Annals of the Institute of Statistical Mathematics manuscript No.
(will be inserted by the editor)

On estimation of nonparametric regression models with autoregressive and moving average errors

Qi Zheng [1] · **Yunwei Cui** [2] · **Rongning Wu** [3]

Abstract The nonparametric regression model with correlated random errors is a powerful tool for time series forecasting. We are interested in the estimation of such a model under a random design setup, where the random errors are assumed to follow an autoregressive and moving average (ARMA) process, and the covariates can also be correlated. Instead of estimating the constituent parts of the model in a sequential fashion, we propose a spline-based method to estimate the mean function and the parameters of the ARMA process jointly based on a least squares method. We establish the desirable asymptotic properties of the proposed approach under mild regularity conditions. Extensive simulation studies demonstrate that our proposed method performs well and generates strong evidence supporting the established theoretical results. The proposed method provides a new addition to the arsenal of tools for nonparametric models for serially correlated data. We further illustrate the practical usefulness of our method by modeling and forecasting the weekly natural gas scraping data for the state of Iowa.

Keywords nonparametric model with correlated errors · oracally efficient estimation · τ-mixing · splines

1 Introduction

The linear regression model with autoregressive and moving average errors (RegARMA) frequently arises in modeling real-life time series data (see, e.g., Greenhouse et al. 1987; Miaou 1990; Wu and Wang 2012; Ganesh et al. 2021), due to its relatively simple structure and straightforward inference. However, there are severe drawbacks in model goodness of fit, if the postulated linearity in the model does not match the nature of the data. To alleviate the problem, researchers proposed to use nonparametric functions

[1] Department of Bioinformatics & Biostatistics, University of Louisville, Louisville, KY, USA
E-mail: qi.zheng@louisville.edu

[2] Department of Mathematics, Towson University, Towson, MD, USA
E-mail: ycui@towson.edu

[3] Zicklin School of Business, Baruch College, New York, NY, USA
E-mail: Rongning.Wu@baruch.cuny.edu

to replace the parametric linearity part of RegARMA (Hart and Wehrly 1986; Kohn et al. 1992; Opsomer et al. 2001; Hall and Keilegom 2003; Durbán and Currie 2003; Krivobokova and Kauermann 2007; Liang and Jing 2009; Lee et al. 2010; Shao and Yang 2011; Qiu et al. 2013; Shao and Yang 2017, among others) and hence considered

$$Y_t = g_0(X_t) + \epsilon_t, \qquad t = 1, \cdots, n, \tag{1}$$

where $g_0(\cdot)$ is a smooth regression function such that $g_0(x) = E[Y_t|X_t = x]$, Y_t's are outcome variables, X_t's are covariates, and ϵ_t's are independent from X_t's and constitute a correlated random error process. It is worth mentioning that although we only consider a single covariate throughout the paper for the simplicity of presentation, our proposed method can be readily extended to multivariate cases by following a general additive model (Hastie and Tibshirani 1990) or a partially linear single-index model (Carroll et al. 1997).

Kernel methods (see, e.g., Fan 1993) and spline approximations (see, e.g., De Boor and De Boor 1978) are the two commonly used approaches in nonparametric regression. They both can be applied to estimate $g_0(\cdot)$. The existing literature devoted to studying kernel regression under model (1) includes Truong (1991); Roussas and Tran (1992); Roussas et al. (1992); Hart (1994); Tran et al. (1996); Hall and Keilegom (2003); Liang and Jing (2009); Lee et al. (2010), among others. Compared with kernel methods, the spline method is more practically appealing due to its easiness in computation and better performance reported in related literature (see, e.g., Shao and Yang 2011). However, the related development based on spline approximations in studying (1) has been relatively sparse (Truong-Van and Bru 2001; Shao and Yang 2017), which is mainly due to several theoretical obstacles encountered in the course of establishing the asymptotic properties. First, as the time series data is correlated, the techniques for consistency of spline regression estimates in independent data (see, e.g., Huang 2003) cannot be directly used without further adaptation for serially correlated data. Second, due to the fact that the number of spline basis functions is required to increase as n increases, special treatment is needed to address the inherent high-dimensional issues, otherwise many powerful techniques in time series, such as uniform ergodic theorem (see, e.g., Straumann and Mikosch 2006) and martingale central limit theorem (see, e.g., Hall and Heyde 2014), are not applicable. Third, due to the MA component, the objective function usually is not convex (Davis and Dunsmuir 1997), making it challenging to examine the limiting distribution of estimators of ARMA parameters. It is by no means a trivial task to tackle these challenges for model (1).

To mitigate the theoretical challenges, Truong-Van and Bru (2001) investigates the spline methods on model (1) by assuming $\{X_t\}$ to be deterministic (so-called fixed design setup). More specifically, they considered almost equally spaced time points and studied the asymptotic properties of penalized least squares estimation under a set of regularity conditions. When $\{X_t\}$ are random (so-called random design setup), Shao and Yang (2017) established the asymptotic properties for a two-step model fitting procedure based on B-splines and maximum likelihood estimators for ARMA processes. First, $g_0(\cdot)$ is fitted by ordinary nonparametric methods with $\{\epsilon_t\}$ being treated as uncorrelated; second, by assuming the consistency of spline regression estimates in the first step, an ARMA process is estimated based on the residuals. Serra et al. (2018) proposed a method to estimate both mean function and autocovariance function based on assumptions of Gaussian noise process and further developed confidence sets for the mean function that quantify the uncertainty of the estimator. To our knowledge, the

consistency of regression function estimates and the asymptotic distribution of ARMA parameters has not been fully addressed in the literature for a random design setup.

In contrast to the existing two-step procedure developed by Shao and Yang (2017), this paper aims to develop a spline-based method that estimates the mean function $g_0(\cdot)$ and the parameters of the ARMA process simultaneously under model (1) with a random design setup. Under the assumption that the random errors, $\{\epsilon_t\}$, are generated by a classic ARMA process, we investigate the random design that even allows covariates $\{X_t\}$ to be be weakly dependent.To rise to the theoretical challenges, we novelly employ Bernstein's inequality for mixing sequences (Merlevède et al. 2011) to establish proper probability bounds and show that the bounds hold even when the number of basis functions increases. As a consequence, our theoretical development not only works for nonparametric spline estimation but can be adapted to high dimensional linear ARMA models. Finally, we utilize the approach developed by Davis and Dunsmuir (1997) and rigorously demonstrate that our spline estimator of $g_0(\cdot)$ is uniformly consistent and the estimator of the parameters of the ARMA process achieves asymptotic normality.

In fact, our estimator of the ARMA parameters performs oracally efficiently, in that it is asymptotically as efficient as if the true function $g_0(\cdot)$ is known and ARMA models are fitted to the real ARMA errors $\{\epsilon_t\}$ (see, e.g., Shao and Yang 2011; Qiu et al. 2013; Shao and Yang 2017). Thus, our proposed method enjoys a broad applicable scope in time series data, where the covariates are often dependent in practice. Another key feature of this paper is that while Shao and Yang (2017)'s oracle efficiency result is achieved based on the assumption that the estimator of $g_0(\cdot)$ is consistent (see assumption (c) in Shao and Yang (2017)), our work completely alleviates this assumption by directly establishing the consistency property of the proposed estimator. Moreover, we can show that the proposed estimator of $g_0(\cdot)$ achieves the optimal global convergence rate of nonparametric models (see, e.g., Stone 1980).

The remainder of this paper is organized as follows. In section 2, we introduce the model setup and propose an estimation method based on least squares estimation with spline approximation. In section 3, the asymptotic properties of the proposed estimation method are established. In section 4, a comprehensive simulation study is performed to evaluate the developed results. In Section 5, we illustrate the practical usage of our method by analyzing a natural gas data set obtained for the state of Iowa. All proofs are relegated to the Appendix.

2 Model setup and estimation method

In the sequel, ϵ_t in (1) is assumed to follow an ARMA(p, q) process, that is

$$\epsilon_t - \sum_{i=1}^{p} \phi_{i*}\epsilon_{t-i} = \zeta_t + \sum_{j=1}^{q} \theta_{j*}\zeta_{t-j}, \tag{2}$$

where $\phi_{i*}, \theta_{j*} \in \mathbb{R}, i = 1, \cdots, p, j = 1, \cdots, q$. Let B denote the backshift operator, such that $B(\epsilon_t) = \epsilon_{t-1}$. The ARMA$(p, q)$ process (2) satisfies $\boldsymbol{\phi}_*(B)\epsilon_t = \boldsymbol{\theta}_*(B)\zeta_t$, where $\boldsymbol{\phi}_*(z) = 1 - \sum_{i=1}^{p} \phi_{i*}z^i$ is the AR polynomial and $\boldsymbol{\theta}_*(z) = 1 + \sum_{j=1}^{q} \theta_{j*}z_{t-j}$ is the MA polynomial. We further denote $(\phi_{1*}, \cdots, \phi_{p*})^\top$ and $(\theta_{1*}, \cdots, \theta_{q*})^\top$ by $\boldsymbol{\phi}_*$ and $\boldsymbol{\theta}_*$, respectively.

Let $\mathbf{B}_n(u) = (B_1(u), \cdots, B_{J_n}(u))^\top$ be a set of κth order normalized B-spline basis functions with knot sequences $\{\tau_s\}$, where $\{\tau_s\}$ satisfy $\tau_1 = \cdots = \tau_\kappa < \tau_{\kappa+1} < \cdots < \tau_{J_n} < \tau_{J_n+1} = \cdots = \tau_{J_n+\kappa}$. Following the literature of spline estimators (see e.g. Zhou et al. 1998; Huang 2003), we require

$$\frac{\max_{\kappa \le s \le J_n} \tau_{s+1} - \tau_s}{\min_{\kappa \le s \le J_n} \tau_{s+1} - \tau_s} < C,$$

uniformly in n, to investigate the asymptotic properties of our proposed estimators. Throughout the rest of the paper, we use C to represent an unspecified positive constant whose value may vary. In addition, we may suppress the dependence of J_n and $\mathbf{B}_n(\cdot)$ on n for notation simplicity, when there is no confusion. The unknown function $g_0(\cdot)$ then can be approximated by B-spline functions:

$$g_0(u) \approx g_*(u) = \sum_{j=1}^{J} \beta_{j*} B_j(u) = \boldsymbol{\beta}_*^\top \mathbf{B}(u),$$

where $\boldsymbol{\beta}_* = (\beta_{1*}, \cdots, \beta_{J*})^\top$ minimizes $\|g_0(u) - \boldsymbol{\beta}^\top \mathbf{B}(u)\|_\infty$ with respect to $\boldsymbol{\beta}$. According to De Boor and De Boor (1978), $\Delta := \|g_*(u) - g_0(u)\|_\infty \le C_0 J^{-\alpha}$ for some constant C_0, under Condition (C1) in Section 3, where α is defined in Condition (C1). Let $\mathbf{W}_t = \mathbf{B}(X_t)$. Then (1) can be approximated by

$$Y_t \approx \boldsymbol{\beta}_*^\top \mathbf{W}_t + \epsilon_t, \qquad t = 1, \cdots, n. \tag{3}$$

Denote $(\boldsymbol{\beta}^\top, \boldsymbol{\phi}^\top, \boldsymbol{\theta}^\top)^\top$ by $\boldsymbol{\xi}$. Let $\boldsymbol{\xi}_* = (\boldsymbol{\beta}_*^\top, \boldsymbol{\phi}_*^\top, \boldsymbol{\theta}_*^\top)^\top$ denote the true regression, AR, and MA coefficients of the model (3). We define $\epsilon_t(\boldsymbol{\beta}) = 1\{t > 0\}(Y_t - \boldsymbol{\beta}^\top \mathbf{W}_t)$ and $\zeta_t(\boldsymbol{\xi}) = 1\{t > 0\}(\epsilon_t(\boldsymbol{\beta}) - \sum_{i=1}^p \phi_i \epsilon_{t-i}(\boldsymbol{\beta}) - \sum_{j=1}^q \theta_j \zeta_{t-j}(\boldsymbol{\xi}))$, where $1\{\cdot\}$ is an indicator function. We propose to obtain $\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\phi}}^\top, \hat{\boldsymbol{\theta}}^\top)^\top$, the least squares estimator of $\boldsymbol{\xi}$, by minimizing $\mathcal{L}_n(\boldsymbol{\xi}) = \sum_{t=1}^n \zeta_t^2(\boldsymbol{\xi})$. Consequently, $\hat{g}(\cdot)$, the estimator of $g_0(\cdot)$, is $\hat{\boldsymbol{\beta}}^\top \mathbf{B}(\cdot)$.

In the next section, we establish the asymptotic properties of $\hat{g}$ and $(\hat{\boldsymbol{\phi}}^\top, \hat{\boldsymbol{\theta}}^\top)^\top$.

3 Asymptotic Properties

We begin with imposing some necessary notations. Let $\|A\|_q$ denote the L_q norm of A, where A can be a vector, matrix, or function. In particular, $\|A\|$ denotes the L_2 norm of A. We also adopt the empirical process notations as follows: for a generic variable Z and function f, $\mathbb{G}_n(f) = \mathbb{G}_n(f(Z_i)) := n^{-1/2} \sum_{i=1}^n (f(Z_i) - E[f(Z_i)])$ and $\mathbb{E}_n f(Z_i) := n^{-1} \sum_{i=1}^n f(Z_i)$.

The following regularity conditions are needed to facilitate our technical derivations:

(C1) $g_0(\cdot) \in \mathcal{C}_D^{(\alpha)}(\mathcal{X})$, where $\mathcal{C}_D^{(\alpha)}(\mathcal{X})$ is the collection of the continuous functions $g : \mathcal{X} \to \mathbb{R}$ on a bounded set $\mathcal{X} \in \mathbb{R}$ with the αth derivative $\|g^{(\alpha)}\|_\infty \le D$, for some integer $\alpha \ge 2$ and $D > 0$. Without loss of generality, we assume $\mathcal{X} = [0, 1]$.

