Computational Statistics and Data Analysis 180 (2023) 107682

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Dimension reduction in time series under the presence of conditional heteroscedasticity

Murilo da Silva, T.N. Sriram, Yuan Ke *

Department of Statistics, University of Georgia, Athens, GA 30602, USA

ARTICLE INFO

Article history:
Received 2 May 2022
Received in revised form 6 December 2022
Accepted 7 December 2022
Available online 14 December 2022

Keywords:
Angular representation
Dimension reduction
Heteroscedasticity
Iterative estimation
Nadaraya-Watson smoother
Time series

ABSTRACT

Consider a time series, where the conditional mean is assumed to be an unknown function of linear combinations of past p observations and the conditional variance is assumed to be an unknown function of linear combinations of past q squared residuals. The linear combinations are assumed to contain all the necessary information about the time series that is available through the conditional mean and conditional variance, respectively. Nadaraya-Watson kernel smoother is used to estimate the unknown mean and variance function and an iterative approach is proposed to estimate the parameter matrices associated with the linear combinations. The estimators are shown to be consistent. To overcome computational challenges and provide numerical stability, a novel angular representation of parameter matrices is introduced. The numerical performance of the proposed method on forecasting the conditional mean is assessed by simulations studies. A real data of Brazilian Real (BRL)/U.S. Dollar Exchange Rate is analyzed. For the BRL/USD series, the estimated linear combinations yield a better time series model than an AR-ARCH model in terms of out-of-sample forecasts.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

In time series models, the forecast of the current value, x_t, based upon the past information is simply the conditional mean of x_t, which depends upon the past values, $\{x_{t-1}, \ldots, x_1\}$, of a series. Traditional econometric models assume that the conditional variance of x_t stays constant at any given time point and does not depend on the past values. However, in applied economics, there are examples where the conditional variance of x_t is larger for some time points (or a range of time points) than for others, leading to so-called heteroscedasticity. In financial applications, where the dependent variable is the return on an asset or portfolio and the variance of the return represents the risk level of those returns, some time periods may be riskier than others in that the magnitude of variance of the return at some times is greater than at others. There are also instances where the risky times are not scattered randomly across data; instead, there is a degree of autocorrelation in the riskiness of financial returns.

To handle the issue of heteroscedasticity in time series, Engle (1982) proposed a new class of models called Autoregressive Conditionally Heteroscedastic (ARCH) models, where the conditional variance depends upon the past values of the series. He also illustrated the usefulness of ARCH models in economics and finance. Bollerslev (1986) generalized the purely autoregressive ARCH model to an autoregressive-moving average model called the Generalized Autoregressive Conditional

* Corresponding author.
E-mail address: Yuan.Ke@uga.edu (Y. Ke).

https://doi.org/10.1016/j.csda.2022.107682
0167-9473/© 2022 Elsevier B.V. All rights reserved.

Heteroscedastic (GARCH) model. The ARCH and GARCH models are widely used for modeling heteroscedastic time series, where the goal is to provide a volatility measure that can be used in financial decisions concerning risk analysis, portfolio selection and derivative pricing; see Engle (2001).

Literature (Masry and Tjøstheim, 1995) proposed a general nonlinear system of the ARCH type and considered nonparametric estimation of the conditional mean function and the conditional variance function characterizing the system. Following the single-indexing idea (e.g. Ichimura, 1993; Xia et al., 2002a) extended the ARCH model of Engle (1982) to a flexible form called single-index volatility models and focused on estimating only the unknown variance function and the associated single-index coefficient. In related literature, Fan and Yao (1998) considered efficient estimation of conditional variance functions in stochastic regression and, more generally, Fan and Yao (2008) provided modern parametric and nonparametric methods for analyzing nonlinear time series data, and Guo et al. (2017) studied dynamic structure for high dimensional covariance matrices. Recently, in a regression set up, Ma and Zhu (2019) proposed a class of locally efficient semiparametric estimators to simultaneously estimate the central mean subspace and central variance subspace with the help of a parameterization strategy. Although their work and our work share some similarities, there are differences in the two approaches. First, we study out-of-sample prediction of the conditional mean function in the presence of conditional heteroscedasticity. Second, we adopt a fully nonparametric iterative estimation of the conditional mean and the variance functions without any model assumption.

Motivated by the work of Li et al. (2003) and Yin and Cook (2005), during the past decade, some authors have extended the theory of sufficient dimension reduction to time series. Without specifying a model, Park et al. (2010) developed the notion of *Time Series Central Subspace* (TSCS), which represents a reduction in the dimension of $\boldsymbol{X}_{t-1} = (x_{t-1}, ..., x_{t-p})^\intercal$ for a pre-specified p such that the conditional distribution of $x_t|\boldsymbol{X}_{t-1}$ is same as the conditional distribution of $x_t|(\boldsymbol{\Phi}_1^\intercal \boldsymbol{X}_{t-1}, ..., \boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_{t-1})$ for some known $d < p$. This is equivalent to saying that x_t is conditionally independent of $\boldsymbol{X}_{t-1}$ given $\boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_{t-1}$, where $\boldsymbol{\Phi}_d = (\Phi_1, \ \ldots, \ \Phi_d)$. They estimated the $p \times d$ matrix $\boldsymbol{\Phi}_d$ nonparametrically by maximizing an estimating function based on Kullback-Leibler divergence and showed that the estimator of $\boldsymbol{\Phi}_d$ is strongly consistent when p and d are known. Park et al. (2009) proposed the notion of *Central Mean Subspace* for time series x_t which represents a reduction in the dimension of $\boldsymbol{X}_{t-1}$, where all the information in the conditional mean $E(x_t|\boldsymbol{X}_{t-1})$ is contained in $E(x_t|\boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_{t-1})$. They estimated $\boldsymbol{\Phi}_d$ by minimizing the residual sum of squares based on a Nadaraya–Watson smoother of the conditional mean function. Once again, they showed that the estimator of $\boldsymbol{\Phi}_d$ is strongly consistent when p and d are known.

Assuming that the conditional mean of x_t is zero, Park and Samadi (2014) developed a notion of *Central Variance Subspace* (TSCS) for the squared series, $z_t = x_t^2$, which represents a reduction in the dimension of $\boldsymbol{Z}_{t-1} = (z_{t-1}, ..., z_{t-p})^\intercal$ for a known p, where all the information in $E(z_t|\boldsymbol{Z}_{t-1})$ is contained in $E[z_t|(\Gamma_1^\intercal \boldsymbol{Z}_{t-1}, ..., \Gamma_d^\intercal \boldsymbol{Z}_{t-1})]$. To estimate $\boldsymbol{\Gamma} = (\Gamma_1, ..., \Gamma_d)$, they used the same approach as in Park et al. (2009). For the square series, z_t, Park and Sriram (2017) considered reduction in the dimension of $\boldsymbol{Z}_{t-1} = (z_{t-1}, ..., z_{t-p})^\intercal$ for a pre-specified p such that the conditional distribution of $z_t|\boldsymbol{Z}_{t-1}$ is same as the conditional distribution of $z_t|(\Gamma_1^\intercal \boldsymbol{Z}_{t-1}, ..., \Gamma_d^\intercal \boldsymbol{Z}_{t-1})$. To estimate $\boldsymbol{\Gamma}$, they proposed a robust estimation methodology based on Density Power Divergences (DPD) that is indexed by a tuning parameter $\alpha \in [0, 1]$ which yields a continuum of estimators, $\{\widehat{\boldsymbol{\Gamma}}_\alpha; \alpha \in [0, 1]\}$, where α controls the trade-off between robustness and efficiency of the DPD estimators.

The aforementioned literature on sufficient dimension reduction for time series focuses either on the reduction in the dimension of $\boldsymbol{X}_{t-1} = (x_{t-1}, ..., x_{t-p})^\intercal$ or on the reduction in the dimension of $\boldsymbol{Z}_{t-1} = (x_{t-1}^2, ..., x_{t-p}^2)^\intercal$ when the conditional mean of x_t is assumed to be zero. Incidentally, the general set up in Park et al. (2010) allowed them to consider a simulation model (see Model 1 in Section 4 of Park et al. (2010)) where both the conditional mean and variance are (nonlinear) functions depending on $\boldsymbol{X}_{t-1}$, and their TSCS approach was successful in estimating the unknown dimensions in the mean and the variance. Therefore, without assuming a model, it is of interest to develop an approach to reduce the dimension in the unknown conditional mean function as well as the conditional variance function of a time series, when x_t is conditionally heteroscedastic as in an AR-ARCH model.

Recently, Park and Samadi (2020) proposed an approach called *Two-stage Central Subspace* (similar to the TSCS approach of Park et al. (2010)) which *separately* estimates the dimensions in the mean and the variance functions. More specifically, their first-stage assumes that x_t is conditionally independent of $\boldsymbol{X}_{t-1}$ given $\boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_{t-1}$, and estimates $\boldsymbol{\Phi}_d$ for a fixed p and d with $d < p$. In the second-stage, they construct a residual series, $\varepsilon_t = x_t - E(x_t|\boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_{t-1})$ based on the unknown $\boldsymbol{\Phi}_d$, set $\boldsymbol{\varepsilon}_{t-1}^2 = (\varepsilon_{t-1}^2, ..., \varepsilon_{t-q}^2)^\intercal$ for some $q \geq 1$, assume that ε_t^2 is conditionally independent of $\boldsymbol{\varepsilon}_{t-1}^2$ given $\boldsymbol{\Gamma}_{\widetilde{d}}^\intercal \boldsymbol{\varepsilon}_{t-1}^2$, and estimate the $q \times \widetilde{d}$ matrix $\boldsymbol{\Gamma}_{\widetilde{d}} = (\Gamma_1, \ \ldots, \ \Gamma_{\widetilde{d}})$ for a fixed q and $\widetilde{d}$ with $\widetilde{d} < q$. They also prove the consistency of estimators of $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$. A major drawback of Park and Samadi (2020)'s work is that their residual series, ε_t, is not observable because it depends on the unknown parameter matrix $\boldsymbol{\Phi}_d$, which makes the estimation of $\boldsymbol{\Gamma}_{\widetilde{d}}$ questionable! In fact, Park and Samadi (2020) recognize this drawback in their article and say that, "*This procedure may not be a very efficient approach to estimate both the mean and variance functions*." While they do mention that in applications one should use an estimated residual that is observable, their theoretical result for the conditional variance function holds only for an unobservable residual series. Thus, the work of Park and Samadi (2020) does not satisfactorily address the problem of estimation of the dimensions in the conditional mean and the variance functions based on observable quantities.

In this article, we assume that the conditional mean is $E(x_t|\mathbf{X}_{t-1}) = f(\boldsymbol{\Phi}_d^\intercal \mathbf{X}_{t-1})$ for an unknown, possibly nonlinear, function $f(\cdot)$ and $\boldsymbol{\Phi}_d$ defined above. We also assume that the conditional variance is $V(x_t|\mathbf{X}_{t-1}) = g(\boldsymbol{\Gamma}_{\widetilde{d}}^\intercal \boldsymbol{\varepsilon}_{t-1}^2)$ for an unknown, possibly nonlinear, function $g(\cdot)$ and for $\boldsymbol{\Gamma}_{\widetilde{d}}$ and $\boldsymbol{\varepsilon}_{t-1}^2$ defined above. Without assuming a parametric model for x_t, we use a

Nadaraya-Watson smoother of $f(\cdot)$ and find an initial estimator, $\widehat{\boldsymbol{\Phi}}_0$, of $\boldsymbol{\Phi}_d$ by minimizing the residual sum of squares. This step focuses only on the estimation of the mean function. We then construct the estimated residuals, $\widehat{\varepsilon}_t = x_t - \widehat{f}_n(\widehat{\boldsymbol{\Phi}}_0^{\intercal}\boldsymbol{X}_{t-1})$, use a Nadaraya-Watson smoother of $g(\cdot)$ and find an estimator, $\widehat{\boldsymbol{\Gamma}}$, of $\boldsymbol{\Gamma}_{\widetilde{d}}$ by minimizing a sum of squared errors involving $\widehat{\varepsilon}_t^2$. This step focuses only on the estimation of the variance function. Finally, we use $\widehat{\boldsymbol{\Gamma}}$, the Nadaraya-Watson smoother of $f(\cdot)$ and $g(\cdot)$, respectively, and obtain a revised estimator $\widehat{\boldsymbol{\Phi}}$ of $\boldsymbol{\Phi}_d$ by minimizing a weighted residual sum of squares. All these details are given in the subsequent sections of the article.

In Section 2, we introduce the notations and assumptions that are used throughout the article. In Section 3.1, we define the Nadaraya-Watson (N-W) smoother for the mean and the variance functions, respectively, and state two lemmas concerning the N-W smoother that are needed to prove our main theorems. In Section 3.2, we define an iterative estimation procedure to estimate the parameter matrices, $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ and state the three main theorems of the article. The theorems are proved in Appendix B. While Section 3.3 introduces a new angular representation for parameter matrices to overcome computational challenges, Section 3.4 discusses selection of hyperparameters $p, q, d,$ and, $\widetilde{d}$. To assess the performance of the estimators numerically, simulation studies are carried out in Section 4 and a real data analysis of the Brazilian Real (BRL)/U.S. Dollar Exchange Rate series is carried out in Section 5. Concluding remarks are given in Section 6.

2. Notations and assumptions

As before, let $\{x_t; t \geq 1\}$ denote a univariate time series and $\boldsymbol{X}_{t-1} = (x_{t-1}, ..., x_{t-p})^{\intercal}$ for some $p \geq 1$. We assume that all the information in the conditional mean $E(x_t|\boldsymbol{X}_{t-1})$ is contained in $E\left(x_t|\Phi_1^T\boldsymbol{X}_{t-1}, ..., \Phi_d^T\boldsymbol{X}_{t-1}\right)$ for some $d \in [1, p]$. Specifically, we assume that there exists a $p \times d$ matrix $\boldsymbol{\Phi}_d = (\Phi_1, \ \ldots, \ \Phi_d)$ such that

$$E(x_t|\mathbf{X}_{t-1}) = E(x_t|\boldsymbol{\Phi}_d^{\intercal}\mathbf{X}_{t-1}) \equiv f(\boldsymbol{\Phi}_d^{\intercal}\mathbf{X}_{t-1}), \tag{2.1}$$

where $f(\cdot)$ is an unknown and possibly nonlinear function. We also allow $\{x_t\}$ to be conditionally heteroscedastic. Let $\varepsilon_t = x_t - f(\boldsymbol{\Phi}_d^{\intercal}\mathbf{X}_{t-1})$ and $\boldsymbol{\varepsilon}_{t-1}^2 = (\varepsilon_{t-1}^2, ..., \varepsilon_{t-q}^2)^{\intercal}$ for some $q \geq 1$. We assume that there exists a $q \times \widetilde{d}$ matrix $\boldsymbol{\Gamma}_{\widetilde{d}} = (\Gamma_1, \ \ldots, \ \Gamma_{\widetilde{d}})$ such that

$$V(x_t|\mathbf{X}_{t-1}) = E(\varepsilon_t^2|\boldsymbol{\varepsilon}_{t-1}^2) = E(\varepsilon_t^2|\boldsymbol{\Gamma}_{\widetilde{d}}^{\intercal}\boldsymbol{\varepsilon}_{t-1}^2) \equiv g(\boldsymbol{\Gamma}_{\widetilde{d}}^{\intercal}\boldsymbol{\varepsilon}_{t-1}^2), \tag{2.2}$$

where $g(\cdot)$ is an unknown, possibly nonlinear, function. Our methodological development first assumes that the hyperparameters p, q, d and $\widetilde{d}$ are pre-specified integers that satisfy $d \ll p < n$ and $\widetilde{d} \ll q < n$ for dimension reduction purposes. However, in Section 3.4, we will consider the case of unknown p, q, d and $\widetilde{d}$. In this paper, we do not pursue the research direction to allow these hyperparameters to diverge with n.

