta})\widehat{\boldsymbol{\theta}}\mathbf{1}_\mathbf{n}' - \xi\widetilde{\mathbf{K}}\widehat{\boldsymbol{\theta}}\mathbf{1}_\mathbf{n}' \\ \mathbf{q}' \end{pmatrix}\left(\Lambda\mathbf{A}_\tau, \widehat{\sigma}_e^2\mathbf{q}\right). \tag{2.49}$$

The number of basis functions m, penalty parameters ξ,λ_1,λ_2 are determined by minimizing BICm, BICp or GIC.

2.5. Sampling Properties

We now study the asymptotic properties of the estimate resulting from the penalized least-square function (2.27).

First we establish the convergence rate of the penalized profile least-square estimator. Assume that penalty functions $p'_{\lambda_{1j}}(\cdot)$ and $p'_{\lambda_{2j}}(\cdot)$ are negative and nondecreasing with $p'_{\lambda_{1j}}(0) = p'_{\lambda_{2j}}(0) = 0$. Let $\boldsymbol{\beta}_0$ and $\mathbf{w}_0$ denote the true values of $\boldsymbol{\beta}$ and $\mathbf{w}$, respectively. Also let

$$\begin{aligned} a_{1n} &= \max_j\left\{\left|p'_{\lambda_{1j}}(|\beta_{j0}|)\right| : \beta_{j0}\neq 0\right\}, \qquad a_{2n} = \max_j\left\{\left|p'_{\lambda_{2j}}(|w_{j0}|)\right| : w_{j0}\neq 0\right\}, \\ b_{1n} &= \max_j\left\{\left|p''_{\lambda_{1j}}(|\beta_{j0}|)\right| : \beta_{j0}\neq 0\right\}, \qquad b_{2n} = \max_j\left\{\left|p''_{\lambda_{2j}}(|w_{j0}|)\right| : w_{j0}\neq 0\right\}. \end{aligned} \tag{2.50}$$

Theorem 2.2. *Under the conditions of Theorem 2.1, if a_{1n}, b_{1n}, a_{2n}, and b_{2n} tend to 0 as $n\to\infty$, then with probability tending to 1, there exist local minimizers $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{w}}$ of $\mathcal{L}(\boldsymbol{\beta},\mathbf{w})$ such that $\|\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{\mathrm{SGLSE}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2}+a_{1n})$ and $\|\widehat{\mathbf{w}}^{\mathrm{HT}}_{\mathrm{SLOSE}} - \mathbf{w}_0\| = O_p(n^{-1/2}+a_{2n})$.*

Theorem 2.2 demonstrates how the rate of convergence of the penalized least-square estimator $\widehat{\boldsymbol{\theta}}^{\mathrm{HT}}_{\mathrm{SGLSE}} = (\widehat{\boldsymbol{\beta}}^{\prime\mathrm{HT}}_{\mathrm{SGLSE}}, \widehat{\mathbf{w}}^{\prime\mathrm{HT}}_{\mathrm{SGLSE}})'$ of $\mathcal{L}(\boldsymbol{\theta})$ depends on λ_{ij} for $i=1,2$. To achieve the root n convergence rate, we have to take λ_{ij} small enough so that $a_n = O_p(n^{-1/2})$.

Next we establish the oracle property for the penalized least-square estimator. Let $\boldsymbol{\beta}_{S_1}$ consist of all nonzero components of $\boldsymbol{\beta}_0$ and let $\boldsymbol{\beta}_{N_1}$ consist of all zero components. Let $\mathbf{w}_{S_2}$ consist of all nonzero components of $\mathbf{w}_0$ and let $\mathbf{w}_{N_2}$ consist of all zero components. Let

$$\begin{aligned} \widehat{\mathbf{x}}(t)'\boldsymbol{\beta}_0 &= \widehat{\mathbf{x}}'_{S_1}\boldsymbol{\beta}_{S_1} + \widehat{\mathbf{x}}'_{N_1}(t)\boldsymbol{\beta}_{N_1} = \widehat{\mathbf{x}}_{S_1}(t)'\boldsymbol{\beta}_{S_1}, \\ \boldsymbol{\phi}(t)'\mathbf{w}_0 &= \phi'_{S_2}\mathbf{w}_{S_2} + \phi'_{N_2}(t)\mathbf{w}_{N_2} = \phi_{S_2}(t)'\mathbf{w}_{S_2}. \end{aligned} \tag{2.51}$$

Write

$$\begin{aligned} \mathbf{b}_\beta &= \left(p'_{\lambda_{1n}}(|\beta_{10}|)\mathrm{sgn}(\beta_{10}),\ldots,p'_{\lambda_{S_1n}}(|\beta_{S_10}|)\mathrm{sgn}(\beta_{S_10})\right)', \\ \mathbf{b}_w &= \left(p'_{\lambda_{2n}}(|w_{10}|)\mathrm{sgn}(w_{10}),\ldots,p'_{\lambda_{S_2n}}(|w_{S_20}|)\mathrm{sgn}(w_{S_20})\right)'. \end{aligned} \tag{2.52}$$

Further, let $\widehat{\boldsymbol{\beta}}_1$ consist of the first S_1 components of $\widehat{\boldsymbol{\beta}}$ and let $\widehat{\boldsymbol{\beta}}_2$ consist of the last $d - S_1$ components of $\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{\mathrm{SGLSE}}$. Let $\widehat{\mathbf{w}}_1$ consist of the first S_2 components of $\widehat{\mathbf{w}}$ and let $\widehat{\mathbf{w}}_2$ consist of the last $m - S_2$ components of $\widehat{\mathbf{w}}^{\mathrm{HT}}_{\mathrm{SGLSE}}$.

Theorem 2.3. *Assume that for $j = 1,\ldots,d$ and $k = 1,\ldots,m$, one has $\lambda_1 \to 0$, $\sqrt{n}\lambda_1 \to \infty$, $\lambda_2 \to 0$ and $\sqrt{n}\lambda_2 \to \infty$. Assume that the penalty functions $p'_{\lambda_1}(|\beta_j|)$ and $p'_{\lambda_2}(|w_k|)$ satisfy*

$$\begin{gathered}\liminf_{n\to\infty}\liminf_{\beta_j\to 0+}\frac{p'_{\lambda_1}(\beta_j)}{\lambda_1} > 0,\\ \liminf_{n\to\infty}\liminf_{\omega_k\to 0+}\frac{p'_{\lambda_2}(\omega_k)}{\lambda_2} > 0.\end{gathered} \tag{2.53}$$

If $a_{1n} = a_{2n} = O_p(n^{-1/2})$ then, under the conditions of Theorem 2.1, with probability tending to 1, the root n consistent local minimizers $\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{\mathrm{SGLSE}} = (\widehat{\boldsymbol{\beta}}'_1, \widehat{\boldsymbol{\beta}}'_2)'$ and $\widehat{\mathbf{w}}^{\mathrm{HT}}_{\mathrm{SGLSE}} = (\widehat{\mathbf{w}}'_1, \widehat{\mathbf{w}}'_2)'$ in Theorem 2.2 must satisfy the following:

(1) *(sparsity)* $\widehat{\boldsymbol{\beta}}_2 = \widehat{\mathbf{w}}_2 = \mathbf{0}$;

(2) *(asymptotic normality)*

$$\begin{gathered}\sqrt{n}(\mathbf{I}_{S_1} + \boldsymbol{\Sigma}_{\lambda_1}(\boldsymbol{\beta}))\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\mathbf{I}_{S_1} + \boldsymbol{\Sigma}_{\lambda_1}(\boldsymbol{\beta}))^{-1}\mathbf{b}_\beta\right) \longrightarrow N_{S_1}\left(0, \boldsymbol{\Sigma}^{-1}_{1(1)}\right),\\ \sqrt{n}(\mathbf{I}_{S_2} + \boldsymbol{\Sigma}_{\lambda_1}(\mathbf{w}))\left(\widehat{\mathbf{w}}_1 - \mathbf{w}_{10} + (\mathbf{I}_{S_2} + \boldsymbol{\Sigma}_{\lambda_2}(\mathbf{w}) + \xi\mathbf{K})^{-1}\mathbf{b}_w\right) \longrightarrow N_{S_2}\left(0, \boldsymbol{\Sigma}^{-1}_{2(1)}\right)\end{gathered} \tag{2.54}$$

Here $\boldsymbol{\Sigma}^{-1}_{1(1)}$ and $\boldsymbol{\Sigma}^{-1}_{2(1)}$ consist of the first S_1 and S_2 rows and columns of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ defined in Theorem 2.1, respectively.

3. Numerical Simulations

We now assess the performance of semiparametric estimators of the proposed in previous section via simulations. We generate simulation data from the model

$$y_i = \alpha(t_i) + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \tag{3.1}$$

where $\alpha(t_i) = \exp(-3(i/n))\sin(3\pi i/n)$, $\boldsymbol{\beta} = (3, 1.5, 0, 0.2, 0, 0, 0)'$ and $\varepsilon(t)$ is a Gaussian AR(1) process with autoregressive coefficient ρ. We used the radial basis function network modeling to fit the trend component. We simulate the covariate vector x from a normal distribution with mean 0 and $\mathrm{cov}(x_i, x_j) = 0.5^{|i-j|}$. In each case, the autoregressive coefficient is set to 0, 0.25, 0.5 or 0.75 and the sample size n is set to 50, 100 or 200. Figure 1 depicts some examples of simulation data.

We compare the effectiveness of our proposed procedure (PLS + HT) with an existing procedure (PLS). We also compare the performance of the information criteria BICm, GIC and BICp for evaluating the models. As discussed in Section 3, the proposed procedure (PLS + HT) excludes basis functions as well as explanatory variables.

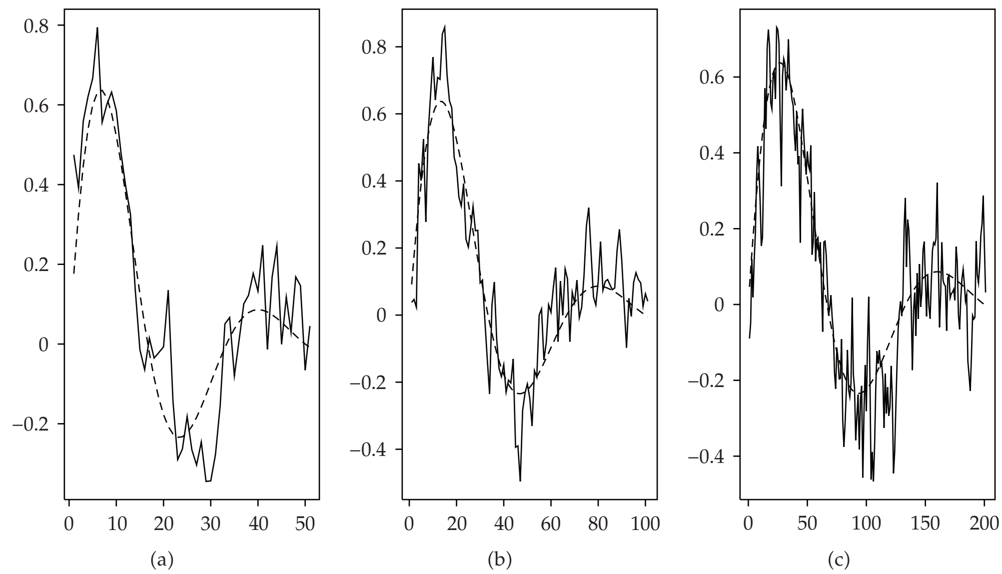

Figure 1: Simulation data with (a) $n = 50$ and $\rho = 0.5$, (b) $n = 100$ and $\rho = 0.5$, (c) $n = 200$ and $\rho = 0.5$. The dotted lines represent $\alpha(t)$; the solid lines $\alpha(t) + \varepsilon(t)$.