(C2) The polynomials $\boldsymbol{\phi}_*(z)$ and $\boldsymbol{\theta}_*(z)$ have no common roots, and their roots lie outside the unit circle in the complex plane.

(C3) $\{\zeta_t\}_{t=1}^n$ and $\{X_t\}_{t=1}^n$ are independent. ζ_t's are independent and identically distributed (i.i.d.) with $E[\zeta_t] = 0$ and $E[\zeta_t^2] = \sigma^2$. In addition, ζ_t satisfies the Bernstein's condition, that is, $E[|\zeta_t|^k] \le k! C_B^k / 2$, for some large $C_B > 0$ and $k \ge 1$.

(C4) $\{X_t\}$ is a strictly stationary sequence of absolutely continuous random variables and $\{X_t\}$ is β-mixing with coefficients $\beta(k) \leq 2\exp(-d_1 k^{\gamma_1})$ for any positive k, where $d_1, \gamma_1 > 0$. We refer the definition of β-mixing to Volkonskii and Rozanov (1959). Let

$$\boldsymbol{\Gamma} = E\left[\frac{\boldsymbol{\phi}_*(B)}{\boldsymbol{\theta}_*(B)}\mathbf{W}_t\left(\frac{\boldsymbol{\phi}_*(B)}{\boldsymbol{\theta}_*(B)}\mathbf{W}_t\right)^\top\right].$$

The smallest eigenvalue of $\boldsymbol{\Gamma}$ is bounded below by $\lambda_{\min}J^{-1}$, for some constant $\lambda_{\min} > 0$.

Remark 1 Condition (C1) is commonly assumed in the spline smoothing literature (e.g. Zhou et al. 1998), which gives smoothness conditions of the nonparametric functions. It is worth mentioning that Condition (C1) implicitly requires the order of splines $\kappa \leq \alpha + 2$. Condition (C2) is often adopted in the literature studying ARMA models (see, e.g., Zinde-Walsh and Galbraith 1991; Davis and Dunsmuir 1997; Wu and Wang 2012). It implies that $\{\zeta_t\}$ is the unique causal-invertible stationary solution.

Remark 2 By Condition (C2), we also have $\boldsymbol{\phi}_*(z)/\boldsymbol{\theta}_*(z) = \sum_{i=0}^{\infty}\pi_{*i}z^i$, $\boldsymbol{\phi}_*^{-1}(z) = \sum_{i=0}^{\infty}\rho_{*i}z^i$, and $\boldsymbol{\theta}_*^{-1}(z) = \sum_{i=0}^{\infty}\nu_{*i}z^i$, where $\pi_{*0} = \rho_{*0} = \nu_{*0} = 1$, $|\pi_{i*}|, |\rho_{i*}|, |\nu_{i*}| \leq C_1 r^i, i \geq 1$, for some $C_1 > 0$ and $0 < r < 1$. We denote $C_1\sum_{i=0}^{\infty}r^i$ by C_2.

Remark 3 Condition (C3) and (C4) are technical conditions. The β-mixing condition allows $\{X_t\}$ to be weakly dependent, which relaxes the widely assumed independent condition and would help improve the applicability of our proposed method. The β-mixing condition can be satisfied if $\{X_t\}$ is an ARMA process, under some mild conditions (Mokkadem 1988). The Bernstein condition is imposed to circumvent the difficulty of establishing the consistency of $\hat{g}$ under the increasing dimensionality of the splines approximation and the dependent covariates. The two conditions are often met in practice. In fact, Bernstein's condition implies that ζ_t is sub-exponential, which is weaker than the commonly used sub-Gaussian condition in the studies with increasing dimensionalities (see e.g., Chernozhukov et al. 2013; Van de Geer et al. 2014).

By Stone (1986), the smallest and largest eigenvalues of $E\left[\mathbf{W}_t\mathbf{W}_t^\top\right]$ are bounded below and above by $\lambda_{\min}J^{-1}$ and $\lambda_{\max}J^{-1}$, respectively, where $0 < \lambda_{\min} < \lambda_{\max} < \infty$. By Proposition 4, we can show that the largest eigenvalue of $\boldsymbol{\Gamma}$ is bounded by $C_2^2\lambda_{\max}J^{-1}$. Thus, the eigenvalue assumption in Condition (C4) can be viewed as an adapted version of the ARMA process.

Let $T(\mathbf{h}) = \mathcal{L}_n(\boldsymbol{\xi}_* + \mathbf{h}) - \mathcal{L}_n(\boldsymbol{\xi}_*)$, where $\mathbf{h} = (\mathbf{h}_1^\top, \mathbf{h}_2^\top, \mathbf{h}_3^\top)^\top \in \mathbb{R}^{J+p+q}$, and $\mathbf{h}_1, \mathbf{h}_2$, and $\mathbf{h}_3$ are vectors of size J, p, and q, respectively. Further, let $\widehat{\mathbf{h}}$ denote a local minimizer of $T(\mathbf{h})$. Then minimizing $\mathcal{L}_n(\boldsymbol{\xi})$ with respect to $\boldsymbol{\xi}$ is equivalent to minimizing $T(\mathbf{h})$ with respect to $\mathbf{h}$ and $\hat{\boldsymbol{\xi}} = \boldsymbol{\xi}_* + \widehat{\mathbf{h}}$.

Theorem 1 *Suppose Conditions (C1)–(C4) hold. If $J \sim n^{1/(2\alpha+1)}$, there exists a local minimizer of $T(\mathbf{h})$, $\widehat{\mathbf{h}}$, such that $\widehat{\mathbf{h}} \to_p 0$.*

Since $\hat{\boldsymbol{\xi}} = \boldsymbol{\xi}_* + \widehat{\mathbf{h}}$, Theorem 1 indicates that there exists a a local minimizer $\hat{\boldsymbol{\xi}}$ of $\mathcal{L}_n(\boldsymbol{\xi})$, such that $\hat{\boldsymbol{\xi}}$ is consistent to $\boldsymbol{\xi}_*$. This immediately implies the following corollary.

Corollary 1 *Under the same conditions as in Proposition 2,*

$$E\left[\left(\hat{g}(X_t) - g_0(X_t)\right)^2\right] = O_p\left(n^{-2\alpha/(2\alpha+1)}\right),$$

where α is defined in Condition (1).

Corollary 1 shows that even if dependent covariates and non-Gaussian ARMA errors are present, the proposed estimator $\hat{g}$ still achieves optimal global convergence rate of nonparametric models (Stone 1980). In particular, if $\alpha = 2$, the convergence rate of $\hat{g}$ is $n^{-2/5}$.

The next theorem summarizes the asymptotic distribution of $(\widehat{\boldsymbol{\phi}}^\top, \widehat{\boldsymbol{\theta}}^\top)^\top$ as desired. We define $\mathbf{Q}_t = (\mathbf{Q}_{t1}^\top, \mathbf{Q}_{t2}^\top, \mathbf{Q}_{t3}^\top)^\top$, such that $\mathbf{Q}_{t1} = \phi_*(B)\boldsymbol{\theta}_*^{-1}(B)\mathbf{W}_t, \mathbf{Q}_{t2} = \boldsymbol{\phi}_*^{-1}(B)(\zeta_{t-1}, \cdots, \zeta_{t-p})^\top$, and $\mathbf{Q}_{t3} = \boldsymbol{\theta}_*^{-1}(B)(\zeta_{t-1}, \cdots, \zeta_{t-q})^\top$.

Theorem 2 *Under the same conditions as in Theorem 1,*

$$\sqrt{n}\begin{pmatrix} \widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_* \\ \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_* \end{pmatrix} \to_d N(0, \sigma^2 \boldsymbol{\Sigma}^{-1}), \qquad \textit{where} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} E\begin{bmatrix}\mathbf{Q}_{t2}\mathbf{Q}_{t2}^\top\end{bmatrix} & E\begin{bmatrix}\mathbf{Q}_{t2}\mathbf{Q}_{t3}^\top\end{bmatrix} \\ E\begin{bmatrix}\mathbf{Q}_{t3}\mathbf{Q}_{t2}^\top\end{bmatrix} & E\begin{bmatrix}\mathbf{Q}_{t3}\mathbf{Q}_{t3}^\top\end{bmatrix} \end{pmatrix}.$$

Comparing Theorem 2 to the classic results for estimating ARMA processes (cf. chapter 8, Brockwell and Davis 1991), one can see that the proposed method for model (1) produces "oracally" efficient estimators of the ARMA parameters $(\boldsymbol{\phi}_*^\top, \boldsymbol{\theta}_*^\top)^\top$, in the sense that $(\widehat{\boldsymbol{\phi}}^\top, \widehat{\boldsymbol{\theta}}^\top)^\top$ are asymptotically indistinguishable from the estimators when $g_0(\cdot)$ is known a priori (Shao and Yang 2017).

Theorems 1 and 2 only require the asymptotic order of the number of knots. In our numerical studies, we chose equally spaced quantiles in $\mathcal{X}$ as knots to avoid the complication of determining the knot locations. Regarding the number of knots and the ARMA orders p and q, there are several approaches to determine. For instance, we can employ the Bayesian Information Criterion (see, e.g., Brockwell and Davis 1991), or we can develop a penalized $\mathcal{L}_n(\boldsymbol{\xi})$ by incorporating the LASSO type penalties (Tibshirani 1996). However, the investigation of the validity of those methods is beyond the scope of this work and will be examined in our future research.

4 Simulation study

We conduct a simulation study to attest to the validity of the asymptotic results. An assortment of model setups and parameter configurations satisfying the regularity conditions of Theorems 1 and 2 are adopted. In particular, three different smooth functions are defined on $\mathcal{X} = [0, 1]$ with

$$\begin{cases} f_1(X_t) = 1 - 6X_t + 36X_t^2 - 53X_t^3 + 22X_t^5; \\ f_2(X_t) = \sin(2\pi X_t) + 2X_t^2; \\ f_3(X_t) = \arctan(5X_t - 5/2) - X_t^2/3. \end{cases}$$

The covariate variable X_t's are generated from AR(1) models with normal innovations first and then rescaled within $\mathcal{X} = [0, 1]$. For the ARMA processes (2), various innovations are investigated, including normal distributions. The proposed estimation method is tested for three different sample sizes, $n = 500$, $n = 1000$, and $n = 2000$. We assume that the order of the ARMA model for ϵ_t are known a priori as in the previous works (e.g., Qiu et al. 2013); yet for practical purpose, we propose using Theorem 2 to assist in selecting a proper order of ARMA process for which the details are given in the next section. The inner knots for $\mathbf{B}(X_t)$ are constructed as eight equally spaced quantiles in $\mathcal{X}$ for $n = 500$, and nine equally spaced quantiles in $\mathcal{X}$ for $n = 1000$ and $n = 2000$. For each model setting, we repeat simulations 1000 times and report the

sample mean and sample standard deviation of the parameter estimates. In light of Theorem 2, the theoretical approximation for the standard deviation of each model setting is computed and presented for comparison purposes.

Overall, the simulation study indicates a strong congruence between the observed estimates and the asymptotic properties for the ARMA parameters. The theory established in this paper is versatile, in the sense that the asymptotic properties are developed under the condition that ζ_t is not merely just normally distributed. In addition, empirically, we found that models with ζ_t that does not satisfy Condition (C3) still comply satisfactorily with the theoretical asymptotic results. We specifically present simulation summaries for models with ζ_t being $t-$distributed in Table 1 through Table 6. Among the reported results, the proposed method was tested for various serial correlation structures for ϵ_t, as introduced by ARMA$(1,1)$, AR(2), and MA(2) models. One can see that, in Table 1 through Table 3, the mean values of the estimates of the ARMA parameters are close to the corresponding true parameter values. It is evident that the related empirical sample standard deviations are in high agreement with their theoretical approximations, which are given in the parentheses. Moreover, comparing results across different values for n, both the decreasing differences between the means and their corresponding true parameter values, and the decreasing standard deviations of the estimates imply that the proposed estimation method becomes more and more accurate and precise as the sample path length increases.

To further gauge the performance of the proposed method, we also adopt two metrics to calculate the discrepancy between $g_0(\cdot)$ and $\hat{g}(\cdot)$

$$\rho(g_0,\hat{g}) = \int_0^1 (g_0(x)-\hat{g}(x))^2 dx, \quad \rho_{19}(g_0,\hat{g}) = \int_{0.1}^{0.9} (g_0(x)-\hat{g}(x))^2 dx.$$

The metric $\rho(g_0,\hat{g})$ is intended to measure the overall performance over $[0,1]$. Since it is well-known that nonparametric smooth function estimation may encounter difficulties and generate larger deviates from the true function in the vicinity of boundary points compared to regions closer to the center, $\rho_{19}(g_0,\hat{g})$ is used to measure the difference over the interval $[0.1,0.9]$ to obtain a more comprehensive evaluation of $\hat{g}(\cdot)$.

In the simulation study, corresponding to each case presented in Table 1 through Table 3, we also calculated ρ and ρ_{19} which were presented in Table 4 through Table 6, wherein the proposed method is labeled as 'one-step' and the sequential method is labeled as 'sequential'. The obtained results demonstrate the consistency property of $\hat{g}(\cdot)$, in that as sample path length n increases, ρ and ρ_{19} both decrease, indicating a diminishing difference between g_0 and $\hat{g}$. Additionally, it is evident that the proposed method dominates the sequential method. Among 108 pairwise comparisons between 'one step' method and 'sequential' method, the proposed method outperformed the sequential method for 105 times. More specifically, the proposed method yielded smaller values than those of the sequential method for all 54 cases in terms of ρ, which indicates better boundary performance; the proposed method yielded smaller values than those of the sequential method in terms of ρ and ρ_{19} for all 72 cases, when sample path size is relatively small, i.e. $n = 500$ and $n = 1000$, indicating better performance for small sample size. The only 3 cases where the sequential method performed better are located in Table 5. All 3 cases are related to AR(2) models with large sample size, $n = 2000$. All in all, the simulation study unequivocally validates the proposed method as a valuable supplementary tool to the existing sequential method.