The unknown coefficient matrices can be defined through population estimating functions as:

$$\boldsymbol{\Gamma}_{\widetilde{d}} = \underset{\boldsymbol{s}}{\operatorname{argmin}}\ E\left[\left(\varepsilon_t^2 - g(\boldsymbol{s}^{\intercal}\boldsymbol{\varepsilon}_{t-1}^2)\right)^2\right], \tag{2.3}$$

$$\text{and} \quad \boldsymbol{\Phi}_d = \underset{\boldsymbol{r}}{\operatorname{argmin}}\ E\left[\frac{\left(x_t - f(\boldsymbol{r}^{\intercal}\boldsymbol{X}_{t-1})\right)^2}{g(\boldsymbol{\Gamma}_{\widetilde{d}}^{\intercal}\boldsymbol{\varepsilon}_{t-1}^2)}\right]. \tag{2.4}$$

The goal of this article is to estimate the unknown functions $f(\cdot)$, $g(\cdot)$, and the coefficient matrices $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$. Since $f(\cdot)$ and $g(\cdot)$ are unknown, both $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ are not identifiable. To illustrate the non-identifiability, let us consider the following time series model:

$$x_t = 0.5 + (0.5\, x_{t-1} + 0.1\, x_{t-3}) + 0.16\, x_{t-2}^2 + \varepsilon_t,$$
$$\text{and} \quad \varepsilon_t = \sqrt{0.5 + 0.2\, \varepsilon_{t-1}^2 + 0.1\, \varepsilon_{t-2}^2}\, e_t,$$

where $\{e_t\}$ is an i.i.d. sequence with $E(e_t) = 0$ and $0 < E(e_t^2) = \sigma^2 < \infty$. We can set the conditional mean and the variance functions as

$$f_1(z_1, z_2) = 0.5 + z_1 + z_2^2 \quad \text{and} \quad g(z) = 0.5 + z.$$

Then, the parameter matrices are

$$\boldsymbol{\Phi}_{1,d} = \begin{bmatrix} 0.5 & 0 & 0.1 \\ 0 & 0.4 & 0 \end{bmatrix}^{\intercal} \quad \text{and} \quad \boldsymbol{\Gamma}_{\widetilde{d}} = [0.2, 0.1]^{\intercal}.$$

Note that we can also set $f_2(z_1, z_2) = 0.5 + 0.5\ z_1 + z_2^2$ and

$$\boldsymbol{\Phi}_{2,d} = \begin{bmatrix} 1.0 & 0 & 0.2 \\ 0 & 0.4 & 0 \end{bmatrix}^{\intercal},$$

such that $f_1(\boldsymbol{\Phi}_{1,d}^{\intercal}\mathbf{X}_{t-1}) = f_2(\boldsymbol{\Phi}_{2,d}^{\intercal}\mathbf{X}_{t-1})$.

Note that we can always find $\boldsymbol{\Phi}_d$ (or $\boldsymbol{\Gamma}_{\widetilde{d}}$) such that its columns are normalized. Then, although we cannot fully identify the parameter matrix, the space spanned by its columns is identifiable. Besides, we can eliminate the ambiguity by imposing additional identification conditions. In this paper, we set the first element of each column of $\boldsymbol{\Phi}_d$ (or $\boldsymbol{\Gamma}_{\widetilde{d}}$) to be positive. Then, we can sort the columns of $\boldsymbol{\Phi}_d$ in descending order according to the first element of each column. If there are ties, we refer to the second element and so on. Such identification conditions can be achieved by replacing $\boldsymbol{\Phi}_d$ with $\boldsymbol{\Phi}_d\boldsymbol{P}$ where $\boldsymbol{P}$ is a signed column permutation matrix that satisfies $\boldsymbol{P}^{\intercal}\boldsymbol{P} = \boldsymbol{I}_d$. Without loss of generality, we denote the parameter matrices $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ as their identifiable versions in the rest of the paper. We refer to Luo et al. (2014) and Park and Samadi (2020) for more discussions on similar identifiability problems.

Finally, we introduce the following notations that will be used throughout this paper. Given a vector $\boldsymbol{v}$, we denote its vector norm as $\|\boldsymbol{v}\|$. For a matrix $\boldsymbol{A}$, we write its spectral-norm as $\|\boldsymbol{A}\|$. If $\boldsymbol{A}$ is a square matrix, we denote its determinant as $|\boldsymbol{A}|$.

3. Estimation methodology

In order to estimate the parameter matrices $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ defined in (2.3) and (2.4), respectively, we need to first estimate the functions $f(\cdot)$ and $g(\cdot)$ nonparametrically. Next, we define the Nadaraya-Watson smoother for the functions $f(\cdot)$ and $g(\cdot)$.

3.1. Nadaraya-Watson smoother

Suppose we have a random sample of multivariate data where $\boldsymbol{y} = (y_1, \ldots, y_n)^{\intercal}$ is an n-dimensional response vector and $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^{\intercal}$ is a $n \times d$ matrix of explanatory variables satisfying

$$y_i = m(\boldsymbol{z}_i) + e_i \quad \text{for} \quad i = 1, \ldots, n,$$

where $m(\cdot)$ is an unknown regression function. Then, the classical Nadaraya-Watson smoother (Nadaraya, 1964; Watson, 1964) of $m(\cdot)$ is defined as

$$\widehat{m}_{\lambda}(\boldsymbol{z}_k) = \frac{\sum_{i=1}^{n} K\left(\boldsymbol{z}_k - \boldsymbol{z}_i, \boldsymbol{\lambda}_n\right) y_i}{\sum_{i=1}^{n} K\left(\boldsymbol{z}_k - \boldsymbol{z}_i, \boldsymbol{\lambda}_n\right)},$$

where $K(\cdot, \boldsymbol{\lambda}_n)$ is a kernel function and $\boldsymbol{\lambda}_n$ is a d-dimensional bandwidth vector. Here, we follow the suggestions in Silverman (1986), Scott (1992) and Park et al. (2009) and assume a product kernel function defined by

$$K\left(\boldsymbol{z}_k - \boldsymbol{z}_i, \boldsymbol{\lambda}_n\right) = \left(n \prod_{j=1}^{d} \lambda_{nj}\right)^{-1} \prod_{j=1}^{d} G\left(\frac{z_{k,j} - z_{i,j}}{\lambda_{nj}}\right) \quad \text{with} \quad \lambda_{nj} = s_j \left[\frac{4}{(d+2)n}\right]^{1/(4+d)}, \tag{3.1}$$

where $z_{k,j}$ is the j^{th} component of the d-dimensional vector $\boldsymbol{z}_k$, G is a univariate kernel function (e.g. Gaussian kernel), s_j is the sample standard deviation of the j^{th} column of $\boldsymbol{Z}$.

This leads us to estimate the conditional mean function $f(\cdot)$ by

$$\widehat{f}_n(\boldsymbol{r}^{\intercal}\boldsymbol{X}_{t-1}) = \frac{\sum_{i=p+1}^{n} K\left(\boldsymbol{r}^{\intercal}\boldsymbol{X}_{t-1} - \boldsymbol{r}^{\intercal}\boldsymbol{X}_{i-1}, \boldsymbol{a}_n\right) x_i}{\sum_{j=p+1}^{n} K\left(\boldsymbol{r}^{\intercal}\boldsymbol{X}_{t-1} - \boldsymbol{r}^{\intercal}\boldsymbol{X}_{j-1}, \boldsymbol{a}_n\right)}, \tag{3.2}$$

where $K(\cdot)$ is the kernel function defined in (3.1), $\boldsymbol{X}_{t-1} = (x_{t-1}, \ldots, x_{t-p})^{\intercal}$ with $t > p$, $\boldsymbol{r}$ is a $p \times d$ coefficient matrix, and $\boldsymbol{a}_n = (a_{n1}, \ldots, a_{nd})^{\intercal}$ is a d-dimensional bandwidth vector.

Similarly, we propose to estimate the conditional variance function $g(\cdot)$ by

$$\widehat{g}_n(\boldsymbol{s}^{\intercal}\boldsymbol{\varepsilon}_{t-1}^2) = \frac{\sum_{i=p+q+1}^{n} K\left(\boldsymbol{s}^{\intercal}\boldsymbol{\varepsilon}_{t-1}^2 - \boldsymbol{s}^{\intercal}\boldsymbol{\varepsilon}_{i-1}^2, \boldsymbol{b}_n\right) \varepsilon_i^2}{\sum_{j=p+q+1}^{n} K\left(\boldsymbol{s}^{\intercal}\boldsymbol{\varepsilon}_{t-1}^2 - \boldsymbol{s}^{\intercal}\boldsymbol{\varepsilon}_{j-1}^2, \boldsymbol{b}_n\right)}, \tag{3.3}$$

where $K(\cdot)$ is the kernel function defined in (3.1), $\boldsymbol{\varepsilon}_{t-1}^2 = (\varepsilon_{t-1}^2, \ldots, \varepsilon_{t-q}^2)^{\intercal}$ with $t > p + q$, $\boldsymbol{s}$ is a $q \times \widetilde{d}$ coefficient matrix, and $\boldsymbol{b}_n = (b_{n1}, \ldots, b_{n\widetilde{d}})^{\intercal}$ is a $\widetilde{d}$-dimensional bandwidth vector.

Let $f^{(1)}(\boldsymbol{r}^{\intercal}\boldsymbol{X}_{t-1})$ and $g^{(1)}(\boldsymbol{s}^{\intercal}\boldsymbol{\varepsilon}_{t-1}^2)$ denote the first order derivatives of $f(\boldsymbol{r}^{\intercal}\boldsymbol{X}_{t-1})$ and $g(\boldsymbol{s}^{\intercal}\boldsymbol{\varepsilon}_{t-1}^2)$, respectively. We then define the estimators of $f^{(1)}(\boldsymbol{r}^{\intercal}\boldsymbol{X}_{t-1})$ and $g^{(1)}(\boldsymbol{s}^{\intercal}\boldsymbol{\varepsilon}_{t-1}^2)$, respectively, as

$$\widehat{f}_n^{(1)}(\boldsymbol{r}^\top \boldsymbol{X}_{t-1}) = \frac{\sum_{i=p+1}^{n} K^{(1)}\left(\boldsymbol{r}^\top \boldsymbol{X}_{t-1} - \boldsymbol{r}^\top \boldsymbol{X}_{i-1}, \boldsymbol{a}_n\right) x_i}{\sum_{j=p+1}^{n} K\left(\boldsymbol{r}^\top \boldsymbol{X}_{t-1} - \boldsymbol{r}^\top \boldsymbol{X}_{j-1}, \boldsymbol{a}_n\right)}, \tag{3.4}$$

$$\text{and} \quad \widehat{g}_n^{(1)}(\boldsymbol{s}^\top \boldsymbol{e}_{t-1}^2) = \frac{\sum_{i=p+q+1}^{n} K^{(1)}\left(\boldsymbol{s}^\top \boldsymbol{e}_{t-1}^2 - \boldsymbol{s}^\top \boldsymbol{e}_{i-1}^2, \boldsymbol{b}_n\right) \varepsilon_i^2}{\sum_{j=p+q+1}^{n} K\left(\boldsymbol{s}^\top \boldsymbol{e}_{t-1}^2 - \boldsymbol{s}^\top \boldsymbol{e}_{j-1}^2, \boldsymbol{b}_n\right)}, \tag{3.5}$$

where $K^{(1)}(\cdot)$ is the first order derivative of the kernel function $K(\cdot)$ defined in (3.1). It should be mentioned that $\widehat{f}_n^{(1)}$ and $\widehat{g}_n^{(1)}$ are not part of the iterative estimation procedure described in Section 3.2. However, these estimators do play a key role in establishing the asymptotic properties of the estimators of the parameter matrices $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ defined in (2.3) and (2.4), respectively.

In the following two lemmas, we state some large sample properties of the aforementioned Nadaraya-Watson estimator and the first order derivative estimators. These are also of independent interest. The assumptions needed for the Lemmas along with some justification are stated in Appendix A.

Lemma 3.1. *Suppose that the assumptions A1 to A5 stated in Appendix A hold. Then, the Nadaraya-Watson estimators defined in (3.2) and (3.3) satisfy*

$$\sup_{\boldsymbol{r}^\top \boldsymbol{X} \in \mathbb{R}^d} |\widehat{f}_n(\boldsymbol{r}^\top \boldsymbol{X}) - f(\boldsymbol{r}^\top \boldsymbol{X})| = O_p\left(\left[\frac{\ln(n-p)}{n-p}\right]^{2/(d+4)}\right),$$

$$\textit{and} \quad \sup_{\boldsymbol{s}^\top \boldsymbol{e}^2 \in \mathbb{R}^{\widetilde{d}}} |\widehat{g}_n(\boldsymbol{s}^\top \boldsymbol{e}^2) - g(\boldsymbol{s}^\top \boldsymbol{e}^2)| = O_p\left(\left[\frac{\ln(n-p-q)}{n-p-q}\right]^{2/(\widetilde{d}+4)}\right),$$

as $n \to \infty$.

Remark 3.1. Lemma 3.1 establishes the rate of uniform convergence in probability for the Nadaraya-Watson estimators proposed in (3.2) and (3.3). When we set p and q to be fixed and choose $d = \widetilde{d} = 1$, the rates in Lemma 3.1 follow $(n^{-1}\ln n)^{2/5}$ which is the optimal nonparametric rate for i.i.d. univariate data proved in Stone (1982). The proof of Lemma 3.1 is similar to the proof of Theorem 8 in Hansen (2008), and hence we omit the proof.

Lemma 3.2. *Suppose that the assumptions A1 to A5 stated in Appendix A hold. Then, the first order derivative estimators defined in* (3.4) *and* (3.5) *satisfy*

$$\sup_{\boldsymbol{r}^\top \boldsymbol{X} \in \mathbb{R}^d} |\widehat{f}_n^{(1)}(\boldsymbol{r}^\top \boldsymbol{X}) - f^{(1)}(\boldsymbol{r}^\top \boldsymbol{X})| \to 0,$$

$$\textit{and} \quad \sup_{\boldsymbol{s}^\top \boldsymbol{e}^2 \in \mathbb{R}^{\widetilde{d}}} |\widehat{g}_n^{(1)}(\boldsymbol{s}^\top \boldsymbol{e}^2) - g^{(1)}(\boldsymbol{s}^\top \boldsymbol{e}^2)| \to 0,$$

with probability 1 as $n \to \infty$.

