DIMENSION REDUCTION IN TIME SERIES UNDER THE PRESENCE OF CONDITIONAL HETEROSCEDASTICITY

By

MURILO MASSARU DA SILVA

(Under the direction of T. N. Sriram and Yuan Ke)

ABSTRACT

We consider a univariate discrete-time series $\{x_t; t \geq 1\}$, where the conditional mean of x_t is assumed to be an unknown function of linear combinations of past observations and the conditional variance of x_t is also assumed to be an unknown function of linear combinations of past squared residuals. These linear combinations are such that they contain all the necessary information about x_t that is available from the conditional mean and conditional variance, respectively.

We have developed an iterative estimation approach, which, in the first step, uses the Nadaraya-Watson estimator of the unknown (conditional) mean function and minimizes a sum of squared error to estimate the parameter associated with the linear combinations of past observations. This initial estimator is then used to form the observed residuals. In the second-step, our estimation approach once again uses a Nadaraya-Watson estimator of the unknown (conditional) variance function and minimizes a sum of squares to estimate the parameter associated with the linear combination of past observed residuals. In the third step, a revised estimate of the parameter associated with the linear combinations of past observations is obtained by minimizing an appropriately weighted sum of squares. This iterative process of estimation is then repeated until convergence of the estimates.

We have theoretically shown that the iterative estimators obtained in the above manner are consistent as sample size tends to infinity. Our theoretical results are validated through

comprehensive simulation studies. Furthermore, we have applied the iterative estimation procedure to the task of forecasting the BRL/USD Exchange Rate. For this data, we have demonstrated that the estimated linear combinations can be used to generate competitive forecasts of the series as compared to those generated using an AR-ARCH model.

Finally, to overcome some of the computational challenges, we have developed a new parametrization technique in order to guarantee that the numerical optimization is more efficient. The advantages of this new parametrization are that: 1) it ensures that the constraints imposed are fully met, 2) it makes convergence more frequent, 3) it reduces the computational time, and 4) it makes the optimization feasible to a wider range of algorithms and software.

INDEX WORDS: Heteroskedasticity, Mean function, Nadaraya-Watson Estimator, Time-Series, Dimension Reduction.

DIMENSION REDUCTION IN TIME SERIES UNDER THE PRESENCE OF CONDITIONAL HETEROSCEDASTICITY

By

MURILO MASSARU DA SILVA

B.S., Universidade Federal de Uberlândia, Brazil, 2011

M.S., Universidade Federal da Paraíba, Brazil, 2013

M.S., University of Georgia, 2016.

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2020

DIMENSION REDUCTION IN TIME SERIES UNDER THE PRESENCE OF CONDITIONAL HETEROSCEDASTICITY

By

MURILO MASSARU DA SILVA

Major Professor:	T.N. Sriram
Co-advisor:	Yuan Ke
Committee:	Shuyang Bai
	Ping Ma
	Cheolwoo Park

Electronic Version Approved:

Ron Walcott
Interim Dean of Graduate School
The University of Georgia
August 2020

Dedication

To my son, Gustavo.

Acknowledgments

I would like to thank my advisor Dr. T.N. Sriram and co-advisor Dr. Yuan Ke for their significant contribution to this research. I also thank the committee members for all the suggestions and discussions throughout the development of this dissertation. I thank the UGA Statistics Department for all the help and support during my entire doctoral training. Last but not the least, the present study was also partially supported by CAPES, Coordenacao de Aperfeicoamento de Pessoal de Nivel Superior, Brazil.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Time series analysis has been an active area of research for many decades. An intrinsic nature of a time series is that the observations are correlated. This severely restricts the direct applicability of many traditional statistical methodologies that are primarily suited for analyzing independent and identically distributed (i.i.d.) data. Unique challenges posed by time series data sets have given rise to two broad approaches: *the time domain approach* and the *frequency domain approach*. While there are many useful parametric and nonparametric methods for analyzing time series data, there is a never-ending quest to build new methodologies to analyze time series data that arise in a variety of fields such as economics, meteorology, engineering, geophysics, social, and environmental science.

A fundamental task of time series analysis is inference about the conditional distribution or the conditional moments (e.g., mean and/or variance) of the current value given its past values. This not only helps unveil underlying dynamics of the series, but also enables meaningful forecast of future values of the series. Nonstandard features such as nonlinearity, asymmetric cycles, and conditional heteroscedasticity observed in many real time series have prompted researchers to look beyond the realm of linear time series models, resulting in the development of nonlinear time series analysis.

In this dissertation, we consider a univariate discrete-time series $\{x_t; t \geq 1\}$, where the conditional mean of x_t is assumed to be an unknown function of linear combinations of past observations and the conditional variance of x_t is also assumed to be an unknown function of linear combinations of past squared residuals. These two linear combinations are such that they contain all the necessary information about x_t that is available from the conditional mean and conditional variance, respectively. The goal of this research is to first estimate the (conditional) mean and variance functions nonparametrically, and then find estimates of

the parameters associated with the linear combinations of past observations or the squared residuals.

Chapter 2 discusses the relevant literature on dimension reduction in time series. Chapter 3 describes our research problem in detail and introduces the notations that will be used throughout the dissertation. The proposed iterative estimation approach, selection criteria, and the theoretical results are stated in Chapter 4. A new parametrization technique to guarantee that our numerical optimization is more efficient is discussed in Chapter 5. A comprehensive set of simulation studies covering various scenarios are presented in Chapter 6. In Chapter 7, we apply our iterative estimation approach to an important financial time series: The Brazilian Real/U.S. Dollar (BRL/USD) Exchange Rate. Lastly, we summarize our main findings in Chapter 8. All the theoretical results stated in Chapter 4 are proved in an Appendix given at the end of the dissertation.

Chapter 2

Literature Review

Suppose $\{x_t; t \geq 1\}$ is a time series. Since forecasting future values of a time series is of interest, it is useful to make inference about the conditional distribution of x_t given all its the past values, $\{x_{t-1}, \ldots, x_1\}$. However, in many instances, one determines a value of $p \geq 1$ to make inference about the conditional distribution of $x_t|\boldsymbol{X}_{t-1}$, where $\boldsymbol{X}_{t-1} = (x_{t-1}, ..., x_{t-p})^\intercal$. Also, if p is known, then we may only need a few linear combinations of $\boldsymbol{X}_{t-1}$ in the final model (Xia and Li (1999), Xia et al. (1999) and Xia et al. (2002)) in order to forecast x_t. When the lag p is not known, Ng and Perron (2005) discuss diagnostic ways and estimation methods for selecting p before starting any inference or prediction.

Park et al. (2010) considered the problem of finding finitely many linear combinations, $\{\Phi_1^T \boldsymbol{X}_{t-1}, ..., \Phi_d^T \boldsymbol{X}_{t-1}\}, d \leq p$, such that the conditional distribution of $x_t|\boldsymbol{X}_{t-1}$ is same as the conditional distribution of $x_t|(\Phi_1^T \boldsymbol{X}_{t-1}, ..., \Phi_d^T \boldsymbol{X}_{t-1})$ without specifying a model. This is equivalent to finding a $p \times d$ matrix $\boldsymbol{\Phi}_d = (\Phi_1, \ldots, \Phi_d)$ such that

$$x_t \perp\!\!\!\perp \boldsymbol{X}_{t-1} | \boldsymbol{\Phi}_d^T \boldsymbol{X}_{t-1} \tag{2.1}$$

that is to say, x_t is conditionally independent of $\boldsymbol{X}_{t-1}$ given $\boldsymbol{\Phi}_d^T \boldsymbol{X}_{t-1}$. Therefore, the $p \times 1$ vector $\boldsymbol{X}_{t-1}$ can be replaced by the $d \times 1$ vector $\boldsymbol{\Phi}_d^T \boldsymbol{X}_{t-1}$ $(d \leq p)$ without loss of information. This represents a useful reduction in the dimension of $\boldsymbol{X}_{t-1}$, where all the information in $\boldsymbol{X}_{t-1}$ about x_t is contained in the d-linear combinations. Park et al. (2010) defined a subspace spanned by the columns of $\boldsymbol{\Phi}_d$ as the so-called *Time Series Central Subspace* and nonparametrically estimated it by maximizing (with respect to h) the Kullback-Leibler divergence between the joint density, $p(h^\intercal \boldsymbol{X}_{t-1}, x_t)$, and the product of the marginal densities, $p(x_t)p(h^\intercal \boldsymbol{X}_{t-1})$, which quantifies the dependence of x_t on $h^\intercal \boldsymbol{X}_{t-1}$. They showed that their nonparametric estimator of $\boldsymbol{\Phi}_d$ is consistent when p and d are known. In addition,

for unknown d and p, they proposed a consistent estimator of d and a graphical method to determine p. Finally, they also presented simulation studies and a data analysis of the well-known *Wolf Yearly Sunspot data* to illustrate their dimension reduction theory for time series. For more details, see Park et al. (2010).

The notion of dimension reduction in time series extends beyond what Park et al. (2010) defined as the *Time Series Central Subspace*. For instance, Li and Shedden (2002) developed a method analogous to Principal Components analysis which identifies a small number of stochastic time series components such that each series from a large ensemble is represented by a weighted sum of series-specific realizations of the components. Becker and Fried (2003) proposed a dynamic version of Sliced Inverse Regression aiming exploratory analysis of high-dimensional multivariate time series. Hall and Yao (2005) discussed a similar dimension reduction approach applicable to time series in which, for a random variable Y and a random d-vector X, the authors proposed to estimate the distribution of $Y|\theta^{\intercal}X$ instead of $Y|X$, where θ is a unit vector. Another dimension reduction approach relies on identifying series that are similar to each other. Thus, selecting a few representative series might be more useful than modeling based on the entire dataset. See Agrawal et al. (1993), Wu et al. (1996), Chan and Fu (1999), Keogh et al. (2002), Chakrabarti et al. (2002) and Lin et al. (2003) for different approaches to perform the similarity search.

In another article, Park et al. (2009) proposed a notion of *Central Mean Subspace* for time series $\{x_t; t \geq 1\}$ which does not require specification of a model but seeks to find a $p \times d$ matrix $\boldsymbol{\Phi}_d$, $d \leq p$, so that the $d \times 1$ vector $\boldsymbol{\Phi}_d^{\intercal}\boldsymbol{X}_{t-1}$ includes all the information about x_t that is available from $E(x_t|\boldsymbol{X}_{t-1})$. This represents a useful 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 $E(x_t|\boldsymbol{\Phi}_d^{\intercal}\boldsymbol{X}_{t-1})$. For known p and d, they estimated $\boldsymbol{\Phi}_d$ through a Nadaraya-Watson kernel smoother and established the strong consistency of their estimator. In addition, they proposed estimation of d and p using a modified Schwarz Bayesian Criterion (SBC), if either of d and p is unknown. Finally, they examined the performance of all the estimators extensively through a variety of simulations and provided a new analysis of the well-known *Canadian lynx data*. The estimation approach in Park et al. (2009) is analogous to that proposed in Cook and Li (2002) for regression.

The estimation based on the notion of *Central Mean Subspace* depends on an appropriate choice of a nonparametric regression estimator for the time series framework. In fact, the field of nonlinear time series analysis includes many possible choices of parametric and nonparametric modeling. For instance, Engle (1982) defined the widely known Autoregressive Conditional Heteroscedasticity (ARCH) model, which was generalized later by Bollerslev (1986). See Tjøstheim (1994), Tiao and Tsay (1994) and Fan and Yao (2003) for a thorough review of most popular parametric and nonparametric nonlinear models for time series. More recently, Lee and Shao (2018) proposed a new methodology for dimension reduction for multivariate time series, seeking a contemporaneous linear transformation based on a martingale difference divergence matrix (MDDM) such that the transformed series can be separated into two parts. One part is defined to be conditionally dependent on the past, whereas the other is conditionally independent.

Park (2011) extended the idea of *Central Mean Subspace* to a bivariate time series framework, where the expectation of the response depends both on its past values and on the present and past values of a covariate. More specifically, suppose $\{y_t, x_t\}$ is a bivariate time series, where x_t is a covariate series. Let $\mathbf{Z}_t = (y_{t-1}, ..., y_{t-p}, x_t, x_{t-1}, ..., x_{t-k})$ for $p \geq 1$ and $k \geq 1$. The goal of his proposed methodology was to find a $(p + k + 1) \times d$ matrix $\boldsymbol{\Omega}_d$ such that $d \leq p + k + 1$ and $y_t \perp\!\!\!\perp \mathbf{Z}_t | \boldsymbol{\Omega}_d^{\intercal} \mathbf{Z}_t$. His article also extends the theoretical results of Park et al. (2009) for the bivariate case and constructs their estimator based on the minimization of the mean squared error, for the case when p, k and q are known. Park (2011) also proposes a criteria similar to the SBC to select p and q when they are assumed to be unknown. Lastly, Park (2011) performed simulation experiments and two real-data analyses to verify how well his approach is capable of estimating the true parameters and how the predictions of future values compare to those from a parametric transfer function model.