Of note, we remark that the residuals for the sequential method which are fed into the ARMA estimation process have a mean equal to zero, since they are obtained by

a regression procedure. As such, they naturally meet the specification of model (1), $E(\epsilon_t) = 0$. However, in order to enforce the condition for the proposed method, an additional sum-to-zero constraint should be imposed for $Y_t - \hat{g}(X_t)$. We simply center all basis functions of X_t and the response values Y_t by their average values, i.e. the column mean is subtracted from each column of the basis functions and the mean of Y_t is subtracted from each value of Y_t, such that all linear combinations of the basis functions and Y_t have mean zero. The upshot is that we solve the model identification problem for numerical implementation of model (1) by a sum-to-zero condition. The initial values of the ARMA process for the optimization process are rendered by random numbers over $[0, 1]$, and all parameters for the basis functions are set to be one. An implementation of the proposed method by R turns out to consistently converge to estimates that are close to the true ARMA parameters in the simulation study. For further details of the numerical implementation of the proposed method, interested readers are invited to visit `https://github.com/cui-yun-wei/trend-model` to play with the code. Based on extensive simulation studies, it appears that the numerical implementation runs efficiently and stably, no matter whether the least squares loss function is convex (for AR random errors) or non-convex (for MA or ARMA random errors). Additionally, we highly suggest that Y_t and X_t of real-world data set should be re-scaled or standardized to achieve comparable magnitude before being fed into the computer routine for model fitting.

Table 1 Estimation of the parameters of ARMA$(1, 1)$ process, when X_t's satisfy the conditions of Theorem 3.1, and innovations ζ_t's have a t distribution with degrees of freedom ν.

	$(\phi, \theta) = (0.6, 0.3)$ $\nu = 3$ $g_0(X_t) = f_1(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\hat{\phi}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\theta}$
mean	0.5907	0.3160	0.5991	0.3058	0.5989	0.3024
s.d.	0.0489	0.0583	0.0325	0.0386	0.0232	0.0272
	(0.0469)	(0.0559)	(0.0332)	(0.0396)	(0.0235)	(0.0280)
	$(\phi, \theta) = (0.6, 0.3)$ $\nu = 3$ $g_0(X_t) = f_2(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\hat{\phi}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\theta}$
mean	0.5934	0.3182	0.5957	0.3081	0.5987	0.3026
s.d.	0.0505	0.0621	0.0327	0.0393	0.0238	0.0285
	(0.0469)	(0.0559)	(0.0332)	(0.0396)	(0.0235)	(0.0280)
	$(\phi, \theta) = (0.6, 0.3)$ $\nu = 3$ $g_0(X_t) = f_3(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\hat{\phi}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\theta}$
mean	0.5911	0.3151	0.5969	0.3061	0.5983	0.3045
s.d.	0.0478	0.0585	0.0327	0.0407	0.0240	0.0290
	(0.0469)	(0.0559)	(0.0332)	(0.0396)	(0.0235)	(0.0280)
	$(\phi, \theta) = (0.2, -0.5)$ $\nu = 3$ $g_0(X_t) = f_1(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\hat{\phi}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\theta}$
mean	0.2064	-0.5230	0.2048	-0.5123	0.2001	-0.5041
s.d.	0.1428	0.1273	0.0922	0.0814	0.0675	0.0606
	(0.1315)	(0.1162)	(0.0930)	(0.0822)	(0.0657)	(0.0581)
	$(\phi, \theta) = (0.2, -0.5)$ $\nu = 3$ $g_0(X_t) = f_2(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\hat{\phi}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\theta}$
mean	0.2185	-0.5320	0.2025	-0.5113	0.2026	-0.5057
s.d.	0.1388	0.1256	0.0956	0.0859	0.0660	0.0581
	(0.1315)	(0.1162)	(0.0930)	(0.0822)	(0.0657)	(0.0581)
	$(\phi, \theta) = (0.2, -0.5)$ $\nu = 3$ $g_0(X_t) = f_3(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\hat{\phi}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\theta}$
mean	0.2022	-0.5165	0.2035	-0.5098	0.2062	-0.5099
s.d.	0.1444	0.1300	0.1009	0.0901	0.0646	0.0562
	(0.1315)	(0.1162)	(0.0930)	(0.0822)	(0.0657)	(0.0581)

Table 2 Estimation of the parameters of AR(2) process, when X_t's satisfy the conditions of Theorem 3.1, and innovations ζ_t's have a t distribution with degrees of freedom ν.

	$(\phi_1,\phi_2)=(0.4,0.2)$ $\nu=3$ $g_0(X_t)=f_1(X_t)$					
	$n=500$		$n=1000$		$n=2000$	
	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$
mean	0.4002	0.1948	0.3996	0.1965	0.3992	0.1986
s.d.	0.0451	0.4610	0.03144	0.0337	0.0234	0.0237
	(0.0438)	(0.0438)	(0.0310)	(0.0310)	(0.0219)	(0.0219)
	$(\phi_1,\phi_2)=(0.4,0.2)$ $\nu=3$ $g_0(X_t)=f_2(X_t)$					
	$n=500$		$n=1000$		$n=2000$	
	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$
mean	0.4007	0.1935	0.3980	0.1969	0.3990	0.1988
s.d.	0.0451	0.0461	0.0309	0.0307	0.0219	0.0222
	(0.0438)	(0.0438)	(0.0310)	(0.0310)	(0.0219)	(0.0219)
	$(\phi_1,\phi_2)=(0.4,0.2)$ $\nu=3$ $g_0(X_t)=f_3(X_t)$					
	$n=500$		$n=1000$		$n=2000$	
	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$
mean	0.4007	0.1936	0.3993	0.1995	0.3963	0.1991
s.d.	0.0448	0.0450	0.0311	0.0312	0.0226	0.0225
	(0.0438)	(0.0438)	(0.0310)	(0.0310)	(0.0219)	(0.0219)
	$(\phi_1,\phi_2)=(0.5,0.1)$ $\nu=3$ $g_0(X_t)=f_1(X_t)$					
	$n=500$		$n=1000$		$n=2000$	
	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$
mean	0.4827	0.0976	0.4995	0.0965	0.4990	0.0989
s.d.	0.0458	0.0467	0.0321	0.0316	0.0227	0.0229
	(0.0445)	(0.0445)	(0.0314)	(0.0314)	(0.0222)	(0.0222)
	$(\phi_1,\phi_2)=(0.5,0.1)$ $\nu=3$ $g_0(X_t)=f_2(X_t)$					
	$n=500$		$n=1000$		$n=2000$	
	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$
mean	0.5002	0.0958	0.4992	0.0977	0.4991	0.0979
s.d.	0.0451	0.0473	0.0331	0.0324	0.0232	0.0237
	(0.0445)	(0.0445)	(0.0314)	(0.0314)	(0.0222)	(0.0222)
	$(\phi_1,\phi_2)=(0.5,0.1)$ $\nu=3$ $g_0(X_t)=f_3(X_t)$					
	$n=500$		$n=1000$		$n=2000$	
	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$
mean	0.4987	0.0924	0.5009	0.0961	0.4995	0.0987
s.d.	0.0453	0.0461	0.0324	0.0318	0.0234	0.0230
	(0.0445)	(0.0445)	(0.0314)	(0.0314)	(0.0222)	(0.0222)

Table 3 Estimation of the parameters of MA(2) process, when X_t's satisfy the conditions of Theorem 3.1, and innovations ζ_t's have a t distribution with degrees of freedom ν.

	$(\theta_1,\theta_2)=(0.4,0.2)$ $\nu=3$ $g_0(X_t)=f_1(X_t)$					
	$n=500$		$n=1000$		$n=2000$	
	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$
mean	0.4046	0.2025	0.4030	0.2031	0.4014	0.2014
s.d.	0.0439	0.0457	0.0322	0.0317	0.0227	0.0208
	(0.0438)	(0.0438)	(0.0310)	(0.0310)	(0.0219)	(0.0219)
	$(\theta_1,\theta_2)=(0.4,0.2)$ $\nu=3$ $g_0(X_t)=f_2(X_t)$					
	$n=500$		$n=1000$		$n=2000$	
	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$
mean	0.4028	0.2026	0.4021	0.2002	0.4010	0.2007
s.d.	0.0452	0.0453	0.0312	0.0315	0.0216	0.0222
	(0.0438)	(0.0438)	(0.0310)	(0.0310)	(0.0219)	(0.0219)
	$(\theta_1,\theta_2)=(0.4,0.2)$ $\nu=3$ $g_0(X_t)=f_3(X_t)$					
	$n=500$		$n=1000$		$n=2000$	
	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$
mean	0.4046	0.2025	0.4029	0.2010	0.4020	0.2009
s.d.	0.0455	0.0465	0.0312	0.0316	0.0222	0.0218
	(0.0438)	(0.0438)	(0.0310)	(0.0310)	(0.0219)	(0.0219)
	$(\theta_1,\theta_2)=(-0.2,-0.4)$ $\nu=3$ $g_0(X_t)=f_1(X_t)$					
	$n=500$		$n=1000$		$n=2000$	
	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$
mean	-0.2140	-0.4155	-0.2072	-0.4071	-0.2036	-0.4034
s.d.	0.0470	0.0454	0.0305	0.0300	0.0201	0.0207
	(0.0410)	(0.0410)	(0.0290)	(0.0290)	(0.0205)	(0.0205)
	$(\theta_1,\theta_2)=(-0.2,-0.4)$ $\nu=3$ $g_0(X_t)=f_2(X_t)$					
	$n=500$		$n=1000$		$n=2000$	
	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$
mean	-0.2163	-0.4160	-0.2073	-0.4092	-0.2042	-0.4030
s.d.	0.0470	0.0436	0.0307	0.0296	0.0210	0.0207
	(0.0410)	(0.0410)	(0.0290)	(0.0290)	(0.0205)	(0.0205)
	$(\theta_1,\theta_2)=(-0.2,-0.4)$ $\nu=3$ $g_0(X_t)=f_3(X_t)$					
	$n=500$		$n=1000$		$n=2000$	
	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$
mean	-0.2150	-0.4158	-0.2083	-0.4070	-0.2034	-0.4032
s.d.	0.0450	0.0457	0.0300	0.0295	0.0207	0.0206
	(0.0410)	(0.0410)	(0.0290)	(0.0290)	(0.0205)	(0.0205)

Table 4 Comparing $\hat{g}(\cdot)$ and $g_0(\cdot)$, when ϵ_t's follow an ARMA(1, 1) process, X_t's satisfy the conditions of Theorem 3.1, and innovations ζ_t's have a t distribution with degrees of freedom ν.

	$(\phi, \theta) = (0.6, 0.3)$ $\nu = 3$ $g_0(X_t) = f_1(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.3193	0.1550	0.2056	0.0798	0.1608	0.0447
sequential	0.7797	0.3345	0.5655	0.1867	0.4633	0.1193
	$(\phi, \theta) = (0.6, 0.3)$ $\nu = 3$ $g_0(X_t) = f_2(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.3018	0.1387	0.2584	0.0959	0.1326	0.0462
sequential	0.7461	0.3255	0.7346	0.2885	0.4343	0.1303
	$(\phi, \theta) = (0.6, 0.3)$ $\nu = 3$ $g_0(X_t) = f_3(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.3003	0.1370	0.2106	0.0770	0.1628	0.0482
sequential	0.8628	0.3414	0.5499	0.1940	0.4890	0.1214
	$(\phi, \theta) = (0.2, -0.5)$ $\nu = 3$ $g_0(X_t) = f_1(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.2079	0.0719	0.1440	0.0401	0.1133	0.0232
sequential	0.2365	0.0820	0.1710	0.0489	0.1342	0.0285
	$(\phi, \theta) = (0.2, -0.5)$ $\nu = 3$ $g_0(X_t) = f_2(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.2033	0.0727	0.1804	0.0527	0.1056	0.0237
sequential	0.2329	0.0838	0.2067	0.0545	0.1261	0.0283
	$(\phi, \theta) = (0.2, -0.5)$ $\nu = 3$ $g_0(X_t) = f_3(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.2022	0.0679	0.1478	0.0422	0.1403	0.0270
sequential	0.2444	0.0843	0.1715	0.0513	0.1667	0.0314

Table 5 Comparing $\hat{g}(\cdot)$ and $g_0(\cdot)$, when ϵ_t's follow an AR(2) process, X_t's satisfy the conditions of Theorem 3.1, and innovations ζ_t's have a t distribution with degrees of freedom ν.