Remark 3.2. Lemma 3.2 provides uniform consistency results for the first order derivative estimators defined in (3.4) and (3.5). The proof of Lemma 3.2 directly follows from the proof of Theorem 2 in Mack and Müller (1989), and hence we omit the proof.

3.2. An iterative estimation procedure

Since the population estimating functions (2.3) and (2.4) are interdependent, we propose an iterative estimation procedure to construct an estimator $\widehat{\boldsymbol{\Phi}}_{d,n}$ of $\boldsymbol{\Phi}_d$ and an estimator $\widehat{\boldsymbol{\Gamma}}_{\widetilde{d},n}$ of $\boldsymbol{\Gamma}_{\widetilde{d}}$. For ease of presentation, we will suppress the subscripts and denote $\widehat{\boldsymbol{\Phi}} = \widehat{\boldsymbol{\Phi}}_{d,n}$ and $\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\Gamma}}_{\widetilde{d},n}$. The estimation procedure contains the following three key steps.

Step 1: Initial estimation.

In this step, disregard the conditional heteroscedasticity for the moment and denote $\boldsymbol{\Phi}_0$ as an approximation of $\boldsymbol{\Phi}_d$ defined by

$$\boldsymbol{\Phi}_0 = \underset{\boldsymbol{r} \in \mathbb{R}^{p \times d}}{\operatorname{argmin}} \; E\left[\left(x_t - f(\boldsymbol{r}^\top \boldsymbol{X}_{t-1})\right)^2\right].$$

Then, we consider a residual sum of squares which is the sample version of $E\left[\left(x_t - f(\boldsymbol{r}^\top \boldsymbol{X}_{t-1})\right)^2\right]$ and define an initial estimator of $\boldsymbol{\Phi}_d$ as

$$\widehat{\boldsymbol{\Phi}}_0 = \underset{\boldsymbol{r}\in\mathbb{R}^{p\times d}}{\operatorname{argmin}} \sum_{t=p+1}^{n} \left(x_t - \widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1})\right)^2, \quad \text{such that} \quad \boldsymbol{r}^\intercal \boldsymbol{r} = \boldsymbol{I}_d, \tag{3.6}$$

where $\widehat{f}_n(\cdot)$ is the Nadaraya-Watson estimator defined in (3.2) and $\boldsymbol{r}^\intercal \boldsymbol{r} = \boldsymbol{I}_d$ is an orthonormal constraint. Notice that the solution of (3.6) is not unique as one can change the sign of each column of $\widehat{\boldsymbol{\Phi}}_0$ and permute the columns. Therefore, we follow the identification conditions discussed in Section 2 to eliminate the ambiguity. We follow similar identification conditions for the other two estimators to be defined in the subsequent steps. Empirically, the choice of identification conditions does not affect the performance of the proposed estimation procedure. We will corroborate this statement via our numerical results.

Step 2: Estimation of $\boldsymbol{\Gamma}_{\widetilde{d}}$.

Next, we estimate the parameter matrix $\boldsymbol{\Gamma}_{\widetilde{d}}$. For this, we use the initial estimator $\widehat{\boldsymbol{\Phi}}_0$ of $\boldsymbol{\Phi}_d$ to compute the residuals

$$\widehat{\varepsilon}_t = x_t - \widehat{f}_n(\widehat{\boldsymbol{\Phi}}_0^\intercal \boldsymbol{X}_{t-1}), \quad t = p+q+1, \ \ldots, \ n. \tag{3.7}$$

Then, we consider a residual sum of squares which is the sample version of the population estimating function in (2.3) and define an estimator of $\boldsymbol{\Gamma}_{\widetilde{d}}$ as

$$\widehat{\boldsymbol{\Gamma}} = \underset{\boldsymbol{s}\in\mathbb{R}_+^{q\times\widetilde{d}}}{\operatorname{argmin}} \sum_{t=p+q+1}^{n} \left(\widehat{\varepsilon}_t^2 - \widehat{g}_n(\boldsymbol{s}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right)^2, \quad \text{such that} \quad \boldsymbol{s}^\intercal \boldsymbol{s} = \boldsymbol{I}_{\widetilde{d}}, \tag{3.8}$$

where $\widehat{g}_n(\cdot)$ is the Nadaraya-Watson estimator defined in (3.3). Denote s_{ij} the (i,j)-th entry of $\boldsymbol{s}$. The constraint $\boldsymbol{s} \in \mathbb{R}_+^{q\times\widetilde{d}}$ requires all elements of $\boldsymbol{s}$ to be non-negative, i.e. $s_{ij} \geq 0$ for $i \in [1,q]$ and $j \in [1,\widetilde{d}]$.

Step 3: Estimation of $\boldsymbol{\Phi}_d$.

In the third step, we take the conditional heteroscedasticity into account and define a revised estimator of $\boldsymbol{\Phi}_d$. More specifically, we minimize a weighted residual sum of squares which is the sample version of the population estimating function in (2.4) and define a revised estimator $\widehat{\boldsymbol{\Phi}}$ of $\boldsymbol{\Phi}_d$ as:

$$\widehat{\boldsymbol{\Phi}} = \underset{\boldsymbol{r}\in\mathbb{R}^{p\times d}}{\operatorname{argmin}} \sum_{t=p+q+1}^{n} \frac{\left(x_t - \widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1})\right)^2}{\widehat{g}_n(\widehat{\boldsymbol{\Gamma}}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)}, \quad \text{such that} \quad \boldsymbol{r}^\intercal \boldsymbol{r} = \boldsymbol{I}_d, \tag{3.9}$$

where $\widehat{\boldsymbol{\Gamma}}$ is the estimator of $\boldsymbol{\Gamma}_{\widetilde{d}}$ as defined in (3.8).

Additionally, we can replace $\widehat{\boldsymbol{\Phi}}_0$ in (3.7) with the updated estimator $\widehat{\boldsymbol{\Phi}}$ in (3.9) and obtain a new set of fitted residuals. This then yields a revised estimate of $\boldsymbol{\Gamma}_{\widetilde{d}}$ via (3.8), which in turn yields a revised estimate of $\boldsymbol{\Phi}_d$ via (3.9). We can then iterate steps 2 and 3 in the estimation procedure until convergence. This iterative process is summarized in Algorithm 1.

Algorithm 1 Iterative Estimation Procedure.

Input: A univariate time series $\{x_t; t \geq 1\}$, a dissimilarity metric $d(\cdot,\cdot)$, a tolerance parameter ϵ, and a number of maximum iterations M.
Initialization: Compute the initial estimate $\widehat{\boldsymbol{\Phi}}_0$ according to equation (3.6). Compute the residuals $\widehat{\varepsilon}_t$ according to equation (3.7). Compute $\widehat{\boldsymbol{\Gamma}}$ from the fitted residuals, according to equation (3.8). Compute $\widehat{\boldsymbol{\Phi}}$, using $\widehat{\boldsymbol{\Gamma}}$ to generate weights, according to equation (3.9).
for $m = 1, \ \ldots, \ M$ **do**
  Let $\widehat{\boldsymbol{\Phi}}_{OLD} = \widehat{\boldsymbol{\Phi}}$.
  Recalculate the residuals based on $\widehat{\boldsymbol{\Phi}}_{OLD}$.
  Compute $\widehat{\boldsymbol{\Gamma}}$ from the new fitted residuals, according to equation (3.8).
  Re-estimate $\widehat{\boldsymbol{\Phi}}$, using the new $\widehat{\boldsymbol{\Gamma}}$ to generate weights, according to equation (3.9).
  if $d(\widehat{\boldsymbol{\Phi}}_{OLD}, \widehat{\boldsymbol{\Phi}}) < \epsilon$ **then**
    break
  end if
end for
Output: $\widehat{\boldsymbol{\Phi}}$ and $\widehat{\boldsymbol{\Gamma}}$.

Theorem 3.1 *(Initial estimator). Suppose that the assumptions A1 to A5 stated in Appendix A hold. Then the initial estimator defined in (3.6) satisfies*

$$\|\widehat{\boldsymbol{\Phi}}_0 - \boldsymbol{\Phi}_0\| = O_p\left(\left[\frac{\ln(n-p)}{n-p}\right]^{2/(d+4)}\right), \quad \text{as } n \to \infty. \tag{3.10}$$

Theorem 3.2 *(Variance estimator). Suppose that assumptions A1–A5 hold. The variance parameter vector estimator defined in (3.8) satisfies*

$$\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_{\widetilde{d}}\| = O_p\left(\left[\frac{\ln(n-p-q)}{n-p-q}\right]^{2/(\widetilde{d}+4)}\right), \quad \text{as } n \to \infty. \tag{3.11}$$

Theorem 3.3 *(Final estimator). Suppose that assumptions A1–A5 hold. The final estimator defined in (3.9) satisfies*

$$\|\widehat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_d\| = O_p\left(\left[\frac{\ln(n-p)}{n-p}\right]^{2/(d+4)}\right), \quad \text{as } n \to \infty. \tag{3.12}$$

Remark 3.3. Theorems 3.1–3.3 establish the rate of convergence in probability for the estimators proposed in Section 3.2. When we set p and q to be fixed and choose $d = \widetilde{d} = 1$, the rate in Theorems 3.1–3.3 is $(n^{-1}\ln n)^{2/5}$ which is the optimal nonparametric rate for i.i.d. univariate data proved in Stone (1982). Our results are obtained for a multivariate and strong-mixing data. The proofs of the above theorems are presented in Appendix B.

3.3. Angular representation for parameter matrices

The optimization problems in (3.6), (3.8) and (3.9) involve a matrix orthonormal constraint. Existing literature (e.g. Edelman et al., 1998; Bai et al., 2000; Wen and Yin, 2013; Sato et al., 2019) propose to solve such problems by exploiting Newton and conjugate gradient algorithms on Stiefel and Grassmann manifolds. However, these differential geometry optimization methods pose computational challenges to the proposed estimation procedure as the derivatives of estimating functions may admit complicated forms. Recently, Park and Samadi (2020) suggested to tackle the matrix orthonormal constraint by a sequential quadratic programming algorithm. Unfortunately, this algorithm is numerically unstable when the parameter matrix of interest has more than one column.

To address the aforementioned computational challenges, we propose a novel angular representation for the parameter matrix which converts the matrix orthonormal constraint into a sequence of linear systems. The proposed angular representation approach avoids calculating the derivatives of estimating functions and improves the numerical stability. We illustrate this approach as follows. Let $d \le p$ and $\boldsymbol{\Theta}$ be a $p \times d$ orthonormal parameter matrix defined as

$$\boldsymbol{\Theta} = (\Theta_1, \ \ldots, \ \Theta_d) = \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1d} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{p1} & \theta_{p2} & \cdots & \theta_{pd} \end{bmatrix},$$

such that $\boldsymbol{\Theta}^\intercal\boldsymbol{\Theta} = \mathbf{I}_d$. Since $\boldsymbol{\Theta}$ is full rank, each column of $\boldsymbol{\Theta}$ has at least one unique nonzero entry. In the following discussion, we assume that $\theta_{k_i i} > 0$ for some $k_i \in \{1, \ \ldots, \ p\}$ and $i = 1, \ \ldots, \ d$. We can always guarantee this assumption by column scaling which does not affect our estimation procedure.

Then, for the (j, i)-th entry of $\boldsymbol{\Theta}$, we can define an angle α_{ji} as

$$\tan(\alpha_{ji}) = \frac{\theta_{ji}}{\theta_{k_i i}} \quad \text{or} \quad \theta_{ji} = \theta_{k_i i}\tan(\alpha_{ji}), \quad \text{for } j = 1, \ \ldots, \ p \text{ and } i = 1, \ \ldots, \ d. \tag{3.13}$$

Since all the columns of $\boldsymbol{\Theta}$ are normalized, we have

$$1 = \sum_{j=1}^{p}\theta_{ji}^2 = \theta_{k_i i}^2 + \sum_{j\neq k_i}\left\{[\theta_{k_i i}\tan(\alpha_{ji})]^2\right\} = \theta_{k_i i}^2\sum_{j=1}^{p}\tan(\alpha_{ji})^2, \quad \text{for } i = 1, \ \ldots, \ d,$$

where the last equation uses the fact that $\tan(\alpha_{k_i i}) = 1$. Denote $C_i = \left[\sum_{j=1}^{p}\left\{\tan(\alpha_{ji})^2\right\}\right]^{1/2}$, we can rewrite (3.13) as

$$\theta_{ji} = \frac{\tan(\alpha_{ji})}{C_i}. \tag{3.14}$$

The parameter matrix $\boldsymbol{\Theta}$ is column-wise orthogonal if and only if $\Theta_i^\intercal\Theta_{i'} = 0$ for $1 \le i' < i \le d$. This is equivalent to require

$$\sum_{j=1}^{p}\theta_{ji}\theta_{ji'} = 0 \quad \text{or} \quad \theta_{k_{i'}i} = \frac{-\sum_{j\neq k_{i'}}\theta_{ji}\theta_{ji'}}{\theta_{k_{i'}i'}}.$$

This together with (3.14) and $\tan(\alpha_{k_{i'}i'}) = 1$ imply that

$$\tan(\alpha_{k_{i'}i}) = -\sum_{j\neq k_{i'}}\tan(\alpha_{ji})\tan(\alpha_{ji'}) \quad \text{and} \quad \sum_{j=1}^{p}\tan(\alpha_{ji})\tan(\alpha_{ji'}) = 0. \tag{3.15}$$

Similar to (3.15), we can represent the orthogonal condition for the i-th column of $\boldsymbol{\Theta}$ as a system of $(i-1)$ linear equations

Table 1
Sample statistics of S over 100 replications.

S	Mean	SD	Min	Median	Max
Original	1.103	0.036	1.015	1.099	1.191
Angular	0.000	0.000	0.000	0.000	0.000

Sample statistics of closeness between $\widehat{\boldsymbol{\Phi}}^{\mathsf{T}}\widehat{\boldsymbol{\Phi}}$ and $\boldsymbol{I}$ over 100 replications. Mean and SD stand for sample mean and sample standard deviation. Min, Median, and Max stand for minimum, median and maximum values in the sample.

$$\begin{cases} \sum_{j=k_1,\ldots,k_{i-1}} \tan(\alpha_{ji})\tan(\alpha_{j1}) & = -\sum_{j\neq k_1,\ldots,k_{i-1}} \tan(\alpha_{ji})\tan(\alpha_{j1}), \\ \qquad\qquad\vdots & \\ \sum_{j=k_1,\ldots,k_{i-1}} \tan(\alpha_{ji})\tan(\alpha_{j(i-1)}) & = -\sum_{j\neq k_1,\ldots,k_{i-1}} \tan(\alpha_{ji})\tan(\alpha_{j(i-1)}). \end{cases} \tag{3.16}$$

By sequentially solving the linear systems defined in (3.16) for $i = 2, \ldots, d$, we represent the $p \times d$ orthonormal matrix $\boldsymbol{\Theta}$ by $dp - \frac{d(d+1)}{2}$ angles. This angular representation approach addresses the long pending numerical stability issue in Park et al. (2009), Park et al. (2010), Park (2011), Park and Sriram (2017) and Park and Samadi (2020). Further, the proposed angular representation may be of independent interest as it is applicable to other optimization problems with an orthonormal matrix constraint or can be used to sample orthonormal random matrices.