First we assess the performance of $\widehat{\alpha}(t)$ by the square root of average squared errors (RASE_α):

$$\text{RASE}_\alpha = \sqrt{n_{\text{grid}}^{-1} \sum_{k=1}^{n_{\text{grid}}} \{\widehat{\alpha}(t_k) - \alpha(t_k)\}^2}, \tag{3.2}$$

where $\{t_k, k = 1, \ldots, n_{\text{grid}}\}$ are the grid points at which the baseline function $\alpha(\cdot)$ is estimated. In our simulation, we use $n_{\text{grid}} = 200$. Table 1 shows the means and standard deviations of RASE_α for $\rho = 0, 0.25, 0.5, 0.75$ based on 500 simulations. RASE_α increases as the autoregressive coefficient increases but decreases as the sample size increases. From Table 1, we see that the proposed procedure (PLS + HT) works better than PLS and that models evaluated by BICp work better than those based on BICm or GIC.

Then the performance of $\widehat{\boldsymbol{\beta}}$ is assessed by the square root of average squared errors (RASE_β):

$$\text{RASE}_\beta = \sqrt{\frac{1}{d} \sum_{i=1}^{d} \left(\widehat{\beta}_i - \beta_i\right)^2}. \tag{3.3}$$

The means and standard deviations of RASE_β for $\rho = 0, 0.25, 0.5, 0.75$ based on 500 simulations are shown in Table 2. We can see that the proposed procedure (PLS + HT) works better than the existing procedure. There is almost no change in RASE_β as the autoregressive coefficient changes (unlike the procedure of You and Chen [10]), whereas RASE_β depends strongly on the information, BICp works the best among the criteria. We can also confirm the consistency of the estimator, that is RASE_β decreases as the sample size increases.

Table 1: Means (standard deviations) of RASE_α.

		$\rho = 0.0$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$
$n = 50$					
PLS	BICm	0.069 (0.013)	0.081(0.017)	0.114 (0.037)	0.232 (0.115)
	GIC	0.047 (0.013)	0.062 (0.015)	0.106 (0.037)	0.229 (0.120)
	BICp	0.042 (0.010)	0.060 (0.019)	0.103 (0.039)	0.226 (0.124)
PLS + HT	BICm	0.061 (0.040)	0.070 (0.021)	0.101 (0.038)	0.226 (0.103)
	GIC	0.053 (0.017)	0.068 (0.020)	0.101 (0.034)	0.218 (0.097)
	BICp	0.046 (0.015)	0.060 (0.019)	0.093 (0.034)	0.214 (0.101)
$n = 100$					
PLS	BICm	0.041(0.008)	0.052 (0.012)	0.080 (0.025)	0.172 (0.080)
	GIC	0.034 (0.008)	0.044 (0.011)	0.074 (0.026)	0.170 (0.080)
	BICp	0.036 (0.010)	0.044 (0.010)	0.070 (0.024)	0.163 (0.079)
PLS + HT	BICm	0.042 (0.008)	0.051 (0.016)	0.080 (0.024)	0.172 (0.079)
	GIC	0.040 (0.015)	0.048 (0.016)	0.073 (0.024)	0.168 (0.078)
	BICp	0.037 (0.011)	0.041 (0.011)	0.068 (0.023)	0.158 (0.075)
$n = 200$					
PLS	BICm	0.029 (0.005)	0.040 (0.016)	0.058 (0.018)	0.129 (0.056)
	GIC	0.025 (0.008)	0.033 (0.010)	0.056 (0.018)	0.125 (0.057)
	BICp	0.029 (0.006)	0.031 (0.007)	0.050 (0.015)	0.114 (0.052)
PLS + HT	BICm	0.030 (0.005)	0.040 (0.016)	0.058 (0.019)	0.127 (0.053)
	GIC	0.027 (0.009)	0.033 (0.011)	0.054 (0.015)	0.123 (0.054)
	BICp	0.019 (0.008)	0.028 (0.009)	0.047 (0.018)	0.109 (0.048)

Table 2: Means (standard deviations) of RASE_β.

		$\rho = 0.0$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$
$n = 50$					
PLS	BICm	0.022 (0.007)	0.023 (0.007)	0.022 (0.007)	0.020 (0.007)
	GIC	0.021 (0.006)	0.023 (0.007)	0.023 (0.010)	0.021 (0.007)
	BICp	0.021(0.006)	0.022 (0.007)	0.022 (0.009)	0.020 (0.007)
PLS + HT	BICm	0.011(0.005)	0.013 (0.007)	0.012 (0.007)	0.010 (0.005)
	GIC	0.010 (0.004)	0.013 (0.007)	0.013 (0.009)	0.011(0.006)
	BICp	0.010 (0.004)	0.011 (0.005)	0.011 (0.006)	0.010 (0.005)
$n = 100$					
PLS	BICm	0.014 (0.004)	0.014 (0.004)	0.014 (0.005)	0.012 (0.004)
	GIC	0.013 (0.004)	0.014 (0.004)	0.013 (0.004)	0.012 (0.004)
	BICp	0.014 (0.004)	0.014 (0.004)	0.013 (0.004)	0.011 (0.004)
PLS + HT	BICm	0.007 (0.003)	0.008 (0.004)	0.007 (0.004)	0.006 (0.003)
	GIC	0.007 (0.003)	0.008 (0.004)	0.007 (0.003)	0.006 (0.003)
	BICp	0.007 (0.003)	0.007 (0.003)	0.006 (0.003)	0.006 (0.003)
$n = 200$					
PLS	BICm	0.009 (0.003)	0.009 (0.003)	0.009 (0.003)	0.007 (0.002)
	GIC	0.009 (0.003)	0.009 (0.003)	0.008 (0.003)	0.007 (0.002)
	BICp	0.009 (0.003)	0.009 (0.003)	0.008 (0.002)	0.007 (0.002)
PLS + HT	BICm	0.004 (0.002)	0.005 (0.002)	0.005 (0.002)	0.005 (0.002)
	GIC	0.005 (0.002)	0.005 (0.002)	0.005 (0.002)	0.004 (0.002)
	BICp	0.005 (0.002)	0.005 (0.002)	0.004 (0.002)	0.005 (0.002)

Table 3: Means (standard deviations) of first step ahead prediction errors.

		$\rho = 0.0$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$
$n = 50$					
PLS	BICm	0.136 (0.115)	0.150 (0.116)	0.140 (0.120)	0.158 (0.117)
	GIC	0.111 (0.088)	0.127 (0.097)	0.134 (0.098)	0.149 (0.122)
	BICp	0.111 (0.088)	0.127 (0.097)	0.131 (0.095)	0.149 (0.122)
PLS + HT	BICm	0.121 (0.096)	0.106 (0.086)	0.119 (0.092)	0.139 (0.112)
	GIC	0.094 (0.071)	0.118 (0.093)	0.126 (0.094)	0.139 (0.112)
	BICp	0.095 (0.071)	0.116 (0.092)	0.124 (0.093)	0.139 (0.112)
$n = 100$					
PLS	BICm	0.101 (0.086)	0.105 (0.082)	0.130 (0.112)	0.145 (0.124)
	GIC	0.090 (0.070)	0.101 (0.078)	0.105 (0.082)	0.137 (0.109)
	BICp	0.091 (0.070)	0.096 (0.072)	0.105 (0.092)	0.137 (0.109)
PLS + HT	BICm	0.097 (0.082)	0.096 (0.078)	0.098 (0.088)	0.140 (0.162)
	GIC	0.084 (0.063)	0.091 (0.071)	0.103 (0.081)	0.130 (0.111)
	BICp	0.084 (0.063)	0.091 (0.071)	0.103 (0.081)	0.130 (0.111)
$n = 200$					
PLS	BICm	0.091 (0.070)	0.105 (0.081)	0.114 (0.087)	0.174 (0.129)
	GIC	0.087 (0.068)	0.095 (0.072)	0.102 (0.077)	0.139 (0.114)
	BICp	0.086 (0.068)	0.095 (0.072)	0.102 (0.077)	0.139 (0.114)
PLS + HT	BICm	0.084 (0.066)	0.090 (0.069)	0.091 (0.068)	0.123 (0.096)
	GIC	0.083 (0.063)	0.090 (0.069)	0.098 (0.076)	0.126 (0.100)
	BICp	0.082 (0.063)	0.092 (0.070)	0.098 (0.076)	0.126 (0.100)

The first step ahead prediction error (PE), which is defined as

$$\mathrm{PE} = \sqrt{\left(\hat{y}_{n+1} - y_{n+1|n}\right)^2}, \tag{3.4}$$

is also investigated. Table 3 shows the means and standard errors of PE for $\rho = 0, 0.25, 0.5, 0.75$ based on 500 simulations. The PE increases as the autoregressive coefficient increases, but the PE decreases as the sample size increases. From Table 3, we see that PLS + HT works better than the existing procedures and there is almost no difference in the PE depending on the information criteria. The models evaluated by BICm perform well for large sample sizes.

The means and standard deviations of the number and deviation of basis functions are shown in Tables 4 and 5. The BICp gives a smaller number of basis functions than the other information criteria. The models evaluated by BICp also give smaller standard deviations of the number of basis functions. The models determined by BICp tend to choose larger deviations of basis functions than those based on BICm and GIC. The number of basis functions increases gradually as the sample size or ρ increase. From Table 4, it appears that the number of basis functions does not depend on the sample size n. From Table 5, it also appears that the deviations of basis functions do not depend on the sample size n and ρ.

We now compare the performance of our procedure with existing procedures in terms of the reduction of model complexity. Table 6 shows simulation results of the means and standard deviations of the number of parameters excluded ($\beta = 0$ or $w = 0$) by the proposed procedure. The results indicate that the proposed procedure reduces model complexity. From Table 6, It appears that the models determined by BICp tend to exclude fewer parameters

Table 4: Means (standard deviations) of the number of basis functions.

	$\rho = 0.0$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$
$n = 50$				
BICm	7.87 (1.38)	8.85 (1.13)	8.76 (1.24)	8.82 (1.17)
GIC	8.06 (1.44)	8.75 (1.27)	8.84 (1.20)	8.84 (1.24)
BICp	6.02 (0.14)	6.15 (0.53)	6.17 (0.37)	6.21 (0.48)
$n = 100$				
BICm	7.98 (1.31)	8.83 (1.17)	8.71 (1.30)	8.71 (1.30)
GIC	8.01 (1.37)	8.91 (1.18)	8.67 (1.29)	8.95 (1.20)
BICp	6.20 (0.50)	6.22 (0.44)	6.31 (0.60)	6.35 (0.66)
$n = 200$				
BICm	7.93 (1.33)	8.18 (1.44)	8.25 (1.48)	8.20 (1.39)
GIC	8.11 (1.35)	8.11 (1.52)	8.39 (1.41)	8.55 (1.37)
BICp	6.15 (0.66)	6.22 (0.73)	6.46 (1.03)	6.93 (1.43)

Table 5: Means (standard deviations) of the deviations of basis functions.

	$\rho = 0.0$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$
$n = 50$				
BICm	0.10 (0.02)	0.10 (0.02)	0.10 (0.02)	0.10 (0.02)
GIC	0.11 (0.03)	0.10 (0.03)	0.10 (0.03)	0.10 (0.03)
BICp	0.14 (0.02)	0.18 (0.03)	0.16 (0.03)	0.16 (0.03)
$n = 100$				
BICm	0.10 (0.02)	0.09 (0.02)	0.09 (0.02)	0.09 (0.03)
GIC	0.11 (0.03)	0.09 (0.02)	0.10 (0.03)	0.09 (0.02)
BICp	0.15 (0.02)	0.15 (0.04)	0.15 (0.03)	0.13 (0.03)
$n = 200$				
BICm	0.10 (0.02)	0.11 (0.03)	0.11 (0.03)	0.10 (0.03)
GIC	0.11 (0.03)	0.12 (0.04)	0.11 (0.04)	0.10 (0.03)
BICp	0.15 (0.03)	0.17 (0.02)	0.16 (0.03)	0.14 (0.04)

and give smaller standard deviations for the number of parameters excluded. This is due to the selection of a smaller number of basis functions compared to the selection based on the other criteria (see Table 4). There is almost no dependence of the number of excluded parameters on ρ. The models evaluated by BICp give a larger number of excluded parameters as the sample size increases. On the other hand, the models evaluated by BICm or GIC give a smaller number of excluded parameters as the sample size increases.