	$(\phi_1, \phi_2) = (0.4, 0.2)$ $\nu = 3$ $g_0(X_t) = f_1(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.2215	0.1044	0.2028	0.0878	0.1983	0.0751
sequential	0.4083	0.1532	0.3086	0.0996	0.2429	0.0538
	$(\phi_1, \phi_2) = (0.4, 0.2)$ $\nu = 3$ $g_0(X_t) = f_2(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.1848	0.0919	0.1478	0.0534	0.1395	0.0381
sequential	0.4028	0.1590	0.3173	0.0899	0.2500	0.0556
	$(\phi_1, \phi_2) = (0.4, 0.2)$ $\nu = 3$ $g_0(X_t) = f_3(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.1873	0.0965	0.1534	0.0706	0.1575	0.0593
sequential	0.3782	0.1455	0.3234	0.0913	0.2811	0.0594
	$(\phi_1, \phi_2) = (0.5, 0.1)$ $\nu = 3$ $g_0(X_t) = f_1(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.2279	0.1096	0.2248	0.0936	0.2057	0.0617
sequential	0.4556	0.1734	0.3723	0.1240	0.2798	0.0436
	$(\phi_1, \phi_2) = (0.5, 0.1)$ $\nu = 3$ $g_0(X_t) = f_2(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.1877	0.0892	0.1514	0.0539	0.1417	0.0413
sequential	0.4290	0.1688	0.3150	0.1027	0.2789	0.0623
	$(\phi_1, \phi_2) = (0.5, 0.1)$ $\nu = 3$ $g_0(X_t) = f_3(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.2016	0.1035	0.1710	0.0792	0.1625	0.0574
sequential	0.4633	0.1848	0.3252	0.1011	0.3475	0.0396

Table 6 Comparing $\hat{g}(\cdot)$ and $g_0(\cdot)$, when ϵ_t's follow an MA(2) process, X_t's satisfy the conditions of Theorem 3.1, and innovations ζ_t's have a t distribution with degrees of freedom ν.

	$(\theta_1, \theta_2) = (0.4, 0.2)$ $\nu = 3$ $g_0(X_t) = f_1(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.2328	0.0884	0.1837	0.0515	0.1612	0.0297
sequential	0.3109	0.1141	0.2489	0.0687	0.2234	0.0400
	$(\theta_1, \theta_2) = (0.4, 0.2)$ $\nu = 3$ $g_0(X_t) = f_2(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.2471	0.0916	0.1918	0.0598	0.1371	0.0317
sequential	0.3262	0.1171	0.2548	0.0763	0.1990	0.0427
	$(\theta_1, \theta_2) = (0.4, 0.2)$ $\nu = 3$ $g_0(X_t) = f_3(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.2520	0.0931	0.2124	0.0549	0.1465	0.0328
sequential	0.3296	0.1199	0.2788	0.0717	0.1966	0.0435
	$(\theta_1, \theta_2) = (-0.2, -0.4)$ $\nu = 3$ $g_0(X_t) = f_1(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.1877	0.0536	0.1185	0.0316	0.1032	0.0182
sequential	0.2760	0.0825	0.1963	0.0522	0.1640	0.0302
	$(\theta_1, \theta_2) = (-0.2, -0.4)$ $\nu = 3$ $g_0(X_t) = f_2(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.1605	0.0536	0.1389	0.0325	0.1017	0.0183
sequential	0.2459	0.0850	0.2189	0.0325	0.1670	0.0302
	$(\theta_1, \theta_2) = (-0.2, -0.4)$ $\nu = 3$ $g_0(X_t) = f_3(X_t)$					
	$n = 500$		$n = 1000$		$n = 2000$	
	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$	$\rho(g_0, \hat{g})$	$\rho_{19}(g_0, \hat{g})$
one step	0.1801	0.0530	0.1349	0.0339	0.1056	0.0178
sequential	0.2791	0.0827	0.2187	0.0564	0.1796	0.0295

5 A real data example

Compared with other fossil fuel options, natural gas is a cleaner and more efficient energy source. The use of natural gas in the U.S. has steadily increased in the last decade. According to the U.S. Energy Information Administration (EIA), natural gas accounts for 35.4% of total U.S. primary energy production in the year 2021. The rise in production contributed to a decline in natural gas prices, which in turn has induced growth in natural gas utilization. About half of the households in the U.S. use natural gas for space heating and water heating. Being used to monitor natural gas transportation and gauge natural gas consumption, the pipeline scrape data comprise the records of the total amount of natural gas entering a state through the interstate pipelines and are obtained from the interstate pipeline Electronic Bulletin Boards as mandated by Order NO. 636 for the capacity release program of the U.S. Federal Energy Regulatory Commission.

Natural gas consumption is highly sensitive to weather impacts, especially for the regions which use natural gas as their primary heating source. Since natural consumption and natural gas scrape data are highly correlated, it is of great interest to study the natural gas scrape data. To model and forecast scrape data, one of the main covariate variables is the Heating Degree Days (HDD) which measures how long and how much the exterior temperature is below a predetermined reference temperature (called the base temperature, usually 65 degrees Fahrenheit). Denote the temperature in a day at time t by $T(t)$, where t represents the elapsed time in hours past midnight. The daily HDD can be computed by the following two formulas, as either the total for a day in

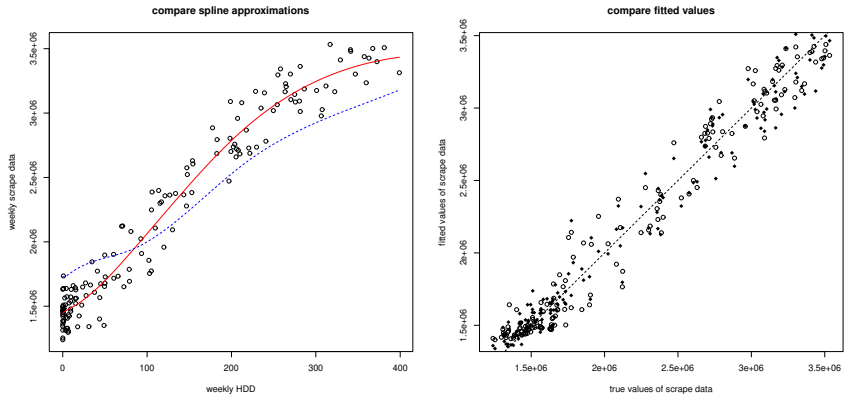
degree × *hours* or the 24-hour average in *degrees*

$$\int_0^{24} (65 - T(t))^+ \, dt, \quad \text{or} \quad \frac{1}{24} \int_0^{24} (65 - T(t))^+ \, dt,$$

where $(65 - T(t))^+ = \max\{65 - T(t), 0\}$. Aggregating the daily HDD of a week gives weekly HDD.

The weekly scrape data from year 2013 through year 2016 were obtained for the state of Iowa. Meanwhile, the corresponding weekly HDD data in *degree* × *hours* were computed based on the hourly temperature profile of the Des Moines International Airport weather station, which were downloaded from the Integrated Surface Database of National Centers for Environmental Information. We are concerned with modeling and forecasting the weekly scrape (data), Y_t, based on the weekly HDD (data), X_t, of Iowa. The whole data are split into two parts. The first 157 weeks, covering the years 2013 through 2015 entirely, constitute the training data to fit proper regression models of Y_t on X_t. The remaining data in 2016, totally 52 weeks, is set aside as testing data for model evaluation.

Fig. 1 Left: the solid line represents the fitted spline approximation function $\widetilde{g}(X_t)$, and the dashed line represents the fitted spline approximation function $\hat{g}(X_t)$; Right: the circles represent fitted values by the sequential method, and the solid diamonds represent the fitted values by the proposed method; the dashed straight line is the 45° degree line going through the origin.



A scatter plot (Figure 1) for the two variables demonstrates a curvilinear relationship. Based on the X_t values in the training set, spline basis functions, $\mathbf{B}(X_t)$, are constructed. The effects of X_t on Y_t are assumed to be appropriately represented by a smooth function, denoted by $g_0(X_t)$. We first obtain an approximation of the smooth function $g_0(X_t)$ by the ordinary multiple linear regression of Y_t on $\mathbf{B}(X_t)$ and denote the approximation by $\widetilde{g}(X_t)$ with $\widetilde{g}(X_t) = \widetilde{\boldsymbol{\beta}}^\top \mathbf{B}(X_t)$, where $\widetilde{\boldsymbol{\beta}}$ represents the estimates for the regression coefficients. The ensuing model diagnostics show that there exists significant serial correlation among the residuals $\tilde{\epsilon}_t = Y_t - \widetilde{g}(X_t)$, and spline approximation alone cannot adequately address the dynamics of the data generating

mechanism of Y_t. Therefore, (1) appears to be a more appropriate model due to its ability to accommodate the serial correlation in the data.

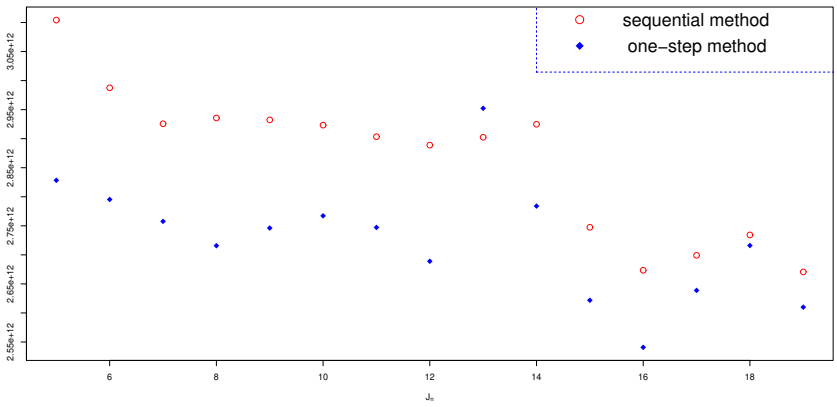
We try to fit model (1) with an ARMA type random error process as defined by (2). We consider the two-step procedure which fits the model in a sequential way (Box et al. 2016; Shao and Yang 2017) and the proposed method (one-step method). For the sequential procedure, first we fit $\widetilde{g}(X_t)$ by ordinary multiple linear regression, and then estimate the ARMA process based on residuals $\tilde{\epsilon}_t$. In contrast, the proposed method estimates the spline approximation and ARMA process jointly through a single optimization. We use $\hat{g}(X_t)$ to denote the spline approximation of $g_0(X_t)$ derived by the optimal solution of the proposed method.

The knots for $\mathbf{B}(X_t)$ are chosen as equally spaced quantile points in the range of X_t. We employ a heuristic to determine a proper value for the number of knots, i.e., J_n should be chosen such that the total sum of squares of the residuals is minimized for the fitted model among all choices of J_n. We implement a grid search with $J_n \in \{5, 6, \ldots, 18, 19\}$. It turns out the value of J_n induces a comparable impact on both the sequential method and the one-step method, in that both sums of squares of the residuals exhibit a similar decreasing pattern. For the sequential method, as the number of knots increases, the total sum of squared residuals (SSR) decreases first and then stabilizes around 2.86×10^{12}; after J_n reaches 15, SSR continues to show a trend of minor but steady decrease from knot to knot. In a similar fashion, the SSR by the one-step method also decreases first and then stabilizes around 2.72×10^{12}. Compared with the sequential method, the SSR by the one-step method has a smaller average value, but exhibits a bigger variation among different values of J_n, e.g., for $J_n = 13$, the value spikes above 2.95×10^{12}, but for $J_n = 16$, it plummets to 2.55×10^{12}. We pick $J_n = 7$ and $J_n = 8$, since each of them marks the smallest number of knots by which the SSR for the sequential method or for the one-step method, after experiencing a series of steady decreases, begins to touch the region of a stable level (the so-called elbow point).

To develop a proper model for the training data, we implement the automatic procedure of `auto.arima` to pick preliminary values for the order of ARMA process for the sequential method. The procedure suggests an ARMA(2, 2) model with $(\hat{\phi}_1, \hat{\phi}_2, \hat{\theta}_1, \hat{\theta}_2) = (-0.6118, -0.2972, 0.8892, 0.6276)$, for $J_n = 7$. The time series model appears to be overfitted since the associated standard errors are given by $(0.1884, 0.1818, 0.1559, 0.1443)$ and $\hat{\phi}_2$ is not significantly different from zero. We also fit the training data with $J_n = 7$ and ARMA(2, 2) by the proposed least squares method which yields $(\hat{\phi}_1, \hat{\phi}_2, \hat{\theta}_1, \hat{\theta}_2) = (0.3371, 0.5345, 0.3352, -0.0988)$. Next, we invoke Theorem 2 to obtain the standard errors of the above estimates by assuming that all regularity conditions are satisfied. It shows that the ARMA(2,2) model by the one-step method is overfitted since the standard errors are given as $(1.1416, 1.0604, 1.1503, 0.3296)$. By repeatedly using standard errors to detect overfitted models, both the sequential method and the one-step method finally entertain with AR(1) models. Specifically, the sequential method settles with $\hat{\phi} = 0.2865$ (se $= 0.076$) and the one-step method settles with $\hat{\phi} = 0.8987$ (se $= 0.035$). Both models pass the Ljung-Box test at five different lags. For the case $J_n = 8$, a similar situation arises, as `auto.arima` suggests an AR(4) with most of the fitted parameters being not significantly different from zero. Once again, through a process of trial and error, both methods favor AR(1) models, i.e., $\hat{\phi} = 0.2772$ (se $= 0.077$) for the sequential method, and $\hat{\phi} = 0.8916$ (se $= 0.036$). Despite $J_n = 7$ being a strong competitor, our preference tilt towards to $J_n = 8$ with AR(1), since in Figure 2, the one-step method usually produces a smaller SSR than that by the sequential method,

and $J_n = 8$ scores a significant reduction in the total sum of squared errors for the one-step method.

Fig. 2 The circles represent the total sum of squared residuals for the sequential method; the solid dots represent the total sum of squared residuals for the proposed method.