We end this section with a simulation study to demonstrate how the angular representation incorporates the orthonormal constraint in optimization and improves numerical stability. For this study, we consider the following nonlinear regression model: Suppose

$$y_i = \left((1/\sqrt{5})(x_{1,i} + 2x_{4,i})\right)^2 + (1/\sqrt{13}(-2x_{1,i} + 2x_{2,i} - 2x_{3,i} + x_{4,i}) + \varepsilon_i,\ i = 1,\ \ldots,\ n.$$

We generate the errors, $\{\varepsilon_i\}$ from i.i.d. $N(0, 1)$, but the covariates, $\{x_{1,i}, \ldots, x_{4,i}\}$ are generated from i.i.d. Uniform$(0, 1)$. We define the parameter matrix associated with this model as

$$\boldsymbol{\Phi} = \begin{bmatrix} \frac{1}{\sqrt{5}} & 0 & 0 & \frac{2}{\sqrt{5}} \\ \frac{-2}{\sqrt{13}} & \frac{2}{\sqrt{13}} & \frac{2}{\sqrt{13}} & \frac{1}{\sqrt{13}} \end{bmatrix}^{\mathsf{T}}.$$

In this simulation study, we let the sample size $n = 1000$ and replicate the simulation 100 times. In each replication, we aim to estimate $\boldsymbol{\Phi}$ subject to an orthonormal constraint $\widehat{\boldsymbol{\Phi}}^{\mathsf{T}}\widehat{\boldsymbol{\Phi}} = \boldsymbol{I}$. First, we estimate $\boldsymbol{\Phi}$ by a popular nonlinear optimization algorithm named Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm (Byrd et al., 1995). Then, we estimate $\boldsymbol{\Phi}$ by L-BFGS plus the angular representation method proposed above. We denote the first method as Original and the second method as Angular. For both methods, the columns of the estimated coefficient matrix are normalized to have a unit length. Then, we measure the closeness between $\widehat{\boldsymbol{\Phi}}^{\mathsf{T}}\widehat{\boldsymbol{\Phi}}$ and the identity matrix by computing

$$S = \sum_{i=1}^{d}\sum_{j=1}^{d} \left| \left[\widehat{\boldsymbol{\Phi}}^{\mathsf{T}}\widehat{\boldsymbol{\Phi}}\right]_{i,j} - \boldsymbol{I}_{i,j} \right|,$$

where $\boldsymbol{A}_{i,j}$ is the (i, j)-th entry of a matrix $\boldsymbol{A}$.

The numerical results measuring the closeness between $\widehat{\boldsymbol{\Phi}}^{\mathsf{T}}\widehat{\boldsymbol{\Phi}}$ and $\boldsymbol{I}$ over 100 replications along with some summary statistics are presented in Table 1. It is clear from the values in Table 1 that the estimators obtained by the Original method consistently violate the orthonormal constraint, even when $d = 2$. In contrast, the estimators provided by Angular perfectly satisfy the orthonormal constraint.

3.4. Selection of hyperparameters

The estimating functions (2.3) and (2.4) involve four hyperparameters, $p, d, q,$ and $\widetilde{d}$. The lag parameters p and q indicate how far back we should look in the history of x_t and ε_t, while the dimension parameters d and $\widetilde{d}$ specify how many linear combinations of the past variables we need. The selection of these hyperparameters plays an important role in modeling the time series of interest. However, for real data, we rarely have accurate prior information to guide us on the selection of these parameters. To this end, we introduce a data-driven method to select the hyperparameters.

Denote the objective functions in (3.6) and (3.8) as

$$S_{n,0}(\boldsymbol{r}) = \sum_{t=p+1}^{n} \left(x_t - \widehat{f}_n(\boldsymbol{r}^{\mathsf{T}}\boldsymbol{X}_{t-1})\right)^2 \quad \text{and} \quad G_n(\boldsymbol{s}) = \sum_{t=p+q+1}^{n} \left(\widehat{\varepsilon}_t^2 - \widehat{g}_n(\boldsymbol{s}^{\mathsf{T}}\widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right)^2,$$

where $\widehat{f}_n(\cdot)$ and $\widehat{g}_n(\cdot)$ are the Nadaraya-Watson estimators defined in (3.2) and (3.3). We follow the idea in Park et al. (2009) and select the pair (p, d) as the minimizer of a Modified Schwarz Bayesian information Criterion (MSBC) defined by

$$(\widehat{p}, \widehat{d}) = \underset{p,d \in \mathbb{Z}_+^*}{\operatorname{argmin}} \left\{ (n-p) \ln \left[S_{n,0}(\widehat{\boldsymbol{\Phi}}_{p,d})/(n-p) \right] \right\} + d^2 p \ln (n-p), \tag{3.17}$$

where $\mathbb{Z}_+^*$ is the set of all positive integers and $\widehat{\boldsymbol{\Phi}}_{p,d}$ is the estimator of $\boldsymbol{\Phi}_d$ for a given p. Similarly, we select the pair $(q, \widetilde{d})$ as the minimizer of another modified MSBC defined as follows

$$(\widehat{q}, \widehat{\widetilde{d}}) = \underset{q,\widetilde{d} \in \mathbb{Z}_+^*}{\operatorname{argmin}} \left\{ (n-p-q) \ln \left[G_n(\widehat{\boldsymbol{\Gamma}}_{q,\widetilde{d}})/(n-p-q) \right] \right\} + \widetilde{d}^2 q \ln (n-p-q), \tag{3.18}$$

where $\widehat{\boldsymbol{\Gamma}}_{q,\widetilde{d}}$ is the estimator of $\boldsymbol{\Gamma}_{\widetilde{d}}$ for a given q.

In simulation studies, however, since we will know the true values of the hyperparameters in the assumed model, while determining an estimate of a hyperparameter, say p, we will fix the remaining hyperparameters at their true value and determine $\widehat{p}$ by minimizing the quantity in (3.17) over p for a fixed d.

4. Simulation studies

In this section, we present three simulation studies to investigate the finite sample performance of the proposed estimators. The models we consider for the simulations are of the type given by

$$\begin{aligned} & x_t = f(\boldsymbol{\Phi}_{\boldsymbol{d}}^{\intercal} \boldsymbol{X}_{t-1}) + \varepsilon_t, \\ \text{and} \quad & \varepsilon_t = \left[\sqrt{g(\boldsymbol{\Gamma}_{\widetilde{d}}^{\intercal} \boldsymbol{\varepsilon}_{t-1}^2)} \right] e_t, \end{aligned}$$

where we consider different specifications of the functions $f(\cdot)$ and $g(\cdot)$, the $p \times d$ parameter matrix $\boldsymbol{\Phi}_d$, and the $q \times \widetilde{d}$ parameter matrix $\boldsymbol{\Gamma}_{\widetilde{d}}$. In the first two simulation models, the errors $\{e_t\}$ are assumed to be independent and normally distributed with a constant variance. Whereas, in the third simulation model, we consider normal errors and gross-error contaminated normal errors.

The optimizations are done based on the parametrization method proposed in Section 3.3, which overcomes some computational challenges and improves the numerical stability. Besides, we apply the *fmincon* function in *Matlab* to estimate $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$. To search the parameter space thoroughly and avoid the local minimums, we initialize the procedure over 50 randomly generated initial values.

In our simulation studies, we use two measures to assess the accuracy of our estimates. The first measure is defined as the sample mean of the *Vector Correlation Coefficient* (Hotelling, 1936; Ye and Weiss, 2003) over M Monte Carlo simulations:

$$\rho(\widehat{\boldsymbol{\Theta}}) = \frac{1}{M} \sum_{m=1}^{M} |\widehat{\boldsymbol{\Theta}}^{\intercal} \boldsymbol{\Theta} \boldsymbol{\Theta}^{\intercal} \widehat{\boldsymbol{\Theta}}|^{1/2}, \tag{4.1}$$

where $\boldsymbol{\Theta}$ is a parameter matrix and $\widehat{\boldsymbol{\Theta}}$ is an estimate of $\boldsymbol{\Theta}$. Note that $0 \le \rho(\widehat{\boldsymbol{\Theta}}) \le 1$, where the higher values of $\rho(\widehat{\boldsymbol{\Theta}})$ imply that the estimated values are closer to the truth. The second measure is defined as the sample mean of the criterion introduced in Xia et al. (2002b) over M Monte Carlo simulations:

$$m^2(\widehat{\boldsymbol{\Theta}_i}) = \frac{1}{M} \sum_{m=1}^{M} \left|\left| (\boldsymbol{I} - \boldsymbol{\Theta} \boldsymbol{\Theta}^{\intercal}) \widehat{\boldsymbol{\Theta}_i} \right|\right|^2, \tag{4.2}$$

where $\widehat{\boldsymbol{\Theta}_i}$ is an estimate for the i^{th} column of $\boldsymbol{\Theta}$ and $\boldsymbol{I}$ is an identity matrix. Here, $0 \le m^2(\widehat{\boldsymbol{\Theta}_i}) \le 1$, and this measure approaches zero when the estimated column is close to the truth. Throughout this section, we consider $\boldsymbol{\Theta} = \boldsymbol{\Phi}_d$ or $\boldsymbol{\Gamma}_{\widetilde{d}}$ and $\widehat{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Phi}}$ or $\widehat{\boldsymbol{\Gamma}}$.

In our simulations studies, we choose sample sizes $n = 100, 300$ and 1000. For each sample size, we compute the estimates using the iterative estimation procedure and evaluate the accuracy measures $\rho(\widehat{\boldsymbol{\Theta}})$ and $m^2(\widehat{\boldsymbol{\Theta}_i})$ over $M = 100$ Monte Carlo simulations. The hyperparameters are selected by MSBC as defined in Section 3.4. Recall that, when selecting one hyperparameter, we fix all the others to avoid the interactive effects.

Model 1: Consider the following conditionally heteroscedastic time series model with a nonlinear mean function and a linear variance function:

$$\begin{aligned} & x_t = 3 - (1/\sqrt{3})(x_{t-1} + x_{t-3} + x_{t-6}) + cos\{(\pi/2)(1/\sqrt{2})(x_{t-2} - x_{t-4})\} + \varepsilon_t, \\ \text{with} \quad & \varepsilon_t = \left[\sqrt{1 + (.1)(\varepsilon_{t-1}^2 + \varepsilon_{t-2}^2)} \right] e_t \quad \text{and} \quad e_t \sim N(0,1). \end{aligned}$$

Table 2
Estimation results for Model 1 over 100 replications.

n	$\rho(\widehat{\boldsymbol{\Phi}})$	$m^2(\widehat{\boldsymbol{\Phi}}_1)$	$m^2(\widehat{\boldsymbol{\Phi}}_2)$	$\rho(\widehat{\boldsymbol{\Gamma}})$	$m^2(\widehat{\boldsymbol{\Gamma}})$
100	0.354	0.291	0.847	0.838	0.483
300	0.450	0.165	0.736	0.856	0.453
1,000	0.695	0.088	0.297	0.907	0.351

The estimation accuracy measures $\rho(\cdot)$ (the larger the better) and $m^2(\cdot)$ (the smaller the better) are defined in (4.1) and (4.2), respectively. Besides, $\widehat{\boldsymbol{\Phi}}_1$ and $\widehat{\boldsymbol{\Phi}}_2$ are the first and the second columns of $\widehat{\boldsymbol{\Phi}}$.

Table 3
Hyperparameter selection for Model 1 over 100 replications.

n	Count	p	q	d	$\widetilde{d}$
100	Under	19	57	98	0
	Correct	79	1	0	99
	Over	2	42	2	1
300	Under	0	48	87	0
	Correct	90	3	0	92
	Over	10	29	13	8
1,000	Under	0	25	0	0
	Correct	91	33	1	55
	Over	9	42	99	45

The rows Under, Correct and Over report the number of times that the selected hyperparameters are smaller than, equal to, or larger than the truth over 100 Monte Carlo simulations.

Further, we set the parameter matrix and vector of interest as

$$\boldsymbol{\Phi}_d = \begin{bmatrix} \frac{1}{\sqrt{3}} & 0 & \frac{1}{\sqrt{3}} & 0 & 0 & \frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{2}} & 0 & \frac{-1}{\sqrt{2}} & 0 & 0 \end{bmatrix}^{\mathsf{T}} \quad \text{and} \quad \boldsymbol{\Gamma}_{\widetilde{d}} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^{\mathsf{T}}.$$

In Model 1, the true lag values are $p = 6$ and $q = 2$, and the number of dimensions are $d = 2$ and $\widetilde{d} = 1$, respectively. The first column of $\boldsymbol{\Phi}_d$ corresponds to a linear component of the mean function, i.e. $E(x_t|\boldsymbol{X}_{t-1})$, and the second column of $\boldsymbol{\Phi}_d$ corresponds to a nonlinear component of the mean function.

We compute the estimators of $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ based on the iterative estimation procedure introduced in Section 3.2. Denote $\widehat{\boldsymbol{\Phi}}_1$ and $\widehat{\boldsymbol{\Phi}}_2$ as the first and the second columns of $\widehat{\boldsymbol{\Phi}}$, respectively. The results in Table 2 report the accuracy measures ρ and m^2 based on 100 Monte Carlo simulations. According to Table 2, the ρ values increase as the sample size increases. Also, the ρ values are relatively higher for the estimation of $\boldsymbol{\Gamma}_{\widetilde{d}}$ than those for $\boldsymbol{\Phi}_d$. We also observe that the m^2 values are smaller for estimating the first column of $\boldsymbol{\Phi}_d$ than that for the second column of $\boldsymbol{\Phi}_d$. This suggests that the linear component of the mean function is better estimated than the nonlinear component, which in turn may have affected the ρ values for the estimation of $\boldsymbol{\Phi}_d$.

In Table 3, we report the number of times the true lags p and q and the true dimensions d and $\widetilde{d}$ are selected by MSBC defined in Section 3.4. As shown in Table 3, the MSBC method accurately estimates p and $\widetilde{d}$ in a large proportion of replications. However, the estimation of q and d are not ideal.

In Fig. 1, we create a 3-D scatter plot with an overlay surface to visualize the conditional mean function, i.e. $f(\cdot)$ as defined in (3.2). This overlay plot provides a useful tool in practice to visually identify the functional relationship between $\boldsymbol{X}_t$ and the two predictors, $\widehat{\boldsymbol{\Phi}}_1^{\mathsf{T}}\boldsymbol{X}_{t-1}$ and $\widehat{\boldsymbol{\Phi}}_2^{\mathsf{T}}\boldsymbol{X}_{t-1}$. In this example, the fitted surface clearly shows a linear relationship between $\boldsymbol{X}_t$ and $\widehat{\boldsymbol{\Phi}}_1^{\mathsf{T}}\boldsymbol{X}_{t-1}$ and a wave-shaped relationship between $\boldsymbol{X}_t$ and $\widehat{\boldsymbol{\Phi}}_2^{\mathsf{T}}\boldsymbol{X}_{t-1}$, which are in line with the model setup.