Table 7 shows the means and standard deviations of the number of basis functions excluded as $w = 0$ by the proposed procedure. From Table 7 it appears that the models evaluated by BICp tend to exclude fewer basis functions than those based on GIC and BIC. Again this is due to the selection of a smaller number of basis functions (see Table 4). The models determined by BICp also give smaller standard deviations of the number of basis functions than the other criteria. There is almost no dependence of the number of basis functions on ρ.

Table 8 shows the means and standard deviations of the number of basis functions excluded as $\beta = 0$ by the proposed procedure. The number of β which values really 0 was five. From Table 8 we see that the proposed procedure gives nearly five. The models determined

Table 6: Means (standard deviations) of the number of parameters excluded.

		$\rho = 0.0$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$
$n = 50$					
PLS + HT	BICm	7.715 (0.915)	6.910 (1.087)	7.300 (1.364)	6.888 (1.343)
	GIC	8.345 (1.568)	7.404 (1.850)	7.620 (1.715)	7.337 (1.598)
	BICp	4.950 (0.419)	5.020 (0.502)	5.070 (0.492)	5.092 (0.421)
$n = 100$					
PLS + HT	BICm	7.506 (0.784)	7.334 (1.251)	5.698 (0.772)	5.460 (0.700)
	GIC	7.916 (1.239)	7.718 (1.435)	5.906 (0.919)	5.740 (0.866)
	BICp	4.990 (0.184)	5.076 (0.332)	5.092 (0.316)	5.086 (0.327)
$n = 200$					
PLS + HT	BICm	7.062 (0.723)	5.594 (0.744)	5.544 (0.736)	5.460 (0.702)
	GIC	7.450 (1.116)	5.764 (0.847)	5.656 (0.864)	5.586 (0.802)
	BICp	5.008 (0.109)	5.152 (0.359)	5.162 (0.385)	5.086 (0.356)

Table 7: Means (stnadard deviations) of the number of basis functions excluded.

	$\rho = 0.0$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$
$n = 50$				
BICm	3.52 (2.29)	4.21 (2.23)	3.98 (1.60)	3.96 (1.49)
GIC	3.74 (2.15)	4.40 (1.90)	4.18 (1.51)	4.26 (1.46)
BICp	1.03 (0.22)	1.20 (0.60)	1.28 (0.54)	1.24 (0.49)
$n = 100$				
BICm	3.35 (2.19)	4.49 (2.04)	3.78 (1.58)	3.95 (1.60)
GIC	3.67 (2.15)	4.62 (1.84)	3.91 (1.53)	4.30 (1.60)
BICp	1.06 (0.31)	1.78 (0.96)	1.31 (0.60)	1.36 (0.66)
$n = 200$				
BICm	3.64 (2.13)	3.26 (1.71)	3.26 (1.71)	3.61 (1.60)
GIC	3.86 (2.02)	3.43 (1.81)	3.65 (1.69)	3.89 (1.76)
BICp	1.12 (0.34)	1.23 (0.75)	1.46 (1.03)	1.93 (1.44)

by BICp give results more close to five and smaller standard deviations of the number of basis functions than the other criteria. The number of basis functions approaches five as the sample size increases. The standard deviations of the number of basis functions excluded decrease as ρ increases. These results indicate that the proposed procedure reduces model complexity.

Table 9 shows the percentage of times that various β_i were estimated as being zero. As for the parameters $\beta_j \neq 0,\ j = 1, 2, 5$, these parameters were not estimated zero for every simulations, we omit to show the corresponding results on Table 9. The results indicate that the proposed procedure excludes insignificant variables and selects significant variables. It can be seen that the proposed procedure gives a better performance as the sample size increases and that BICp is superior to the other criteria.

4. Real Data Analysis

In this section we present the consequence of analyzing the real-time series data using proposed procedure. We use two data in this study; the data about the spirit consumption data in United Kingdom and the association between fertility and female employment in Japan.

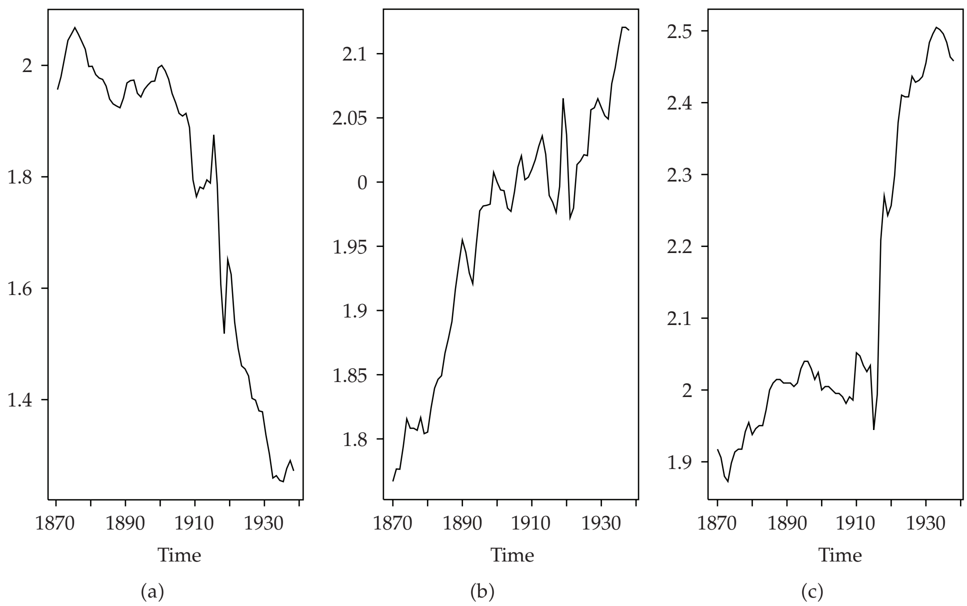

Figure 2: Application data-set: (a): y_i = log(the annual per capita consumption of spirits); (b): x_{i1} = log(per capita income); (c): x_{i2} = log(price of spirits).

Table 8: Means (standard deviations) of the number of explanatory variables excluded.

	$\rho = 0.0$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$
$n = 50$				
BICm	4.14 (1.60)	4.15 (1.63)	4.69 (0.92)	4.79 (0.74)
GIC	4.28 (1.47)	4.35 (1.41)	4.70 (0.89)	4.72 (0.87)
BICp	4.97 (0.21)	4.95 (0.26)	4.97 (0.23)	4.99 (0.14)
$n = 100$				
BICm	4.15 (1.59)	4.17 (1.55)	4.72 (0.92)	4.77 (0.87)
GIC	4.22 (1.51)	4.29 (1.47)	4.77 (0.84)	4.65 (1.03)
BICp	4.98 (0.14)	4.95 (0.26)	5.00 (0.04)	5.00 (0.06)
$n = 200$				
BICm	4.14 (1.59)	4.78 (0.82)	4.78 (0.82)	4.72 (0.86)
GIC	4.16 (1.55)	4.68 (1.01)	4.75 (0.88)	4.66 (1.04)
BICp	4.99 (0.11)	4.99 (0.15)	5.00 (0.00)	5.00 (0.04)

4.1. The Spirit Consumption Data in the United Kingdom

We now illustrate our theory through an application to spirit consumption data for the United Kingdom from 1870 to 1938. The data-set can be found in Fuller [26, page 523]. In this data-set, the dependent variable y_i is the logarithm of the annual per capita consumption of spirits. The explanatory variables x_{i1} and x_{i2} are the logarithms of per capita income and the price of spirits, respectively, and $x_{i3} = x_{i1}x_{i2}$. Figure 2 shows that there is a change-point at the start of the First World War (1914). Therefore, we prepare a variable z: $z = 0$ from 1870 to 1914 and

Table 9: Percentage of times β_i is estimated as zero.

		$\beta_3 = 0$	$\beta_4 = 0$	$\beta_6 = 0$	$\beta_7 = 0$	$\beta_8 = 0$
$n = 50$	$\rho = 0$					
	BICm	0.84	0.83	0.82	0.83	0.82
	GIC	0.87	0.85	0.87	0.86	0.83
	BICp	1.00	0.99	0.99	1.00	1.00
	$\rho = 0.25$					
	BICm	0.83	0.83	0.84	0.83	0.83
	GIC	0.86	0.86	0.86	0.89	0.87
	BICp	0.99	0.99	0.99	0.99	0.98
	$\rho = 0.50$					
	BICm	0.95	0.93	0.94	0.94	0.93
	GIC	0.94	0.94	0.93	0.94	0.95
	BICp	0.99	0.99	1.00	1.00	0.99
	$\rho = 0.75$					
	BICm	0.96	0.96	0.95	0.95	0.97
	GIC	0.94	0.93	0.95	0.94	0.96
	BICp	1.00	1.00	1.00	1.00	1.00
$n = 100$	$\rho = 0$					
	BICm	0.83	0.83	0.84	0.82	0.82
	GIC	0.85	0.84	0.85	0.84	0.84
	BICp	1.00	0.99	1.00	0.99	1.00
	$\rho = 0.25$					
	BICm	0.83	0.84	0.83	0.82	0.85
	GIC	0.87	0.85	0.88	0.85	0.85
	BICp	0.99	0.99	0.99	0.99	1.00
	$\rho = 0.50$					
	BICm	0.95	0.93	0.95	0.95	0.94
	GIC	0.96	0.95	0.94	0.96	0.95
	BICp	1.00	1.00	1.00	1.00	1.00
	$\rho = 0.75$					
	BICm	0.96	0.95	0.95	0.95	0.95
	GIC	0.93	0.94	0.94	0.92	0.92
	BICp	1.00	1.00	1.00	1.00	1.00
$n = 200$	$\rho = 0$					
	BICm	0.92	0.93	0.92	0.91	0.94
	GIC	0.94	0.94	0.94	0.95	0.95
	BICp	1.00	1.00	1.00	1.00	0.99

Table 9: Continued.

	$\beta_3 = 0$	$\beta_4 = 0$	$\beta_6 = 0$	$\beta_7 = 0$	$\beta_8 = 0$
$\rho = 0.25$					
BICm	0.95	0.94	0.94	0.95	0.93
GIC	0.94	0.94	0.94	0.94	0.93
BICp	1.00	1.00	1.00	1.00	1.00
$\rho = 0.50$					
BICm	0.97	0.95	0.95	0.96	0.95
GIC	0.96	0.95	0.95	0.94	0.95
BICp	1.00	1.00	1.00	1.00	1.00
$\rho = 0.75$					
BICm	0.96	0.94	0.95	0.95	0.93
GIC	0.93	0.93	0.93	0.94	0.94
BICp	1.00	1.00	1.00	1.00	1.00

Table 10: Estimated Coefficients for Model 4.1.

	PLS estimators	SE	PLS + HT estimators	SE
β_1	−0.653	3.080	0	
β_2	−1.121	5.962	0	
β_3	1.842	9.164	0	
β_4	3.570	3.761	2.395	0.421
β_5	−2.553	4.455	0	
β_6	−1.25	7.763	−2.411	0.524

$z = 1$ from 1915 to 1933. From this we derive another three explanatory variables: $x_{i4} = x_{i1}z$, $x_{i5} = x_{i2}z$, and $x_{i6} = x_{i3}z$. We consider the semiparametric model:

$$y_i = \alpha(t_i) + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \quad i = 1, \ldots, 69. \tag{4.1}$$

We computed the basis function estimate for α using the existing procedure (PLS) and the proposed procedure (PLS + HT) with BICp. The resulting estimates and standard errors (SEs) of $\boldsymbol{\beta}$ are given in Table 10. The selected number of basis function is seven with one excluded basis function and the spread parameter s is estimated as 0.12. Therefore, we obtain the model

$$\widehat{y}_i = \widehat{\alpha}(t_i) + 2.395x_{i4} - 2.411x_{i5}, \quad i = 1, \ldots, 69. \tag{4.2}$$

The results indicate that the proposed procedure excludes some variable and reduces model complexity. Table 10 shows that PLS + HT selects only β_4 and β_6. That indicates possible interactions between consumption and income and between consumption and income×price after 1915. Consumption increases as income increases; however, as income increases and the price also increases, consumption decreases. We plot the estimated trend curve, residuals and autocorrelations functions in Figures 3 to 5. The residual mean square is 1.7×10^{-4}.