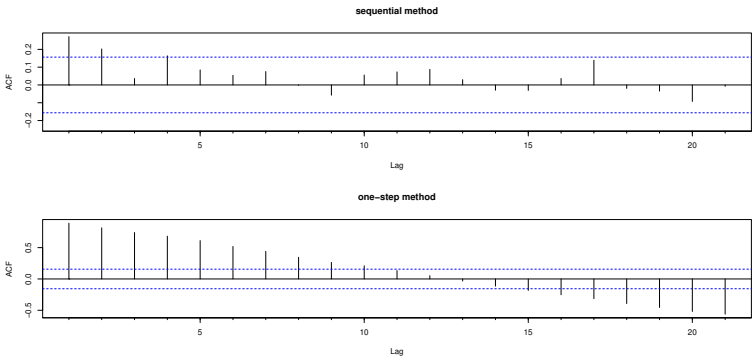


Finally, we fit model (1) with an AR(1) type random error processes for the training data by the sequential method and the proposed method respectively with $J_n = 8$. From Figure 1, it is evident that a significant difference exists between $\widetilde{g}(X_t)$ and $\hat{g}(X_t)$. At first sight, it might appear to be counterintuitive for $\hat{g}(X_t)$, the fitted smooth function by the one-step method, to differ substantially from $\widetilde{g}(X_t)$, the fitted smooth function by the sequential method (see the blue dashed curve and the solid red curve of Figure 1). However, the overall fitted values (the fitted curve plus the fitted ARMA process) by the sequential method and the proposed method are more closely in line with each other. While the sequential method subjugates the ARMA process to address the remnant of fitting a spline approximation to the data, our method possesses the advantage of being able to balance the trade-offs between the spline approximation and the ARMA model. The scatter plot of the fitted values vs the true values of Y_t, Figure 1, shows that the fitted model by the proposed method has a better fit, since the fitted values are distributed more compactly around the 45° line going through the origin. In Figure 3, we plot the ACF graphs for the remnants $Y_t - \tilde{g}(X_t)$ and $Y_t - \hat{g}(X_t)$. It shows that spline approximation alone cannot adequately address the dynamics of the scrape data, since both ACF graphs show the existence of unaccounted autocorrelations. At the same time, we note that in Figure 4 both the sequential method and the one-step method adequately solve the issue, in that the generated residuals seem to be uncorrelated, especially for one-step method, as all the ACF ticks are within the 95% confidence interval.

We continue to perform an evaluation of the model ($J_n = 8$, AR(1)) on the basis of a rolling forecasting origin method for successively 52 weeks for the year 2016. The forecasting origin is initially placed on January 2nd, 2016. All the data, from December 30, 2012, to the forecasting origin, are assumed to be known and fitted by model (1). Subsequently, one week ahead forecast for the scrape data is generated

by using the fitted models and the true value of HDD. Then the forecasting origin is moved forward one week at a time to fit new models and generate new forecasts, until it reaches December 24, 2016. Altogether, 52 forecasts are made for both the sequential method and the proposed least squares method. Let Y_t denote the true scrape value in week t of the the testing set and $\widehat{Y}_t$ denote its forecast value. We examine the forecasting accuracy based on the following forecast error metrics, i.e., Absolute Deviation (AD), Mean Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE), and Root Mean Square Deviation (RMSD), with AD $= \left|Y_t - \widehat{Y}_t\right|$ for week t, MAD $= \sum_{t=1}^{52} \left|Y_t - \widehat{Y}_t\right| /52$, MAPE $= \sum_{t=1}^{52} \left|\left(Y_t - \widehat{Y}_t\right) /Y_t\right| /52$, and RMSD $= \sqrt{\sum_{t=1}^{52} \left(Y_t - \widehat{Y}_t\right)^2 /52}$ (see, e.g., Armstrong and Collopy 1992; Bowerman et al. 2005; Hyndman et al. 2008; Petropoulos et al. 2022). We find that our one-step method performs better than the sequential method in terms of MAD, MAPE, and RMSD. More specifically, it achieves 109312.5 thousand cubic feet (MCF), 0.0472 , and 154919.0 MCF respectively for MAD, MAPE, and RMSD, while the sequential method achieves 129235.5 MCF, 0.0543, and 187598.5 MCF. The sequential method produces a smaller maximum AD, 547000.3 MCF, than that by the one-step method, 642692.5 MCF. A further investigation shows that both maximum values happen in week 48, from November 27th, 2016 to December 3rd, 2016. The plot for the ordered AD, Figure 5, shows that the one-step method generally produces more accurate forecast than that by the sequential method, except for the last point which corresponds to the forecast in week 48. If the highest point of Figure 5 is removed, one-step method outperforms sequential methods for the next 18 highest values of AD.

Fig. 3 Top: the ACF for $Y_t - \tilde{g}(X_t)$ values which are the estimates for ϵ_t in (1) given by the sequential method; Bottom: the ACF for $Y_t - \hat{g}(X_t)$ values which are the estimates for ϵ_t in (1) given by the proposed method.

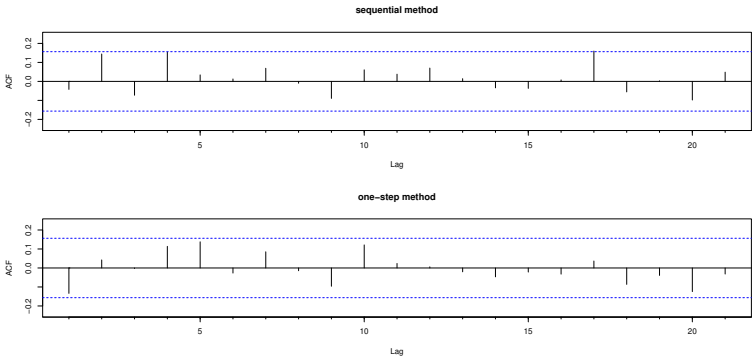


During the previous forecasting test, for the sequential method, the ARMA orders are fixed at $(1, 0)$. Aiming at more flexibility, we implement the sequential method with `auto.arima` being used to pick the order of ARMA process. During the 52 testing weeks, it achieves 131607.2 MCF for MAD, 0.0547 for MAPE, 621101.3 MCF for the maximum AD, and 191983.1 MCF for RMSD. For completeness, we also assess the

performance of $\widetilde{g}(X_t)$, the spline approximation without time series consideration. It shows that $\widetilde{g}(X_t)$ is outperformed by both the sequential method and the proposed method, demonstrating the necessity of using a time series component in scrape data modeling. Specifically, if time series dynamics are not addressed in the testing data and only spline approximation is used, MDA, MAPE, and maximum AD are equal to 174298.9 MCF, 0.0712, and 770422.2 MCF.

Moreover, to produce a comprehensive picture of the forecasting capacity of the one-step method, we repeat the test across different J_n values. The forecasting test results are summarized in Table 7. A row by row inspection of Table 7 makes it clear that our proposed method with an AR(1) process dominates the metrics MAD, MAPE, and RMSD, in that it almost always generates the best results for each J_n value. The results demonstrate the efficacy of Theorem 2 for model selection purpose. Also from an applied perspective, the heuristic to choose J_n by its elbow point appears to work well.

Fig. 4 Top: the ACF for residuals, estimates for ζ_t in (2), given by the sequential method; Bottom: the ACF for residuals, estimates for ζ_t in (2), given by the proposed method.

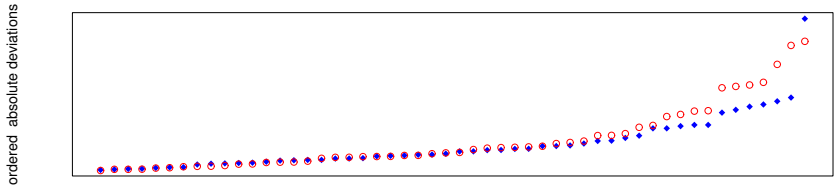


In summary, we fit a real-world data set with model (1) by using the proposed method and show that the proposed method is effective in model fitting. We invoke Theorem 2 to choose a proper ARMA order. The technique greatly facilitates model selection, and its effectiveness is established in the forecasting test. Rather than separating spline approximation and ARMA models into two arenas of model fitting and relegating ARMA models to a second place, the one-step method allows spline approximations and ARMA models to compete on an equal basis on the same stage to minimize the sum of squared errors. The proposed method opens a new possibility and offers new insight for nonparametric model (1). It readily lends itself as an addition to the arsenal of tools available for nonparametric time series analysis.

Table 7 Forecast accuracy for different models; Regression refers to spline approximation without addressing the serial correlation issue; Two-step refers to sequential method with auto.arima being used to choose ARMA order; Two-step AR(1) refers to sequential method with AR(1); and One-step AR(1) refers to the proposed method with AR(1).

		Regression	Two-step	Two-step AR(1)	One-step AR(1)
$J_n = 5$	MAD	177517.3	130700.2	127577.3	110803.0
	MAPE	0.0734	0.0558	0.0541	0.0481
	max AD	733710.5	535115.6	503171.3	644047.8
	RMSD	245891.7	186652.6	182356.9	156313.5
$J_n = 7$	MAD	174353.2	131383.8	128225	110215.1
	MAPE	0.07140065	0.0546	0.0540	0.0478
	max AD	762511.9	622775.3	553074.1	658089.4
	RMSD	246153.2	192506.8	186318.0	155665.3
$J_n = 8$	MAD	174298.9	131607.2	129235.5	109312.5
	MAPE	0.0712	0.0547	0.0543	0.0472
	max AD	770422.2	621101.3	547000.3	642692.5
	RMSD	246971.2	191983.1	187598.5	154919.0
$J_n = 11$	MAD	175420.4	138837.7	133504.9	115376.8
	MAPE	0.0717	0.0583	0.0560	0.0499
	max AD	759285.9	655971.6	579935.8	604635.9
	RMSD	250072.8	200472.2	193994.1	162019.7
$J_n = 14$	MAD	181497.5	141109.2	143740.2	114578.4
	MAPE	0.0755	0.0599	0.0619	0.0482
	max AD	755049.7	674575.4	588503.0	694568
	RMSD	254529.6	201237.5	200988.8	164707.3
$J_n = 17$	MAD	184547.3	143267.8	147126.7	113950.1
	MAPE	0.0762	0.0602	0.0626	0.0485
	max AD	786012.7	655668.4	572125.8	719067.2
	RMSD	256977.5	206435.5	205303.8	164544.7

Fig. 5 The plot for the ordered absolute deviations: the circles represent the sequential method, and the solid diamonds represent the one-step method.



6 Concluding remarks

We propose a method to estimate the mean function and the parameters of the ARMA process based on least squares estimation with spline approximation under a random design setup. Our proposed approach relaxes the independent covariates assumption and allows them to be weakly dependent. Utilizing results in empirical processes in

mixing sequence, we establish the consistency and asymptotic normality of the resulting estimator. Our numerical analysis, including both simulation studies and the examination of Iowa natural gas scrape data, shows that the proposed method provided excellent model fitting and forecasting ability and the performance supports our theoretical results.

Acknowledgements The natural gas scrape data were obtained by the second author through the research contract (Grant 5040224) between Applied Mathematics Laboratory of Towson University and Exelon Generation Company LLC. This work was partially supported by the National Institutes of Health grant R03AG067611 and R21AG070659, and the National Science Foundation grant DMS-1952486.

Appendix

A. The main results and proofs

We present the main results in this section. The proofs for Theorem 1 and Theorem 2 are provided. The remaining proofs are all relegated to the supplementary materials.

As $\mathcal{L}_n(\boldsymbol{\xi}) = \sum_{t=1}^{n} \zeta_t^2(\boldsymbol{\xi})$ is not convex with respect to $\boldsymbol{\xi}$, due to the MA component $\boldsymbol{\theta}$, in order to study the asymptotic property of $\hat{\boldsymbol{\xi}}$, we employ a second-order Taylor's expansion of $\zeta_t(\boldsymbol{\xi})$ around $\boldsymbol{\xi}_*$ (Davis and Dunsmuir 1997) : $\zeta_t(\boldsymbol{\xi}) \approx \zeta_t(\boldsymbol{\xi}_*) - \mathbf{D}_t^\top(\boldsymbol{\xi}_*)(\boldsymbol{\xi}-\boldsymbol{\xi}_*) - (\boldsymbol{\xi}-\boldsymbol{\xi}_*)^\top \mathbf{H}_t(\boldsymbol{\xi}_*)(\boldsymbol{\xi}-\boldsymbol{\xi}_*)/2$, where $\mathbf{D}_t(\boldsymbol{\xi}) = -\partial\zeta_t(\boldsymbol{\xi})/\partial\boldsymbol{\xi}$ and $\mathbf{H}_t(\boldsymbol{\xi}) = -\partial^2\zeta_t(\boldsymbol{\xi})/(\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^\top)$.

We decompose $\mathbf{D}_t(\boldsymbol{\xi})$ as $(\mathbf{D}_{t1}(\boldsymbol{\xi}), \mathbf{D}_{t2}(\boldsymbol{\xi}), \mathbf{D}_{t3}(\boldsymbol{\xi}))^\top$, such that $\mathbf{D}_{t1}(\boldsymbol{\xi}) = -\partial\zeta_t(\boldsymbol{\xi})/\partial\boldsymbol{\beta}$, $\mathbf{D}_{t2}(\boldsymbol{\xi}) = -\partial\zeta_t(\boldsymbol{\xi})/\partial\boldsymbol{\phi}$, and $\mathbf{D}_{t3}(\boldsymbol{\xi}) = -\partial\zeta_t(\boldsymbol{\xi})/\partial\boldsymbol{\theta}$, and partition $\mathbf{H}_t(\boldsymbol{\xi})$ as follows:

$$\mathbf{H}_t(\boldsymbol{\xi}) = \begin{pmatrix} \mathbf{H}_{t,11}(\boldsymbol{\xi}) & \mathbf{H}_{t,12}(\boldsymbol{\xi}) & \mathbf{H}_{t,13}(\boldsymbol{\xi}) \\ \mathbf{H}_{t,21}(\boldsymbol{\xi}) & \mathbf{H}_{t,22}(\boldsymbol{\xi}) & \mathbf{H}_{t,23}(\boldsymbol{\xi}) \\ \mathbf{H}_{t,31}(\boldsymbol{\xi}) & \mathbf{H}_{t,32}(\boldsymbol{\xi}) & \mathbf{H}_{t,33}(\boldsymbol{\xi}) \end{pmatrix}$$

where $\mathbf{H}_{t,11}(\boldsymbol{\xi}) = -\partial^2\zeta_t(\boldsymbol{\xi})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top$ is a zero $J\times J$ matrix, $\mathbf{H}_{t,12}(\boldsymbol{\xi}) = -\partial^2\zeta_t(\boldsymbol{\xi})/\partial\boldsymbol{\beta}\partial\boldsymbol{\phi}^\top$ is a $J\times p$ matrix, $\mathbf{H}_{t,13}(\boldsymbol{\xi}) = -\partial^2\zeta_t(\boldsymbol{\xi})/\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}^\top$ is a $J\times q$ matrix, $\mathbf{H}_{t,21}(\boldsymbol{\xi}) = \mathbf{H}_{t,12}^\top(\boldsymbol{\xi})$, $\mathbf{H}_{t,22}(\boldsymbol{\xi}) = -\partial^2\zeta_t(\boldsymbol{\xi})/\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^\top$ is a zero $p\times p$ matrices, $\mathbf{H}_{t,23}(\boldsymbol{\xi}) = -\partial^2\zeta_t(\boldsymbol{\xi})/\partial\boldsymbol{\phi}\partial\boldsymbol{\theta}^\top$ is a $p\times q$ matrix, $\mathbf{H}_{t,31}(\boldsymbol{\xi}) = \mathbf{H}_{t,13}^\top(\boldsymbol{\xi})$, $\mathbf{H}_{t,32}(\boldsymbol{\xi}) = \mathbf{H}_{t,23}^\top(\boldsymbol{\xi})$, and $\mathbf{H}_{t,33}(\boldsymbol{\xi}) = -\partial^2\zeta_t(\boldsymbol{\xi})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ is a $q\times q$ matrix.