For many time series arising in economics and finance, the magnitude of the error component is associated with the magnitude of past errors, leading to heteroskedasticity. Park and Sriram (2017) considered an instance where Z_t is the square of a time series Y_t whose conditional mean is zero. Without specifying a model for Y_t, they assumed that there exists a $p \times 1$ parameter vector $\boldsymbol{\Phi}$ such that the conditional distribution of $Z_t | \boldsymbol{Z}_{t-1}$ is the same as that of $Z_t | \boldsymbol{\Phi}^T \boldsymbol{Z}_{t-1}$, where $\boldsymbol{Z}_{t-1} = (Z_{t-1}, \ldots, Z_{t-p})^T$ for some lag $p \geq 1$. Consequently, the

conditional variance of Y_t is some function of $\mathbf{\Phi}^\intercal \boldsymbol{Z}_{t-1}$. To estimate $\mathbf{\Phi}$, they proposed a robust estimation methodology based on Density Power Divergences (DPD) proposed by Basu et al. (1998). The DPD is indexed by a tuning parameter $\alpha \in [0, 1]$, which yields a continuum of estimators, $\{\widehat{\mathbf{\Phi}}_\alpha; \alpha \in [0, 1]\}$, where α controls the trade-off between robustness and efficiency of the DPD estimator. In a related article, Iaci and Sriram (2013) demonstrated that multivariate association measures based on this tuned DPD approach robustly recover both linear and nonlinear dependence between multiple sets of random vectors. For each α, Park and Sriram (2017) show that $\widehat{\mathbf{\Phi}}_\alpha$ is strongly consistent. They also developed data-dependent criteria for the selection of optimal α and lag p in practice. Furthermore, they illustrated the usefulness of their DPD methodology via simulation studies for ARCH-type models, where the errors are drawn from a gross-error contamination model and the conditional variance is a linear and/or nonlinear function of $\mathbf{\Phi}^\intercal \boldsymbol{Z}_{t-1}$. Finally, they analyzed the *Chicago Board Options Exchange Dow Jones Volatility Index data* and showed that their DPD approach yields viable models for the conditional variance, which are as good or superior to ARCH/GARCH models and two other divergence-based models in terms of *in-sample* and *out-of-sample* forecasts. A different approach for the same problem was introduced on Park and Samadi (2014), which estimates the unknown variance function by a Kernel smoother, analogously to the Park et al. (2009) estimation of the mean function .

Recently, Park and Samadi (2019) introduced the idea of applying this dimension reduction approach to both mean and variance functions of an univariate time series y_t. First, they demonstrate that their estimator based on Kullback-Lieber divergence is consistent in estimating the parameter matrix ($\mathbf{\Phi}$) that represents the linear combinations of the past values of y_t that contains all information about its conditional mean. Their approach is a special case of Luo et al. (2014) general paradigm of sufficient dimension reduction in regression. Park and Samadi (2019) assume that the series y_t is generated by the addition of a true conditional mean function and some White Noise term x_t. They prove that if x_t is truly observable, then their estimator would also be consistent to estimate the linear combination of past squared errors which contain all information about the conditional variance of y_t. For practical applications, the authors suggest using the mean model residuals as if they were the true error series.

Chapter 3

Research Problem and Main Goal

As defined earlier, let $\{x_t; t \geq 1\}$ represent a univariate time series and $\boldsymbol{X}_{t-1} = (x_{t-1}, ..., x_{t-p})^\intercal$ for some $p \geq 1$. We assume that there exists d linear combinations, $\{\Phi_1^T \boldsymbol{X}_{t-1}, ..., \Phi_d^T \boldsymbol{X}_{t-1}\}, d \leq p$ of $\mathbf{X}_{t-1}$, which contains all the necessary information about x_t that is available from $E(x_t|\boldsymbol{X}_{t-1})$. Specifically, we assume that for some $p \times d$ matrix $\boldsymbol{\Phi}_d = (\Phi_1, \ldots, \Phi_d)$ we have that

$$E(x_t|\mathbf{X}_{t-1}) = E(x_t|\boldsymbol{\Phi}_d^\intercal \mathbf{X}_{t-1}) = f(\boldsymbol{\Phi}_d^\intercal \mathbf{X}_{t-1}) \tag{3.1}$$

If $f(\cdot)$ is linear, then we would have a mean function similar to an autoregressive (AR) series of order p. However, as in Park et al. (2009), here f is assumed to be an unknown, possibly nonlinear, function.

Next, let $\varepsilon_t = x_t - f(\boldsymbol{\Phi}_d^\intercal \mathbf{X}_{t-1})$. We also assume the existence of $\widetilde{d}$ linear combinations of $\boldsymbol{\varepsilon}_{t-1}^2 = (\varepsilon_{t-1}^2, ..., \varepsilon_{t-q}^2)^\intercal$, for $\widetilde{d} \leq q$, that contains all the necessary information about x_t that is available from $V(x_t|\mathbf{X}_{t-1})$. Specifically, we assume that there exists a $q \times \widetilde{d}$ matrix $\boldsymbol{\Gamma}_{\widetilde{d}} = (\Gamma_1, \ldots, \Gamma_{\widetilde{d}})$ such that

$$V(x_t|\mathbf{X}_{t-1}) = E(\varepsilon_t^2|\boldsymbol{\varepsilon}_{t-1}^2) = E(\varepsilon_t^2|\boldsymbol{\Gamma}_{\widetilde{d}}^\intercal \boldsymbol{\varepsilon}_{t-1}^2) = g(\boldsymbol{\Gamma}_{\widetilde{d}}^\intercal \boldsymbol{\varepsilon}_{t-1}^2). \tag{3.2}$$

where g is assumed to be an unknown, possibly nonlinear, function. Therefore, the series $\{x_t\}$ is conditionally heteroskedastic by assumption.

As an example, consider the following model defined by

$$\begin{aligned} x_t &= 2 + (0.3\ x_{t-1} + 0.5\ x_{t-3}) + x_{t-2}^2 + \varepsilon_t, \\ \varepsilon_t &= \sqrt{1 + 0.2\ \varepsilon_{t-1}^2 + 0.1\ \varepsilon_{t-2}^2}\ e_t \end{aligned} \tag{3.3}$$

where $\{e_t\}$ is an i.i.d. sequence with $E(e_t) = 0 < E(e_t^2) = \sigma^2 < \infty$. For this model, the true parameter matrices are:

$$\boldsymbol{\Phi}_d = \begin{bmatrix} 0.3 & 0 \\ 0 & 1 \\ 0.5 & 0 \end{bmatrix},$$

and

$$\boldsymbol{\Gamma}_{\widetilde{d}} = [0.2, 0.1]^\intercal.$$

Consequently, $f(z_1, z_2) = 2 + z_1 + z_2^2$, and $g(z) = 1 + z$. Since $f(\cdot)$ and $g(\cdot)$ are unknown, both $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ are not identifiable. For instance, it is possible to find different choices for $\boldsymbol{\Phi}_d$ and $f(\cdot)$ which can yield the same mean function. To see this, let

$$\boldsymbol{\Phi}_{2,d} = \begin{bmatrix} 0 & 0.3 \\ 2 & 0 \\ 0 & 0.5 \end{bmatrix}$$

and $f_2(z_1, z_2) = 2 + (z_1/2)^2 + z_2$.

It is trivial to check that $E(x_t|\mathbf{X}_{t-1}) = f_2(\boldsymbol{\Phi}_{2,d}^\intercal \mathbf{X}_{t-1})$. Note that we can always find $\boldsymbol{\Phi}_{2,d}$ and $\boldsymbol{\Gamma}_{2,\widetilde{d}}$ such that its columns are normalized. Therefore, without loss of generality, from now on we refer to the true parameter matrices $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ as their normalized versions. Even though we cannot identify the parameter matrices, the space spanned by its columns is identifiable. See Luo et al. (2014) and Park and Samadi (2019) for a detailed discussion of identifiability for similar research problems.

We can define that

$$\boldsymbol{\Gamma}_{\widetilde{d}} = \underset{\boldsymbol{s}}{\operatorname{argmin}} \; E\left[\left(\varepsilon_t^2 - g(\boldsymbol{s}^\intercal \boldsymbol{\varepsilon}_{t-1}^2)\right)^2\right], \tag{3.4}$$

$$\text{and} \quad \boldsymbol{\Phi}_d = \underset{\boldsymbol{r}}{\operatorname{argmin}} \; E\left[\frac{\left(x_t - f(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1})\right)^2}{g(\boldsymbol{\Gamma}_{\widetilde{d}}^\intercal \boldsymbol{\varepsilon}_{t-1}^2)}\right]. \tag{3.5}$$

The main goal of this research is to find estimators $\widehat{\boldsymbol{\Phi}}_{d,n}$ and $\widehat{\boldsymbol{\Gamma}}_{\widetilde{d},n}$ which converge to $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ respectively as $n \to \infty$, without assuming a model for x_t.

Chapter 4

Estimation and Theoretical Results

In order to estimate the parameters $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ from the conditional mean function and the conditional variance function defined in (3.1) and (3.2), respectively, we need to first estimate the functions f and g nonparametrically. Section 4.1 defines the Nadaraya-Watson estimators for the functions f and g, respectively. Section 4.2 discusses an iterative estimation approach and defines the estimators $\widehat{\boldsymbol{\Phi}}_{d,n}$ and $\widehat{\boldsymbol{\Gamma}}_{\widetilde{d},n}$ of $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$, respectively. Henceforth, we will suppress the subscripts and denote $\widehat{\boldsymbol{\Phi}} = \widehat{\boldsymbol{\Phi}}_{d,n}$ and $\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\Gamma}}_{\widetilde{d},n}$. Theoretical consistency results for the two estimators along with all the required assumptions and Lemmas are stated in Section 4.3. Section 4.4 discusses sparse estimation of the parameter matrices, while Section 4.5 discusses simultaneous estimation of $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$. In Section 4.6, we consider the case when p, q, d, and $\widetilde{d}$ are unknown and describe selection/estimation of these quantities.

4.1 Nadaraya-Watson Estimator

In Chapter 3, we assumed that the mean function f and the variance function g were unknown and possibly nonlinear. We will now estimate these functions nonparametrically. Opsomer et al. (2001) raise caution when using conventional models in nonparametric regressiom where the errors are not independent. The authors illustrate that ignoring the dependence structure of the errors can mislead the data-driven methods [1] of selecting the bandwidth parameters to overfit the data. Moreover, they show that inappropriately choosing the bandwidth parameter may induce a spurious autocorrelation of the residuals. Hence, the nonparametric fitting is a delicate task in the context of our problem.

[1] e.g. Cross-Validation

We begin by describing a general Nadaraya-Watson estimator as in Nadaraya (1964) and Watson (1964). Suppose we have a random sample of multivariate data in which $\boldsymbol{y}$ is a n-dimensional response vector and $\boldsymbol{Z}$ is a $n \times d$ explanatory variables matrix satisfying $y_i = m(\boldsymbol{z}_i) + e_i$ for $i = 1, \ldots, n$, where $\boldsymbol{z}_i$ is the i^{th} row of $\boldsymbol{Z}$ and $m(\cdot)$ is an unknown function. Then, the Nadaraya-Watson estimator of the unknown regression function is defined as:

$$\widehat{m}_\lambda(\boldsymbol{z}_k) = \frac{\sum_{i=1}^{n} K\big(\boldsymbol{z}_k - \boldsymbol{z}_i, \boldsymbol{\lambda}\big) y_i}{\sum_{i=1}^{n} K\big(\boldsymbol{z}_k - \boldsymbol{z}_i, \boldsymbol{\lambda}\big)} \tag{4.1}$$

where K is a kernel function and $\lambda > 0$ is the bandwidth. In our approach, we use a Gaussian Kernel defined as:

$$K\left(\boldsymbol{z}_k - \boldsymbol{z}_i, \boldsymbol{a}_n\right) = \left(n \prod_{j=1}^{d} a_{nj}\right)^{-1} \prod_{j=1}^{d} G\left(\frac{z_{kj} - z_{ij}}{a_{nj}}\right) \tag{4.2}$$

where G is a univariate kernel function, $a_{nj} = s_j \left[\frac{4}{(d+2)n}\right]^{1/(4+d)}$, and s_j is the sample standard deviation of the j^{th} column of $\boldsymbol{Z}$. The choice of the bandwidth parameters is the same as in Park et al. (2009), following what was suggested in Silverman (1986) and Scott (1992). In the context of our problem, we estimate the mean function of x_t as:

$$\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) = \frac{\sum_{i=p+1}^{n} K\big(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1} - \boldsymbol{r}^\intercal \boldsymbol{X}_{i-1}, a_{nj}\big) x_i}{\sum_{i=p+1}^{n} K\big(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1} - \boldsymbol{r}^\intercal \boldsymbol{X}_{i-1}, a_{nj}\big)} \tag{4.3}$$

where $\boldsymbol{X}_{t-1} = (x_{t-1}, ..., x_{t-p})^\intercal$, $\boldsymbol{r}$ is a fixed $p \times d$ matrix and $t > p$. Similarly, we estimate the mean function of ε_t^2 defined in Section 3 as:

$$\widehat{g}_n(\boldsymbol{s}^\intercal \boldsymbol{\varepsilon}_{t-1}^2) = \frac{\sum_{i=p+q+1}^{n} K\big(\boldsymbol{s}^\intercal \boldsymbol{\varepsilon}_{t-1}^2 - \boldsymbol{s}^\intercal \boldsymbol{\varepsilon}_{i-1}^2, a_{nj}\big) \varepsilon_i^2}{\sum_{j=p+q+1}^{n} K\big(\boldsymbol{s}^\intercal \boldsymbol{\varepsilon}_{t-1}^2 - \boldsymbol{s}^\intercal \boldsymbol{\varepsilon}_{i-1}^2, a_{nj}\big)} \tag{4.4}$$

where $\boldsymbol{\varepsilon}_{t-1}^2 = (\varepsilon_{t-1}^2, ..., \varepsilon_{t-q}^2)^\intercal$ for $\widetilde{d} \leq q$, $\boldsymbol{s}$ is a fixed $q \times \widetilde{d}$ matrix and $t > p + q$.