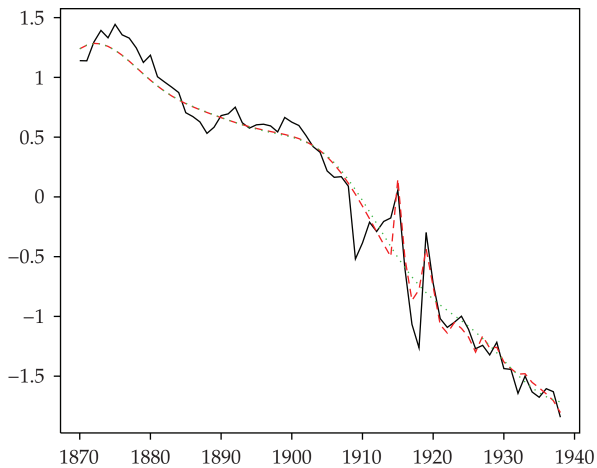

Figure 3: Plots of estimated curves. The solid line represents y; the dotted lines are the estimates of y; the dashed lines are the estimated curve $\widehat{\alpha}$.

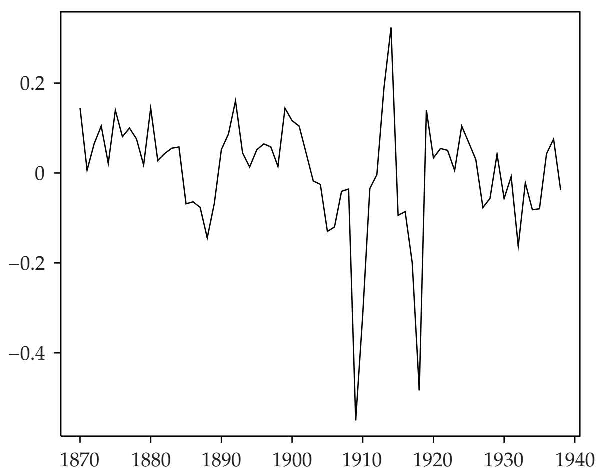

Figure 4: Plot of residuals.

You and Chen [10] used the following semiparametric partially linear regression model:

$$y_i = \alpha(t_i) + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad i = 1, \ldots, 69. \tag{4.3}$$

The semiparametric least-square (SLS) regression gives $\widehat{y}_i = \widehat{\alpha}(t_i) + 0.65x_{i1} - 0.95x_{i2}$. The residual mean square is 2.2×10^{-4}, which is more than that of our SGLSE fit. For a fair comparison, we use model (4.3) to revise You and Chen's estimation. Our semiparametric generalized least-square gives $\widehat{y}_i = \widehat{\alpha}(t_i) - 0.71x_{i2}$. The result indicates that x_{i1} is insignificant in model (4.3).

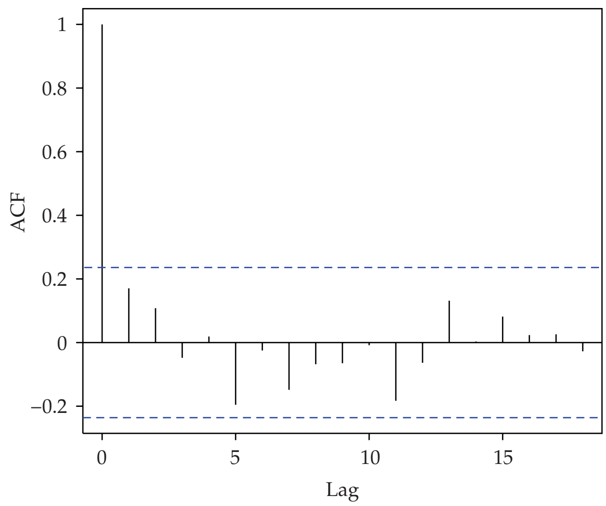

Figure 5: ACF plot of residuals.

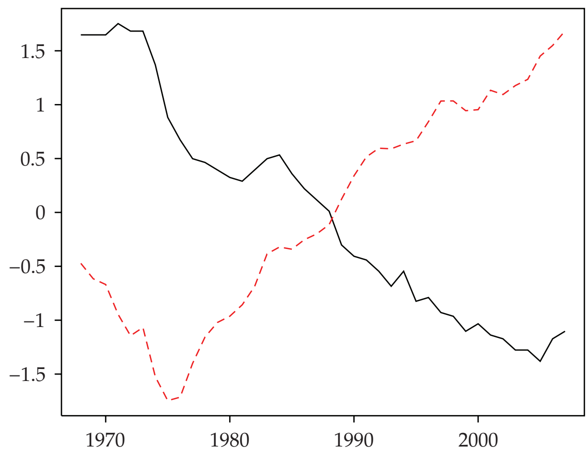

Figure 6: Plots of standardized total fertility rate and female labor force participation rate for women aged 15 to 49 in Japan, 1968–2007. The solid line represents standardized TRF; the dotted lines are standardized FLP.

4.2. The Association between Fertility and Female Employment in Japan

Recent literature finds that in OECD countries the cross-country correlation between the total fertility rate and the female labor force participation rate, which until the beginning of the 1980s had a negative value, has since acquired a positive value. See for example, Brewster and Rindfuss [27], Ahn and Mira [28], and Engelhardt et al. [29]. This result is often interpreted as evidence for a changing sign in the time series association between fertility and female employment within OECD countries.

However, OECD countries, including Japan, have different cultural backgrounds. We investigate whether or not the relation between the total fertility rate (TFR) and the female labor force participation rate (FLP) has changed in Japan from a negative value to a positive value. This application challenges previous findings and could be good news for policy

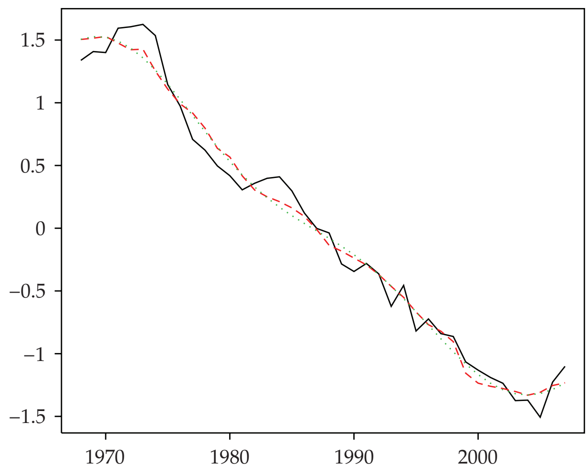

Figure 7: Plots of estimated curves, the solid line represents y; the dotted lines are the estimated curves of y; the dashed lines are the estimated curves of α.

Table 11: Estimated coefficients for Model 4.4.

log(FLP)	PLS estimators	SE	PLS + HT estimators	SE
for 1968–1984	−0.32	−1.99	−0.36	−2.16
for 1984–2007	−0.28	−2.00	−0.31	−2.18
for 1968–1979	0.02	0.17	0	
for 1980–1989	0	1.37	0	
for 1990–1999	−0.04	−0.51	0	
for 2000–2007	0.04	0.17	0	

makers, as a positive relationship implies that a rising FLP is associated with an increasing TFR.

Usually, FLP contains all women aged 15 to 64. However, TFR is a combination of fertility rates for ages 15–49, so we use the FLP of women aged 15 to 49 instead of women aged 15 to 64. We take the TFR from 1968 to 2007 in Japan. The estimation is a semiparametric regression of $\log(\mathrm{TFR}_i)$ on $\log(\mathrm{FLP}_i)$. As the law of the Equal Employment Act came into force in 1985, we use the interaction variables "dummy for 1968–1984 $\times \log(\mathrm{TFR})$" ($x_{i2}$) and for 1985–2007 ($x_{i3}$). We also use dummy variables for 1990–1999 and 2000–2007 (x_{i4}, x_{i5}) and consider the semiparametric model

$$\log(\mathrm{TFR})_i = \alpha(t_i) + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \quad i = 1, \dots, 40. \tag{4.4}$$

We applied the existing procedure (PLS) and proposed procedure (PLS + HT) with BIC_p. The resulting estimates and standard errors (SE) of $\boldsymbol{\beta}$ are given in Table 11. Therefore, we obtain the model

$$\hat{y}_i = \hat{\alpha}(t_i) - 0.27x_{i1} - 0.20x_{i2}, \quad i = 1, \dots, 40. \tag{4.5}$$

The residual mean square of PLS + HT is 2.24×10^{-2} and that of PLS is 2.47×10^{-2}. The selected number of basis functions is six with one excluded basis function and the spread parameter s

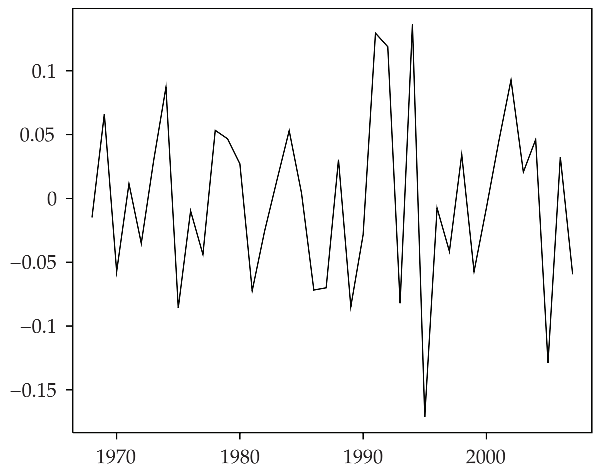

Figure 8: Plot of residuals.

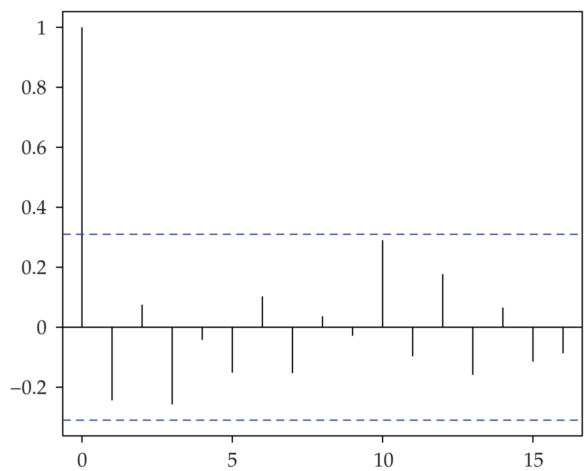

Figure 9: ACF plot of residuals.

is estimated as 0.30. Table 11 shows that PLS + HT selects only $\log(\mathrm{FLP}_i)$ 1968–1984 and 1985–2007. That indicates a negative correlation between TFR and FLP for 1968–2007, especially for 1968–1984, which means TFR decreases as FLP increases. We could not see the positive association in 80s which has been reported in recent studies, see, for example, Brewster and Rindfuss [27], Ahn and Mira [28], and Engelhardt et al. [29]. We plot the estimated trend curve, residuals and autocorrelation functions in Figures 7 to 9.

5. Concluding Remarks

In this article we have proposed variable and model selection procedures for the semi-parametric time series regression. We used the basis functions to fit the trend component. An algorithm of estimation procedures is provided and the asymptotic properties are investigated. From the numerical simulations, we have confirmed that the models determined by the proposed procedure are superior to those based on the existing procedure. They reduce the complexity of models and give good fitting by excluding basis functions and nuisance explanatory variables.

The development here is limited to the case with Gaussian stationary errors, but it seems likely our approach can be extended to the case with non-Gaussian long-range dependent errors, along with the lines explored in recent work by Aneiros-Pérez et al. [30]. However, the efficient estimation for regression parameter is an open question in case of long-range dependence. This is a question we hope to address in future work. We also plan to explore the question of whether the proposed techniques can be extended to the cointegrating regression models with an autoregressive distributed lag framework.

Appendix

Proofs

In this appendix we give the proofs of the theorems in Section 2. We use $\|x\|$ to denote the Euclidian norm of x.