Model 2: Consider the following conditionally heteroscedastic time series model. The mean function admits a complicated nonlinear form but the linear variance function is set to be the same as in Model 1.

$$\begin{aligned} x_t = &-1 + (.4/\sqrt{5})(x_{t-1} + 2x_{t-4}) - cos\{(\pi/2)(1/\sqrt{5})(x_{t-3} + 2x_{t-5})\} + \\ &exp\{-[(1/\sqrt{15})(-2x_{t-1} + 2x_{t-2} - 2x_{t-3} + x_{t-4} - x_{t-5} + x_{t-6})]^2\} + \varepsilon_t, \end{aligned}$$

$$\text{with} \quad \varepsilon_t = \left[\sqrt{1 + (.1)(\varepsilon_{t-1}^2 + \varepsilon_{t-2}^2)}\right] e_t \quad \text{and} \quad e_t \sim N(0,1).$$

Further, we set the parameter matrix and vector of interest as

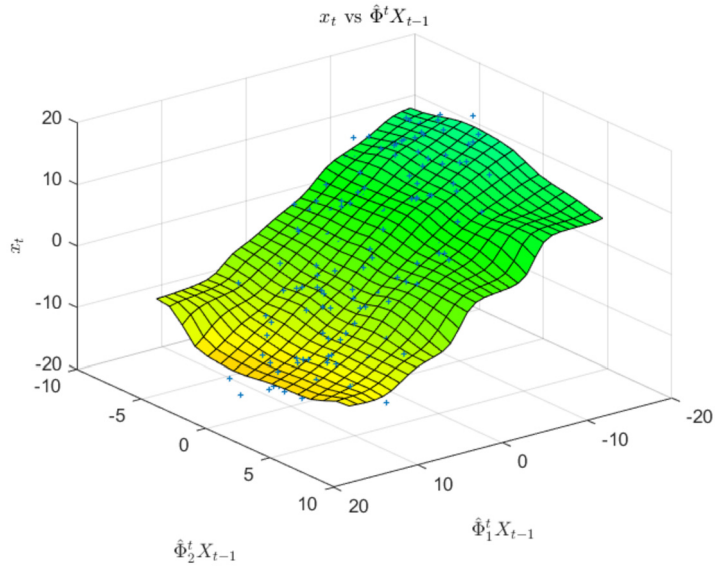

Fig. 1. 3-D scatter plot with an overlay surface of the fitted conditional mean function for Model 1.

Table 4
Estimation results for Model 2 over 100 replications.

n	$\rho(\widehat{\boldsymbol{\Phi}})$	$m^2(\widehat{\boldsymbol{\Phi}}_1)$	$m^2(\widehat{\boldsymbol{\Phi}}_2)$	$m^2(\widehat{\boldsymbol{\Phi}}_3)$	$\rho(\widehat{\boldsymbol{\Gamma}})$	$m^2(\widehat{\boldsymbol{\Gamma}})$
100	0.725	0.353	0.199	0.401	0.804	0.534
300	0.971	0.152	0.104	0.128	0.795	0.557
1,000	0.991	0.087	0.071	0.066	0.824	0.508

The estimation accuracy measures $\rho(\cdot)$ (the larger the better) and $m^2(\cdot)$ (the smaller the better) are defined in (4.1) and (4.2), respectively. Besides, $\widehat{\boldsymbol{\Phi}}_1$, $\widehat{\boldsymbol{\Phi}}_2$ and $\widehat{\boldsymbol{\Phi}}_3$ are the first, the second and the third columns of $\widehat{\boldsymbol{\Phi}}$.

$$\boldsymbol{\Phi}_d = \begin{bmatrix} \frac{1}{\sqrt{5}} & 0 & 0 & \frac{2}{\sqrt{5}} & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{5}} & 0 & 0 & \frac{2}{\sqrt{5}} \\ \frac{-2}{\sqrt{15}} & \frac{2}{\sqrt{15}} & \frac{-2}{\sqrt{15}} & \frac{1}{\sqrt{15}} & \frac{-1}{\sqrt{15}} & \frac{1}{\sqrt{15}} \end{bmatrix}^{\mathsf{T}} \quad \text{and} \quad \boldsymbol{\Gamma}_{\widetilde{d}} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^{\mathsf{T}}.$$

In Model 2, the true lag values are $p = 6$ and $q = 2$, and the number of dimensions are $d = 3$ and $\widetilde{d} = 1$, respectively. The mean function and white noise term in Model 2 follow the simulation Model 3 of Park et al. (2009) and Example 3 of Xia et al. (2002b). However, our Model 2 has an additional conditional variance function that depends on the residuals of x_t.

Table 4 summarizes estimation results for Model 2. Our iterative estimation method estimates $\boldsymbol{\Phi}_d$ well as the ρ values are close to 1 and the m^2 values are close to 0 in all scenarios. The results are comparable to those in Park et al. (2009) (see Table 3, pp. 723), where the error term is homoscedastic. The ρ values for $\widehat{\boldsymbol{\Gamma}}$ also increase to 1 as the sample size increases and the m^2 values for $\widehat{\boldsymbol{\Gamma}}$ are of moderate sizes. Similar to Model 1, MSBC can accurately select p and $\widetilde{d}$ in most replications; see Table 5. However, the selection of q and d is still challenging.

Model 3: Consider the following conditionally heteroscedastic time series model, where the Gaussian error is mildly contaminated by a uniform error:

$$x_t = \sqrt{(3 + (1/\sqrt{3})(x_{t-1} + x_{t-3} + x_{t-6}))^2} + \varepsilon_t,$$

with $\varepsilon_t = \left[\sqrt{(1/\sqrt{10})(3 + \varepsilon_{t-1}^2 + \varepsilon_{t-2}^2)}\right] e_t$ and $e_t \sim 0.95N(0,1) + 0.05U(0,5)$. Here, the parameter matrix and vector of interest are:

$$\boldsymbol{\Phi}_d = \begin{bmatrix} \frac{1}{\sqrt{3}} & 0 & \frac{1}{\sqrt{3}} & 0 & 0 & \frac{1}{\sqrt{3}} \end{bmatrix}^{\mathsf{T}} \quad \text{and} \quad \boldsymbol{\Gamma}_{\widetilde{d}} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^{\mathsf{T}}.$$

In Model 3, the true lag values are $p = 6$ and $q = 2$, and the number of dimensions are $d = 1$ and $\widetilde{d} = 1$, respectively. The estimation results of Model 3 are summarized in Table 6. Due to the automatic robustness introduced by the Nadaraya-Watson estimator, the proposed estimation method maintains high ρ values and relatively low m^2 values in all scenarios. This suggests that the proposed method can be applied to heavy-tailed time series, which are ubiquitous in economics and finance.

Table 5
Hyperparameter selection for Model 2 over 100 replications.

n	Count	p	q	d	$\widetilde{d}$
100	Under	66	95	100	0
	Correct	32	3	0	100
	Over	2	2	0	0
300	Under	0	92	100	0
	Correct	100	6	0	100
	Over	0	2	0	0
1,000	Under	0	89	48	0
	Correct	100	7	52	100
	Over	0	4	0	0

The rows Under, Correct and Over report the number of times that the selected hyperparameters are smaller than, equal to, or larger than the truth over 100 Monte Carlo simulations.

Table 6
Estimation results for Model 3 over 100 replications.

n	$\rho(\widehat{\boldsymbol{\Phi}})$	$m^2(\widehat{\boldsymbol{\Phi}})$	$\rho(\widehat{\boldsymbol{\Gamma}})$	$m^2(\widehat{\boldsymbol{\Gamma}})$
100	0.924	0.347	0.857	0.448
300	0.965	0.222	0.872	0.479
1,000	0.989	0.111	0.911	0.341

The estimation accuracy measures $\rho(\cdot)$ (the larger the better) and $m^2(\cdot)$ (the smaller the better) are defined in (4.1) and (4.2), respectively.

5. Analysis of the BRL/USD exchange rate series

Central Banks around the world aim to guarantee that their national currency is reasonably stable and trustworthy. Economic agents need to be confident that this financial asset will keep its value compared to the other products that it can be traded for. If a currency is too volatile, any task which requires planning becomes too uncertain, affecting major investment decisions. However, stability by itself is not enough to yield a strong currency. For instance, many countries have adopted legal measures of broad price control without observing the desired outcome.

During the second half of the twentieth century, Brazil experienced an accelerating inflationary process. During this time, several economic measures were attempted, including the adoption of new currencies, to control the general price increase. An important Brazilian inflation index is the General Price Index - Overall Supply (IGP-OG) that can be accessed at the Institute for Applied Economic Research (IPEA) databases.[1] From the plot of monthly IGP-OG in Fig. 2, it is clear that it was only after adopting the Real Currency (BRL) that the inflation rate stabilized to a reasonable level.

Besides IGP-OG, it is also of interest to study the impact of adopting the new currency in terms of its relation with the other major currencies.

For instance, the Brazilian Real/U.S. Dollar (BRL/USD) exchange rate is an important index for the Brazilian economy as it influences key economic features, such as, competitiveness of exports, production costs and investment returns.

In this section, we implement the iterative estimation approach proposed in Section 3.2 to analyze the monthly BRL/USD Foreign Exchange Rate series from January 1999 to December 2019; see Fig. 3 for a plot of this time series. Although the IPEA database provides observations from this series prior to January 1999, we only analyze the time period after the adoption of the floating exchange rate regime when the currency price fluctuated mostly according to market forces instead of government fixation. Note that, from the adoption of the BRL as the official currency in July 1994 until early January 1999, the fixed exchange regime was in place, resulting in no variation in the series during this period. Therefore, we do not include this period in our data analysis.

We split the time series into a training data and a test data. The training data ranges from 01/1999 to 10/2015 (202 observations) and the test data ranges from 11/2015 to 12/2019 (50 observations). For the training data, we will build an AR-ARCH model and a semi-parametric time series model using our iterative estimation approach. Then, we will use the test data to assess the performance of the fitted time series models based on the accuracy of out-of-sample forecasts.

5.1. AR-ARCH model

A benchmark in conditionally heteroscedastic parametric modeling is the family of AR(p)-ARCH(q) models. We use the training data for the monthly BRL/USD Exchange Rate series (01/1999 to 10/2015) to find a suitable AR-ARCH model. Our

[1] https://www.ipea.gov.br.

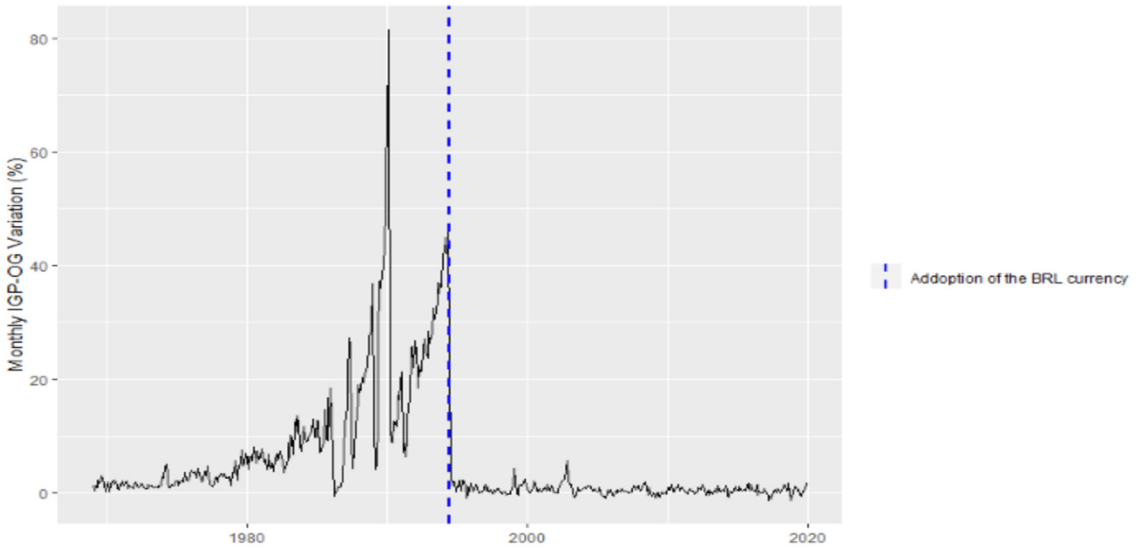

Fig. 2. Monthly IGP-OG variation (%) in Brazil.

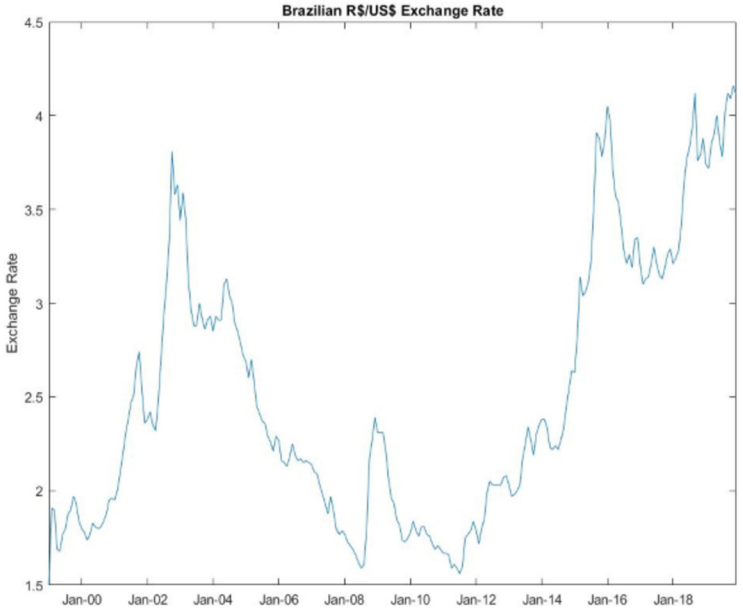

Fig. 3. Monthly BRL/USD exchange rate between Jan 1999 and Dec 2019.

goal is to use an out-of-sample forecast measure to compare the performance of the fitted AR-ARCH model to the time series model fitted using our iterative estimation approach.

To fit an AR-ARCH model, the first step is to find appropriate lags p and q for the mean and variance functions, respectively. To this end, we fit all possible AR-ARCH models by a grid search for p and q values using the *estimate* function in MATLAB and compute the Schwarz Bayesian Criterion (SBC) value for each pair of p and q. Table 7 gives the SBC values for various choices of p and q. We observe from Table 7 that the SBC criterion is minimized when $p = 4$ and $q = 1$. Then, the fitted AR(4)-ARCH(1) model for the training data is as follows

$$\widehat{x}_t = -0.023455 + 1.6059x_{t-1} - 0.90397x_{t-2} + 0.55746x_{t-3} - 0.26837x_{t-4} + \widehat{\varepsilon}_t, \tag{5.1}$$

with $\widehat{\varepsilon}_t = \sqrt{\widehat{h}_t}\, e_t$, $\widehat{h}_t = 0.0029 + 0.7695\widehat{\varepsilon}^2_{t-1}$ and $e_t \sim N(0,1)$.

Besides, all the estimated coefficients are significant at the 1% level, except for the intercept in the mean function.

Table 7
Lags selection for AR(p)-ARCH(q) by SBC.