4.2 Estimating Function

We adopt a iterative approach in estimation that first obtains a preliminary estimate of $\boldsymbol{\Phi}_d$ without taking the conditional variance into account. More specifically, our initial estimator

of $\boldsymbol{\Phi}_d$ is obtained by minimizing the objective function

$$\widetilde{S}_{n,0}(\boldsymbol{r}) = \sum_{t=p+1}^{n} \Big(x_t - \widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) \Big)^2 \tag{4.5}$$

with respect to $\boldsymbol{r}$ such that $\boldsymbol{r}^\intercal \boldsymbol{r} = \boldsymbol{I}_d$, where $\widehat{f}_n(\cdot)$ is the Nadaraya-Watson estimator defined in (4.3). We define our initial estimate $\widehat{\boldsymbol{\Phi}}_0$ as:

$$\widehat{\boldsymbol{\Phi}}_0 = \underset{\boldsymbol{r}}{\operatorname{argmin}} \sum_{t=p+1}^{n} \Big(x_t - \widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1}) \Big)^2, \quad \textit{such that} \quad \boldsymbol{r}^\intercal \boldsymbol{r} = \boldsymbol{I}_d, \tag{4.6}$$

where $\boldsymbol{r}^\intercal \boldsymbol{r} = \boldsymbol{I}_d$ is an identification condition. Notice that the solution of (4.6) is not unique as one can flip the sign of each column of $\widehat{\boldsymbol{\Phi}}_0$ and permute the columns. Regarding the theoretical analysis, one can avoid the ambiguity by imposing additional identification conditions. For example, we can set the first element of each column of $\boldsymbol{\Phi}$ to be positive. Besides, we can sort the columns of $\boldsymbol{\Phi}$ 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 always be achieved by replacing $\boldsymbol{\Phi}$ with $\boldsymbol{\Phi}\boldsymbol{P}$ where $\boldsymbol{P}$ is a signed column permutation matrix that satisfies $\boldsymbol{P}^\intercal \boldsymbol{P} = \boldsymbol{I}_d$. Empirically, this non-unique issue does not affect the performance of the proposed procedure, which is further corroborated via extensive numerical studies. We follow similar identification arguments for the other two estimators defined in this subchapter.

Next, we use the initial estimator of $\boldsymbol{\Phi}_d$ to compute the residuals

$$\widehat{\varepsilon}_t = x_t - \widehat{f}_n(\widehat{\boldsymbol{\Phi}}_0^\intercal \boldsymbol{X}_{t-1}). \tag{4.7}$$

The second step in our estimation consists of estimating the parameter vector $\boldsymbol{\Gamma}_{\widetilde{d}}$. This is done by minimizing the objective function

$$\widetilde{G}_n(\boldsymbol{s}) = \sum_{t=p+q+1}^{n} \Big(\widehat{\varepsilon}_t^2 - \widehat{g}_n(\boldsymbol{s}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) \Big)^2, \tag{4.8}$$

with respect to $\boldsymbol{s}$ such that $\boldsymbol{s}^\intercal \boldsymbol{s} = \boldsymbol{I}_{\widetilde{d}}$, where $\widehat{g}_n(\cdot)$ is the Nadaraya-Watson estimator defined in (4.4). We define our estimator of the variance parameter vector as:

$$\widehat{\boldsymbol{\Gamma}} = \underset{\boldsymbol{s}}{\operatorname{argmin}} \sum_{t=p+q+1}^{n} \left(\widehat{\varepsilon}_t^2 - \widehat{g}_n(\boldsymbol{s}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right)^2, \quad \textit{such that} \quad \boldsymbol{s}^\intercal \boldsymbol{s} = \boldsymbol{I}_{\widetilde{d}}, \tag{4.9}$$

where $s_{ij} \geq 0 \; \forall \; i \in [1, q], \; j \in [1, \widetilde{d}]$. Once again, the restriction that $\boldsymbol{s}^\intercal \boldsymbol{s} = \boldsymbol{I}_{\widetilde{d}}$ is necessary to guarantee identifiability. For obvious reasons, we also restrict all elements of $\boldsymbol{s}$ to be non-negative.

Finally, we propose a revised estimate of $\boldsymbol{\Phi}_d$ by minimizing a weighted sum of squares. More specifically, our revised estimator of the mean function parameter matrix $\boldsymbol{\Phi}_d$ is defined by:

$$\widehat{\boldsymbol{\Phi}} = \underset{\boldsymbol{r}}{\operatorname{argmin}} \sum_{t=p+q+1}^{n} \frac{\left(x_t - \widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1})\right)^2}{\widehat{g}_n(\widehat{\boldsymbol{\Gamma}}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)}, \quad \textit{such that} \quad \boldsymbol{r}^\intercal \boldsymbol{r} = \boldsymbol{I}_d. \tag{4.10}$$

where $\widehat{\boldsymbol{\Gamma}}$ is as defined in (4.9). Note that the estimator in (4.10) allows us to estimate $\boldsymbol{\Phi}_d$ while taking into account the presence of conditional heteroskedasticity. Additionally, we can replace $\widehat{\boldsymbol{\Phi}}_0$ in (4.7) with our revised estimate $\widehat{\boldsymbol{\Phi}}$ in (4.10) and obtain new set of fitted residuals. This then yields a revised estimate of $\boldsymbol{\Gamma}_{\widetilde{d}}$ via (4.9), which in turn yields a revised estimate of $\boldsymbol{\Phi}_d$ via (4.10). We can then iterate this estimation process until convergence of the estimates.

In order to determine the choices of $\boldsymbol{r}$ and $\boldsymbol{s}$ which will minimize the estimation functions defined above we use the *fmincon* function in *Matlab* since it is capable of handling multidimensional optimization under the necessary restrictions described in equations (4.6), (4.9), and (4.10). For additional discussion on the numerical optimization, see Chapter 5.

4.3 Consistency Theorems

4.3.1 Notations and assumptions

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 $\widetilde{d} \leq q$ 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 (4.3) with $\boldsymbol{r} = \boldsymbol{\Phi}_d$. Further, we define $\boldsymbol{\Phi}_0$ as the minimizer of the population counterpart of (4.6), i.e.

$$\boldsymbol{\Phi}_0 = \underset{\boldsymbol{r}}{\operatorname{argmin}} \; E\Big[\big(x_t - f(\boldsymbol{r}^\intercal \boldsymbol{X}_{t-1})\big)^2\Big].$$

Next, we state the assumptions for technical lemmas and the main theorems stated in this section.

Assumptions:

(A1) $(\boldsymbol{X}_t, \boldsymbol{\varepsilon}_t)$, $t \geq 0$ is strictly stationary and strong mixing with mixing coefficient $\alpha(m) \leq Am^{-\beta}$, 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 $\boldsymbol{\Gamma}_{\tilde{d}}^\intercal \boldsymbol{\varepsilon}_{t-1}^2$ are bonded and bounded away from zero on their supports which are closed intervals. Also, there is some $t^* < \infty$ such that for all $t \geq t^*$

$$\sup_{a_0, a_t} E\left(|\boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_0 \boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_t| \Big| \boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_0 = a_0, \; \boldsymbol{\Phi}_d^\intercal \boldsymbol{X}_t = a_t\right) p_t(a_0, a_t) < \infty$$
$$\text{and} \quad \sup_{b_0, b_t} E\left(|\boldsymbol{\Gamma}_{\tilde{d}}^\intercal \boldsymbol{\varepsilon}_0^2 \boldsymbol{\Gamma}_{\tilde{d}}^\intercal \boldsymbol{\varepsilon}_t^2| \; \Big| \boldsymbol{\Gamma}_{\tilde{d}}^\intercal \boldsymbol{\varepsilon}_0^2 = b_0, \; \boldsymbol{\Gamma}_{\tilde{d}}^\intercal \boldsymbol{\varepsilon}_t^2 = b_t\right) q_t(b_0, b_t) < \infty,$$

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 $\{\boldsymbol{\Gamma}_{\tilde{d}}^\intercal \boldsymbol{\varepsilon}_0^2, \; \boldsymbol{\Gamma}_{\tilde{d}}^\intercal \boldsymbol{\varepsilon}_t^2\}$.

(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

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

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 $\int uK(u)du = 0$, $\int u^2K(u)du < \infty$, and the Fourier transformation of $K(u)$ is absolutely integrable.

Remark 4.3.1 . 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.

4.3.2 Main results

Theorem 4.3.1 (Initial estimator). Suppose that assumptions A1 – A5 hold and choose $a_n = O(n^{-1/5})$. The initial estimator defined in (4.6) satisfies

$$\|\widehat{\boldsymbol{\Phi}}_0 - \boldsymbol{\Phi}_0\| = O_p\left(\left[\frac{\ln(n-p)}{n-p}\right]^{2/5}\right) \tag{4.11}$$

with probability approaching one as $n \to \infty$.

Theorem 4.3.2 (Variance estimator). Suppose that assumptions A1 – A5 hold and choose $a_n = O(n^{-1/5})$. The variance parameter vector estimator defined in (4.9) satisfies

$$\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_{\widetilde{d}}\| = O_p\left(\left[\frac{\ln(n-p-q)}{n-p-q}\right]^{2/5}\right) \tag{4.12}$$

with probability approaching one as $n \to \infty$.

Theorem 4.3.3 (Final estimator). Suppose that assumptions A1 – A5 hold and choose $a_n = O(n^{-1/5})$. The final estimator defined in (4.10) satisfies

$$\|\widehat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_d\| = O_p\left(\left[\frac{\ln(n-p)}{n-p}\right]^{2/5}\right) \tag{4.13}$$

with probability approaching one as $n \to \infty$.

Remark 4.3.2 . Theorems 4.3.1–4.3.3 deliver the convergence in probability results for the estimators proposed on Section 4.2. When p and q are fixed, the rates in Theorems 4.3.1–4.3.3 follow $(n^{-1}\ln n)^{2/5}$ which is the optimal nonparametric rate for i.i.d. univariate data proved in Stone (1982). Our results are obtained with multivariate and strong-mixing data. The proof of the above theorems are presented in an Appendix given at the end of the dissertation.

4.3.3 Technical lemmas

Lemma 4.3.1 . Suppose that assumptions A1 – A5 hold and choose $a_n = O(n^{-1/5})$. Then, the Nadaraya-Watson estimators defined in (4.3) and (4.4) satisfy

$$\sup_{\boldsymbol{r}\in\mathbb{R}^d, \boldsymbol{X}\in\mathbb{R}^d} |\widehat{f}_n(\boldsymbol{r}^\intercal \boldsymbol{X}) - f(\boldsymbol{r}^\intercal \boldsymbol{X})| = O_p\left(\left[\frac{\ln(n-p)}{n-p}\right]^{2/5}\right)$$

$$\text{and} \quad \sup_{\boldsymbol{s}\in\mathbb{R}^{\tilde{d}}, \boldsymbol{\varepsilon}^2\in\mathbb{R}^{\tilde{d}}} |\widehat{g}_n(\boldsymbol{s}^\intercal \boldsymbol{\varepsilon}^2) - g(\boldsymbol{s}^\intercal \boldsymbol{\varepsilon}^2)| = O_p\left(\left[\frac{\ln(n-p-q)}{n-p-q}\right]^{2/5}\right).$$

Remark 4.3.3 . Lemma 4.3.1 establishes the rate for uniform convergence in probability for the Nadaraya-Watson estimators proposed in (4.3) and (4.4). When p and q are fixed, the rates in Lemma 4.3.1 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 4.3.1 is similar to the proof of Theorem 8 in Hansen (2008), and hence we omit the proof.