Let $\mathbf{a}_{\tau,n} = (a_{1,n}, \ldots, a_{\tau,n})'$ be the infeasible estimator for $\mathbf{a}_\tau = (a_1, \ldots, a_\tau)'$ constructed using OLS methods. That is $\mathbf{a}_{\tau,n} = (a_{1,n}, a_{2,n}, \ldots, a_{\tau,n})' = (\mathbf{E}_\tau'\mathbf{E}_\tau)^{-1}\mathbf{E}_\tau'\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} = (\varepsilon_{\tau+1}, \ldots, \varepsilon_n)'$ and $\mathbf{E}_\tau = [\varepsilon_{i,j}]$, $i = 1, \ldots, n$, $j = 1, \ldots, \tau$ with $\varepsilon_{i,j} = \varepsilon_{i-j-\tau}$. For ease of notation, we set $\widehat{a}_{j,n} = a_{j,n} = 0$ for $j > \tau$, and $\widehat{a}_{0,n} = a_{0,n} = 1$. We write $\Gamma(k)$ for $\operatorname{cov}(\varepsilon_0, \varepsilon_k)$. Then we can construct the infeasible estimate $\mathbf{V}$ using $\mathbf{a}_{\tau,n}$ and $\Gamma(k), k = 0, \ldots, \tau$. The following lemma states that the estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{w}}$ given in Theorem 2.1 have asymptotically normal distributions.

Lemma .1. *Under the assumptions of Theorem 2.1, one has*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \underset{D}{\rightarrow} N\left(0, \boldsymbol{\Sigma}_1^{-1}\right), \tag{A.1}$$

$$\sqrt{n}(\widehat{\mathbf{w}} - \mathbf{w}) \underset{D}{\rightarrow} N\left(0, \boldsymbol{\Sigma}_2^{-1}\right), \tag{A.2}$$

where $\boldsymbol{\Sigma}_1^{-1}$ and $\boldsymbol{\Sigma}_2^{-1}$ are defined in Theorem 2.1.

Proof of Lemma .1. From model (2.6), $\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\mathbf{w}$ can be written as

$$\begin{aligned}
\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\mathbf{w} &= \mathbf{y} - \mathbf{B}\mathbf{w} + \left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{X}}\boldsymbol{\beta}\right) - \left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{X}}\boldsymbol{\beta}\right) - \mathbf{X}\boldsymbol{\beta} \\
&= \left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{X}}\boldsymbol{\beta}\right) + \left(\widetilde{\mathbf{X}} - \mathbf{X}\right)\boldsymbol{\beta} - (\widetilde{\mathbf{y}} - \mathbf{y}) - \mathbf{B}\mathbf{w} \\
&= \left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{X}}\boldsymbol{\beta}\right) + \mathbf{S}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \mathbf{B}\mathbf{w} \\
&= \left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{X}}\boldsymbol{\beta}\right) - \mathbf{B}(\mathbf{w} - \widetilde{\mathbf{w}}),
\end{aligned} \tag{A.3}$$

where $\tilde{\mathbf{y}}, \tilde{\mathbf{X}}$, and $\tilde{\mathbf{w}}$ are given by $\tilde{\mathbf{y}} = (\mathbf{I}-\mathbf{S})\mathbf{y}$, $\tilde{\mathbf{X}} = (\mathbf{I}-\mathbf{S})\mathbf{X}$, and $\tilde{\mathbf{w}} = (\mathbf{B}'\mathbf{B}+n\xi\mathbf{K})^{-1}\mathbf{B}'\boldsymbol{\varepsilon}$, respectively. Hence $\tilde{\mathbf{w}}$ can be expressed without using $\boldsymbol{\beta}$. Then the minimization function $\mathcal{L}(\boldsymbol{\beta}, \mathbf{w})$ in (2.17) can be written as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, \mathbf{w}) &= \frac{1}{2}\Big\{\left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right)'\mathbf{V}^{-1}\left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right) - 2(\tilde{\mathbf{w}} - \mathbf{w})'\mathbf{B}'\mathbf{V}^{-1}\left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right) \\ &\qquad + (\mathbf{w} - \tilde{\mathbf{w}})'\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}(\mathbf{w} - \tilde{\mathbf{w}})\Big\} + \alpha\mathbf{w}'\mathbf{K}\mathbf{w} \\ &\equiv I_1(\boldsymbol{\beta}) + I_2(\boldsymbol{\beta}, \mathbf{w}) + I_3(\mathbf{w}) + I_4(\mathbf{w}).\end{aligned} \tag{A.4}$$

First we consider asymptotic normality for $\hat{\mathbf{w}}$, using the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{B}\mathbf{w}_0 + \boldsymbol{\varepsilon}. \tag{A.5}$$

The estimators $\hat{\mathbf{w}}$ minimize the function $\mathcal{L}(\boldsymbol{\beta}, \mathbf{w})$, which yields

$$\begin{aligned}\frac{\partial\mathcal{L}(\boldsymbol{\beta}, \mathbf{w})}{\partial(\mathbf{w})} &= I_2'(\boldsymbol{\beta}, \mathbf{w}) + I_3'(\mathbf{w}) + I_4'(\mathbf{w}) \\ &= -\mathbf{B}'\mathbf{V}^{-1}\left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right) + \mathbf{B}'\mathbf{V}^{-1}\mathbf{B}(\mathbf{w} - \mathbf{w}_0) \\ &\qquad + \mathbf{B}'\mathbf{V}^{-1}\mathbf{B}(\mathbf{w}_0 - \tilde{\mathbf{w}}) + 2n\xi\mathbf{K}(\mathbf{w} - \mathbf{w}_0) + n\xi\mathbf{K}\mathbf{w}_0.\end{aligned} \tag{A.6}$$

Then the minimization of this quadratic function is given by

$$\begin{aligned}\hat{\mathbf{w}} - \mathbf{w}_0 &= \left(\mathbf{B}'\mathbf{V}^{-1}\mathbf{B} + n\xi\mathbf{K}\right)^{-1}\mathbf{B}'\mathbf{V}^{-1}\Big\{\left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right) - \mathbf{B}(\tilde{\mathbf{w}} - \mathbf{w}_0) - n\xi\mathbf{V}\mathbf{B}^{-1}\mathbf{K}\mathbf{w}_0\Big\} \\ &= \left(\mathbf{B}'\mathbf{V}^{-1}\mathbf{B} + n\xi\mathbf{K}\right)^{-1}\mathbf{B}'\mathbf{V}^{-1}\left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right) \\ &\quad + \left(\mathbf{B}'\mathbf{V}^{-1}\mathbf{B} + n\xi\mathbf{K}\right)^{-1}\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}(\mathbf{w}_0 - \tilde{\mathbf{w}}) \\ &\quad - n\xi\left(\mathbf{B}'\mathbf{V}^{-1}\mathbf{B} + n\xi\mathbf{K}\right)^{-1}\mathbf{B}'\mathbf{V}^{-1}\mathbf{V}\mathbf{B}^{-1}\mathbf{K}\mathbf{w}_0 \\ &\equiv A_1 + A_2 + A_3.\end{aligned} \tag{A.7}$$

We now deal with A_1, A_2, and A_3. First we evaluate A_1. From the expansion $(\mathbf{A} + a\mathbf{B})^{-1} = \mathbf{A}^{-1} - a\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} + O(a^2)$, we can see that

$$\begin{aligned}
A_1 &= \left(\mathbf{B}'\mathbf{V}^{-1}\mathbf{B} + n\xi\mathbf{K}\right)^{-1}\mathbf{B}'\mathbf{V}^{-1}\boldsymbol{\varepsilon} \\
&= \left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n} + \xi\mathbf{K}\right)^{-1}\frac{\mathbf{B}'\mathbf{V}^{-1}\boldsymbol{\varepsilon}}{n} \\
&= \left\{\left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)^{-1}\mathbf{K}\mathbf{w}_0 + \xi\left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)^{-1}\mathbf{K}\left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)^{-1}\mathbf{K}\mathbf{w}_0 + O\left(\xi^2\right)\right\}\frac{\mathbf{B}'\mathbf{V}^{-1}}{n}\boldsymbol{\varepsilon} \\
&= \left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)^{-1}\frac{\mathbf{B}'\mathbf{V}^{-1}}{n}\boldsymbol{\varepsilon} - \xi\left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)^{-1}\mathbf{K}\left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)^{-1}\frac{\mathbf{B}'\mathbf{V}^{-1}}{n}\boldsymbol{\varepsilon} + O\left(\xi^2\right)\frac{\mathbf{B}'\mathbf{V}^{-1}}{n}\boldsymbol{\varepsilon} \\
&= \left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)^{-1}\frac{\mathbf{B}'\mathbf{V}^{-1}}{n}\boldsymbol{\varepsilon} + O(\xi) + O\left(\xi^2\right).
\end{aligned} \tag{A.8}$$

Similarly, we obtain

$$\begin{aligned}
A_2 &= \left(\mathbf{B}'\mathbf{V}^{-1}\mathbf{B} + n\xi\mathbf{K}\right)^{-1}\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}(\mathbf{w}_0 - \tilde{\mathbf{w}}) \\
&= \left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n} + \xi\mathbf{K}\right)^{-1}\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}(\mathbf{w}_0 - \tilde{\mathbf{w}}) \\
&= \left\{\left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)^{-1} - \xi\left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)^{-1}\mathbf{K}\left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)^{-1} + O\left(\xi^2\right)\right\} \\
&\quad\times\left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)(\mathbf{w}_0 - \tilde{\mathbf{w}}) \\
&= (\mathbf{w}_0 - \tilde{\mathbf{w}}) - \xi\left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)^{-1}\mathbf{K}(\mathbf{w}_0 - \tilde{\mathbf{w}}) + O\left(\xi^2\right)\left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)(\mathbf{w}_0 - \tilde{\mathbf{w}}).
\end{aligned} \tag{A.9}$$

Finally, we can evaluate A_3 as follows:

$$\begin{aligned}
A_3 &= -\left(\mathbf{B}'\mathbf{V}^{-1}\mathbf{B} + n\xi\mathbf{K}\right)^{-1}\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}^{-1}n\xi\mathbf{K}\mathbf{w}_0 \\
&= \left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n} + \xi\mathbf{K}\right)^{-1}\xi\mathbf{K}\mathbf{w}_0 \\
&= \left\{\left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)^{-1} - \xi\left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)^{-1}\mathbf{K}\left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)^{-1} + O\left(\xi^2\right)\right\}\xi\mathbf{K}\mathbf{w}_0 \\
&= -\xi\left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)^{-1}\mathbf{K}\mathbf{w}_0 + \xi^2\left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)\mathbf{K}\left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)\mathbf{K}\mathbf{w}_0 + O\left(\xi^3\right)\mathbf{K}\mathbf{w}_0.
\end{aligned} \tag{A.10}$$

We can also observe that the weighted least-square estimates $\tilde{\mathbf{w}}$ have a normal distribution. Hence

$$\tilde{\mathbf{w}} - \mathbf{w}_0 = O_p\left(n^{-1/2}\right). \tag{A.11}$$

If $\xi = O(n^{\eta})$ and $\eta < -1/2$, then A_1, A_2, and A_3 become

$$\begin{aligned} A_1 &= \left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)^{-1} \frac{\mathbf{B}'\mathbf{V}^{-1}}{n}\boldsymbol{\varepsilon} + O(\xi) + O\left(\xi^2\right), \\ A_2 &= (\mathbf{w}_0 - \tilde{\mathbf{w}}) + O(\xi) \times O_p(\mathbf{w}_0 - \tilde{\mathbf{w}}) + O\left(\xi^2\right) \times O_p(\mathbf{w}_0 - \tilde{\mathbf{w}}) = O_p\left(n^{-1/2}\right), \end{aligned} \tag{A.12}$$

and $A_3 = O(\xi) + O(\xi^2) + O(\xi^3) = O(\xi)$. Therefore, (A.7) can be written as

$$\widehat{\mathbf{w}} - \mathbf{w}_0 = \left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n}\right)^{-1} \frac{\mathbf{B}'\mathbf{V}^{-1}}{n}\boldsymbol{\varepsilon} + O_p\left(n^{-1/2}\right). \tag{A.13}$$