SBC	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$
$q=1$	−371.06	−409.09	−406.32	**−413.34**	−408.06	−407.08
$q=2$	−378.97	−409.18	−408.89	−410.58	−406.68	−403.68

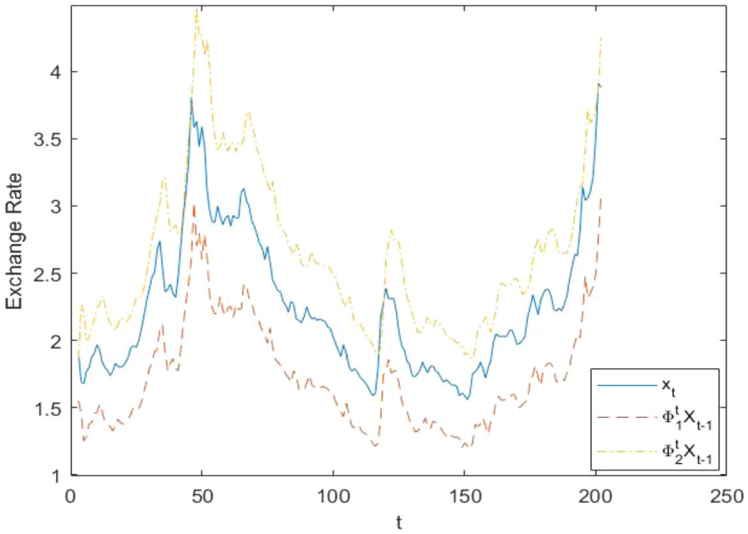

Fig. 4. Time series plot for x_t, $\widehat{\boldsymbol{\Phi}}_1^\top \boldsymbol{X}_{t-1}$ and $\widehat{\boldsymbol{\Phi}}_2^\top \boldsymbol{X}_{t-1}$ using the training data.

5.2. Semi-parametric time series model

In this subsection, we use the iterative estimation approach proposed in Section 3.2 to build a semi-parametric time series model for the training data. First, we use MSBC defined in Section 3.4 to select the lag parameters p and q, and the number of dimensions d and $\widetilde{d}$ for the mean and variance parameter matrices. The estimated parameters and dimensions are $\widehat{p}=2$, $\widehat{q}=1$, $\widehat{d}=2$ and $\widehat{\widetilde{d}}=1$, respectively. Then, we follow the proposed iterative estimation approach to obtain the following parameter estimates:

$$\widehat{\boldsymbol{\Phi}}=\begin{bmatrix} 0.9783 & 0.2071 \\ -0.2071 & 0.9783 \end{bmatrix} \quad \text{and} \quad \widehat{\boldsymbol{\Gamma}}=1.$$

Next, we build a time series model using the estimates above. To be specific, we denote the estimated linear combinations as $\widehat{\boldsymbol{\Phi}}_1^\top \boldsymbol{X}_{t-1}$ and $\widehat{\boldsymbol{\Phi}}_2^\top \boldsymbol{X}_{t-1}$, where $\widehat{\boldsymbol{\Phi}}_k$ denotes the k^{th} column of $\widehat{\boldsymbol{\Phi}}$ for $k=1,2$. As discussed in Section 2, we notice that the true parameter matrices are not identifiable. In this real-data analysis, the optimization of (3.9) will lead to non-unique solutions such that the columns of $\widehat{\boldsymbol{\Phi}}$ cab be switched and/or re-scaled by -1. For instance, the following three estimates of $\boldsymbol{\Phi}$ are equivalent to $\widehat{\boldsymbol{\Phi}}$:

$$\begin{bmatrix} -0.9783 & 0.2071 \\ 0.2071 & 0.9783 \end{bmatrix}, \quad \begin{bmatrix} 0.9783 & -0.2071 \\ -0.2071 & -0.9783 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0.2071 & -0.9783 \\ 0.9783 & 0.2071 \end{bmatrix}.$$

Nonetheless, the identifiability issue will not cause trouble in this study as all equivalent estimates of $\boldsymbol{\Phi}$ will lead to the same linear combinations up to a sign change, which will be equally good for the out-of-sample forecast. Before we move on, we visualize x_t, $\widehat{\boldsymbol{\Phi}}_1^\top \boldsymbol{X}_{t-1}$, and $\widehat{\boldsymbol{\Phi}}_2^\top \boldsymbol{X}_{t-1}$ using the training data in Fig. 4. This time series plot clearly shows the two linear combinations, $\widehat{\boldsymbol{\Phi}}_1^\top \boldsymbol{X}_{t-1}$ and $\widehat{\boldsymbol{\Phi}}_2^\top \boldsymbol{X}_{t-1}$, can capture the dynamics of the response time series x_t.

In the next step, we estimate the conditional mean function $\widehat{f}(\cdot)$ based on $\widehat{\boldsymbol{\Phi}}_1^\top \boldsymbol{X}_{t-1}$ and $\widehat{\boldsymbol{\Phi}}_2^\top \boldsymbol{X}_{t-1}$. To this end, we create a 3-D scatter plot of x_t, $\widehat{\boldsymbol{\Phi}}_1^\top \boldsymbol{X}_{t-1}$ with an overlay surface to visualize the conditional mean function in Fig. 5. The surface, fitted by the Nadaraya-Watson estimation, visually suggests a near linear relationship between x_t and the two linear combinations. This motivates us to fit the following time series model for x_t using the training data:

$$x_t = 0.0274 + 1.3569\,\widehat{\boldsymbol{\Phi}}_1^\top \boldsymbol{X}_{t-1} - 0.0473\,\widehat{\boldsymbol{\Phi}}_2^\top \boldsymbol{X}_{t-1} + \varepsilon_t. \tag{5.2}$$

Denote $\widehat{\varepsilon}_t = x_t - 0.0274 - 1.3569\,\widehat{\boldsymbol{\Phi}}_1^\top \boldsymbol{X}_{t-1} + 0.0473\,\widehat{\boldsymbol{\Phi}}_2^\top \boldsymbol{X}_{t-1}$ as the fitted residual at time t. We then estimate the conditional variance function $\widehat{g}(\widehat{\boldsymbol{\Gamma}}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)$. Here, we visualize $\widehat{\boldsymbol{\varepsilon}}_t^2$ and $\widehat{\boldsymbol{\Gamma}}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2$ using the training data in Fig. 6. This time series plot

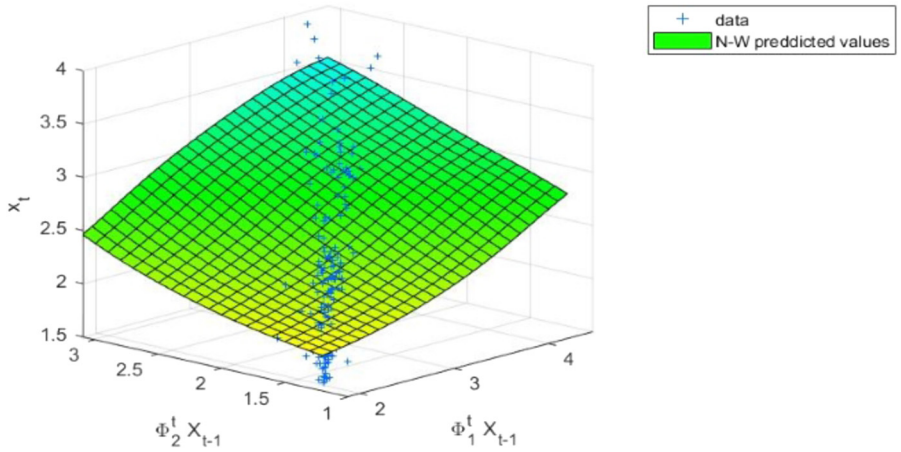

Fig. 5. 3-D scatter plot with an overlay surface of the fitted conditional mean function for the real data analysis.

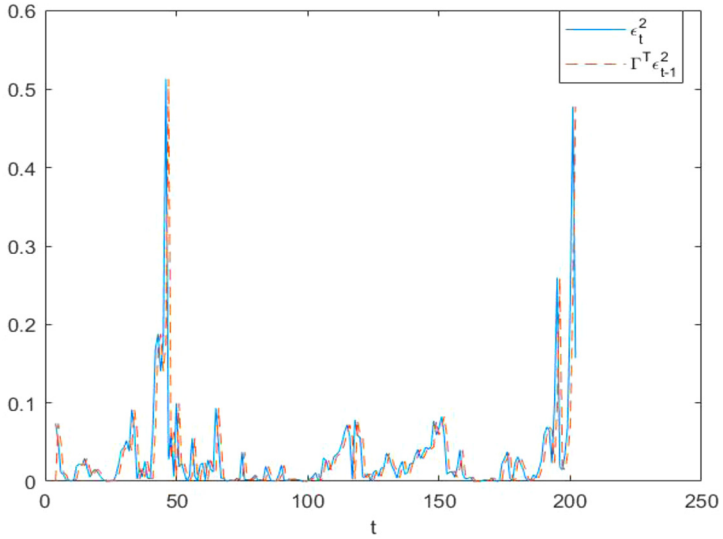

Fig. 6. Time series plot for $\widehat{\varepsilon}_t^2$ and $\widehat{\boldsymbol{\Gamma}}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2$ using the training data.

shows that $\widehat{\boldsymbol{\Gamma}}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2$ captures the dynamics of $\widehat{\boldsymbol{\varepsilon}}_t^2$ quite well. Furthermore, we draw a scatter plot between ε_t^2 and $\widehat{\boldsymbol{\Gamma}}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2$ given in Fig. 7. The blue curve in Fig. 7 is the Nadaraya-Watson estimation for $g(\widehat{\boldsymbol{\Gamma}}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)$ which can be approximated well by a linear function. Therefore, the final model we learned from the training data is as follows.

$$x_t = 0.0274 + 1.3569 \widehat{\boldsymbol{\Phi}}_1^\intercal \boldsymbol{X}_{t-1} - 0.0473 \widehat{\boldsymbol{\Phi}}_2^\intercal \boldsymbol{X}_{t-1} + \widehat{\varepsilon}_t, \tag{5.3}$$

$$\text{with} \quad \widehat{\varepsilon}_t = \sqrt{\widehat{h}_t} e_t, \quad \widehat{h}_t = 0.0158 + 0.5056 \varepsilon_{t-1}^2 \quad \text{and} \quad e_t \sim N(0,1).$$

5.3. Comparison of out-of-sample forecasts

In this subsection, we compute an out-of-sample forecast measure of our semi-parametric time series model defined in (5.3) and compare it with that of the AR(4)-ARCH(1) model defined in (5.1). The test set is the monthly BRL/USD foreign exchange rate collected from November 2015 to December 2019. The forecast performance is assessed by the Mean Squared Prediction Error (MSPE) over the test set, which is defined as

$$\text{MSPE} = n_T^{-1} \sum_{t=1}^{n_T} (x_t - \widehat{x}_t)^2,$$

where n_T is the sample size of the test set and $\widehat{x}_t$ is the predicted value of x_t obtained by either our semi-parametric model or the AR(4)-ARCH(1) model.

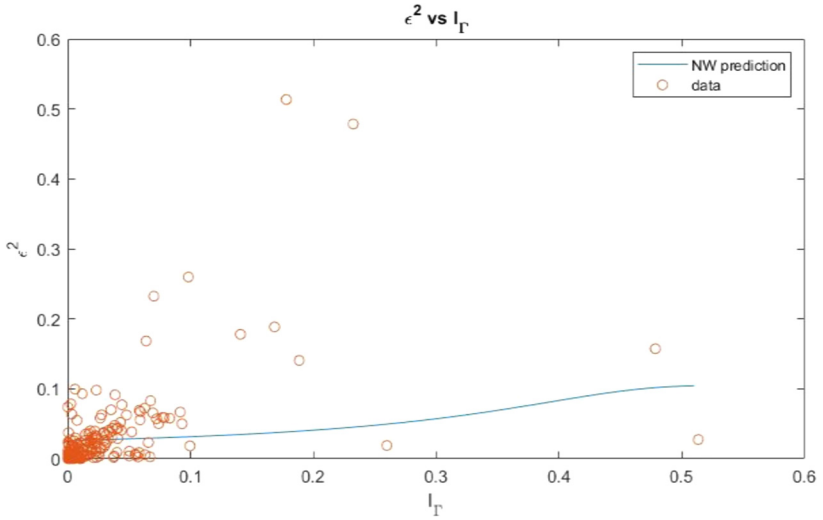

Fig. 7. Scatter plot for ε_t^2 and $\widehat{\boldsymbol{\Gamma}}^{\mathsf{T}}\widehat{\boldsymbol{\varepsilon}}_{t-1}^2$ with the Nadaraya-Watson estimation curve.

Table 8
Comparison of out-of-sample forecast results.

Model	MSPE
Model (5.1): AR(4)-ARCH(1)	0.0157
Model (5.3): Our model	0.0151

The forecast results are summarized in Table 8. According to Table 8, in terms of out-of-sample forecasts, our semi-parametric time series model achieves a better forecast accuracy than the AR(4)-ARCH(1) model. This real data analysis provides a piece of evidence that our iterative estimation method has the penitential to construct an efficient forecast model by utilizing the fitted linear combinations as explanatory variables.

6. Concluding remarks

In this article, we have developed a non-parametric iterative estimation approach for dimension reduction in time series where the reduction in dimension is aimed at both the conditional mean function as well as the conditional variance function of the series. While Nadaraya–Watson kernel smoothers are used to estimate the conditional mean and variance functions, respectively, the parameter matrices associated with reduction in the dimension are estimated iteratively by minimizing a residual sum of squares and a weighted residual sum of squares which are sample versions of the respective population estimating functions. The resulting estimators are shown to be consistent as the sample size tends to infinity.

The optimization problems associated with the estimation of parameter matrices involve matrix orthonormal constraints. Existing literature tackles such problems by using either differential geometry optimization methods or sequential quadratic programming algorithms. Unfortunately, these approaches pose computational challenges and are numerically unstable when the parameter matrix of interest has more than one column. We have proposed a novel angular representation for the parameter matrix which converts the matrix orthonormal constraint into a sequence of linear systems thereby overcoming certain computational challenges and improving numerical stability. We believe that the angular representation is applicable more generally to other optimization problems.

Overall, the theory of dimension reduction in time series in the presence of conditional heteroscedasticity poses many challenges, but a variety of encouraging results presented through our simulation study and the analysis of the Brazilian Real/U.S. Dollar exchange rate series suggest that our iterative estimation approach has great potential for providing a viable and meaningful alternative to traditional AR-ARCH type time series analysis. We hope that this approach will open new avenues in modeling financial and economic time series.

Acknowledgements

The authors are grateful to the associate editor and referees for their insightful comments that have significantly improved the article. Ke acknowledges the support of National Science Foundation (NSF), USA grant DMS-2210468. Sriram's work was supported by the National Science Foundation (NSF), USA grant with award number 1309665.

Appendix A. Assumptions and notations

Assume that d, $\widetilde{d}$, p and q defined earlier are fixed and known numbers. Recall that with $\boldsymbol{X}_{t-1} = (x_{t-1}, ..., x_{t-p})^\intercal$, we assumed that the conditional mean function $\mathrm{E}(x_t|\boldsymbol{X}_{t-1}) = f(\boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_{t-1})$ and the conditional variance function $V(x_t|\boldsymbol{X}_{t-1}) = g(\boldsymbol{\Gamma}_{\widetilde{d}}^\intercal \boldsymbol{\varepsilon}_{t-1}^2)$, where $\boldsymbol{\varepsilon}_{t-1}^2 = (\varepsilon_{t-1}^2, ..., \varepsilon_{t-q}^2)^\intercal$ for $\widetilde{d} \le q$ with $\varepsilon_t = x_t - f(\boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_{t-1})$. Denote $\xi_t = f(\boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_{t-1}) - \widehat{f}_n(\boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_{t-1})$, where $\widehat{f}_n(\boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_{t-1})$ is defined as in (3.2) with $\boldsymbol{r} = \boldsymbol{\Phi}_d$. Next, we state the assumptions for technical lemmas and the main theorems stated in this section.