Let $[\mathbf{A}]_l$ denote the l^{th} element of the vector $\mathbf{A}$. By simple algebra, we obtain that $\mathbf{D}_{t1}(\boldsymbol{\xi}) = \boldsymbol{\theta}^{-1}(B)\boldsymbol{\phi}(B)\mathbf{W}_t$, $[\mathbf{D}_{t2}(\boldsymbol{\xi})]_l = \boldsymbol{\theta}^{-1}(B)\epsilon_{t-l}(\boldsymbol{\beta}), 1 \le l \le p$, $[\mathbf{D}_{t3}(\boldsymbol{\xi})]_l = \boldsymbol{\theta}^{-1}(B)\zeta_{t-l}(\boldsymbol{\xi})$, $1 \le l \le q$,

$$\begin{aligned}
\left[\frac{\partial}{\partial\boldsymbol{\phi}}\left[\frac{\partial\zeta_t(\boldsymbol{\xi})}{\partial\boldsymbol{\beta}}\right]_l\right]_m &= \frac{1}{\boldsymbol{\theta}(B)}\left[\mathbf{W}_{t-m}\right]_l, 1 \le l \le J, 1 \le m \le p, \\
\left[\frac{\partial}{\partial\boldsymbol{\theta}}\left[\frac{\partial\zeta_t(\boldsymbol{\xi})}{\partial\boldsymbol{\beta}}\right]_l\right]_m &= \frac{\boldsymbol{\phi}(B)}{\boldsymbol{\theta}^2(B)}\left[\mathbf{W}_{t-m}\right]_l, 1 \le l \le J, 1 \le m \le q, \\
\left[\frac{\partial}{\partial\boldsymbol{\theta}}\left[\frac{\partial\zeta_t(\boldsymbol{\xi})}{\partial\boldsymbol{\phi}}\right]_l\right]_m &= \frac{\epsilon_{t-l-m}(\boldsymbol{\beta})}{\boldsymbol{\phi}(B)\boldsymbol{\theta}(B)}, 1 \le l \le p, 1 \le m \le q, \text{ and} \\
\left[\frac{\partial}{\partial\boldsymbol{\theta}}\left[\frac{\partial\zeta_t(\boldsymbol{\xi})}{\partial\boldsymbol{\theta}}\right]_l\right]_m &= \frac{2}{\boldsymbol{\theta}^2(B)}\zeta_{t-l-m}(\boldsymbol{\xi}), 1 \le l, m \le q.
\end{aligned}$$

Furthermore, let $\mathbf{V}_t$ be a symmetric matrix of dimension $(J+p+q)\times(J+p+q)$, whose upper triangular elements are given as

$$[\mathbf{V}_t]_{l,m} = \begin{cases} 0 & \text{if } 1 \le l \le m \le J \text{ or } J+1 \le l \le m \le J+p, \\ -\boldsymbol{\theta}_*^{-1}(B)\left[\mathbf{W}_{t-(m-J)}\right]_l & \text{if } 1 \le l \le J, 1 \le m-J \le p, \\ -\boldsymbol{\theta}_*^{-2}(B)\boldsymbol{\phi}_*(B)\left[\mathbf{W}_{t-(m-J-p)}\right]_l & \text{if } 1 \le l \le J, 1 \le m-J-p \le q, \\ -\boldsymbol{\theta}_*^{-1}(B)\boldsymbol{\phi}_*^{-1}(B)\zeta_{t-(l-J)-(m-J-p)} & \text{if } 1 \le l-J \le p, 1 \le m-J-p \le q, \\ -2\boldsymbol{\theta}_*^{-2}(B)\zeta_{t-(l-J-p)-(m-J-p)} & \text{if } J+p+1 \le l \le m \le J+p+q. \end{cases}$$

We partition $\mathbf{V}_t$ as follows:

$$\mathbf{V}_t = \begin{pmatrix} \mathbf{V}_{t,11} & \mathbf{V}_{t,12} & \mathbf{V}_{t,13} \\ \mathbf{V}_{t,21} & \mathbf{V}_{t,22} & \mathbf{V}_{t,23} \\ \mathbf{V}_{t,31} & \mathbf{V}_{t,32} & \mathbf{V}_{t,33} \end{pmatrix},$$

where $\mathbf{V}_{t,11}$ is a $J\times J$ matrix, $\mathbf{V}_{t,12}$ is a $J\times p$ matrix, $\mathbf{V}_{t,13}$ is a $J\times q$ matrix, $\mathbf{V}_{t,22}$ is a $p\times p$ matrix, $\mathbf{V}_{t,23}$ is a $p\times q$ matrix, and $\mathbf{V}_{t,33}$ is a $q\times q$ matrix. By the definition, $\mathbf{V}_{t,11}=\mathbf{0}$ and $\mathbf{V}_{t,22}=\mathbf{0}$.

In addition, let $R_t = (g_0(X_t) - \boldsymbol{\beta}_*^\top \mathbf{B}(X_t))1\{t>0\} = (\epsilon_t(\boldsymbol{\beta}_*) - \epsilon_t)1\{t>0\}$ be the spline approximation error at time t. In the following Proposition 1, we show that $\mathbf{D}_t(\boldsymbol{\xi}_*)$ and $\mathbf{H}_t(\boldsymbol{\xi}_*)$ are well approximated by $\mathbf{Q}_t$ and $\mathbf{V}_t$, respectively.

Proposition 1 *Suppose Conditions (C1) – (C4) hold. There exists some constants δ_1 and δ_2, such that for all $\|\boldsymbol{\beta}-\boldsymbol{\beta}_*\|\le\delta_1, \|(\boldsymbol{\phi}^\top,\boldsymbol{\theta}^\top)-(\boldsymbol{\phi}_*^\top,\boldsymbol{\theta}_*^\top)\|\le\delta_2$,*

(i) $|\zeta_t|\le\eta_t$, $|\zeta_t(\boldsymbol{\xi}_*) - \boldsymbol{\phi}_*(B)\boldsymbol{\theta}_*^{-1}(B)R_t - \zeta_t| \le r^t\eta_0$, $|\zeta(\boldsymbol{\xi})|\le\eta_t + C_2(\Delta+\delta_1)$, *and* $|\zeta_t(\boldsymbol{\xi}) - \zeta_t(\boldsymbol{\xi}_*)| \le C_3\delta_2\eta_t + C_2C_3\delta_2(\delta_1+\Delta) + C_2\delta_1$,

(ii) $\|\mathbf{D}_t(\boldsymbol{\xi})\|_\infty \le \omega_t$, $\mathbf{D}_{t1}(\boldsymbol{\xi}_*) - \mathbf{Q}_{t1} = \mathbf{0}$, *and* $\left\|\left(\mathbf{D}_{t2}^\top(\boldsymbol{\xi}_*), \mathbf{D}_{t3}^\top(\boldsymbol{\xi}_*)\right) - (\mathbf{Q}_{t2}^\top, \mathbf{Q}_{t3}^\top)\right\|_\infty \le r^t\eta_0 + C_2\Delta$,

(iii) $\|\mathbf{H}_t(\boldsymbol{\xi})\|_{\max}\le\omega_t$, $\mathbf{H}_{t,11}(\boldsymbol{\xi}_*) - \mathbf{V}_{t,11} = \mathbf{0}$, *and* $\|\mathbf{H}_t(\boldsymbol{\xi}_*) - \mathbf{V}_t\|_{\max} \le r^t\eta_0 + C_2\Delta$,

where $\eta_t = C_1\sum_{j=0}^\infty r^j\left|\epsilon_{t-j}\right|$, $\omega_t = \max\left\{C_2, r^{-(p+q)}\eta_t + C_2(\Delta+\delta_1)\right\}$, *and C_3 is defined in Lemma 7.*

Proposition 1 indicates that $\mathbf{D}_t(\boldsymbol{\xi}_*)$ and $\mathbf{H}_t(\boldsymbol{\xi}_*)$ can be approximated by $\mathbf{Q}_t$ and $\mathbf{V}_t$, respectively. Moreover, if $\boldsymbol{\xi}$ is sufficiently close to the true parameters $\boldsymbol{\xi}_*$, $\|\mathbf{D}_t(\boldsymbol{\xi})\|_\infty$ and $\|\mathbf{H}_t(\boldsymbol{\xi})\|_{\max}$ are bounded and the difference between $\zeta_t(\boldsymbol{\xi})$ and $\zeta_t(\boldsymbol{\xi}_*)$ is well bounded, too.

To circumvent the non-covexity of $T(\mathbf{h})$ with respect to $\mathbf{h}$, we study a convex objective function

$$T_1(\mathbf{h}) = \sum_{t=1}^n\left[\left(\zeta_t + \frac{\boldsymbol{\phi}_*(B)}{\boldsymbol{\theta}_*(B)}R_t - \mathbf{h}^\top\mathbf{Q}_t\right)^2 - \left(\zeta_t + \frac{\boldsymbol{\phi}_*(B)}{\boldsymbol{\theta}_*(B)}R_t\right)^2\right].$$

To facilitate the investigation of the the property of $T_1(\mathbf{h})$, two extra terms, $T_2(\mathbf{h})$ and $T_3(\mathbf{h})$, are introduced for the theoretical development

$$T_2(\mathbf{h}) = \sum_{t=1}^n\left[\left(\zeta_t + \frac{\boldsymbol{\phi}_*(B)}{\boldsymbol{\theta}_*(B)}R_t - \mathbf{h}^\top\mathbf{Q}_t - \frac{1}{2}\mathbf{h}^\top\mathbf{V}_t\mathbf{h}\right)^2 - \left(\zeta_t + \frac{\boldsymbol{\phi}_*(B)}{\boldsymbol{\theta}_*(B)}R_t\right)^2\right],$$

$$T_3(\mathbf{h}) = \sum_{t=1}^n\left[\left(\zeta_t(\boldsymbol{\xi}_*) - \mathbf{h}^\top\mathbf{D}_t(\boldsymbol{\xi}_*) - \frac{1}{2}\mathbf{h}^\top\mathbf{H}_t(\boldsymbol{\xi}_*)\mathbf{h}\right)^2 - \zeta_t^2(\boldsymbol{\xi}_*)\right],$$

which are to be investigated in the Lemmas 4 – 6 to bridge the gap between $T_1(\mathbf{h})$ and $T(\mathbf{h})$. It is noteworthy that, as these terms involve unknown quantities, such as $\mathbf{Q}_t$ and R_t, they cannot be computed in practice.

In light of Lemmas 4 – 6, we first establish that $T_1(\mathbf{h})$ is an excellent approximation of $T(\mathbf{h})$. Define $\Omega(C) := \{\mathbf{h} : \|\mathbf{h}_1\| \leq CJn^{-1/2}, \| \left(\mathbf{h}_2^\top, \mathbf{h}_3^\top\right) \| \leq CJ^{1/2}n^{-1/2}\}$ for any $C > 0$. We use $\bar{\Omega}(C)$ and $\Omega^c(C)$ to denote the boundary and the complement of $\Omega(C)$, respectively.

Proposition 2 *Suppose Conditions (C1)–(C4) hold. If $J = n^{1/(2\alpha+1)}$, for any $C > 0$,*

$$\sup_{\mathbf{h}\in\Omega(C)} |T_1(\mathbf{h}) - T(\mathbf{h})| \to_p 0.$$

Proposition 2 is inspired by Davis and Dunsmuir (1997). It demonstrates that $T(\mathbf{h})$ can be well approximated by $T_1(\mathbf{h})$ locally. Therefore, we can study the properties of the minimizer of $T_1(\mathbf{h})$ and infer the properties of the minimizer of $T(\mathbf{h})$. We refer to Davis and Dunsmuir (1997) for a detailed discussion. We next show that $T_1(\mathbf{h})$ achieves its minimum in a ball round 0 in the following proposition.

Proposition 3 *Under the same conditions as in Proposition 2, given any $0 < \varepsilon < 1$, there exists some $C_\varepsilon > 0$, such that*

$$P\left(\inf_{\mathbf{h}\in\bar{\Omega}(C_\varepsilon)\bigcup\Omega^c(C_\varepsilon)} T_1(\mathbf{h}) > 1\right) > 1 - \varepsilon.$$

Propositions 2 and 3 together enable us to establish the consistency of $\hat{\mathbf{h}}$ and subsequently $\hat{\boldsymbol{\xi}}$. Hence, the proofs for Theorem 1 and Theorem 2 are in order.