By the law of large numbers and the central limit theorem,

$$\sqrt{n}(\widehat{\mathbf{w}} - \mathbf{w}_0) \underset{D}{\rightarrow} N\left(0, \boldsymbol{\Sigma}_2^{-1}\right). \tag{A.14}$$

Next we deal with the estimators $\widehat{\boldsymbol{\beta}}$. These minimize the function $\mathcal{L}(\boldsymbol{\beta}, \mathbf{w})$, which yields

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta}, \mathbf{w})}{\partial \boldsymbol{\beta}} = I_1'(\boldsymbol{\beta}) + I_2'(\boldsymbol{\beta}, \mathbf{w}) = -\tilde{\mathbf{X}}'\mathbf{V}^{-1}\boldsymbol{\varepsilon} + \tilde{\mathbf{X}}'\mathbf{V}^{-1}\tilde{\mathbf{X}}\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right) + \tilde{\mathbf{X}}'\mathbf{V}^{-1}\mathbf{B}(\mathbf{w} - \tilde{\mathbf{w}}). \tag{A.15}$$

The minimization of this quadratic function is given by

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \boldsymbol{\beta}_0 + \left(\tilde{\mathbf{X}}'\mathbf{V}^{-1}\tilde{\mathbf{X}}\right)^{-1} \left\{\tilde{\mathbf{X}}'\mathbf{V}^{-1}\boldsymbol{\varepsilon} + \tilde{\mathbf{X}}'\mathbf{V}^{-1}\mathbf{B}(\mathbf{w} - \tilde{\mathbf{w}})\right\} \\ &= \boldsymbol{\beta}_0 + \left(\tilde{\mathbf{X}}'\mathbf{V}^{-1}\tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}'\mathbf{V}^{-1}\{\boldsymbol{\varepsilon} + \mathbf{B}(\mathbf{w} - \tilde{\mathbf{w}})\}. \end{aligned} \tag{A.16}$$

If we substitute $\mathbf{w}$ for its estimator $\widehat{\mathbf{w}}_0$, from (A.14) and (A.11), we have

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \boldsymbol{\beta}_0 + \left(\tilde{\mathbf{X}}'\mathbf{V}^{-1}\tilde{\mathbf{X}}\right)^{-1} \left\{\tilde{\mathbf{X}}'\mathbf{V}^{-1}\boldsymbol{\varepsilon} + \tilde{\mathbf{X}}'\mathbf{V}^{-1}\mathbf{B}(\widehat{\mathbf{w}}_0 - \tilde{\mathbf{w}})\right\} \\ &= \boldsymbol{\beta}_0 + \left(\tilde{\mathbf{X}}'\mathbf{V}^{-1}\tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}'\mathbf{V}^{-1}\boldsymbol{\varepsilon} + O_p\left(n^{-1/2}\right). \end{aligned} \tag{A.17}$$

Similarly, by the law of large numbers and the central limit theorem,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \underset{D}{\rightarrow} N\left(0, \boldsymbol{\Sigma}_1^{-1}\right). \tag{A.18}$$

This completes the proof of the lemma. □

Proof of Theorem 2.1. Let the estimator $\widehat{\mathbf{a}}_{\tau,n} = (\widehat{a}_{1,n}, \ldots \widehat{a}_{\tau,n})'$ be the ordinary least-square estimate applied to model (2.18). For the ease of notation, we set $\widehat{a}_{j,n} = a_{j,n} = 0$ for $j > \tau$ and $\widehat{a}_{0,n} = a_{0,n} = 1$. Then we write

$$\widehat{e}_{i,n} = e_i - S_{i,n} + R_{i,n} + Q_{i,n}, \tag{A.19}$$

where

$$\begin{aligned} S_{i,n} &= \sum_{j=0}^{\infty} a_j \left\{ \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \mathbf{x}_{i-j} + (\widehat{\mathbf{w}} - \mathbf{w})' \phi_{i-j} \right\}, \\ R_{i,n} &= \sum_{j=0}^{\tau} (\widehat{a}_{j,n} - a_{j,n}) \left(y_{i-j} - \widehat{\boldsymbol{\beta}}' \mathbf{x}_{i-j} - \widehat{\mathbf{w}}' \phi_{i-j} \right), \\ Q_{i,n} &= \sum_{j=0}^{\infty} (a_{j,n} - a_j) \left(y_{i-j} - \widehat{\boldsymbol{\beta}}' \mathbf{x}_{i-j} - \widehat{\mathbf{w}}' \phi_{i-j} \right). \end{aligned} \tag{A.20}$$

From assumptions (A.1), (A.2), and Lemma .1 we can see that under the assumptions about τ and by the Caucy-Schwarz inequality

$$\begin{aligned} |S_{i,n}| &\leq \sum_{j=0}^{\infty} |a_j| \left| \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \mathbf{x}_{i-j} + (\widehat{\mathbf{w}} - \mathbf{w})' \phi_{i-j} \right| \\ &\leq \left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right\| \left\| \sum_{j=0}^{\infty} a_j \mathbf{x}_{i-j} \right\| + \|\widehat{\mathbf{w}} - \mathbf{w}\| \left\| \sum_{j=0}^{\infty} a_j \phi_{i-j} \right\| = O_p\left(n^{-1/2}\right). \end{aligned} \tag{A.21}$$

Next we evaluate $R_{i,n}$. In An et al. [31, proof of Theorem 5]: it is shown that, under the assumptions about $\tau(n)$,

$$\sum_{j=0}^{\tau} (\widehat{a}_{j,n} - a_{j,n})^2 = o\left(\left(\frac{\log(n)}{n} \right)^{1/2} \right). \tag{A.22}$$

Thus, by the Cauchy-Schwarz inequality

$$|R_{i,n}| \leq \left(\sum_{j=0}^{\tau} (\widehat{a}_{j,n} - a_{j,n})^2 \right)^{1/2} \left(\sum_{j=0}^{\tau} \left(y_{i-j} - \widehat{\boldsymbol{\beta}}' \mathbf{x}_{i-j} - \widehat{\mathbf{w}}' \phi_{i-j} \right)^2 \right)^{1/2}, \tag{A.23}$$

which yields $R_{i,n} = o((\log(n)/n)^{1/4}) O_p(\tau^{1/2}) = o_p(1)$. Finally, we evaluate $Q_{i,n}$. By the extended Baxter inequality from Bühlmann [32, proof of Theorem 3.1], we have

$$\sum_{j=0}^{\infty} |a_{j,n} - a_j| \leq C \sum_{j=\tau+1}^{\infty} |a_j|. \tag{A.24}$$

Notice that $y_{i-j} - \widehat{\boldsymbol{\beta}}'\mathbf{x}_{i-j} - \widehat{\mathbf{w}}'\boldsymbol{\phi}_{i-j} = e_{i,n}$. Since $e_{i,n}$ is a stationary and invertible process whose linear process coefficients satisfy the given summability assumptions, we have for some $M > 0$,

$$|Q_{i,n}| \leq M\sum_{j=0}^{\infty}\left|a_{j,n} - a_j\right| \leq M\sum_{j=\tau+1}^{\infty}\left|a_j\right| = o_p(1). \tag{A.25}$$

From the above decomposition and evaluation, we can see that

$$\underline{\mathbf{y}} - \underline{\mathbf{X}}\boldsymbol{\beta} - \underline{\mathbf{B}}\mathbf{w} = \mathbf{y}_{\widehat{\tau}} - \mathbf{X}_{\widehat{\tau}}\widehat{\boldsymbol{\beta}} - \mathbf{B}_{\widehat{\tau}}\widehat{\mathbf{w}} + o_p(1). \tag{A.26}$$

Therefore, in order to prove the second equation in Theorem 2.1 we just need to show

$$\begin{aligned}
\frac{1}{n}\left(\widetilde{\mathbf{X}}'\widehat{\mathbf{V}}_\tau^{-1}\widetilde{\mathbf{X}} - \widetilde{\mathbf{X}}'\mathbf{V}_\tau^{-1}\widetilde{\mathbf{X}}\right) &= O_p\left(\left(\frac{\tau}{n}\right)^{1/2}\right),\\
\frac{1}{n}\left(\mathbf{B}'\widehat{\mathbf{V}}_\tau^{-1}\mathbf{B} - \mathbf{B}'\mathbf{V}_\tau^{-1}\mathbf{B}\right) &= O_p\left(\left(\frac{\tau}{n}\right)^{1/2}\right),\\
\frac{1}{\sqrt{n}}\widetilde{\mathbf{X}}'\left(\widehat{\mathbf{V}}_\tau^{-1} - \mathbf{V}_\tau^{-1}\right)\boldsymbol{\varepsilon} &= O_p\left(\left(\frac{\tau}{n}\right)^{1/2}\right),\\
\frac{1}{\sqrt{n}}\mathbf{B}'\left(\widehat{\mathbf{V}}_\tau^{-1} - \mathbf{V}_\tau^{-1}\right)\boldsymbol{\varepsilon} &= O_p\left(\left(\frac{\tau}{n}\right)^{1/2}\right).
\end{aligned} \tag{A.27}$$

To see the above results are true, let $y_{\tau,i}$ be the ith element $\mathbf{y}_\tau$ of model (2.20). We have for $\widehat{\mathbf{T}}_{\tau,i}$ (the ith row of $\widehat{\mathbf{T}}_\tau$), $\widetilde{\mathbf{X}}_{\tau,i}$ (the ith column of $\widetilde{\mathbf{X}}_\tau$), and $\mathbf{B}_{\tau,i}$ (the ith column of $\mathbf{B}_\tau$)

$$\begin{aligned}
\widehat{e}_i &= \widehat{\mathbf{T}}_{\tau,i}\boldsymbol{\varepsilon} = \mathbf{e}_i + \sum_{j=1}^{\tau}\left(\widehat{a}_j - a_j\right)\varepsilon_{i-j},\\
\widetilde{X}_{\widehat{\tau},ij} &= \widehat{\mathbf{T}}_{\tau,j}\cdot\widetilde{\mathbf{X}}_{\tau,\cdot i} = \widetilde{X}_{\tau,ij} + \sum_{j=1}^{\tau}a_j\widetilde{X}_{i-j,i} + \sum_{j=1}^{\tau}\left(\widehat{a}_j - a_j\right)\widetilde{X}_{i-j,i}\\
&\equiv \widetilde{X}_{\tau,ij} + \sum_{j=1}^{\tau}\left(\widehat{a}_j - a_j\right)\widetilde{X}_{i-j,i},\\
\widehat{B}_{\widehat{\tau},ij} &= \widehat{\mathbf{T}}_{\tau,j}\cdot\mathbf{B}_{\tau,\cdot i} = B_{\tau,ij} + \sum_{j=1}^{\tau}a_jB_{i-j,i} + \sum_{j=1}^{\tau}\left(\widehat{a}_j - a_j\right)B_{i-j,i}\\
&\equiv B_{\tau,ij} + \sum_{j=1}^{\tau}\left(\widehat{a}_j - a_j\right)B_{i-j,i}
\end{aligned} \tag{A.28}$$

for $i = \tau+1, \tau+2, \ldots, n$, with similar expressions holding for $i = 1, 2, \ldots, \tau$. By (A.26) and the fact that $\|\widehat{\mathbf{a}}_\tau - \mathbf{a}\| = O_p((\tau/n)^{1/2})$ (see Xiao et al. [22]), it follows that

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{e}_i\widetilde{X}_{\widehat{\tau},ij} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}e_i\widetilde{X}_{\tau,ij} + O_p\left(\left(\frac{\tau}{n}\right)^{1/2}\right),\\
\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_{\widehat{\tau},ij}\widetilde{X}_{\widehat{\tau},ik} &= \frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_{\tau,ij}\widetilde{X}_{\tau,ik} + O_p\left(\left(\frac{\tau}{n}\right)^{1/2}\right),\\
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{e}_i\widehat{B}_{\tau,ij} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}e_iB_{\tau,ij} + O_p\left(\left(\frac{\tau}{n}\right)^{1/2}\right),\\
\frac{1}{n}\sum_{i=1}^{n}\widehat{B}_{\tau,ij}\widehat{B}_{\tau,ik} &= \frac{1}{n}\sum_{i=1}^{n}B_{\tau,ij}B_{\tau,ik} + O_p\left(\left(\frac{\tau}{n}\right)^{1/2}\right).
\end{aligned} \tag{A.29}$$