Denote $f^{(1)}(\boldsymbol{r}^\intercal \boldsymbol{x})$ and $g^{(1)}(\boldsymbol{s}^\intercal \boldsymbol{\varepsilon}^2)$ the first order derivatives of $f(\boldsymbol{r}^\intercal \boldsymbol{x})$ and $g(\boldsymbol{s}^\intercal \boldsymbol{\varepsilon}^2)$, respectively. Let $\widehat{f}_n^{(1)}(\boldsymbol{r}^\intercal \boldsymbol{x})$ and $\widehat{g}_n^{(1)}(\boldsymbol{s}^\intercal \boldsymbol{\varepsilon}^2)$ be the first order derivative estimators which are defined

as

$$\widehat{f}_n^{(1)}(\boldsymbol{r}^\intercal \boldsymbol{x}) = \frac{\sum_{i=p+1}^{n} K^{(1)}\left(\frac{\boldsymbol{r}^\intercal \boldsymbol{x} - \boldsymbol{r}^\intercal \boldsymbol{X}_{i-1}}{a_n}\right) x_i}{a_n \sum_{i=p+1}^{n} K\left(\frac{\boldsymbol{r}^\intercal \boldsymbol{x} - \boldsymbol{r}^\intercal \boldsymbol{X}_{i-1}}{a_n}\right)} \tag{4.14}$$

$$\text{and} \quad \widehat{g}_n^{(1)}(\boldsymbol{s}^\intercal \boldsymbol{\varepsilon}^2) = \frac{\sum_{i=p+q+1}^{n} K^{(1)}\left(\frac{\boldsymbol{s}^\intercal \boldsymbol{\varepsilon}^2 - \boldsymbol{s}^\intercal \boldsymbol{\varepsilon}_{i-1}^2}{a_n}\right) \varepsilon_i}{a_n \sum_{i=p+q+1}^{n} K\left(\frac{\boldsymbol{s}^\intercal \boldsymbol{\varepsilon}^2 - \boldsymbol{s}^\intercal \boldsymbol{\varepsilon}_{i-1}^2}{a_n}\right)}, \tag{4.15}$$

where $K^{(1)}(\cdot)$ is the first order derivative of the kernel function $K(\cdot)$.

Lemma 4.3.2 . Suppose that assumptions A1 – A5 hold and choose $a_n = O(n^{-1/5})$. We have

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

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

with probability 1 as $n - p - q \to \infty$.

Remark 4.3.4 . Lemma 4.3.2 provides uniform consistency results for the first order derivative estimators defined in (4.14) and (4.15). The proof of Lemma 4.3.2 directly follows the proof of Theorem 2 in Mack and Müller (1989), and hence we omit the proof.

4.4 Sparse Estimation for the one-dimensional case

A natural question when estimating parameters is distinguishing which ones are statistically different from zero. In many cases, it is considered good practice to remove insignificant variables from the final model. Our problem however, assumes no model, thus the estimated parameters cannot be tested as different from zero or not.

When $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ are vectors ($d = \widetilde{d} = 1$), we propose to check whether sparse versions of $\widehat{\boldsymbol{\Phi}}$ and $\widehat{\boldsymbol{\Gamma}}$ can be more accurate than the estimators defined in Section 4.2. The goal of sparse estimation is to replace small magnitude elements of $\widehat{\boldsymbol{\Phi}}$ (or $\widehat{\boldsymbol{\Gamma}}$) by zero, which might result in a more accurate estimate, when the true parameter vector contains null elements.

We propose to use cross-validation for selecting which elements should be replaced by zero. The first step of Sparse Estimation is to split the time series into two parts, so we define some time t_k s.t. $p+q+1 < t_k < n$. The observations before t_k are taken to be our training data and the remainder is used as test data. Then, $\widehat{\boldsymbol{\Phi}}$ and $\widehat{\boldsymbol{\Gamma}}$ are estimated using only the training portion.

Secondly, we define possible versions of the estimated parameter matrices where some of the elements are replaced by zero. For the mean function case, since $\boldsymbol{\Phi}_d$ has p elements, we create p different versions of $\widehat{\boldsymbol{\Phi}}$ by excluding the smallest magnitude elements one at a time, replacing them by zero and re-normalizing the resulting vector. Similarly, we can also find q versions of $\widehat{\boldsymbol{\Gamma}}$. For instance, let $\widehat{\boldsymbol{\Phi}} = (1/\sqrt{30})(1,-2,5)^\intercal$. Then we would have three possible candidates as the sparse estimate: $(1/\sqrt{30})(1,-2,5)^\intercal$, $(1/\sqrt{29})(0,-2,5)^\intercal$, and $(1/\sqrt{25})(0,0,5)^\intercal$.

Lastly, we use only the test dataset to compute the total squared prediction error for each one of the candidates following equations (4.5) and (4.8). The final Sparse estimates $\widehat{\boldsymbol{\Phi}}_s$ and $\widehat{\boldsymbol{\Gamma}}_s$ are chosen by smallest out-of-sample prediction errors, out of all p (or q) candidates.

This framework cannot be extended to the multidimensional case, since simply substituting elements by zero can violate the columns orthogonality assumption.

4.5 Simultaneous Estimation

In this dissertation, we also investigate the possibility of adopting a simultaneous estimation approach, instead of the iterative estimation approach presented in Section 4.2. Here we propose a new estimating function, similar to that defined in (4.10), but minimization is performed simultaneously.

More specifically, define $\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\phi}_1 \\ \vdots \\ \boldsymbol{\phi}_d \\ \\ \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_{\widetilde{d}} \end{pmatrix}$.

where $\boldsymbol{\phi}_i$ is the i^{th} column of $\boldsymbol{\Phi}_d$ and $\boldsymbol{\gamma}_i$ is the i^{th} column of $\boldsymbol{\Gamma}_{\widetilde{d}}$. Based on this vector of parameters, we can rewrite the mean function of x_t as $\mu_t(\boldsymbol{\Omega})$ and the variance function as $h_t(\boldsymbol{\Omega})$. Now, let us consider the joint estimation of $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$, which is equivalent to the estimation of $\boldsymbol{\Omega}$ by minimizing the following estimating function:

$$T_n(\mathbf{u}) = \sum_{t=p+q+1}^{n} \frac{(x_t - \widehat{\mu}_t(\boldsymbol{u}))^2}{\widehat{h}_t(\boldsymbol{u})/S_{\widehat{h}}(\boldsymbol{u})} \tag{4.16}$$

where $\boldsymbol{u}$ is a vector with $pd + q\widetilde{d}$ elements, $S_{\widehat{h}}(\boldsymbol{u}) = \sum_{t=p+q+1}^{n} \widehat{h}_t(\boldsymbol{u})$. The denominator in (4.16) is divided by $S_{\widehat{h}}(\boldsymbol{u})$, so we have relative weights. Hence, our estimator is based on the weighted least squares rationale, where each data point should have an appropriate amount of relative influence over the weigthed sum of squares. It is worth noticing that in our iterative estimation approach, this scaling is not necessary because the denominator elements are fixed before minimizing the final objective function. Nevertheless, if we did scale the weights in equation (4.10), the optimization would still be the same, as we would be just dividing the objective function by a constant term.

Now, in order to estimate $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ simultaneously, let us rewrite our simultaneous objective function in (4.16) as

$$T_n(\boldsymbol{r}, \boldsymbol{s}) = \sum_{t=p+q+1}^{n} \left[\frac{(x_t - \widehat{f}(\boldsymbol{r}^\intercal \mathbf{X}_{t-1}))^2 \left[\sum_{j=p+q+1}^{n} \widehat{g} \left(\boldsymbol{s}^\intercal \begin{pmatrix} (x_{j-q} - \widehat{f}(\mathbf{r}^\intercal \mathbf{X}_{j-q-1}))^2 \\ \vdots \\ (x_{j-1} - \widehat{f}(\mathbf{r}^\intercal \mathbf{X}_{j-1-1}))^2 \end{pmatrix} \right) \right]}{\widehat{g} \left(\boldsymbol{s}^\intercal \begin{pmatrix} (x_{t-q} - \widehat{f}(\mathbf{r}^\intercal \mathbf{X}_{t-q-1}))^2 \\ \vdots \\ (x_{t-1} - \widehat{f}(\mathbf{r}^\intercal \mathbf{X}_{t-1-1}))^2 \end{pmatrix} \right)} \right] \tag{4.17}$$

where $\widehat{f}(\cdot)$ and $\widehat{g}(\cdot)$ are the Nadaraya-Watson function estimates. Hence, the simultaneous estimates of $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ are:

$$(\widehat{\boldsymbol{\Phi}}_m, \widehat{\boldsymbol{\Gamma}}_m) = \underset{\boldsymbol{r},\boldsymbol{s}}{\operatorname{argmin}}\, T_n(\boldsymbol{r}, \boldsymbol{s}), \quad \textit{such that} \quad \boldsymbol{r}^\intercal \boldsymbol{r} = \boldsymbol{I}_d \quad \textit{and} \quad \boldsymbol{s}^\intercal \boldsymbol{s} = \boldsymbol{I}_{\widetilde{d}}. \tag{4.18}$$

4.6 Selection of Lag Parameters and Dimension of Matrices

4.6.1 Modified Schwarz Bayesian Information Criterion (MSBC)

Unlike in regression, there may not be any prior information on the number of lags p and/or q. The lag selection problem is a common task in time series modeling. The Akaike Information Criterion (AIC) [Akaike (1974)] and the Schwarz Bayesian Information Criterion (SBC) [Schwarz (1978)] are commonly used for lag selection in time series models. See Ng and Perron (2005) for a review and comparison of some most popular criteria.

We aim to develop a data-dependent method for estimating p and q, when both are unknown. In the time series literature, different modifications of the SBC criterion have been proposed for different reasons [see Hannan and Quinn (1979), Broman and Speed (2002) and Zhu et al. (2006)]. As in Park et al. (2009), for fixed d and $\widetilde{d}$, we propose a Modified SBC (MSBC) criterion to estimate the lag parameters:

$$\widehat{p} = \underset{p}{\operatorname{argmin}} \left\{ (n-p) \ln \left[\widetilde{S}_{n,0}(\widehat{\boldsymbol{\Phi}})/(n-p) \right] \right\} + d^2 p \ln (n-p), \tag{4.19}$$

and

$$\widehat{q} = \underset{q}{\operatorname{argmin}} \left\{ (n-p-q) \ln \left[\widetilde{G}_{n}(\widehat{\boldsymbol{\Gamma}})/(n-p-q) \right] \right\} + \widetilde{d}^2 q \ln (n-p-q). \tag{4.20}$$

where $\widetilde{S}_{n,0}(\widehat{\boldsymbol{\Phi}})$ and $\widetilde{G}_{n}(\widehat{\boldsymbol{\Gamma}})$ are as defined in (4.5) and (4.8), respectively.

Similarly, for fixed p and q we propose estimators for d and $\widetilde{d}$:

$$\widehat{d} = \underset{d}{\operatorname{argmin}} \left\{ (n-p) \ln \left[\widetilde{S}_{n,0}(\widehat{\boldsymbol{\Phi}})/(n-p) \right] \right\} + d^2 p \ln (n-p), \tag{4.21}$$

and

$$\widehat{\widetilde{d}} = \underset{\widetilde{d}}{\operatorname{argmin}} \left\{ (n-p-q) \ln \left[\widetilde{G}_{n}(\widehat{\boldsymbol{\Gamma}})/(n-p-q) \right] \right\} + \widetilde{d}^2 q \ln (n-p-q). \tag{4.22}$$

4.6.2 Selection by Cross Validation

In Section 4.4, we introduced the notion of sparse estimation when the lags are known, and the parameter matrices $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ are one-dimensional. Here, we demonstrate how the sparse estimator can also be used to select the lags.

To this end, let us consider the mean function first. Our strategy consists of choosing an arbitrarily large value for the length of $\boldsymbol{\Phi}_d$, namely p_{max}. Then, we compute the sparse estimator $\widehat{\boldsymbol{\Phi}}_s$ as a vector of p_{max} dimension. This estimate will assign zeros to the elements of the mean vector based on cross-validation prediction error as described in Section 4.4. We then observe what is the largest lag with a non-zero element to choose as $\widehat{p}_{CV}$. For example, if $\widehat{\boldsymbol{\Phi}}_s = (1, 0, 0, 1, 0, 0)^\intercal/\sqrt{2}$, then $\widehat{p}_{CV} = 4$. We expect that if the sparse estimator can select the appropriate null elements, then all the unnecessary lags beyond the true p should also be estimated as zeros. Similarly, $\widehat{q}_{CV}$ can be chosen according to the largest lag with a non-zero element of $\widehat{\boldsymbol{\Gamma}}_s$.

Chapter 5

Computational Details and Reparametrization

In order to minimize the objective functions defined in Section 4, our approach requires a multidimensional numerical optimization under a non-linear constraint, for example, $\boldsymbol{\Phi}_d^T \boldsymbol{\Phi}_d = \boldsymbol{I}_d$. It should be pointed out that this restriction poses challenges for the numerical search over the parameter space. For instance, let us assume that $\boldsymbol{\Phi}_d = (\phi_1, \phi_2)^T$. The restriction would imply that the norm of the mean parameter vector equals to one. We now draw the parameter space for $\boldsymbol{\Phi}_d$; see Figure 5.1. The restricted parameter space is exactly the boundary of the circle represented by the solid line. Therefore, any numerical search is extremely delicate, as it requires an algorithm capable of not wandering off to the inside or outside of the circle.

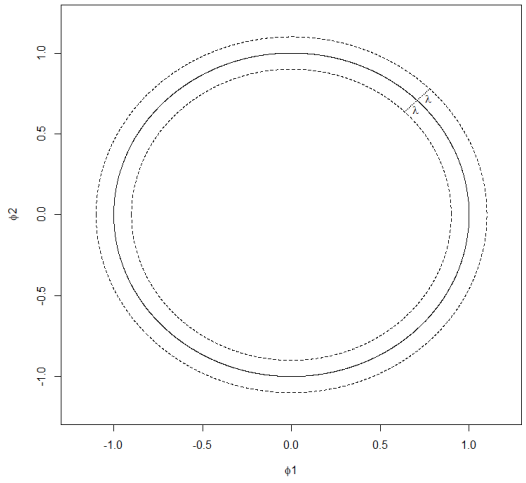

Figure 5.1: Parameter space for $\boldsymbol{\Phi}_d$ when $p = 2$ and $d = 1$.