Proof of Theorem 1: By Proposition 3, given any $0 < \varepsilon < 1$, there exists some C_ε, such that

$$P\left(\inf_{\mathbf{h}\in\bar{\Omega}(C_\varepsilon)\bigcup\Omega^c(C_\varepsilon)} T_1(\mathbf{h}) > 1\right) > 1 - \varepsilon.$$

Under the event $\{\inf_{\mathbf{h}\in\bar{\Omega}(C_\varepsilon)\bigcup\Omega^c(C_\varepsilon)} T_1(\mathbf{h}) > 1\}$, we claim that there exists a local minimizer of $T(\mathbf{h})$, $\widehat{\mathbf{h}}$, which satisfies $\widehat{\mathbf{h}} \in \Omega(C_\varepsilon)$ but $\widehat{\mathbf{h}} \notin \bar{\Omega}(C_\varepsilon)$. Suppose the claim is not true. We can find a $\mathbf{h}_a \in \bar{\Omega}(C_\varepsilon)$, such that $T(\mathbf{h}_a) = \min_{\mathbf{h}\in\Omega(C_\varepsilon)} T(\mathbf{h})$.

By Proposition 2, for any $C > 0$, $\sup_{\mathbf{h}\in\Omega(C)} |T_1(\mathbf{h}) - T(\mathbf{h})| \to_p 0$. Choose C as C_ε. Then $0 \geq T(\mathbf{h}_a) - T(\mathbf{0}) \to_p T_1(\mathbf{h}_a) - T_1(\mathbf{0}) = T_1(\mathbf{h}_a) > 1$. Contradiction! Therefore, for any $0 < \varepsilon < 1$, there exists C_ε, such that $\widehat{\mathbf{h}} \in \Omega(C_\varepsilon)$ with probability at least $1 - \varepsilon$.

Given any $\mathbf{h} \in \Omega(C_\varepsilon)$, $E\left[\mathbf{h}_1^\top \mathbf{W}_t \mathbf{W}_t^\top \mathbf{h}_1\right] \leq \lambda_{\max} J^{-1} \left(C_\varepsilon^2 J^2 n^{-1}\right) = \lambda_{\max} C_\epsilon^2 J n^{-1}$. Noting that $\hat{\boldsymbol{\xi}} = \boldsymbol{\xi}_* + \widehat{\mathbf{h}}$, with probability at least $1 - \varepsilon$,

$$\begin{aligned} &E\left[\left(\hat{g}(X_t) - g_0(X_t)\right)^2\right] \leq 2E\left[\left(\hat{g}(X_t) - g_*(X_t)\right)^2\right] + 2E\left[\left(g_*(X_t) - g_0(X_t)\right)^2\right] \\ &= E\left[\widehat{\mathbf{h}}_1^\top \mathbf{W}_t \mathbf{W}_t^\top \widehat{\mathbf{h}}_1\right] + 2C_0^2 J^{-2\alpha} \leq 2\lambda_{\max} C_\epsilon^2 J n^{-1} + 2C_0^2 J^{-2\alpha}. \end{aligned}$$

Thus, $E\left[\left(\hat{g}(X_t) - g_0(X_t)\right)^2\right] = O_p(Jn^{-1} + J^{-2\alpha}) = O_p\left(n^{-2\alpha/(2\alpha+1)}\right)$. This completes the proof of Theorem 1. □

Proof of Theorem 2: In the proof of Theorem 1, we have shown that for any $0 < \varepsilon < 1$, there exists C_ε, such that $\widehat{\mathbf{h}} \in \Omega(C_\varepsilon)$ with probability at least $1 - \varepsilon$. Thus, we restrict our attention to the event that $\widehat{\mathbf{h}} \in \Omega(C_\varepsilon)$.

We consider that $S(\mathbf{b}_2, \mathbf{b}_3) := T_1((\widehat{\mathbf{h}}_1^\top, \mathbf{b}_2^\top/\sqrt{n}, \mathbf{b}_3^\top/\sqrt{n})^\top) - T_1((\widehat{\mathbf{h}}_1^\top, \mathbf{0}^\top, \mathbf{0}^\top)^\top)$. It is easily seen that

$$\begin{aligned} S(\mathbf{b}_2, \mathbf{b}_3) = &\sum_{t=1}^{n} \left(\frac{\mathbf{b}_2^\top \mathbf{Q}_{t2}}{\sqrt{n}} + \frac{\mathbf{b}_3^\top \mathbf{Q}_{t3}}{\sqrt{n}} \right)^2 - 2 \sum_{t=1}^{n} \zeta_t \left(\frac{\mathbf{b}_2^\top \mathbf{Q}_{t2}}{\sqrt{n}} + \frac{\mathbf{b}_3^\top \mathbf{Q}_{t3}}{\sqrt{n}} \right) \\ &- 2 \sum_{t=1}^{n} \frac{\boldsymbol{\phi}_*(B)}{\boldsymbol{\theta}_*(B)} R_t \left(\frac{\mathbf{b}_2^\top \mathbf{Q}_{t2}}{\sqrt{n}} + \frac{\mathbf{b}_3^\top \mathbf{Q}_{t3}}{\sqrt{n}} \right) \\ &+ 2 \sum_{t=1}^{n} \widehat{\mathbf{h}}_1^\top \mathbf{Q}_{t1} \left(\frac{\mathbf{b}_2^\top \mathbf{Q}_{t2}}{\sqrt{n}} + \frac{\mathbf{b}_3^\top \mathbf{Q}_{t3}}{\sqrt{n}} \right) \end{aligned}$$

By Lemma 3, we obtain that

$$\begin{aligned} &\sum_{t=1}^{n} \left[\left(\frac{\mathbf{b}_2^\top \mathbf{Q}_{t2}}{\sqrt{n}} + \frac{\mathbf{b}_3^\top \mathbf{Q}_{t3}}{\sqrt{n}} \right)^2 - 2\zeta_t \left(\frac{\mathbf{b}_2^\top \mathbf{Q}_{t2}}{\sqrt{n}} + \frac{\mathbf{b}_3^\top \mathbf{Q}_{t3}}{\sqrt{n}} \right) \right] \\ \to_d\ &\sigma^2 \left(\mathbf{b}_2^\top, \mathbf{b}_3^\top \right) \boldsymbol{\Sigma} \left(\mathbf{b}_2, \mathbf{b}_3 \right) - 2 \left(\mathbf{b}_2^\top, \mathbf{b}_3^\top \right) N(0, \sigma^2 \boldsymbol{\Sigma}), \end{aligned} \tag{4}$$

over $\|(\mathbf{b}_2^\top, \mathbf{b}_3^\top)\| \leq C$ for any $C > 0$.

According to Condition (C3), $\{\zeta_t\}$ and $\{X_t\}$ are independent. Hence, $\{R_t\}$ and $\{(\mathbf{Q}_{t2}, \mathbf{Q}_{t3})\}$ are independent. As $|R_t| \leq \Delta \leq C_0 J^{-\alpha}$ and hence $|\boldsymbol{\phi}_*(B)\boldsymbol{\theta}_*^{-1}(B)R_t| \leq C_0 C_2 J^{-\alpha} \to 0$, by the same arguments as used for Lemma 3, we can show that

$$\sup_{\|(\mathbf{b}_2^\top, \mathbf{b}_3^\top)\| \leq C} 2 \left| \sum_{t=1}^{n} \frac{\boldsymbol{\phi}_*(B)}{\boldsymbol{\theta}_*(B)} R_t \left(\frac{\mathbf{b}_2^\top \mathbf{Q}_{t2}}{\sqrt{n}} + \frac{\mathbf{b}_3^\top \mathbf{Q}_{t3}}{\sqrt{n}} \right) \right| = o_p(1). \tag{5}$$

The independence between $\{\zeta_t\}$ and $\{X_t\}$ again implies the independence between $\mathbf{Q}_{t1}$ and $(\mathbf{Q}_{t2}, \mathbf{Q}_{t3})$, $\mathbf{b}_1^\top \mathbf{Q}_{t1}$. Thus, $E\left[\mathbf{h}_1^\top \mathbf{Q}_{t1} \left(\mathbf{b}_2^\top \mathbf{Q}_{t2} + \mathbf{b}_3^\top \mathbf{Q}_{t3} \right) \right] = 0$, as $E\left[\mathbf{Q}_{t2}\right] = E\left[\mathbf{Q}_{t3}\right] = 0$. Noting that $\|\widehat{\mathbf{h}}_1\| \leq C_\varepsilon J n^{-1/2}$, it follows from Lemma 2 that

$$\begin{aligned} 2 \left| \sum_{t=1}^{n} \widehat{\mathbf{h}}_1^\top \mathbf{Q}_{t1} \left(\frac{\mathbf{b}_2^\top \mathbf{Q}_{t2}}{\sqrt{n}} + \frac{\mathbf{b}_3^\top \mathbf{Q}_{t3}}{\sqrt{n}} \right) \right| &= C_\varepsilon J n^{-1/2} O_p\left(7(p^{1/2} + q^{1/2}) \sqrt{C_4 J \log n} \right) \\ &= o_p(1). \end{aligned} \tag{6}$$

Combining (4), (5), and (6) together yield that

$$S(\mathbf{b}_2, \mathbf{b}_3) \to_d \sigma^2 \left(\mathbf{b}_2^\top, \mathbf{b}_3^\top \right) \boldsymbol{\Sigma} \left(\mathbf{b}_2, \mathbf{b}_3 \right) - 2 \left(\mathbf{b}_2^\top, \mathbf{b}_3^\top \right) N(0, \sigma^2 \boldsymbol{\Sigma})$$

over $\|(\mathbf{b}_2^\top, \mathbf{b}_3^\top)\| \leq C$ for any $C > 0$.

Following from Lemmas 4– 6, we have uniformly over $\|(\mathbf{b}_2^\top, \mathbf{b}_3^\top)\| \leq C$ for any $C > 0$.

$$T((\widehat{\mathbf{h}}_1^\top, \mathbf{b}_2^\top/\sqrt{n}, \mathbf{b}_3^\top/\sqrt{n})^\top) - T((\widehat{\mathbf{h}}_1^\top, \mathbf{0}^\top, \mathbf{0}^\top)^\top) \to_p S(\mathbf{b}_2, \mathbf{b}_3).$$

Noting that $N(0, \sigma^2 \boldsymbol{\Sigma}^{-1})$ is the minimizer of the random process which $S(\mathbf{b}_2, \mathbf{b}_3)$ converges to, by Lemma 2.2 and Remark 1 in Davis et al. (1992), there exists $(\widehat{\mathbf{b}}_2^\top, \widehat{\mathbf{b}}_2^\top)$, a local minimizer of $T((\widehat{\mathbf{h}}_1^\top, \mathbf{b}_2^\top/\sqrt{n}, \mathbf{b}_3^\top/\sqrt{n})^\top) - T((\widehat{\mathbf{h}}_1^\top, \mathbf{0}^\top, \mathbf{0}^\top)^\top)$, such that $(\widehat{\mathbf{b}}_2^\top, \widehat{\mathbf{b}}_3^\top)^\top \to_d N(0, \sigma^2 \boldsymbol{\Sigma}^{-1})$.

Since $\widehat{\mathbf{h}}$ is the minimizer of $T(\mathbf{h})$, $(\widehat{\mathbf{h}}_2^\top, \widehat{\mathbf{h}}_2^\top)$ must also be the minimizer of

$$T((\widehat{\mathbf{h}}_1^\top, \mathbf{h}_2^\top, \mathbf{h}_3^\top)^\top) - T((\widehat{\mathbf{h}}_1^\top, \mathbf{0}^\top, \mathbf{0}^\top)^\top).$$

We thus have $\sqrt{n}(\widehat{\mathbf{h}}_2^\top, \widehat{\mathbf{h}}_2^\top) = (\widehat{\mathbf{b}}_2^\top, \widehat{\mathbf{b}}_2^\top)$ and $\sqrt{n}(\widehat{\mathbf{h}}_2^\top, \widehat{\mathbf{h}}_3^\top)^\top \to_d N(0, \sigma^2 \boldsymbol{\Sigma}^{-1})$. This completes the proof of Theorem 2. □

B. Preliminary proposition and lemmas

Next, we present the technical proposition and lemmas that are used in the proofs of our theorems and corollaries. The proofs of the proposition and lemmas are relegated to supplementary materials.

Proposition 4 *If Condition (C4) is satisfied,*

$$\sup_{\|\mathbf{h}_1\|=1, \|(\boldsymbol{\phi}^\top, \boldsymbol{\theta}^\top) - (\boldsymbol{\phi}_*^\top, \boldsymbol{\theta}_*^\top)\| \le \delta_2} \mathbf{h}_1^\top E\left[\left(\frac{\boldsymbol{\phi}(B)}{\boldsymbol{\theta}(B)}\mathbf{W}_t\right)\left(\frac{\boldsymbol{\phi}(B)}{\boldsymbol{\theta}(B)}\mathbf{W}_t^\top\right)\right]\mathbf{h}_1 \le \lambda_{\max} J^{-1} C_2^2,$$

where δ_2 is chosen as in Proposition 1.

Lemma 1 *Suppose Condition (C3) holds. Then*
(i) $P(|\zeta_t| > v) \le 2\exp\left(\frac{-v^2}{2(C_B^2 + C_B v)}\right)$ *and*
(ii) $E[\left|\sum_{i=0}^\infty a_i \zeta_{t-i}\right|^k] \le \left(\sum_{i=0}^\infty |a_i|\right)^k k! C_B^k/2$, *for any a sequence* $\{a_t, t \ge 0\}$ *and* $k \ge 1$.