Therefore, using the expansion $(\mathbf{A} + a\mathbf{B})^{-1} = \mathbf{A}^{-1} - a\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} + O(a^2)$ and from (A.17), (A.13) and (A.27), we have

$$\begin{aligned}
\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{SGLSE}} - \boldsymbol{\beta}_0\right) &= \left(\frac{\widetilde{\mathbf{X}}'\widehat{\mathbf{V}}^{-1}\widetilde{\mathbf{X}}}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\left\{\widetilde{\mathbf{X}}'\widehat{\mathbf{V}}^{-1}\boldsymbol{\varepsilon} + O_p\left(n^{-1/2}\right)\right\}\right)\\
&= \left(\frac{\widetilde{\mathbf{X}}'\mathbf{V}^{-1}\widetilde{\mathbf{X}}}{n} + O_p\left(\left(\frac{\tau}{n}\right)^{1/2}\right)\right)^{-1}\\
&\quad\times\left(\frac{\widetilde{\mathbf{X}}'\mathbf{V}^{-1}}{\sqrt{n}}\boldsymbol{\varepsilon} + O_p\left(\left(\frac{\tau}{n}\right)^{1/2}\right) + O_p\left(n^{-1}\right)\right)\\
&= \sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + O_p\left(\left(\frac{\tau}{n}\right)^{1/2}\right),\\
\sqrt{n}(\widehat{\mathbf{w}}_{\mathrm{SGLSE}} - \mathbf{w}_0) &= \left(\frac{\mathbf{B}'\widehat{\mathbf{V}}^{-1}\mathbf{B}}{n}\right)^{-1}\left\{\frac{1}{\sqrt{n}}\left(\mathbf{B}'\widehat{\mathbf{V}}^{-1}\boldsymbol{\varepsilon} + O_p\left(n^{-1/2}\right)\right)\right\}\\
&= \left(\frac{\mathbf{B}'\mathbf{V}^{-1}\mathbf{B}}{n} + O_p\left(\left(\frac{\tau}{n}\right)^{1/2}\right)\right)^{-1}\left(\frac{\mathbf{B}'\mathbf{V}^{-1}}{\sqrt{n}}\boldsymbol{\varepsilon} + O_p\left(\left(\frac{\tau}{n}\right)^{-1/2}\right) + O_p\left(n^{-1}\right)\right)\\
&= \sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + O_p\left(\left(\frac{\tau}{n}\right)^{1/2}\right).
\end{aligned} \tag{A.30}$$

This completes the proof of Theorem 2.1. □

Proof of Theorem 2.2. We write $\alpha_n = n^{-1/2} + a_n$. It is sufficient to show that, for any given $\zeta > 0$, there exist large constants C such that

$$P\left\{ \inf_{\|\mathbf{u}\|=C} \mathcal{S}(\boldsymbol{\theta}_0 + \alpha_n \mathbf{u}) \geq \mathcal{S}(\boldsymbol{\theta}_0) \right\} \leq 1 - \zeta. \tag{A.31}$$

This implies, with probability at least $1 - \zeta$, that there exists a local minimizer in the balls $\{\boldsymbol{\theta}_0 + \alpha_n \mathbf{u} : \|\mathbf{u}\| \leq C\}$. Define

$$D_n(\mathbf{u}) = \mathcal{S}(\boldsymbol{\theta}_0 + \alpha_n \mathbf{u}) - \mathcal{S}(\boldsymbol{\theta}_0). \tag{A.32}$$

Note that $p_{\lambda_{jn}}(0) = 0$ and that $p_{\lambda_{jn}}(|\theta_j|)$ is nonnegative, so

$$\begin{aligned} D_n(\mathbf{u}) \geq\ & n^{-1}\{l(\boldsymbol{\theta}_0 + \alpha_n \mathbf{u}) - l(\boldsymbol{\theta}_0)\} \\ & + \sum_{j=1}^{d+m} \left\{ p_{\lambda_{jn}}\left(\left|\theta_{j0} + \alpha_n u_j\right|\right) - p_{\lambda_{jn}}\left(\left|\theta_{j0}\right|\right) \right\} \\ & + \xi(\boldsymbol{\theta}_0 + \alpha_n \mathbf{u})' \widetilde{\mathbf{K}} (\boldsymbol{\theta}_0 + \alpha_n \mathbf{u}) - \xi \boldsymbol{\theta}_0' \widetilde{\mathbf{K}} \boldsymbol{\theta}_0, \end{aligned} \tag{A.33}$$

where $l(\boldsymbol{\theta})$ is the first term of (2.7) and $\widetilde{\mathbf{K}}$ is defined in (2.47). Now

$$\begin{aligned} & \frac{1}{2} n^{-1}\{l(\boldsymbol{\theta}_0 + \alpha_n \mathbf{u}) - l(\boldsymbol{\theta})\} \\ & \quad = \frac{\alpha_{1n}^2}{2} \mathbf{u}_1' \left\{ \frac{\mathbf{B}'\mathbf{V}\mathbf{B}}{n} + o_p(1) \right\} \mathbf{u}_1 - \alpha_{1n} \mathbf{u}_1' \left\{ \frac{\mathbf{B}'\mathbf{V}}{n} \boldsymbol{\varepsilon} + o_p\left(n^{-1/2}\right) \right\} \\ & \quad + \frac{\alpha_{2n}^2}{2} \mathbf{u}_2' \left\{ \frac{\widetilde{\mathbf{X}}'\mathbf{V}^{-1}\widetilde{\mathbf{X}}}{n} + o_p(1) \right\} \mathbf{u}_2 - \alpha_{2n} \mathbf{u}_2' \left\{ \frac{\widetilde{\mathbf{X}}'\mathbf{V}^{-1}}{n} \boldsymbol{\varepsilon} + o_p\left(n^{-1/2}\right) \right\}. \end{aligned} \tag{A.34}$$

Note that $\mathbf{B}'\widehat{\mathbf{V}}^{-1}\mathbf{B}/n \to \boldsymbol{\Sigma}_1$, $\mathbf{B}'\widehat{\mathbf{V}}^{-1}\boldsymbol{\varepsilon}/n \to \boldsymbol{\xi}_1$, $\widetilde{\mathbf{X}}'\widehat{\mathbf{V}}^{-1}\widetilde{\mathbf{X}}/n \to \boldsymbol{\Sigma}_2$, and $\widetilde{\mathbf{X}}'\widehat{\mathbf{V}}^{-1}\boldsymbol{\varepsilon}/n \to \boldsymbol{\xi}_2$ are finite positive definite matrices in probability. So the first term in the right side of (A.34) is of order $O_p(C_1^2\alpha_{1n}^2)$, and the second term is of order $O_p(C_1 n^{-1/2}\alpha_{1n}) = O_p(C\alpha_{1n}^2)$. Similarly, the third term of (A.34) is of order $O_p(C_2^2\alpha_{2n}^2)$ and the fourth term is order $O_p(C_2 n^{-1/2}\alpha_{2n}^2)$. Furthermore,

$$\sum_{j=1}^{m} \left\{ p_{\lambda_{j1n}}\left(\left|w_{j0} + \alpha_{1n} u_j\right|\right) - p_{\lambda_{j1n}}\left(\left|w_{j0}\right|\right) \right\}, \tag{A.35}$$

$$\sum_{j=1}^{d} \left\{ p_{\lambda_{j2n}}\left(\left|\beta_{j0} + \alpha_{2n} u_j\right|\right) - p_{\lambda_{j2n}}\left(\left|\beta_{j0}\right|\right) \right\}, \tag{A.36}$$

are bounded by

$$\begin{aligned}
\sqrt{m}\alpha_{1n}a_{1n}\|\mathbf{u}\| + \alpha_{1n}^2 b_{1n}\|\mathbf{u}\|^2 &= C\alpha_n^2\left(\sqrt{m} + b_{1n}C\right),\\
\sqrt{d}\alpha_{2n}a_{2n}\|\mathbf{u}\| + \alpha_n^2 b_{2n}\|\mathbf{u}\|^2 &= C\alpha_{2n}^2\left(\sqrt{d} + b_{2n}C\right)
\end{aligned} \tag{A.37}$$

by the Taylor expansion and the Cauchy-Schwarz inequality. As $b_n \to 0$, the first term on the right side of (A.34) will dominate (A.35) and (A.36) as well as the second term on the right side of (A.34), by taking C sufficiently large. Hence (A.31) holds for sufficiently large C. This completes the proof of the theorem. □

Lemma .2. *Under the conditions of Theorem 2.3, with probability tending 1, for any given $\boldsymbol{\beta}$ and $\mathbf{w}$, satisfying $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = \|\mathbf{w}_1 - \mathbf{w}_{10}\| = O_p(n^{-1/2})$ and any constant C,*

$$\mathcal{S}\left\{\left(\boldsymbol{\beta}_1', \mathbf{0}'\right)', \left(\mathbf{w}_1', \mathbf{0}'\right)\right\} = \min_{\|\beta_2\| \le C_1 n^{-1/2}, \|\mathbf{w}_2\| \le C_2 n^{-1/2},} \mathcal{S}\left\{\left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2'\right)', \left(\mathbf{w}_1', \mathbf{w}_2'\right)'\right\}. \tag{A.38}$$

Proof. We show that with probability tending to 1, as $n \to \infty$, for any $\boldsymbol{\beta}_1$ and $\mathbf{w}_1$ satisfying $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = \|\mathbf{w}_1 - \mathbf{w}_{10}\| = O_p(n^{-1/2})$, $\|\boldsymbol{\beta}_2\| \le C_1 n^{-1/2}$, and $\|\mathbf{w}_2\| \le C_2 n^{-1/2}$, $\partial l(\boldsymbol{\beta}, \mathbf{w})/\partial \beta_j$ and β_j have the same signs for $\beta_j \in (-C_1 n^{-1/2}, C_1 n^{-1/2})$, for $j = S_1 + 1, \ldots, d$. Also $\partial l(\boldsymbol{\beta}, \mathbf{w})/\partial w_j$ and w_j have the same signs for $w_j \in (-C_2 n^{-1/2}, C_2 n^{-1/2})$, for $j = S_2 + 1, \ldots, m$. Thus the minimization is attained at $\boldsymbol{\beta}_2 = \mathbf{w}_2 = 0$.

For $\beta_j \neq 0$ and $j = S_1 + 1, \ldots, d$,

$$\frac{\partial \mathcal{S}(\boldsymbol{\beta})}{\partial \beta_j} = l_j'(\beta) + np_{\lambda_{j2n}}\left(|\beta_j|\right)\operatorname{sgn}(\beta_j), \tag{A.39}$$

where $l_j'(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta})/\partial \beta_j$. By the proof of Theorem 2.1,

$$l_j'(\boldsymbol{\beta}) = -n\left\{\hat{\xi}_{2j} - (\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\boldsymbol{\Sigma}_{1j} + O_p\left(n^{-1/2}\right)\right\}, \tag{A.40}$$

where ξ_{2j} is the jth component of $\boldsymbol{\xi}_{2n}$ and $\widehat{\boldsymbol{\Sigma}}_{1j}$ is the jth column of $\widehat{\boldsymbol{\Sigma}}_1$. From $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$, $n^{-1}l_j'(\boldsymbol{\beta})$ is of order $O_p(n^{-1/2})$. Therefore,

$$\frac{\partial \mathcal{S}(\boldsymbol{\beta})}{\partial \beta_j} = n\lambda_{j2n}\left\{\lambda_{j1n}p_{\lambda_{j1n}}'\left(|\beta_j|\right)\operatorname{sgn}(\beta_j) + O_p\left(n^{-1/2}/\lambda_{1n}\right)\right\}. \tag{A.41}$$

Because $\liminf_{n\to\infty}\liminf_{\beta_j \to 0+}\lambda_{j1n}^{-1}p_{\lambda_{j1n}}'(|\beta_j|) > 0$ and $n^{-1/2}\lambda_{j1n} \to 0$, the sign of the derivative is completely determined by that of β_j.