In practice, we can define a small tolerance value as to how close we want to satisfy the restriction (represented by λ in Figure 5.1), so the algorithm can examine points between the two dashed lines around the solid line. This approach is problematic from the computational perspective because of the identifiability issue, as it is possible to find infinite number of equivalent vectors inside the tolerance area. Needless to say, the computational problems are compounded for the case when the number of lags is larger than 2, where the appropriate parameter space would be the boundary of a hyper-sphere.

A major computational problem arises when we consider parameter matrices with more than one column. Not only we require that each column has a unit norm, but also need the columns to be orthogonal to each other. In this case, the parameter space cannot be actually drawn, and the numerical search becomes much more challenging. Park and Samadi (2019)[2] mention that the Sequential Quadratic Programming algorithm is capable of incorporating this kind of constraints into the optimization problem. However, in practice, adopting this algorithm does not solve the convergence issues found when running the simulation codes of Park et al. (2009).

Based on the aforementioned computational challenges, we propose a new parametrization in order to guarantee that the numerical optimization is more efficient. The advantages of this new representation are that: 1) it ensures that the constraints imposed are fully met, 2) it makes convergence more frequent, 3) it reduces the computational time, and 4) it makes the optimization feasible to a wider range of algorithms and software.

5.1 Angular Representation for Parameter Vectors

Let $\boldsymbol{\Theta} = (\theta_1, ..., \theta_p)^\intercal$. Without loss of generality, let us assume that $\theta_k > 0$ for some $k \in \{1, ..., p\}$. Now, for θ_j, $j \in \{1, ..., p\}$, but $j \neq k$, we define the angle α_j as the following:

$$\tan(\alpha_j) = \frac{\theta_j}{\theta_k}. \tag{5.1}$$

Figure 5.2 illustrates the formation of the angle α_j depending on the value of θ_j. By knowing this angle, we know how large θ_j is compared to θ_k and also its sign.

[2]In fact, the authors only present estimation results for the case when $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ are vectors.

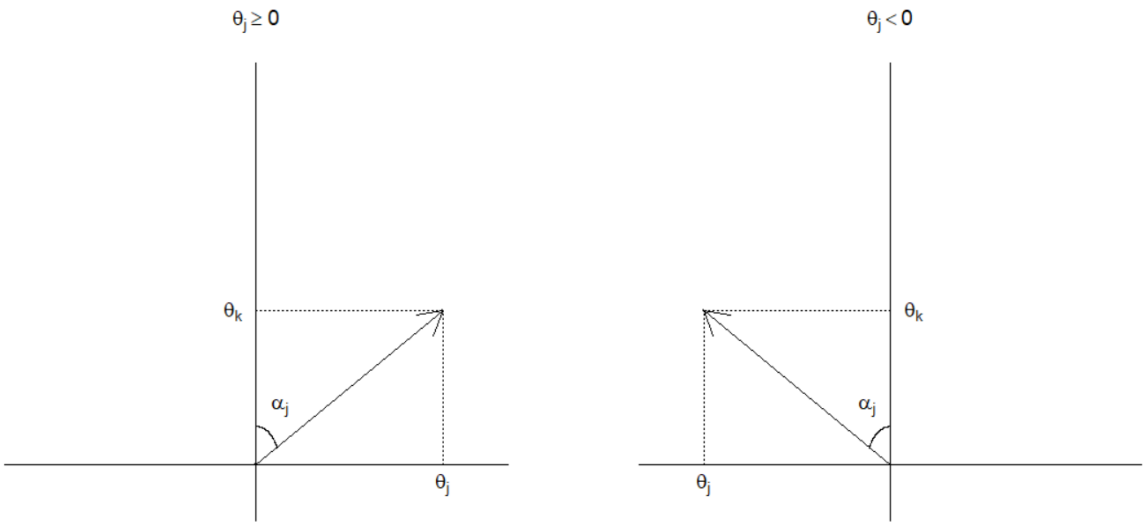

Figure 5.2: Angle between two elements of a vector.

From (5.1), note that we can write θ_j as:

$$\theta_j = \theta_k \tan(\alpha_j). \tag{5.2}$$

For the case considered here, the orthonormality restriction reduces to $\boldsymbol{\Theta}^\intercal\boldsymbol{\Theta} = 1$, thus

$$\theta_k^2 + \sum_{j \neq k} \theta_j^2 = 1 \tag{5.3}$$

According to (5.2), we have that

$$\begin{aligned} \theta_k^2 + \sum_{j \neq k} \left\{ [\theta_k \tan(\alpha_j)]^2 \right\} &= 1 \\ \theta_k^2 \left[1 + \sum_{j \neq k} \left\{ \tan(\alpha_j)^2 \right\} \right] &= 1 \\ \theta_k^2 = \frac{1}{\left[1 + \sum_{j \neq k} \left\{ \tan(\alpha_j)^2 \right\} \right]}&. \end{aligned} \tag{5.4}$$

Since $\theta_k > 0$ by assumption, we have

$$\theta_k = \frac{1}{\left[1 + \sum_{j \neq k} \left\{ \tan(\alpha_j)^2 \right\}\right]^{1/2}} \tag{5.5}$$

Based on the reparametrization (5.5), we can now freely choose every α_j for $j \neq k$, and the resulting vector $\boldsymbol{\Theta}$ will have unit length. Conversely, every unit length vector, for which some $\theta_k > 0$, can be represented by a $(p-1)$-dimensional vector of angles for fixed k.

In our research problem, the assumption that some $\theta_k > 0$ is not problematic since our nonparametric estimator is scale invariant. In this sense, for our estimation approach, we only require that $\theta_k \neq 0$ for some k, which is always guaranteed for vectors with positive length.

The main advantage of using this angular representation of our parameter vector relies on the computational aspect. The unit length constraint is non-linear and requires an optimization algorithm capable of handling this non-linearity. Furthermore, the parameter space based on this restriction is a p-dimensional hyper-sphere with unit radius in which only the boundary points are feasible. Therefore, even if we start the algorithm from an appropriate initial vector, it is harder to enforce that every upcoming search step will satisfy the constraint. However, by using the angular representation, we can fix some k and optimize our function based on the $(p-1)$-dimensional vector of angles that can freely vary inside a hypercube such that $-\frac{\pi}{2} < \alpha_j < \frac{\pi}{2}$ for $j \neq k$. This approach allows us to guarantee a vector of unit length, regardless of the choice of the angles. This representation also allows us to use a wider class of optimization algorithms, since we only require simple boundary restrictions.

In practical applications, if k is unknown, we can run optimizations for $k = 1, .., p$ and select a setting that performs the best. We should note that, for computation purposes, even if we select a wrong value of k for which $\theta_k = 0$, it is still possible to get a numerical approximation of the true parameter vector. For example, suppose we want to represent the vector $\boldsymbol{\Theta} = (\frac{1}{\sqrt{2}})(1, 0, 1)^\intercal$ using the proposed angular parametrization, and we end up choosing $k = 2$. The original vector could not be theoretically reparametrized, but if we choose α_1 and α_3 to be very close to $\pi/2$, then we would obtain a close approximation of $\boldsymbol{\Theta}$ where θ_2 is close to zero, and the remaining elements are close to $\frac{1}{\sqrt{2}}$.

5.2 Angular Representation for Parameter Matrices

We now consider a $p \times d$ parameter matrix $\mathbf{\Theta}$ such that $d \leq p$:

$$\mathbf{\Theta} = \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1d} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{p1} & \theta_{p2} & \cdots & \theta_{pd} \end{bmatrix}$$

Let $\mathbf{\Theta}_i$ denote the i^{th} column of $\mathbf{\Theta}$ and assume that $\mathbf{\Theta}^\intercal\mathbf{\Theta} = \mathbf{I}_d$, i.e. for $i = 1, ..., d$ and $i \neq i'$:

$$\mathbf{\Theta}_i^\intercal\mathbf{\Theta}_i = 1 \quad \& \quad \mathbf{\Theta}_i^\intercal\mathbf{\Theta}_{i'} = 0. \tag{5.6}$$

We require the existence of a vector $(k_1, ..., k_d)^\intercal$ such that $\theta_{k_i i} \neq 0$ & $k_i \neq k_{i'}$ for $i \neq i'$. If this vector does not exist, then $\mathbf{\Theta}$ cannot be of full rank, and thus orthonormal. Since our estimation procedure is scale invariant, we can assume without loss of generality that $\theta_{k_i i} > 0$. Then

$$\begin{aligned} \frac{\theta_{ji}}{\theta_{k_i i}} &= \tan(\alpha_{ji}). \\ \theta_{ji} &= \theta_{k_i i} \tan(\alpha_{ji}). \end{aligned} \tag{5.7}$$

Since $\sum_{j=1}^{p} \theta_{ji}^2 = 1$, then from (5.5) we have that

$$\theta_{k_i i} = \frac{1}{\left[1 + \sum_{j \neq k_i} \left\{ \tan(\alpha_{ji})^2 \right\}\right]^{1/2}} = \frac{1}{\left[\sum_{j=1}^{p} \left\{ \tan(\alpha_{ji})^2 \right\}\right]^{1/2}}. \tag{5.8}$$

Let $C_i = \left[\sum_{j=1}^{p} \left\{ \tan(\alpha_{ji})^2 \right\}\right]^{1/2}$. From (5.7) and (5.8), we can write

$$\theta_{ji} = \frac{\tan(\alpha_{ji})}{C_i}. \tag{5.9}$$

In order to guarantee orthogonality between the columns of $\mathbf{\Theta}$, it is sufficient to ensure that the i^{th} column is orthogonal to the previous columns for $i = 1, ..., d$. Formally, we want $\mathbf{\Theta}_i \perp \mathbf{\Theta}_{i'}$ for $i' < i$, i.e.:

$$
\begin{aligned}
\sum_{j=1}^{p} \theta_{ji}\theta_{ji'} &= 0 \\
\theta_{k_{i'}i}\theta_{k_{i'}i'} + \sum_{j\neq k_{i'}} \theta_{ji}\theta_{ji'} &= 0 \\
\theta_{k_{i'}i} &= \frac{-\sum_{j\neq k_{i'}} \theta_{ji}\theta_{ji'}}{\theta_{k_{i'}i'}}
\end{aligned}
\tag{5.10}
$$

From (5.9), we can rewrite (5.10) as:

$$
\begin{aligned}
\frac{\tan(\alpha_{k_{i'}i})}{C_i} &= \frac{\frac{-\sum_{j\neq k_{i'}} \tan(\alpha_{ji})\tan(\alpha_{ji'})}{C_iC_{i'}}}{\frac{1}{C_{i'}}} \\
\tan(\alpha_{k_{i'}i}) &= -\sum_{j\neq k_{i'}} \tan(\alpha_{ji})\tan(\alpha_{ji'}).
\end{aligned}
\tag{5.11}
$$

Since $\tan(\alpha_{k_{i'}i'}) = 1$, we have that (5.11) is equivalent to:

$$
\begin{aligned}
\sum_{j=1}^{p} \tan(\alpha_{ji})\tan(\alpha_{ji'}) &= 0 \\
\sum_{j=k_1,\ldots,k_{i-1}} \tan(\alpha_{ji})\tan(\alpha_{ji'}) &= -\sum_{j\neq k_1,\ldots,k_{i-1}} \tan(\alpha_{ji})\tan(\alpha_{ji'})
\end{aligned}
\tag{5.12}
$$

Thus, we can expand (5.12) as a system of $(i-1)$ linear equations:

$$
\begin{cases}
\sum_{j=k_1,\ldots,k_{i-1}} \tan(\alpha_{ji})\tan(\alpha_{j1}) = -\sum_{j\neq k_1,\ldots,k_{i-1}} \tan(\alpha_{ji})\tan(\alpha_{j1}) \\
\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
\sum_{j=k_1,\ldots,k_{i-1}} \tan(\alpha_{ji})\tan(\alpha_{ji-1}) = -\sum_{j\neq k_1,\ldots,k_{i-1}} \tan(\alpha_{ji})\tan(\alpha_{ji-1}).
\end{cases}
\tag{5.13}
$$

Therefore, we can sequentially define each column of $\mathbf{\Theta}$ by ensuring that the elements $(\tan(\alpha_{k_1i}), ..., \tan(\alpha_{k_{i-1}i}))$ are the solution to the system of $(i-1)$ linear equations described by (5.13). By using this representation, we have a total of $dp - \frac{d(d+1)}{2}$ angles to be optimized such that each angle can freely vary within $(-\frac{\pi}{2}, \frac{\pi}{2})$, while the other angles are already theoretically defined. Hence, for a fixed vector $(k_1, ..., k_d)^\intercal$, the numerical optimization can

happen by considering a much smaller number of parameters than if we had to work with all $d \times p$ elements of $\boldsymbol{\Theta}$ simultaneously, while guaranteeing the desired orthonormality for each search point.