Assumptions:

(A1) $(\boldsymbol{X}_t, \boldsymbol{\varepsilon}_t)$, $t \ge 0$ is strictly stationary and strong mixing with mixing coefficient $\alpha(m) \le Am^{-\beta}$, where $A < \infty$. For some $s > 2$, $E|\boldsymbol{X}_t|^s < \infty$ and $E|\boldsymbol{\varepsilon}_t^2|^s < \infty$ and $\beta > (2s-1)/(s-2)$.

(A2) The marginal densities of $\boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_{t-1}$ and $\boldsymbol{\Gamma}_{\widetilde{d}}^\intercal \boldsymbol{\varepsilon}_{t-1}^2$ are bonded and bounded away from zero on their supports which are closed intervals. Also, there is some $t^* < \infty$ such that for all $t \ge t^*$

$$\sup_{a_0,a_t} E\left(|\boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_0 \boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_t| \,\middle|\, \boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_0 = a_0,\ \boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_t = a_t \right) p_t(a_0, a_t) < \infty,$$

$$\text{and} \quad \sup_{b_0,b_t} E\left(|\boldsymbol{\Gamma}_{\widetilde{d}}^\intercal \boldsymbol{\varepsilon}_0^2 \boldsymbol{\Gamma}_{\widetilde{d}}^\intercal \boldsymbol{\varepsilon}_t^2| \,\middle|\, \boldsymbol{\Gamma}_{\widetilde{d}}^\intercal \boldsymbol{\varepsilon}_0^2 = b_0,\ \boldsymbol{\Gamma}_{\widetilde{d}}^\intercal \boldsymbol{\varepsilon}_t^2 = b_t \right) q_t(b_0, b_t) < \infty,$$

where $p_t(a_0, a_t)$ denotes the joint density of $\{\boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_0,\ \boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_t\}$ and $q_t(b_0, b_t)$ denotes the joint density of $\{\boldsymbol{\Gamma}_{\widetilde{d}}^\intercal \boldsymbol{\varepsilon}_0^2,\ \boldsymbol{\Gamma}_{\widetilde{d}}^\intercal \boldsymbol{\varepsilon}_t^2\}$.

(A3) The eigenvalues of $E[\boldsymbol{X}_t \boldsymbol{X}_t^\intercal]$ and $E[\boldsymbol{\varepsilon}_t^2 \boldsymbol{\varepsilon}_t^{2\intercal}]$ are bounded and bounded away from zero.

(A4) The first two derivatives of $f(\cdot)$ and $g(\cdot)$ exist and are continuous on $\mathbb{R}$. Further, $f(\cdot)$ and $g(\cdot)$ satisfy the following Lipschitz continuous conditions

$$|f(u) - f(v)| \le C_f |u - v| \quad \text{and} \quad |g(u) - g(v)| \le C_g |u - v|,$$

where C_f and C_g are two positive Lipschitz constants.

(A5) The kernel function $K(u)$ is compactly supported with bounded second order derivative such that $\int uK(u)du = 0$, $\int u^2 K(u)du < \infty$, and the Fourier transformation of $K(u)$ is absolutely integrable.

Remark A.1. Here, we explain the assumptions made above. (A1) assumes that the serial dependence in the data is strong mixing. The decay rate depends on the moment conditions of $\boldsymbol{X}_t$ and $\boldsymbol{\varepsilon}_t^2$. When $s = \infty$, e.g. $\boldsymbol{X}_t$ is bounded or Gaussian, the condition on the decay parameter simplifies to $\beta > 2$. (A2) requires the marginal densities of $\boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_{t-1}$ and $\boldsymbol{\Gamma}_{\widetilde{d}}^\intercal \boldsymbol{\varepsilon}_{t-1}^2$ to be bounded. It also controls the tail behaviors of the joint densities and conditional expectations with lags greater than t^*. (A1) and (A2) are mild regularity assumptions to study the uniform consistency and convergence rate of the Nadaraya-Watson estimator, see Hansen (2008) and Hong and Linton (2020) among others. (A3) is a bounded eigenvalue condition which is imposed to avoid degenerate covariance and precision matrices. (A4) assumes $f(\cdot)$ and $g(\cdot)$ to be Lipschitz continuous which is commonly assumed in nonparametric regression literature. (A5) contains some smoothness conditions for the kernel function.

Appendix B. Proof of the three theorems in Section 3

B.1. Proof of Theorem 3.1

Let $\delta_n = \left(\frac{\ln(n-p)}{n-p}\right)^{2/(d+4)}$ and η_n be a sequence that diverges with n at an arbitrarily slow rate. In order to prove the convergence rate $\|\widehat{\boldsymbol{\Phi}}_0 - \boldsymbol{\Phi}_0\| = O_p(\delta_n)$, it suffices to show that

$$\inf_{\|\boldsymbol{r} - \boldsymbol{\Phi}_0\| = \eta_n \delta_n} \sum_{t=p+1}^{n} \left(x_t - \widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1})\right)^2 - \sum_{t=p+1}^{n} \left(x_t - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right)^2 > 0,$$

with probability approaching one if $n \to \infty$ and $\eta_n \to \infty$ arbitrarily slowly.

With some calculations, one can show

$$\begin{aligned}
&\sum_{t=p+1}^{n} \left(x_t - \widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1})\right)^2 - \sum_{t=p+1}^{n} \left(x_t - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right)^2 \\
&= \sum_{t=p+1}^{n} \left\{ \widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1})^2 - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})^2 - 2x_t \left[\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1}) \right] \right\}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{t=p+1}^{n} \left[\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right]^2 \\
&\quad - 2 \sum_{t=p+1}^{n} \left[x_t - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right]\left[\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right] \\
&= \sum_{t=p+1}^{n} \left[\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right]^2 - 2 \sum_{t=p+1}^{n} (\varepsilon_t + \xi_t)\left[\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right] \\
&= \sum_{t=p+1}^{n} \left[\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right]^2 - 2 \sum_{t=p+1}^{n} \varepsilon_t\left[\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right] \\
&\quad - 2 \sum_{t=p+1}^{n} \xi_t\left[\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right] \\
&\equiv I_1 - 2I_2 - 2I_3.
\end{aligned}
$$

With mean value theorem, we can re-write I_1 by

$$
\begin{aligned}
I_1 &= \sum_{t=p+1}^{n} \left[\widehat{f}_n^{((1)}(\boldsymbol{r}_*^\intercal \boldsymbol{X}_{t-1})(\boldsymbol{r} - \boldsymbol{\Phi}_0)^\intercal \boldsymbol{X}_{t-1}\right]^2 \\
&= \sum_{t=p+1}^{n} \operatorname{tr}\left\{(\boldsymbol{r} - \boldsymbol{\Phi}_0)^\intercal \left[\widehat{f}_n^{((1)}(\boldsymbol{r}_*^\intercal \boldsymbol{X}_{t-1})^2 \boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\intercal\right](\boldsymbol{r} - \boldsymbol{\Phi}_0)\right\},
\end{aligned}
$$

where $\boldsymbol{r}_*$ is an interior point on the line segment between $\boldsymbol{r}$ and $\boldsymbol{\Phi}_0$, and $\widehat{f}_n^{((1)}(\cdot)$ is the first order derivative of $f^{(1)}(\cdot)$ defined in (3.4).

According to Lemma 3.2 and the optimally of $\boldsymbol{\Phi}_0$, we have

$$
\lim_{n\to\infty} \widehat{f}_n^{((1)}(\boldsymbol{r}_*^\intercal \boldsymbol{X})^2 = f^{((1)}(\boldsymbol{r}_*^\intercal \boldsymbol{X})^2 > f^{((1)}(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X})^2 = 0, \tag{B.1}
$$

which holds uniformly over $\boldsymbol{X} \in \mathbb{R}$. Follow the strong law of large numbers, we have

$$
\lim_{n\to\infty} \frac{1}{n-p} \sum_{t=p+1}^{n} \boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\intercal = E[\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\intercal]. \tag{B.2}
$$

Denote $\lambda_{min}(\boldsymbol{A})$ and $\lambda_{max}(\boldsymbol{A})$ the smallest and the largest eigenvalue of a symmetric matrix $\boldsymbol{A}$, respectively. Given (B.1), (B.2) and Assumption (A3), we can lower bounded I_1 by

$$
\begin{aligned}
I_1 &\geq \|\boldsymbol{r} - \boldsymbol{\Phi}_0\|^2 \sum_{t=p+1}^{n} \widehat{f}_n^{((1)}(\boldsymbol{r}_*^\intercal \boldsymbol{X}_{t-1})^2 \lambda_{min}(\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\intercal) \\
&\geq C_1(n-p)\|\boldsymbol{r} - \boldsymbol{\Phi}_0\|^2(1 + o_p(1)).
\end{aligned} \tag{B.3}
$$

Next, by Cauchy-Schwartz inequality, we can upper bound I_2 by

$$
\begin{aligned}
I_2 &= \sum_{t=p+1}^{n} \varepsilon_t\left[\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right] \\
&\leq \left\{\sum_{t=p+1}^{n} \varepsilon_t^2 \sum_{t=p+1}^{n} \left[\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right]^2\right\}^{1/2} \\
&\leq (n-p)\max_t \mathrm{E}[\boldsymbol{\varepsilon}_t^2](1 + o(1))\left\{\frac{1}{n-p}\sum_{t=p+1}^{n} \left[\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right]^2\right\}^{1/2},
\end{aligned}
$$

where the last line follows the strong law of large numbers and assumption (A1).

With Lemma 3.1 and assumption (A4), we can derive

$$\begin{aligned}
&\left[\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right]^2 \\
&= \Big\{\left[\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - f(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1})\right] - \left[\widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1}) - f(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right] \\
&\quad + \left[f(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - f(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right]\Big\}^2 \\
&\le 4\Big\{\left[\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - f(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1})\right]^2 + \left[\widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1}) - f(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right]^2 \\
&\quad + \left[f(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - f(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right]^2\Big\} \\
&\le C_2(\delta_n^2 + \|\boldsymbol{r} - \boldsymbol{\Phi}_0\|^2),
\end{aligned} \tag{B.4}$$

where C_2 is a large enough positive constant.

With (B.4), we have upper bound I_2 by

$$I_2 \le C_2(n-p)\left(\delta_n^2 + \|\boldsymbol{r} - \boldsymbol{\Phi}_0\|^2\right)^{1/2}(1 + o_p(1)). \tag{B.5}$$

Similarly, we can upper bound I_3 by

$$\begin{aligned}
I_3 &= \sum_{t=p+1}^{n} \xi_t\left[\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right] \\
&\le \left\{\sum_{t=p+1}^{n} \xi_t^2 \sum_{t=p+1}^{n}\left[\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) - \widehat{f}_n(\boldsymbol{\Phi}_0^\intercal \boldsymbol{X}_{t-1})\right]^2\right\}^{1/2} \\
&\le C_3(n-p)\left(\delta_n^2 + \|\boldsymbol{r} - \boldsymbol{\Phi}_0\|^2\right)^{1/2}\delta_n(1 + o_p(1)),
\end{aligned} \tag{B.6}$$

where C_3 is a large enough positive constant and the last inequality follows Lemma 3.1 and (B.4).

Since η_n diverges with n at an arbitrarily slow rate, we have $\|\boldsymbol{r} - \boldsymbol{\Phi}_0\| = \eta_n\delta_n \gg \delta_n$ as n diverges. Therefore, we complete the proof by showing

$$\inf_{\|\boldsymbol{r}-\boldsymbol{\Phi}_0\|=\eta_n\delta_n} (I_1 - 2I_2 - 2I_3) > 0,$$

with probability approaching one if $n \to \infty$ and $\eta_n \to \infty$ arbitrarily slowly. □

B.2. Proof of Theorem 3.2

Let $\delta_n = \left(\frac{\ln(n-p)}{n-p}\right)^{2/(d+4)}$, $\delta_n' = \left(\frac{\ln(n-p-q)}{n-p-q}\right)^{2/(\widetilde{d}+4)}$ and η_n be a sequence that diverges with n at an arbitrarily slow rate. To simplify the presentation, we write $\boldsymbol{\Gamma}_{\widetilde{d}}$ as $\boldsymbol{\Gamma}$ throughout this proof.

In order to prove the convergence rate $\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\| = O_p(\delta_n')$, it suffices to show that

$$\inf_{\|\boldsymbol{s}-\boldsymbol{\Gamma}\|=\eta_n\delta_n'} \sum_{t=p+q+1}^{n}\left(\widehat{\boldsymbol{\varepsilon}}_t^2 - \widehat{g}_n(\boldsymbol{s}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right)^2 - \sum_{t=p+q+1}^{n}\left(\widehat{\boldsymbol{\varepsilon}}_t^2 - \widehat{g}_n(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right)^2 > 0$$

with probability approaching one if $n \to \infty$ and $\eta_n \to \infty$ arbitrarily slowly.

With some calculations, one can show

$$\begin{aligned}
&\sum_{t=p+q+1}^{n}\left(\widehat{\boldsymbol{\varepsilon}}_t^2 - \widehat{g}_n(\boldsymbol{s}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right)^2 - \sum_{t=p+1}^{n}\left(\widehat{\boldsymbol{\varepsilon}}_t^2 - \widehat{g}_n(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right)^2 \\
&= \sum_{t=p+q+1}^{n}\left[\widehat{g}_n(\boldsymbol{s}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]^2 \\
&\quad - 2\sum_{t=p+q+1}^{n}\left[\widehat{\varepsilon}_t^2 - \widehat{g}_n(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]\left[\widehat{g}_n(\boldsymbol{s}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right] \\
&= \sum_{t=p+q+1}^{n}\left[\widehat{g}_n(\boldsymbol{s}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]^2 - 2\sum_{t=p+q+1}^{n}\left[\widehat{\varepsilon}_t^2 - \varepsilon_t^2\right]\left[\widehat{g}_n(\boldsymbol{s}^\intercal \widehat{\boldsymbol{\varepsilon}}_t^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_t^2)\right]
\end{aligned}$$

$$
\begin{aligned}
&- 2 \sum_{t=p+q+1}^{n} \left[\varepsilon_t^2 - g(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]\left[\widehat{g}_n(\boldsymbol{s}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right] \\
&- 2 \sum_{t=p+q+1}^{n} \left[g(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]\left[\widehat{g}_n(\boldsymbol{s}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right] \\
\equiv\, & J_1 - 2J_2 - 2J_3 - 2J_4.
\end{aligned}
$$

Similar to the proof of (B.3), we can lower bound J_1 by

$$J_1 \geq C_4(n-p-q)\|\boldsymbol{s}-\boldsymbol{\Gamma}\|^2(1+o_p(1)),$$

for some positive constant C_4.