For $w_j \neq 0$ and $j = S_1 + 1, \ldots, m$,

$$\frac{\partial \mathcal{S}(\mathbf{w})}{\partial w_j} = l_j'(w) + np_{\lambda_{j1n}}\left(|w_j|\right)\operatorname{sgn}(w_j) + 2n\xi\mathbf{K}\mathbf{w}, \tag{A.42}$$

where $l_j'(\mathbf{w}) = \partial l(\mathbf{w})/\partial w_j$. By the proof of Theorem 2.1,

$$l_j'(\mathbf{w}) = -n\left\{\hat{\xi}_{1j} - (\mathbf{w}-\mathbf{w}_0)'\boldsymbol{\Sigma}_{2j} + O_p\left(n^{-1/2}\right)\right\}, \tag{A.43}$$

where ξ_{1j} is the jth component of $\boldsymbol{\xi}_{1n}$ and $\widehat{\boldsymbol{\Sigma}}_{2j}$ is the jth column of $\widehat{\boldsymbol{\Sigma}}_2$. From $\|\mathbf{w}-\mathbf{w}_0\| = O_p(n^{-1/2})$, $n^{-1}l_j'(\mathbf{w})$ is of order $O_p(n^{-1/2})$. Therefore,

$$\frac{\partial \mathcal{S}(\mathbf{w})}{\partial w_j} = n\lambda_{j2n}\left\{\lambda_{j2n}p'_{\lambda_{j2n}}(|w_j|)\operatorname{sgn}(w_j) + O_p\left(n^{-1/2}/\lambda_{2n}\right)\right\}. \tag{A.44}$$

Because $\liminf_{n\to\infty}\liminf_{w_j\to 0+}\lambda_{j2n}^{-1}p'_{\lambda_{j2n}}(|w_j|) > 0$ and $n^{-1/2}\lambda_{j2n} \to 0$, $n^{-1/2}\lambda_{j2n} \to 0$, the sign of the derivative is completely determined by that of w_j. This completes the proof. □

Proof of Theorem 2.3. Part (a) follows directly from follows by Lemma .2. Now we prove part (b). Using an argument similar to the proof of Theorem 2.1, it can be shown that there exist a $\widehat{\mathbf{w}}_1$ and $\widehat{\boldsymbol{\beta}}_1$ in Theorem 2.3 that are a root-n consistent local minimizer of $\mathcal{S}\{(\mathbf{w}_1', \mathbf{0}')'\}$ and $\mathcal{S}\{(\boldsymbol{\beta}_1', \mathbf{0}')'\}$, satisfying the penalized least-square equations:

$$\begin{aligned} \frac{\partial \mathcal{S}\left(\mathbf{w}_1', \mathbf{0}'\right)'}{\partial \mathbf{w}_1} &= \mathbf{0}, \\ \frac{\partial \mathcal{S}\left(\boldsymbol{\beta}_1', \mathbf{0}'\right)'}{\partial \boldsymbol{\beta}_1} &= \mathbf{0}. \end{aligned} \tag{A.45}$$

Following the proof of Theorem 2.1, we have

$$\begin{aligned} \frac{\partial \mathcal{S}\left(\mathbf{w}_1', \mathbf{0}'\right)'}{\partial w_j} &= n\left[\hat{\boldsymbol{\xi}}_{2(1)} + O_p\left(n^{-1/2}\right) + n\xi\mathbf{K}\mathbf{w} + \left\{\widehat{\boldsymbol{\Sigma}}_{2(1)} + O_p(1)\right\}(\widehat{\mathbf{w}}_1 - \mathbf{w}_{10})\right] \\ &\quad + n\left[\mathbf{b}_n + \boldsymbol{\Sigma}_{\lambda_1}(\mathbf{w})\{1+O_p(1)\}(\widehat{\mathbf{w}}_1 - \mathbf{w}_{10})\right], \\ \frac{\partial \mathcal{S}\left(\boldsymbol{\beta}_1', \mathbf{0}'\right)'}{\partial \beta_j} &= n\left[\hat{\boldsymbol{\xi}}_{1(1)} + O_p\left(n^{-1/2}\right) + \left\{\widehat{\boldsymbol{\Sigma}}_{1(1)} + O_p(1)\right\}\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}\right)\right] \\ &\quad + n\left[\mathbf{b}_n + \boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_{\lambda_2}(\boldsymbol{\beta})\{1+O_p(1)\}\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}\right)\right] \end{aligned} \tag{A.46}$$

where $\hat{\xi}_{1(1)}$ and $\hat{\xi}_{2(1)}$ consist of the first S_j, $j = 1,2,\ldots,S_1$ and $j = 1,2,\ldots,S_2$ components of $\hat{\boldsymbol{\xi}}_1$ and $\hat{\boldsymbol{\xi}}_2$ respectively. Also $\widehat{\boldsymbol{\Sigma}}_{1(1)}$ and $\widehat{\boldsymbol{\Sigma}}_{2(1)}$ consist of the first S_j, $j = 1,2,\ldots,S_1$ and $j = 1,2,\ldots,S_2$ rows and columns of $\widehat{\boldsymbol{\Sigma}}_1$ and $\widehat{\boldsymbol{\Sigma}}_2$, respectively.

Therefore, similar to the proof of Theorem 2.1 and by Slutsky's theorem, it follows that

$$\sqrt{n}(\mathbf{I}_{S_1} + \boldsymbol{\Sigma}_{\lambda_1}(\boldsymbol{\beta}))\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\mathbf{I}_{S_1} + \boldsymbol{\Sigma}_{\lambda_1}(\boldsymbol{\beta}))^{-1}\mathbf{b}_{\beta}\right) \longrightarrow N_{S_1}\left(0, \boldsymbol{\Sigma}_{1(1)}^{-1}\right),$$
$$\sqrt{n}(\mathbf{I}_{S_2} + \boldsymbol{\Sigma}_{\lambda_2}(\mathbf{w}))\left(\widehat{\mathbf{w}}_1 - \mathbf{w}_{10} + (\mathbf{I}_{S_2} + \boldsymbol{\Sigma}_{\lambda_2}(\mathbf{w}) + \xi\mathbf{K})^{-1}\mathbf{b}_{w}\right) \longrightarrow N_{S_2}\left(0, \boldsymbol{\Sigma}_{2(1)}^{-1}\right) \tag{A.47}$$

This completes the proof of Theorem 2.3. □

Acknowledgments

The authors are grateful to two anonymous referees whose probing questions have led to a substantial improvement of the paper. This research was supported by the Norinchukin Bank and Nochu Information System endowed chair of Financial Engineering in the Department of Management Science, Tokyo University of Science.

References

[1] N. S. Altman, "Kernel smoothing of data with correlated errors," *Journal of the American Statistical Association*, vol. 85, pp. 749–759, 1990.
[2] J. D. Hart, "Kernel regression estimation with time series errors," *Journal of the Royal Statistical Society: Series B*, vol. 53, pp. 173–188, 1991.
[3] E. Herrmann, T. Gasser, and A. Kneip, "Choice of bandwidth for kernel regression when residuals are correlated," *Biometrika*, vol. 79, pp. 783–795, 1992.
[4] P. Hall and J. D. Hart, "Nonparametric regression with long-range dependence," *Stochastic Processes and Their Applications*, vol. 36, pp. 339–351, 1990.
[5] B. K. Ray and R. S. Tsay, "Bandwidth selection for kernel regression with long-range dependence," *Biometrika*, vol. 84, pp. 791–802, 1997.
[6] J. Beran and Y. Feng, "Local polynomial fitting with long-memory, short-memory and antipersistent errors," *Annals of the Institute of Statistical Mathematics*, vol. 54, no. 2, pp. 291–311, 2002.
[7] J. Fan and I. Gijbels, *Local Polynomial Modeling and Its Applications*, Chapman and Hall, London, UK, 1996.
[8] J. Fan and Q. Yao, *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer, New York, NY, USA, 2005.
[9] J. T. Gao, "Asymptotic theory for partially linear models," *Communications in Statistics: Theory and Methods*, vol. 22, pp. 3327–3354, 1995.
[10] J. You and G. Chen, "Semiparametric generalized least squares estimation in partially linear regression models with correlated errors," *Journal of Statistical Planning and Inference*, vol. 137, no. 1, pp. 117–132, 2007.
[11] G. Aneiros-Pérez and J. M. Vilar-Fernández, "Local polynomial estimation in partial linear regression models under dependence," *Journal of Statistical Planning and Inference*, vol. 52, pp. 2757–2777, 2008.
[12] S. Konishi and G. Kitagawa, *Information Criteria and Statistical Modeling*, Springer, New York, NY, USA, 2008.
[13] L. Breiman, "Heuristics of instability and stabilization in model selection," *The Annals of Statistics*, vol. 24, no. 6, pp. 2350–2383, 1996.
[14] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.
[15] S. Konishi and G. Kitagawa, "Generalised information criteria in model selection," *Biometrika*, vol. 83, no. 4, pp. 875–890, 1996.
[16] T. J. Hastie and R. Tibshirani, *Generalized Additive Models*, Chapman and Hall, London, UK, 1990.
[17] S. Konishi, T. Ando, and S. Imoto, "Bayesian information criteria and smoothing parameter selection in radial basis function networks," *Biometrika*, vol. 91, no. 1, pp. 27–43, 2004.

[18] Y. Yu and D. Ruppert, "Penalized spline estimation for partially linear single-index models," *Journal of the American Statistical Association*, vol. 97, no. 460, pp. 1042–1054, 2002.
[19] P. J. Green and B. W. Silverman, *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall, London, UK, 1994.
[20] J. Green, "Penalized likelihood for generalized semi-parametric regression models," *International Statistical Review*, vol. 55, pp. 245–259, 1987.
[21] P. Speckman, "Kernel smoothing in partial linear models," *Journal of the Royal Statistical Society: Series B*, vol. 50, pp. 413–436, 1988.
[22] Z. Xiao, O. B. Linton, R. J. Carroll, and E. Mammen, "More efficient local polynomial estimation in nonparametric regression with autocorrelated errors," *Journal of the American Statistical Association*, vol. 98, no. 464, pp. 980–992, 2003.
[23] J. Fan and R. Li, "New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis," *Journal of the American Statistical Association*, vol. 99, no. 467, pp. 710–723, 2004.
[24] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society: Series B*, vol. 58, pp. 267–288, 1996.
[25] R. Tibshirani, "The LASSO method for variable selection in the Cox model," *Statistics in Medicine*, vol. 16, no. 4, pp. 385–395, 1997.
[26] A. Fuller, *Introduction to Statistical Time Series*, John Wiley & Sons, New York, NY, USA, 2nd edition, 1996.
[27] K. L. Brewster and R. R. Rindfuss, "Fertility and women's employment in industrialized countries," *Annual Review of Sociology*, vol. 26, pp. 271–286, 2000.
[28] N. Ahn and P. Mira, "A note on the changing relationship between fertility and female employment rates in developed countries," *Journal of Population Economics*, vol. 15, no. 4, pp. 667–682, 2002.
[29] H. Engelhardt, T. Kögel, and A. Prskawetz, "Fertility and women's employment reconsidered: a macro-level time-series analysis for developed countries, 1960–2000," *Population Studies*, vol. 58, no. 1, pp. 109–120, 2004.
[30] G. Aneiros-Pérez, W. González-Manteiga, and P. Vieu, "Estimation and testing in a partial linear regression model under long-memory dependence," *Bernoulli*, vol. 10, no. 1, pp. 49–78, 2004.
[31] H. Z. An, Z. G. Chen, and E. J. Hannan, "Autocorrelation, autoregression and autoregressive approximation," *The Annals of Statistics*, vol. 10, pp. 926–936, 1982.
[32] P. Bühlmann, "Moving-average representation of autoregressive approximations," *Stochastic Processes and Their Applications*, vol. 60, no. 2, pp. 331–342, 1995.