In practical applications, it might be the case that $\boldsymbol{\Theta}$ has some sparse vectors, so it is good practice to try different possibilities for $(k_1, ..., k_d)^\intercal$ and compare which vector choice provided the best optimization. Similar to the vector case, even if we choose the wrong k_i elements, it is still possible to find a numerically close approximation to the true parameter matrix.

The relevance of this angular approach can be seen when we try to replicate the results of Park et al. (2009). In supplemental material of their article, we can run the "accuracy.m" script on Matlab in order to compute accuracy measures for simulated data based on Model 3, where there is a 6×3 parameter matrix. We note that the reported average measures are based only on the Monte Carlo replications which converged. In our replication of their results, we observe that around 15% of replicates were excluded, which may be problematic for real data applications where you cannot afford to just ignore the estimates that did not converge. Another major problem is the actual validity of the results from simulation experiments. By just excluding replicates with problematic data, the reported results could be biased.

We attempted to reproduce the results of Park et al. (2009) under the proposed reparametrization to assess if it could improve on their numerical convergence issues. All of our simulation replicates through this angular representation converged and the computed average accuracy measures are also roughly the same as those reported on the original paper, but without excluding any potentially problematic series.

This representation also enables us to sample random orthonormal matrices to be used as initial points on the optimization algorithm. For instance, the codes found on the supplemental material of Park et al. (2009) use initial points by combining d random normalized vectors, which are not orthogonal to each other, so optimization starts outside the proper parameter space.

The proposed parametrization is applicable to any optimization problem which requires the parameter matrix to be orthonormal. Therefore, it can be helpful for those who intend

to apply, replicate, or extend the estimation procedure from this thesis and most of our main references such as : Park et al. (2009), Park et al. (2010), Park (2011), Park and Sriram (2017) and Park and Samadi (2019). Naturally, it could also be applied beyond the Time Series area, for research problems with similar restriction setup. For instance, in the area of Design of Experiments, we can use this parametrization to sample random orthonormal designs or to numerically search for a design matrix that minimizes some criterion.

Chapter 6

Simulation Studies and Results

In this chapter, we investigate the accuracy of our proposed estimators via simulation studies. We perform these studies for different models and sample sizes. Our goal is to understand how fast the estimators of the parameter matrices converge to their true values.

Here, we simulate data from the following model with different mean and variance functions; see Section 6.1 for details:

$$\begin{aligned} x_t &= f(\boldsymbol{\Phi}' \boldsymbol{X}_{t-1}) \ + \varepsilon_t, \\ \varepsilon_t &= \left[\sqrt{g(\boldsymbol{\Gamma}' \boldsymbol{\varepsilon}_{t-1}^2)} \right] e_t. \end{aligned} \tag{6.1}$$

where $e_t \overset{i.i.d.}{\sim} N(0,1)$.

We consider simulation models where $f(\cdot)$ and $g(\cdot)$ are nonlinear and the number of lags are relatively large. This allows us to assess the performance of our estimation approach for more complex models. The numerical optimizations to estimate $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\tilde{d}}$ are done by applying the *fmincon* function in *Matlab* to 50 randomly generated initial values. We use this strategy to make sure that our search over the parameter space is more thorough, thus avoiding getting stuck on possible local minimum values. The optimizations are done based on the parametrization proposed on Chapter 5, so the process will carry the described computational advantages.

In all our simulation studies, we use two measures to assess the accuracy of our estimates. The first measure we use is the *Vector Correlation Coefficient* proposed by Ye and Weiss

(2003); also see Hotelling (1936):

$$\rho = |\widehat{\boldsymbol{\Theta}}^{\intercal}\boldsymbol{\Theta}\boldsymbol{\Theta}^{\intercal}\widehat{\boldsymbol{\Theta}}|^{1/2}, \tag{6.2}$$

where $\boldsymbol{\Theta}$ is the true matrix, $\widehat{\boldsymbol{\Theta}}$ is the estimated matrix and $|\boldsymbol{A}|$ denotes the determinant of a matrix $\boldsymbol{A}$. Note that $0 \leq \rho \leq 1$, where higher values of ρ imply that the estimated vector is closer to the true parameter value. Another measure that we use to assess the accuracy of our estimated vectors is based on Xia et al. (2002):

$$m^2 = \left|\left|(I - \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Theta}}^{\intercal})\boldsymbol{\Theta}\right|\right|^2 \tag{6.3}$$

where $||a||^2$ is the Euclidean norm of a vector a and $0 \leq m^2 \leq 1$. This measure, however, approaches zero when the estimated vector is closer to the true value of the parameter value. In our research problem, $\boldsymbol{\Theta} = \boldsymbol{\Phi}_d$ or $\boldsymbol{\Gamma}_{\widetilde{d}}$ and $\widehat{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Phi}}$ or $\widehat{\boldsymbol{\Gamma}}$.

6.1 Models with Parameter Vectors

First, we list five different time series models from which we generate observations:

Model 1:

$$\begin{aligned} x_t &= (1/\sqrt{3})(x_{t-1} + x_{t-3} + x_{t-6}) + \varepsilon_t \\ \varepsilon_t &= \left[\sqrt{h_t}\right] e_t \\ h_t &= 1 + (1/6)(\varepsilon_{t-1}^2 + \varepsilon_{t-2}^2) \\ e_t &\sim N(0,1) \\ \boldsymbol{\Phi}_d &= \left[\frac{1}{\sqrt{3}}, 0, \frac{1}{\sqrt{3}}, 0, 0, \frac{1}{\sqrt{3}}\right]^{\intercal} \\ \boldsymbol{\Gamma}_{\widetilde{d}} &= \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^{\intercal} \end{aligned} \tag{6.4}$$

Note that this is an AR(6)-ARCH(2) model.

Model 2:

$$\begin{aligned} x_t &= cos((\pi/2)(1/\sqrt{2})(x_{t-1} + x_{t-3})) + \varepsilon_t \\ \varepsilon_t &= \left[\sqrt{h_t}\right] e_t \\ h_t &= 1 + (0.1)(\varepsilon_{t-1}^2 + \varepsilon_{t-2}^2) \\ e_t &\sim N(0,1) \\ \boldsymbol{\Phi}_d &= \left[\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}\right]^\intercal \\ \boldsymbol{\Gamma}_{\widetilde{d}} &= \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^\intercal. \end{aligned} \tag{6.5}$$

Model 3:

$$\begin{aligned} x_t &= \sqrt{(3 + (1/\sqrt{3})(x_{t-1} + x_{t-3} + x_{t-6}))^2} + \varepsilon_t \\ \varepsilon_t &= \left[\sqrt{h_t}\right] e_t \\ h_t &= (1/\sqrt{10})(3 + \varepsilon_{t-1}^2 + \varepsilon_{t-2}^2) \\ e_t &\sim N(0,1) \\ \boldsymbol{\Phi}_d &= \left[\frac{1}{\sqrt{3}}, 0, \frac{1}{\sqrt{3}}, 0, 0, \frac{1}{\sqrt{3}}\right]^\intercal \\ \boldsymbol{\Gamma}_{\widetilde{d}} &= \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^\intercal \end{aligned} \tag{6.6}$$

Model 4:

$$\begin{aligned} x_t &= \log\left((3 + (1/\sqrt{3})(x_{t-1} + x_{t-3} + x_{t-6}))^2\right) + \varepsilon_t \\ \varepsilon_t &= \left[\sqrt{h_t}\right] e_t \\ h_t &= (.1)(10 + \varepsilon_{t-1}^2 + \varepsilon_{t-2}^2) \\ e_t &\sim N(0,1) \\ \boldsymbol{\Phi}_d &= \left[\frac{1}{\sqrt{3}}, 0, \frac{1}{\sqrt{3}}, 0, 0, \frac{1}{\sqrt{3}}\right]^\intercal \\ \boldsymbol{\Gamma}_{\widetilde{d}} &= \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^\intercal \end{aligned} \tag{6.7}$$

Model 5:

$$
\begin{aligned}
x_t &= \exp\left(-(3 + (1/\sqrt{16})(x_{t-3} + x_{t-8} + x_{t-20}))\right) + \varepsilon_t \\
\varepsilon_t &= \left[\sqrt{h_t}\right] e_t \\
h_t &= (.1)(10 + \varepsilon_{t-1}^2 + \varepsilon_{t-2}^2) \\
e_t &\sim N(0,1) \\
\boldsymbol{\Phi}_d &= \left[0, 0, \frac{1}{\sqrt{3}}, 0, 0, 0, 0, \frac{1}{\sqrt{3}}, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, \frac{1}{\sqrt{3}}\right]^{\intercal} \\
\boldsymbol{\Gamma}_{\widetilde{d}} &= \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^{\intercal}
\end{aligned}
\tag{6.8}
$$

In each simulation study, we set three different sample sizes: $n = 200$, 600, and $2,000$. For each sample size, we compute three different estimators using the iterative estimation, Sparse estimation and Simultaneous Estimation, respectively, and evaluate the average of the accuracy measures ρ and m^2 based on 100 Monte Carlo replications. Finally, we also report the frequencies, $f_{(p=i)}$ and $f_{(q=i)}$, which count the number of times the i^{th} lag was selected (by the modified SBC criteria or by the Cross-Validation approach under Sparse Estimation). The true dimensions are shown in bold on the summary tables. Note that p (or q) is selected by fixing the other lag at its true value. The estimations based on Cross-Validation are done by splitting the series in half, in which the first part is used as the Training set, and the latter is used as Test data. The simulations results are given next.

Table 6.1: Model 1–Simulation Results under Iterative Estimation

n	$\boldsymbol{\Phi}_d/\boldsymbol{\Gamma}_{\widetilde{d}}$	ρ	m^2	$MSBC(p)$	$MSBC(q)$	$CV_s(p)$	$CV_s(q)$
200	$\widehat{\boldsymbol{\Phi}}$	.9742	.1842	$f_{(p=1)}=0$	$f_{(q=1)}=46$	$f_{(p=1)}=0$	$f_{(q=1)}=38$
	$\widehat{\boldsymbol{\Gamma}}$	.8533	.4705	$f_{(p=2)}=0$	$\boldsymbol{f_{(q=2)}=23}$	$f_{(p=2)}=0$	$\boldsymbol{f_{(q=2)}=24}$
	$\widehat{\boldsymbol{\Phi}}_s$	.9709	.2116	$f_{(p=3)}=0$	$f_{(q=3)}=13$	$f_{(p=3)}=0$	$f_{(q=3)}=17$
	$\widehat{\boldsymbol{\Gamma}}_s$	.8469	.4691	$f_{(p=4)}=0$	$f_{(q=4)}=18$	$f_{(p=4)}=0$	$f_{(q=4)}=21$
				$f_{(p=5)}=0$		$f_{(p=5)}=0$	
				$\boldsymbol{f_{(p=6)}=93}$		$\boldsymbol{f_{(p=6)}=71}$	
				$f_{(p=7)}=5$		$f_{(p=7)}=19$	
				$f_{(p=8)}=2$		$f_{(p=8)}=10$	
600	$\widehat{\boldsymbol{\Phi}}$	.9942	.0987	$f_{(p=1)}=0$	$f_{(q=1)}=36$	$f_{(p=1)}=0$	$f_{(q=1)}=22$
	$\widehat{\boldsymbol{\Gamma}}$	.8827	.4024	$f_{(p=2)}=0$	$\boldsymbol{f_{(q=2)}=27}$	$f_{(p=2)}=0$	$\boldsymbol{f_{(q=2)}=32}$
	$\widehat{\boldsymbol{\Phi}}_s$	.9908	.1195	$f_{(p=3)}=0$	$f_{(q=3)}=16$	$f_{(p=3)}=0$	$f_{(q=3)}=17$
	$\widehat{\boldsymbol{\Gamma}}_s$	.8550	.4557	$f_{(p=4)}=0$	$f_{(q=4)}=21$	$f_{(p=4)}=0$	$f_{(q=4)}=29$
				$f_{(p=5)}=0$		$f_{(p=5)}=0$	
				$\boldsymbol{f_{(p=6)}=91}$		$\boldsymbol{f_{(p=6)}=75}$	
				$f_{(p=7)}=8$		$f_{(p=7)}=16$	
				$f_{(p=8)}=1$		$f_{(p=8)}=9$	
2000	$\widehat{\boldsymbol{\Phi}}$	.9980	.0578	$f_{(p=1)}=0$	$f_{(q=1)}=14$	$f_{(p=1)}=0$	$f_{(q=1)}=5$
	$\widehat{\boldsymbol{\Gamma}}$	.9150	.3281	$f_{(p=2)}=0$	$\boldsymbol{f_{(q=2)}=38}$	$f_{(p=2)}=0$	$\boldsymbol{f_{(q=2)}=45}$
	$\widehat{\boldsymbol{\Phi}}_s$	.9968	.0706	$f_{(p=3)}=0$	$f_{(q=3)}=24$	$f_{(p=3)}=0$	$f_{(q=3)}=18$
	$\widehat{\boldsymbol{\Gamma}}_s$	.8913	.3863	$f_{(p=4)}=0$	$f_{(q=4)}=24$	$f_{(p=4)}=0$	$f_{(q=4)}=32$
				$f_{(p=5)}=0$		$f_{(p=5)}=0$	
				$\boldsymbol{f_{(p=6)}=95}$		$\boldsymbol{f_{(p=6)}=73}$	
				$f_{(p=7)}=5$		$f_{(p=7)}=23$	
				$f_{(p=8)}=0$		$f_{(p=8)}=4$	

Model 1 is an AR(6)-ARCH(2) model, where $f(\cdot)$ and $g(\cdot)$ are linear functions. Here, the performance accuracy of our estimates is assessed for linear mean and variance functions. Clearly, for the estimation of the mean and variance parameter matrices, respectively, the accuracy measures ρ and m^2 reported in Table 6.1 improve as the sample size increases. In general, the sparse estimates were slightly less accurate than the non-sparse ones, which is to be expected because the sparse estimation is only using half of the data to generate its

initial value, whereas the remaining data is used to assign zeros to small elements of the estimated vector.