Lemma 2 *Suppose Conditions (C1) – (C4) hold. There exists some constant $C_4 > 0$ that does not depend on n, such that if $J = O(n^{1/(2\alpha+1)})$,*
(i)

$$P\left(\sup_{\|\mathbf{h}_1\| \le 1} \left|\mathbb{G}_n\left[\left(\mathbf{h}_1^\top \mathbf{Q}_{t1}\right)^2 \zeta_t^2\right]\right| > 7C_2\sqrt{C_4 J \log n}\right) \le 2\exp(-6J\log n).$$

(ii)

$$\sup_{\|\mathbf{h}_1\| \le 1, \mathbf{h}_1 \ne 0} \left|\left(\sigma^2 E\left[\left(\mathbf{h}_1^\top \mathbf{Q}_{t1}\right)^2\right]\right)^{-1} \mathbb{E}_n\left[\left(\mathbf{h}_1^\top \mathbf{Q}_{t1}\right)^2 \zeta_t^2\right]\right| = 1 + o_p(1).$$

(iii)

$$P\left(\sup_{\|\mathbf{h}_1\| \le 1, \|\mathbf{h}_2\| \le 1} n^{-1/2}\left|\mathbb{G}_n\left[\mathbf{h}_1^\top \mathbf{Q}_{t1}\mathbf{Q}_{t2}^\top \mathbf{h}_2\right]\right| > 7p^{1/2}\sqrt{C_4 J n^{-1}\log n}\right)$$
$$\le 2p\exp(-6J\log n).$$
$$P\left(\sup_{\|\mathbf{h}_1\| \le 1, \|\mathbf{h}_3\| \le 1} n^{-1/2}\left|\mathbb{G}_n\left[\mathbf{h}_1^\top \mathbf{Q}_{t1}\mathbf{Q}_{t3}^\top \mathbf{h}_3\right]\right| > 7q^{1/2}\sqrt{C_4 J n^{-1}\log n}\right)$$
$$\le 2q\exp(-6J\log n).$$

Lemma 3 *Suppose Conditions (C1) – (C4) hold. Then,*
(i) $\sup_{\|(\mathbf{h}_2^\top,\mathbf{h}_3^\top)\|\le 1}\left|\mathbb{E}_n\left[(\mathbf{h}_2^\top\mathbf{Q}_{t2}+\mathbf{h}_3^\top\mathbf{Q}_{t3})^2\zeta_t^2\right]-\sigma^2\left(\mathbf{h}_2^\top,\mathbf{h}_3^\top\right)\boldsymbol{\Sigma}\left(\mathbf{h}_2^\top,\mathbf{h}_3^\top\right)^\top\right|\to_{a.s.} 0$,
(ii) $\mathbb{G}_n\left[(\mathbf{h}_2^\top\mathbf{Q}_{t2}+\mathbf{h}_3^\top\mathbf{Q}_{t3})\zeta_t\right]\to_d\left(\mathbf{h}_2^\top,\mathbf{h}_3^\top\right)N(0,\sigma^2\boldsymbol{\Sigma})$, *given any* $(\mathbf{h}_2^\top,\mathbf{h}_3^\top)$ *such that* $\|(\mathbf{h}_2^\top,\mathbf{h}_3^\top)\|\le C$, *for any* $C>0$.
(iii) $\mathbb{G}_n\left[(\mathbf{h}_2^\top\mathbf{Q}_{t2}+\mathbf{h}_3^\top\mathbf{Q}_{t3})\zeta_t\right]\to_d\left(\mathbf{h}_2^\top,\mathbf{h}_3^\top\right)N(0,\sigma^2\boldsymbol{\Sigma})$ *on* $\|(\mathbf{h}_2^\top,\mathbf{h}_3^\top)\|\le C$, *for any* $C>0$.

Lemmas 4 – 6 follow from the steps in Davis and Dunsmuir (1997) and Brockwell and Davis (1991).

According to Proposition 1, $|\zeta_t|\le\eta_t$, $\|\mathbf{Q}_t\|_\infty\le\|\mathbf{Q}_t-\mathbf{D}_t(\boldsymbol{\xi}_*)\|_\infty+\|\mathbf{D}_t(\boldsymbol{\xi}_*)\|_\infty\le r^t\eta_0+C_2\Delta+\omega_t=:\chi_t$, and similarly $\|\mathbf{V}_t\|_{\max}\le\chi_t$. Thus,

$$\left|\mathbf{h}^\top\mathbf{Q}_t\right|\le\|\mathbf{h}_1\mathbf{Q}_{t1}\|+\|\mathbf{h}_2^\top\mathbf{Q}_{t2}+\mathbf{h}_3^\top\mathbf{Q}_{t3}\|\le C_2\|\mathbf{h}_1\|+\chi_t(\sqrt{p}\|\mathbf{h}_2\|+\sqrt{q}\|\mathbf{h}_3\|), \quad (7)$$

$$\begin{aligned}\left|\mathbf{h}^\top\mathbf{V}_t\mathbf{h}\right| &= \left|2\mathbf{h}_2^\top\mathbf{V}_{t,21}\mathbf{h}_1+2\mathbf{h}_3^\top\mathbf{V}_{t,31}\mathbf{h}_1+2\mathbf{h}_3^\top\mathbf{V}_{t,32}\mathbf{h}_2+\mathbf{h}_3^\top\mathbf{V}_{t,33}\mathbf{h}_3\right| \\ &\le 2C_2(\sqrt{p}\|\mathbf{h}_2\|+\sqrt{q}\|\mathbf{h}_3\|)\|\mathbf{h}_1\|+2\sqrt{pq}\chi_t\|\mathbf{h}_2\|\|\mathbf{h}_3\|+q\chi_t\|\mathbf{h}_3\|^2. \quad (8)\end{aligned}$$

Lemma 4 *Suppose Conditions (C1) – (C4) hold. If* $J^2\log n=o(n^{1/2})$, *then for any* $C>0$, $\sup_{\mathbf{h}\in\Omega(C)}|T_1(\mathbf{h})-T_2(\mathbf{h})|\to_p 0$.

Lemma 5 *Suppose Conditions (C1) – (C4) hold. If* $J^{-2\alpha+1/2}=o(n^{-1/2})$, *then for any* $C>0$, $\sup_{\mathbf{h}\in\Omega(C)}|T_2(\mathbf{h})-T_3(\mathbf{h})|\to_p 0$.

Lemma 6 *Suppose Conditions (C1) – (C4) hold. If* $J^2\log n=o(n^{1/2})$, *then for any* $C>0$, $\sup_{\mathbf{h}\in\Omega(C)}|T_3(\mathbf{h})-T(\mathbf{h})|\to_p 0$.

Lemma 7 *Under the same conditions as in Proposition 1, for any sequence* $\{a_t\}, t\ge 1$, *there exists some constant* C_3 *such that*

$$\left|\left(\frac{\boldsymbol{\phi}(B)}{\boldsymbol{\theta}(B)}-\frac{\boldsymbol{\phi}_*(B)}{\boldsymbol{\theta}_*(B)}\right)a_t\right|\le C_3\delta_2\sum_{i=0}^{\infty}r^i|a_{t-i}|,$$

where δ_2 *and* r *are defined in Proposition 1.*

References

Armstrong, J. S., Collopy, F. (1992). Error measures for generalizing about forecasting methods: empirical comparisons. *International Journal of Forecasting*, *8*(1), 69–80.

Bowerman, B. L., O'Connell, R. T., Koehler, A. B. (2005). *Forecasting, time series, and regression: an applied approach. 4th* edition. Boston, MA: Brooks/Cole, Cengage Learning.

Box, G. E., Jenkins, G. M., Reinsel, G. C., Ljung, G. M. (2016). *Time series analysis: forecasting and control. 5th* edition. Hoboken, New Jersey: John Wiley and Sons Inc.

Brockwell, P. J., Davis, R. A. (1991). *Time series: theory and methods, 2nd* edition. Springer Series in Statistics. New York: Springer-Verlag.

Carroll, R. J., Fan, J., Gijbels, I., Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, *92*(438), 477–489.

Chernozhukov, V., Chetverikov, D., Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, *41*(6), 2786–2819.

Davis, R. A., Dunsmuir, W. T. (1997). Least absolute deviation estimation for regression with arma errors. *Journal of Theoretical Probability*, *10*(2), 481–497.

Davis, R. A., Knight, K., Liu, J. (1992). M-estimation for autoregressions with infinite variance. *Stochastic Processes and Their Applications*, *40*(1), 145–180.

De Boor, C., De Boor, C. (1978). *A practical guide to splines*, volume 27. New York: springer-verlag.

Durbán, M., Currie, I. D. (2003). A note on p-spline additive models with correlated errors. *Computational Statistics*, *18*(2), 251–262.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, *21*(1), 196–216.

Ganesh, E., Rajendran, V., Ravikumar, D., Kumar, P. S., Revathy, G., Harivardhan, P. (2021). Detection and route estimation of ship vessels using linear filtering and arma model from AIS data. *International Journal of Oceans and Oceanography*, *15*(1), 1–10.

Greenhouse, J. B., Kass, R. E., Tsay, R. S. (1987). Fitting nonlinear models with arma errors to biological rhythm data. *Statistics in Medicine*, *6*(2), 167–183.

Hall, P., Heyde, C. C. (2014). *Martingale limit theory and its application*. New York: Academic Press, Inc.

Hall, P., Keilegom, I. V. (2003). Using difference-based methods for inference in nonparametric regression with time series errors. *Journal of the Royal Statistical Society. Series B*, *65*(2), 443–456.

Hart, J. D. (1994). Automated kernel smoothing of dependent data by using time series cross-validation. *Journal of the Royal Statistical Society, Series B*, *56*(3), 529–542.

Hart, J. D., Wehrly, T. E. (1986). Kernel regression estimation using repeated measurements data. *Journal of the American Statistical Association*, *81*(396), 1080–1088.

Hastie, T. J., Tibshirani, R. J. (1990). *Generalized additive models*. Boca Raton: Routledge.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, *31*(5), 1600–1635.

Hyndman, R. J., Koehler, A. B., Ord, J. K., Snyder, R. D. (2008). *Forecasting with exponential smoothing: the state space approach*. Berlin: Springer-Verlag.

Kohn, R., Ansley, C. F., Wong, C.-M. (1992). Nonparametric spline regression with autoregressive moving average errors. *Biometrika*, *79*(2), 335–346.

Krivobokova, T., Kauermann, G. (2007). A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, *102*(480), 1328–1337.

Lee, Y. K., Mammen, E., Park, B. U. (2010). Bandwidth selection for kernel regression with correlated errors. *Statistics*, *44*(4), 327–340.

Liang, H.-Y., Jing, B.-Y. (2009). Asymptotic normality in partial linear models based on dependent errors. *Journal of statistical planning and inference*, *139*(4), 1357–1371.

Merlevède, F., Peligrad, M., Rio, E. (2011). A bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, *151*(3-4), 435–474.

Miaou, S.-P. (1990). A stepwise time series regression procedure for water demand model identification. *Water Resources Research*, *26*(9), 1887–1897.

Mokkadem, A. (1988). Mixing properties of arma processes. *Stochastic Processes and Their Applications*, *29*(2), 309–315.

Opsomer, J., Wang, Y., Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science*, *16*(2), 134–153.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., ... Ziel, F. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, *38*(3), 705–871.

Qiu, D., Shao, Q., Yang, L. (2013). Efficient inference for autoregressive coefficients in the presence of trends. *Journal of Multivariate Analysis*, *114*, 40 – 53.

Roussas, G. G., Tran, L. T. (1992). Asymptotic normality of the recursive kernel regression estimate under dependence conditions. *The Annals of Statistics*, *20*(1), 98–120.

Roussas, G. G., Tran, L. T., Ioannides, D. A. (1992). Fixed design regression for time series: Asymptotic normality. *Journal of Multivariate Analysis*, *40*(2), 262–291.

Serra, P., Krivobokova, T., Rosales, F. (2018). Adaptive non-parametric estimation of mean and autocovariance in regression with dependent errors. *arXiv preprint arXiv:1812.06948*.

Shao, Q., Yang, L. (2011). Autoregressive coefficient estimation in nonparametric analysis. *Journal of Time Series Analysis*, *32*(2), 587–597.

Shao, Q., Yang, L. (2017). Oracally efficient estimation and consistent model selection for auto-regressive moving average time series with trend. *Journal of the Royal Statistical Society Series B*, *79*(2), 507–524.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, *8*(6), 1348–1360.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, *14*(2), 590–606.

Straumann, D., Mikosch, T. (2006). Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *The Annals of Statistics*, *34*(5), 2449–2495.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tran, L., Roussas, G., Yakowitz, S., Van, B. T. (1996). Fixed-design regression for linear time series. *The Annals of Statistics*, *24*(3), 975–991.

Truong, Y. K. (1991). Nonparametric curve estimation with time series errors. *Journal of Statistical Planning and Inference*, *28*(2), 167–183.

Truong-Van, B., Bru, N. (2001). Asymptotic normality of spline estimator when the errors are a linear stationary process. *Journal of Nonparametric Statistics*, *13*(5), 741–761.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, *42*(3), 1166–1202.

Volkonskii, V., Rozanov, Y. A. (1959). Some limit theorems for random functions. i. *Theory of Probability & Its Applications*, *4*(2), 178–197.

Wu, R., Wang, Q. (2012). Shrinkage estimation for linear regression with arma errors. *Journal of Statistical Planning and Inference*, *142*(7), 2136–2148.

Zhou, S., Shen, X., Wolfe, D. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, *26*(5), 1760–1782.

Zinde-Walsh, V., Galbraith, J. W. (1991). Estimation of a linear regression model with stationary arma (p, q) errors. *Journal of Econometrics*, *47*(2-3), 333–357.