By Cauchy-Schwartz inequality, we can upper bound J_2, J_3 and J_4 by

$$
\begin{aligned}
J_2 &\leq \left\{ \sum_{t=p+q+1}^{n} \left[\widehat{\varepsilon}_t^2 - \varepsilon_t^2\right]^2 \sum_{t=p+q+1}^{n} \left[\widehat{g}_n(\boldsymbol{s}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]^2 \right\}^{1/2}, \\
J_3 &\leq \left\{ \sum_{t=p+q+1}^{n} \left[\varepsilon_t^2 - g(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]^2 \sum_{t=p+q+1}^{n} \left[\widehat{g}_n(\boldsymbol{s}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]^2 \right\}^{1/2}, \\
\text{and} \quad J_4 &\leq \left\{ \sum_{t=p+q+1}^{n} \left[g(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]^2 \sum_{t=p+q+1}^{n} \left[\widehat{g}_n(\boldsymbol{s}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]^2 \right\}^{1/2}.
\end{aligned}
$$

Similar to the derivation of (B.4), Lemma 3.2 and assumption (A4) yield

$$
\begin{aligned}
&\left[\widehat{g}_n(\boldsymbol{s}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]^2 \\
&\leq 4\Big\{\widehat{g}_n(\boldsymbol{s}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - g(\boldsymbol{s}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]^2 + \left[\widehat{g}_n(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - g(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]^2 \\
&\quad + \left[g(\boldsymbol{s}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - g(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]^2\Big\} \\
&\leq C_5(\delta_n'^2 + \|\boldsymbol{s}-\boldsymbol{\Gamma}\|^2),
\end{aligned} \tag{B.7}
$$

where C_5 is a large enough positive constant.

Following Lemma 3.1 and Theorem 3.1, we can show

$$\sum_{t=p+q+1}^{n} \left[\widehat{\varepsilon}_t^2 - \varepsilon_t^2\right]^2 = \sum_{t=p+q+1}^{n} \left[(\widehat{\varepsilon}_t - \varepsilon_t)(\widehat{\varepsilon}_t + \varepsilon_t)\right]^2 \leq C_5(n-p-q)\delta_n E(\varepsilon_t)(1+o_p(1)). \tag{B.8}$$

Notice the fact that $E(g(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)) = \varepsilon_t^2$, strong law of large numbers and Lemma 3.1 suggest

$$\sum_{t=p+q+1}^{n} \left[\varepsilon_t^2 - g(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]^2 \leq C_5(n-p-q)o(1), \tag{B.9}$$

$$\text{and} \quad \sum_{t=p+q+1}^{n} \left[g(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]^2 \leq C_5(n-p-q)\delta_n'(1+o_p(1)). \tag{B.10}$$

The results in (B.7)–(B.10) imply

$$\max\{J_2, J_3, J_4\} \ll C_5(n-p-q)\|\boldsymbol{s}-\boldsymbol{\Gamma}\|^2, \quad \text{as} \quad n \to \infty.$$

Therefore, we complete the proof since

$$\inf_{\|\boldsymbol{s}-\boldsymbol{\Gamma}\|=\eta_n\delta_n'} (J_1 - 2J_2 - 2J_3 - 2J_4) > 0,$$

with probability approaching one if $n \to \infty$ and $\eta_n \to \infty$ arbitrarily slowly. □

B.3. Proof of Theorem 3.3

Let $\delta_n = \left(\frac{\ln(n-p)}{n-p}\right)^{2/(d+4)}$, $\delta_n' = \left(\frac{\ln(n-p-q)}{n-p-q}\right)^{2/(\widetilde{d}+4)}$ and η_n be a sequence that diverges with n at an arbitrarily slow rate. To simplify the presentation, we write $\boldsymbol{\Phi}_d$ as $\boldsymbol{\Phi}$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ as $\boldsymbol{\Gamma}$ throughout this proof.

In order to prove the convergence rate $\|\widehat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}\| = O_p(\delta_n)$, it suffices to show that

$$\inf_{\|\boldsymbol{r}-\boldsymbol{\Phi}\|=\eta_n\delta_n} \sum_{t=p+q+1}^{n} \frac{\left(x_t - \widehat{f}_n(\boldsymbol{r}^\top \boldsymbol{X}_{t-1})\right)^2}{\widehat{g}_n(\widehat{\boldsymbol{\Gamma}}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)} - \sum_{t=p+q+1}^{n} \frac{\left(x_t - \widehat{f}_n(\boldsymbol{\Phi}^\top \boldsymbol{X}_{t-1})\right)^2}{\widehat{g}_n(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)} > 0, \tag{B.11}$$

with probability approaching one if $\eta_n \to \infty$ arbitrarily slowly.

To keep the presentation neat, we introduce the following notations

$$w_t = \widehat{g}_n(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2), \quad \widehat{w}_t = \widehat{g}_n(\widehat{\boldsymbol{\Gamma}}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2), \quad \Delta_t = w_t - \widehat{w}_t, \quad \text{and} \quad \psi_t = w_t \widehat{w}_t,$$

for $t = p+q+1, \ \ldots, \ n$.

With some calculations, one can show

$$\begin{aligned}
&\sum_{t=p+q+1}^{n} \frac{\left(x_t - \widehat{f}_n(\boldsymbol{r}^\top \boldsymbol{X}_{t-1})\right)^2}{\widehat{g}_n(\widehat{\boldsymbol{\Gamma}}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)} - \sum_{t=p+q+1}^{n} \frac{\left(x_t - \widehat{f}_n(\boldsymbol{\Phi}^\top \boldsymbol{X}_{t-1})\right)^2}{\widehat{g}_n(\boldsymbol{\Gamma}^\top \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)} \\
=& \sum_{t=p+q+1}^{n} \widehat{w}_t^{-1}\left(x_t - \widehat{f}_n(\boldsymbol{r}^\top \boldsymbol{X}_{t-1})\right)^2 - \sum_{t=p+q+1}^{n} w_t^{-1}\left(x_t - \widehat{f}_n(\boldsymbol{\Phi}^\top \boldsymbol{X}_{t-1})\right)^2 \\
=& \sum_{t=p+q+1}^{n} (\widehat{w}_t^{-1} - w_t^{-1})\left(x_t - \widehat{f}_n(\boldsymbol{r}^\top \boldsymbol{X}_{t-1})\right)^2 \\
&+ \sum_{t=p+q+1}^{n} w_t^{-1}\left\{\left(x_t - \widehat{f}_n(\boldsymbol{r}^\top \boldsymbol{X}_{t-1})\right)^2 - \left(x_t - \widehat{f}_n(\boldsymbol{\Phi}^\top \boldsymbol{X}_{t-1})\right)^2\right\} \\
\equiv& K_1 + K_2.
\end{aligned} \tag{B.12}$$

We can show K_1 can be upper bounded by the sum of two terms

$$\begin{aligned}
K_1 =& \sum_{t=p+q+1}^{n} (\widehat{w}_t^{-1} - w_t^{-1})\left(x_t - \widehat{f}_n(\boldsymbol{r}^\top \boldsymbol{X}_{t-1})\right)^2 \\
=& \sum_{t=p+q+1}^{n} \Delta_t \psi_t^{-1}\left(x_t - f_n(\boldsymbol{r}^\top \boldsymbol{X}_{t-1}) + f_n(\boldsymbol{r}^\top \boldsymbol{X}_{t-1}) - \widehat{f}_n(\boldsymbol{r}^\top \boldsymbol{X}_{t-1})\right)^2 \\
=& \sum_{t=p+q+1}^{n} \Delta_t \psi_t^{-1}\left(\boldsymbol{\varepsilon}_t - \xi_t\right)^2 \\
\le& 2\sum_{t=p+q+1}^{n} \Delta_t \psi_t^{-1} \boldsymbol{\varepsilon}_t^2 + 2\sum_{t=p+q+1}^{n} \Delta_t \psi_t^{-1} \xi_t^2 \\
\equiv& K_{11} + K_{12}.
\end{aligned} \tag{B.13}$$

Following assumption (A1), Lemma 3.1, Theorem 3.1 and Theorem 3.2, K_{11} and K_{12} can be upper bounded by

$$K_{11} \le \frac{2}{\min_t \psi_t}\left\{\sum_{t=p+q+1}^{n} \Delta_t^2 \sum_{t=p+q+1}^{n} \boldsymbol{\varepsilon}_t^4\right\}^{1/2} \le C_6(n-p-q)\delta_n'(1+o_p(1)), \tag{B.14}$$

$$\text{and} \quad K_{12} \le \frac{2}{\min_t \psi_t}\left\{\sum_{t=p+q+1}^{n} \Delta_t^2 \sum_{t=p+q+1}^{n} \xi_t^4\right\}^{1/2} \le C_6(n-p-q)\delta_n^2\delta_n'(1+o_p(1)), \tag{B.15}$$

where C_6 is a large enough positive constant.

Next, it is easy to check that the following inequality of K_2 holds for some positive constant C_7

$$\begin{aligned} K_2 &= \sum_{t=p+q+1}^{n} w_t \left\{ \left(x_t - \widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1})\right)^2 - \left(x_t - \widehat{f}_n(\boldsymbol{\Phi}^\intercal \boldsymbol{X}_{t-1})\right)^2 \right\} \\ &\geq \min_t w_t \sum_{t=p+q+1}^{n} \left(x_t - \widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1})\right)^2 - \left(x_t - \widehat{f}_n(\boldsymbol{\Phi}^\intercal \boldsymbol{X}_{t-1})\right)^2 \\ &\geq C_7(n-p-q)\|\boldsymbol{r}-\boldsymbol{\Phi}\|^2(1+o_p(1)), \end{aligned} \tag{B.16}$$

where the last inequality can be proved in a similar fashion as (B.3).

By plugging (B.14), (B.15) and (B.16) back to (B.12), we complete the proof since

$$\inf_{\|\boldsymbol{r}-\boldsymbol{\Phi}\|=\eta_n\delta_n} (K_1 + K_2) > 0,$$

with probability approaching one if $n \to \infty$ and $\eta_n \to \infty$ arbitrarily slowly. □

References

Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., van der Vorst, H., 2000. Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide. SIAM.

Bollerslev, T., 1986. Generalized autoregressive conditional eteroskedasticity. J. Econom. 31 (3), 307–327.

Byrd, R.H., Lu, P., Nocedal, J., Zhu, C., 1995. A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. 16 (5), 1190–1208.

Edelman, A., Arias, T.A., Smith, S.T., 1998. The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. 20 (2), 303–353.

Engle, R., 2001. GARCH 101: the use of ARCH/GARCH models in applied econometrics. J. Econ. Perspect. 15 (4), 157–168.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica, 987–1007.

Fan, J., Yao, Q., 1998. Efficient estimation of conditional variance functions in stochastic regression. Biometrika 85 (3), 645–660.

Fan, J., Yao, Q., 2008. Nonlinear Time Series: Nonparametric and Parametric Methods. Springer Science & Business Media.

Guo, S., Box, J.L., Zhang, W., 2017. A dynamic structure for high-dimensional covariance matrices and its application in portfolio allocation. J. Am. Stat. Assoc. 112 (517), 235–253.

Hansen, B.E., 2008. Uniform convergence rates for kernel estimation with dependent data. Econom. Theory 24 (3), 726–748.

Hong, S.Y., Linton, O., 2020. Nonparametric estimation of infinite order regression and its application to the risk-return tradeoff. J. Econom. 219 (2), 389–424. https://doi.org/10.1016/j.jeconom.2020.03.009.

Hotelling, H., 1936. Relations between two sets of variables. Biometrika 28, 321–377.

Ichimura, H., 1993. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. J. Econom. 58 (1–2), 71–120.

Li, B., Cook, R.D., Chiaromonte, F., 2003. Dimension reduction for the conditional mean in regressions with categorical predictors. Ann. Stat. 31 (5), 1636–1668.

Luo, W., Li, B., Yin, X., 2014. On efficient dimension reduction with respect to a statistical functional of interest. Ann. Stat. 42 (1), 384–412.

Ma, Y., Zhu, L., 2019. Semiparametric estimation and inference of variance function with large dimensional covariates. Stat. Sin. 29 (2), 567–588.

Mack, Y., Müller, H.-G., 1989. Derivative estimation in nonparametric regression with random predictor variable. Sankhyā, Ser. A, 59–72.

Masry, E., Tjøstheim, D., 1995. Nonparametric estimation and identification of nonlinear arch time series strong convergence and asymptotic normality: strong convergence and asymptotic normality. Econom. Theory 11 (2), 258–289.

Nadaraya, E., 1964. On estimating regression. Theory Probab. Appl. 9, 141–142.

Park, J.-H., 2011. Dimension reduction transfer function model. J. Stat. Comput. Simul. 81, 2131–2140. https://doi.org/10.1080/00949655.2010.519704.

Park, J.-H., Samadi, S.Y., 2014. Heteroscedastic modelling via the autoregressive conditional variance subspace. Can. J. Stat. 42 (3), 423–435.

Park, J.-H., Samadi, S.Y., 2020. Dimension reduction for the conditional mean and variance functions in time series. Scand. J. Stat. 47 (1), 134–155. https://doi.org/10.1111/sjos.12405.

Park, J.-H., Sriram, T.N., 2017. Robust estimation of conditional variance of time series using density power divergences. J. Forecast. 36 (6), 703–717. https://doi.org/10.1002/for.2465.

Park, J.-H., Sriram, T.N., Yin, X., 2009. Central mean subspace in time series. J. Comput. Graph. Stat. 18 (3), 717–730. https://doi.org/10.1198/jcgs.2009.08076.

Park, J.-H., Sriram, T.N., Yin, X., 2010. Dimension reduction in time series. Stat. Sin. 20 (2), 747–770. http://www.jstor.org/stable/24309020.

Sato, K., Sato, H., Damm, T., 2019. Riemannian optimal identification method for linear systems with symmetric positive-definite matrix. IEEE Trans. Autom. Control 65 (11), 4493–4508.

Scott, D.W., 1992. Multivariate Density Estimation: Theory, Practice and Visualization. John Wiley & Sons.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall.

Stone, C.J., 1982. Optimal global rates of convergence for nonparametric regression. Ann. Stat., 1040–1053.

Watson, G., 1964. Smooth regression analysis. Sankhyā, Ser. A 26 (4), 359–372.

Wen, Z., Yin, W., 2013. A feasible method for optimization with orthogonality constraints. Math. Program. 142 (1–2), 397–434.

Xia, Y., Tong, H., Li, W.K., 2002a. Single-index volatility models and estimation. Stat. Sin., 785–799.

Xia, Y., Tong, H., Li, W.K., Zhu, L.-X., 2002b. An adaptive estimation of dimension reduction space. J. R. Stat. Soc., Ser. B, Stat. Methodol. 64 (3), 363–410. http://www.jstor.org/stable/3088779.

Ye, Z., Weiss, R.E., 2003. Using the bootstrap to select one of a new class of dimension reduction methods. J. Am. Stat. Assoc. 98 (464), 968–979. https://doi.org/10.1198/016214503000000927.

Yin, X., Cook, R.D., 2005. Direction estimation in single-index regressions. Biometrika 92 (2), 371–384.