We now evaluate our lag selection approaches. For this simulated data, we have that $p = 6$ and $q = 2$. First, we consider choosing the number of lags of the mean parameter vector based on the modified SBC criteria defined in equation (4.19). We search for p over a grid of possible values, taking q fixed as its true value. Then, we check how many times each grid value is selected after 100 replications. Similarly, q is chosen based on equation (4.20). Table 6.1 also summarizes the efficacy of the Cross-Validation selection approach on columns $CV_s(p)$ and $CV_s(q)$, which is described in Section 4.6.

We observe that the modified SBC can frequently select the correct lag for $\boldsymbol{\Phi}_d$ even for smaller sample sizes. However, selecting the correct dimension of $\boldsymbol{\Gamma}_{\widetilde{d}}$ appears to be more challenging. As the sample size increases, the number of correct selections improves, but it might be necessary to have an even larger sample size in order to have reasonably precise selection. Selecting p by sparse estimation and Cross Validation also appears to be useful, but it is not as good as the modified SBC criteria for this model. Differently, choosing q appears to be more efficient on the $CV_(q)$ column.

Table 6.2: Model 1 Simulation Results under Simultaneous Estimation

n	$\boldsymbol{\Phi}_d/\boldsymbol{\Gamma}_{\widetilde{d}}$	ρ	m^2	$MSBC(p)$	$MSBC(q)$
200	$\widehat{\boldsymbol{\Phi}}_m$	.9824	.1749	$f_{(p=1)} = 0$	$f_{(q=1)} = 48$
	$\widehat{\boldsymbol{\Gamma}}_m$	.8960	.3864	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 23}$
				$f_{(p=3)} = 0$	$f_{(q=3)} = 19$
				$f_{(p=4)} = 0$	$f_{(q=4)} = 10$
				$f_{(p=5)} = 0$	
				$\boldsymbol{f_{(p=6)} = 86}$	
				$f_{(p=7)} = 13$	
				$f_{(p=8)} = 1$	
600	$\widehat{\boldsymbol{\Phi}}_m$	.9933	.1066	$f_{(p=1)} = 0$	$f_{(q=1)} = 39$
	$\widehat{\boldsymbol{\Gamma}}_m$	.8876	.3983	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 37}$
				$f_{(p=3)} = 0$	$f_{(q=3)} = 14$
				$f_{(p=4)} = 0$	$f_{(q=4)} = 10$
				$f_{(p=5)} = 0$	
				$\boldsymbol{f_{(p=6)} = 91}$	
				$f_{(p=7)} = 9$	
				$f_{(p=8)} = 0$	
2000	$\widehat{\boldsymbol{\Phi}}_m$	.9980	.0583	$f_{(p=1)} = 0$	$f_{(q=1)} = 21$
	$\widehat{\boldsymbol{\Gamma}}_m$	.9343	.2897	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 61}$
				$f_{(p=3)} = 0$	$f_{(q=3)} = 11$
				$f_{(p=4)} = 0$	$f_{(q=4)} = 7$
				$f_{(p=5)} = 0$	
				$\boldsymbol{f_{(p=6)} = 95}$	
				$f_{(p=7)} = 5$	
				$f_{(p=8)} = 0$	

The results reported in Table 6.2 are once again for Model 1 but using the simultaneous estimation approach. We can observe that $\widehat{\boldsymbol{\Phi}}_m$ appears to be slightly better than the original $\widehat{\boldsymbol{\Phi}}$ estimator for smaller sample sizes. It is interesting to note that the simultaneous estimation of $\boldsymbol{\Gamma}_{\widetilde{d}}$ is consistently more accurate than the iterative estimation procedure. We should stress that $\widehat{\boldsymbol{\Gamma}}$ is obtained through the minimization of total squared prediction error of ε_t^2, whereas $\widehat{\boldsymbol{\Gamma}}_m$ is obtained through an objective function in which the only role of the fitted variance is

to be used as weights on a Weighted Least Squares framework. Therefore, it is intuitive to imagine why $\widehat{\boldsymbol{\Gamma}}$ should converge, but the consistency of $\widehat{\boldsymbol{\Gamma}}_m$ is not as obvious.

We can also observe that the lag selection for the vector $\boldsymbol{\Phi}_d$ under the simultaneous approach seems to be worse than in the iterative estimation, but they did show the same accuracy on the larger sample size scenario. Whereas, the selection of q improves greatly for larger sample sizes, which could be a reflection of using a more accurate estimator according to ρ and m^2.

Table 6.3: Model 2 Simulation Results under Iterative Estimation

n	$\boldsymbol{\Phi}_d/\boldsymbol{\Gamma}_{\widetilde{d}}$	ρ	m^2	$MSBC(p)$	$MSBC(q)$	$CV_s(p)$	$CV_s(q)$
200	$\widehat{\boldsymbol{\Phi}}$	.9943	.0939	$f_{(p=1)}=0$	$f_{(q=1)}=42$	$f_{(p=1)}=3$	$f_{(q=1)}=0$
	$\widehat{\boldsymbol{\Gamma}}$	.8418	.4738	$f_{(p=2)}=0$	$\boldsymbol{f_{(q=2)}=36}$	$f_{1(p=2)}=0$	$\boldsymbol{f_{(q=2)}=05}$
	$\widehat{\boldsymbol{\Phi}}_s$	.9927	.0976	$\boldsymbol{f_{(p=3)}=98}$	$f_{(q=3)}=11$	$\boldsymbol{f_{(p=3)}=90}$	$f_{(q=3)}=17$
	$\widehat{\boldsymbol{\Gamma}}_s$	.8567	.4467	$f_{(p=4)}=2$	$f_{(q=4)}=11$	$f_{(p=4)}=5$	$f_{(q=4)}=78$
				$f_{(p=5)}=0$		$f_{(p=5)}=2$	
600	$\widehat{\boldsymbol{\Phi}}$	.9989	.0420	$f_{(p=1)}=0$	$f_{(q=1)}=09$	$f_{(p=1)}=0$	$f_{(q=1)}=0$
	$\widehat{\boldsymbol{\Gamma}}$	.9018	.3608	$f_{(p=2)}=0$	$\boldsymbol{f_{(q=2)}=27}$	$f_{(p=2)}=0$	$\boldsymbol{f_{(q=2)}=3}$
	$\widehat{\boldsymbol{\Phi}}_s$	.9979	.0491	$\boldsymbol{f_{(p=3)}=99}$	$f_{(q=3)}=35$	$\boldsymbol{f_{(p=3)}=99}$	$f_{(q=3)}=19$
	$\widehat{\boldsymbol{\Gamma}}_s$	.8754	.4110	$f_{(p=4)}=1$	$f_{(q=4)}=39$	$f_{(p=4)}=1$	$f_{(q=4)}=78$
				$f_{(p=5)}=0$		$f_{(p=5)}=0$	
2000	$\widehat{\boldsymbol{\Phi}}$	.9997	.0224	$f_{(p=1)}=0$	$f_{(q=1)}=3$	$f_{(p=1)}=0$	$f_{(q=1)}=0$
	$\widehat{\boldsymbol{\Gamma}}$	.9141	.3419	$f_{(p=2)}=0$	$\boldsymbol{f_{(q=2)}=30}$	$f_{(p=2)}=0$	$\boldsymbol{f_{(q=2)}=0}$
	$\widehat{\boldsymbol{\Phi}}_s$	.9994	.0289	$\boldsymbol{f_{(p=3)}=97}$	$f_{(q=3)}=26$	$\boldsymbol{f_{(p=3)}=99}$	$f_{(q=3)}=15$
	$\widehat{\boldsymbol{\Gamma}}_s$	.9014	.3567	$f_{(p=4)}=3$	$f_{(q=4)}=41$	$f_{(p=4)}=1$	$f_{(q=4)}=85$
				$f_{(p=5)}=0$		$f_{(p=5)}=0$	

Model 2 considers the case when $f(\cdot)$ is non-linear. More specifically, in Model 2, we assume that $E[x_t|\boldsymbol{X}_{t-1}]$ depends on $\boldsymbol{\Phi}'\boldsymbol{X}_{t-1}$ through a cosine function. Here, it is of interest to investigate how this nonlinear mean function influences the accuracy of our estimates. The results in Table 6.3show that it is possible to obtain high accuracy when estimating $\boldsymbol{\Phi}_d$, even for smaller sample size scenarios. Additionally, both $\widehat{\boldsymbol{\Phi}}$ and $\widehat{\boldsymbol{\Gamma}}$ became more accurate as n increases. The sparse estimates are also almost as precise as those obtained using

iterative estimation, which once again improve with larger sample sizes. A comparison of results reported in Table 6.3 with those from Table 6.1 shows that a nonlinear model with less parameters has more accurate mean parameter estimates, which indicate that high-dimensional vectors could have a larger impact on the acuuracy of the estimates than the nonlinearity itself.

Under Model 2, the true lags are $p = 3$ and $q = 2$. For the mean parameter vector case, we observe that both the methods in most cases choose the correct lag across different sample sizes. However, the lag selection for $\boldsymbol{\Gamma}_d$ remains a challenge. There is an apparent convergence to the correct lag under the MSBC criteria, but the Cross-Validation approach consistently overestimates q.

Table 6.4: Model 2 Simulation Results under Simultaneous Estimation

n	$\boldsymbol{\Phi}_d/\boldsymbol{\Gamma}_{\widetilde{d}}$	ρ	m^2	$MSBC(p)$	$MSBC(q)$
200	$\widehat{\boldsymbol{\Phi}}_m$	.9958	.0796	$f_{(p=1)} = 0$	$f_{(q=1)} = 50$
	$\widehat{\boldsymbol{\Gamma}}_m$	.8769	.4029	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 28}$
				$\boldsymbol{f_{(p=3)} = 96}$	$f_{(q=3)} = 13$
				$f_{(p=4)} = 2$	$f_{(q=4)} = 09$
				$f_{(p=5)} = 2$	
600	$\widehat{\boldsymbol{\Phi}}_m$	.9987	.0448	$f_{(p=1)} = 0$	$f_{(q=1)} = 31$
	$\widehat{\boldsymbol{\Gamma}}_m$	.9147	.3365	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 41}$
				$\boldsymbol{f_{(p=3)} = 99}$	$f_{(q=3)} = 16$
				$f_{(p=4)} = 1$	$f_{(q=4)} = 12$
				$f_{(p=5)} = 0$	
2000	$\widehat{\boldsymbol{\Phi}}_m$	.9996	.0261	$f_{(p=1)} = 0$	$f_{(q=1)} = 24$
	$\widehat{\boldsymbol{\Gamma}}_m$	.9197	.3314	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 48}$
				$\boldsymbol{f_{(p=3)} = 100}$	$f_{(q=3)} = 15$
				$f_{(p=4)} = 0$	$f_{(q=4)} = 13$
				$f_{(p=5)} = 0$	

Table 6.4 presents the simultaneous estimation simulation results for Model 2. We can see that $\widehat{\boldsymbol{\Phi}}_m$ is as accurate as the iterative approach estimates, and it also improves as the sample size increases. Additionally, the MSBC criteria in most cases selected the correct

value for p. Similar to the analysis of Model 1 simulations, estimates based on $\widehat{\boldsymbol{\Gamma}}_m$ were consistently better than those from $\widehat{\boldsymbol{\Gamma}}$ and lag selection is also more accurate for the parameter q.

Table 6.5: Model 3 Simulation Results under Iterative Estimation

n	$\boldsymbol{\Phi}_d/\boldsymbol{\Gamma}_{\widetilde{d}}$	ρ	m^2	$MSBC(p)$	$MSBC(q)$	$CV_s(p)$	$CV_s(q)$
200	$\widehat{\boldsymbol{\Phi}}$	.9513	.2863	$f_{(p=1)} = 8$	$f_{(q=1)} = 43$	$f_{(p=1)} = 53$	$f_{(q=1)} = 47$
	$\widehat{\boldsymbol{\Gamma}}$	.8707	.4199	$f_{(p=2)} = 1$	$\boldsymbol{f_{(q=2)} = 27}$	$f_{(p=2)} = 2$	$\boldsymbol{f_{(q=2)} = 20}$
	$\widehat{\boldsymbol{\Phi}}_s$	.8910	.3761	$f_{(p=3)} = 10$	$f_{(q=3)} = 17$	$f_{(p=3)} = 18$	$f_{(q=3)} = 11$
	$\widehat{\boldsymbol{\Gamma}}_s$	.8577	.4393	$f_{(p=4)} = 2$	$f_{(q=4)} = 13$	$f_{(p=4)} = 0$	$f_{(q=4)} = 22$
				$f_{(p=5)} = 0$		$f_{(p=5)} = 3$	
				$\boldsymbol{f_{(p=6)} = 73}$		$\boldsymbol{f_{(p=6)} = 19}$	
				$f_{(p=7)} = 6$		$f_{(p=7)} = 5$	
				$f_{(p=8)} = 0$		$f_{(p=8)} = 0$	
600	$\widehat{\boldsymbol{\Phi}}$	.9815	.1791	$f_{(p=1)} = 0$	$f_{(q=1)} = 27$	$f_{(p=1)} = 1$	$f_{(q=1)} = 18$
	$\widehat{\boldsymbol{\Gamma}}$	.8769	.3985	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 32}$	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 39}$
	$\widehat{\boldsymbol{\Phi}}_s$	.9806	.1715	$f_{(p=3)} = 0$	$f_{(q=3)} = 21$	$f_{(p=3)} = 5$	$f_{(q=3)} = 26$
	$\widehat{\boldsymbol{\Gamma}}_s$	.8528	.4500	$f_{(p=4)} = 0$	$f_{(q=4)} = 20$	$f_{(p=4)} = 0$	$f_{(q=4)} = 17$
				$f_{(p=5)} = 0$		$f_{(p=5)} = 0$	
				$\boldsymbol{f_{(p=6)} = 96}$		$\boldsymbol{f_{(p=6)} = 83}$	
				$f_{(p=7)} = 4$		$f_{(p=7)} = 9$	
				$f_{(p=8)} = 0$		$f_{(p=8)} = 2$	
2000	$\widehat{\boldsymbol{\Phi}}$	.9963	.0818	$f_{(p=1)} = 0$	$f_{(q=1)} = 06$	$f_{(p=1)} = 0$	$f_{(q=1)} = 8$
	$\widehat{\boldsymbol{\Gamma}}$	.9257	.3090	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 39}$	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 49}$
	$\widehat{\boldsymbol{\Phi}}_s$	.9947	.0905	$f_{(p=3)} = 0$	$f_{(q=3)} = 25$	$f_{(p=3)} = 0$	$f_{(q=3)} = 22$
	$\widehat{\boldsymbol{\Gamma}}_s$	.9160	.3364	$f_{(p=4)} = 0$	$f_{(q=4)} = 30$	$f_{(p=4)} = 0$	$f_{(q=4)} = 21$
				$f_{(p=5)} = 0$		$f_{(p=5)} = 0$	
				$\boldsymbol{f_{(p=6)} = 98}$		$\boldsymbol{f_{(p=6)} = 93}$	
				$f_{(p=7)} = 2$		$f_{(p=7)} = 5$	
				$f_{(p=8)} = 0$		$f_{(p=8)} = 2$	

The purpose of Model 3 in our simulation study is to use the same parameter matrices as in Model 1, but $f(\cdot)$ is taken to be nonlinear. We observe on Table 6.5 that the estimates for $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ become more accurate as n increases for both iterative estimates and sparse

estimates. In general, $\widehat{\boldsymbol{\Phi}}_s$ and $\widehat{\boldsymbol{\Gamma}}_s$ are slightly less accurate than the iterative estimates. Additionally, the accuracy of parameter estimates for Model 3 are close to those for Model 1 (Table 6.1) and the difference between them seems to decrease as n increases.

Lag selection also presented a similar behavior as in the linear mean function case. We can observe that large sample sizes are capable of selecting the correct value for p in the vast majority of replicates in both $MSBC(p)$ and $CV_s(p)$ columns of Table 6.5. The selection of dimension of $\boldsymbol{\Gamma}_d$ also has a similar behavior as in the first model scenario. The selection of correct lag improves as the sample size increases, and the selection using sparse estimation seems to be better for larger sample sizes.

Table 6.6: Model 3 Simulation Results under Simultaneous Estimation

n	$\boldsymbol{\Phi}_d/\boldsymbol{\Gamma}_{\widetilde{d}}$	ρ	m^2	$MSBC(p)$	$MSBC(q)$
200	$\widehat{\boldsymbol{\Phi}}_m$	.9436	.3109	$f_{(p=1)} = 13$	$f_{(q=1)} = 59$
	$\widehat{\boldsymbol{\Gamma}}_m$	.8764	.4108	$f_{(p=2)} = 1$	$\boldsymbol{f_{(q=2)} = 28}$
				$f_{(p=3)} = 13$	$f_{(q=3)} = 7$
				$f_{(p=4)} = 1$	$f_{(q=4)} = 6$
				$f_{(p=5)} = 0$	
				$\boldsymbol{f_{(p=6)} = 69}$	
				$f_{(p=7)} = 3$	
				$f_{(p=8)} = 0$	
600	$\widehat{\boldsymbol{\Phi}}_m$	.9822	.1762	$f_{(p=1)} = 0$	$f_{(q=1)} = 43$
	$\widehat{\boldsymbol{\Gamma}}_m$	.9115	.3487	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 37}$
				$f_{(p=3)} = 0$	$f_{(q=3)} = 15$
				$f_{(p=4)} = 0$	$f_{(q=4)} = 5$
				$f_{(p=5)} = 0$	
				$\boldsymbol{f_{(p=6)} = 94}$	
				$f_{(p=7)} = 6$	
				$f_{(p=8)} = 0$	
2000	$\widehat{\boldsymbol{\Phi}}_m$	.9949	.0954	$f_{(p=1)} = 0$	$f_{(q=1)} = 27$
	$\widehat{\boldsymbol{\Gamma}}_m$	.9164	.3336	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 46}$
				$f_{(p=3)} = 0$	$f_{(q=3)} = 17$
				$f_{(p=4)} = 0$	$f_{(q=4)} = 10$
				$f_{(p=5)} = 0$	
				$\boldsymbol{f_{(p=6)} = 99}$	
				$f_{(p=7)} = 1$	
				$f_{(p=8)} = 0$	

According to Table 6.6, the accuracy of simultaneous estimation is similar to the iterative estimation approach. Both estimates appear to converge asymptotically, however, in this simulated model scenario. We could not find evidence that one approach is consistently better than the other for estimating one or both parameter matrices. The selection of p under the MSBC criteria appears to be less efficient under the simultaneous approach for smaller sample sizes. Whereas, the selection of q was more precise for all three sample sizes.

Table 6.7: Model 4 Simulation Results under Iterative Estimation

n	$\boldsymbol{\Phi}_d/\boldsymbol{\Gamma}_{\widetilde{d}}$	ρ	m^2	$MSBC(p)$	$MSBC(q)$	$CV_s(p)$	$CV_s(q)$
200	$\widehat{\boldsymbol{\Phi}}$	.5238	.7987	$f_{(p=1)} = 71$	$f_{(q=1)} = 56$	$f_{(p=1)} = 72$	$f_{(q=1)} = 38$
	$\widehat{\boldsymbol{\Gamma}}$	.8230	.4881	$f_{(p=2)} = 18$	$\boldsymbol{f_{(q=2)} = 21}$	$f_{(p=2)} = 18$	$\boldsymbol{f_{(q=2)} = 22}$
	$\widehat{\boldsymbol{\Phi}}_s$	.3840	.8794	$f_{(p=3)} = 8$	$f_{(q=3)} = 14$	$f_{(p=3)} = 4$	$f_{(q=3)} = 17$
	$\widehat{\boldsymbol{\Gamma}}_s$	.8291	.5044	$f_{(p=4)} = 2$	$f_{(q=4)} = 09$	$f_{(p=4)} = 3$	$f_{(q=4)} = 23$
				$f_{(p=5)} = 0$		$f_{(p=5)} = 1$	
				$\boldsymbol{f_{(p=6)} = 1}$		$\boldsymbol{f_{(p=6)} = 2}$	
				$f_{(p=7)} = 0$		$f_{(p=7)} = 0$	
				$f_{(p=8)} = 1$		$f_{(p=8)} = 0$	
600	$\widehat{\boldsymbol{\Phi}}$	.7059	.6391	$f_{(p=1)} = 40$	$f_{(q=1)} = 32$	$f_{(p=1)} = 74$	$f_{(q=1)} = 22$
	$\widehat{\boldsymbol{\Gamma}}$	.8757	.4039	$f_{(p=2)} = 14$	$\boldsymbol{f_{(q=2)} = 33}$	$f_{(p=2)} = 10$	$\boldsymbol{f_{(q=2)} = 29}$
	$\widehat{\boldsymbol{\Phi}}_s$	.4847	.8226	$f_{(p=3)} = 39$	$f_{(q=3)} = 11$	$f_{(p=3)} = 10$	$f_{(q=3)} = 24$
	$\widehat{\boldsymbol{\Gamma}}_s$	.8678	.4357	$f_{(p=4)} = 2$	$f_{(q=4)} = 24$	$f_{(p=4)} = 1$	$f_{(q=4)} = 25$
				$f_{(p=5)} = 1$		$f_{(p=5)} = 3$	
				$\boldsymbol{f_{(p=6)} = 4}$		$\boldsymbol{f_{(p=6)} = 2}$	
				$f_{(p=7)} = 0$		$f_{(p=7)} = 0$	
				$f_{(p=8)} = 0$		$f_{(p=8)} = 0$	
2000	$\widehat{\boldsymbol{\Phi}}$	.9365	.3273	$f_{(p=1)} = 6$	$f_{(q=1)} = 05$	$f_{(p=1)} = 31$	$f_{(q=1)} = 5$
	$\widehat{\boldsymbol{\Gamma}}$	.9222	.3167	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 50}$	$f_{(p=2)} = 3$	$\boldsymbol{f_{(q=2)} = 40}$
	$\widehat{\boldsymbol{\Phi}}_s$	.8509	.4406	$f_{(p=3)} = 32$	$f_{(q=3)} = 24$	$f_{(p=3)} = 43$	$f_{(q=3)} = 22$
	$\widehat{\boldsymbol{\Gamma}}_s$	.9070	.3492	$f_{(p=4)} = 1$	$f_{(q=4)} = 21$	$f_{(p=4)} = 2$	$f_{(q=4)} = 33$
				$f_{(p=5)} = 0$		$f_{(p=5)} = 1$	
				$\boldsymbol{f_{(p=6)} = 57}$		$\boldsymbol{f_{(p=6)} = 15}$	
				$f_{(p=7)} = 0$		$f_{(p=7)} = 3$	
				$f_{(p=8)} = 0$		$f_{(p=8)} = 2$	

The simulation experiment under Model 4 assumes the same parameter matrices as Model 1 and Model 3, but consider a different nonlinear function. The accuracy measures reported in Table 6.7 show that the estimated parameter vectors converge under both estimation approaches. However, when we compare these results to those from Models 1 and 3, we observe that the estimation accuracy of $\boldsymbol{\Phi}_d$ is much lower, specially for smaller sample sizes.

Therefore, the choice of $f(\cdot)$ can make the estimation of parameters much more challenging, eventually requiring larger sample sizes. Whereas, the accuracy of $\boldsymbol{\Gamma}_{\widetilde{d}}$ estimates are close to those under the previous models.

The smaller accuracy of estimates also influences the lag selection procedures. For small sample sizes, both the $MSBC(p)$ and $CV_s(p)$ columns in Table 6.7 show that in general, p is mostly underestimated, but when the sample size increases, the selected lag start converging to the true value (p=6), where the MSBC criteria seems to be more efficient. Under this simulated model, the selection of q also improves as the sample size increases, but the MSBC criteria seems to perform better than the sparse estimation approach.

Table 6.8: Model 4 Simulation Results under Simultaneous Estimation

n	$\boldsymbol{\Phi}_d/\boldsymbol{\Gamma}_{\widetilde{d}}$	ρ	m^2	$MSBC(p)$	$MSBC(q)$
200	$\widehat{\boldsymbol{\Phi}}_m$	.5263	.7940	$f_{(p=1)} = 52$	$f_{(q=1)} = 53$
	$\widehat{\boldsymbol{\Gamma}}_m$	.8800	.4079	$f_{(p=2)} = 28$	$\boldsymbol{f_{(q=2)} = 25}$
				$f_{(p=3)} = 12$	$f_{(q=3)} = 10$
				$f_{(p=4)} = 4$	$f_{(q=4)} = 12$
				$f_{(p=5)} = 1$	
				$\boldsymbol{f_{(p=6)} = 3}$	
				$f_{(p=7)} = 0$	
				$f_{(p=8)} = 0$	
600	$\widehat{\boldsymbol{\Phi}}_m$	.7770	.5892	$f_{(p=1)} = 47$	$f_{(q=1)} = 39$
	$\widehat{\boldsymbol{\Gamma}}_m$	.8991	.3632	$f_{(p=2)} = 13$	$\boldsymbol{f_{(q=2)} = 43}$
				$f_{(p=3)} = 34$	$f_{(q=3)} = 13$
				$f_{(p=4)} = 1$	$f_{(q=4)} = 5$
				$f_{(p=5)} = 0$	
				$\boldsymbol{f_{(p=6)} = 4}$	
				$f_{(p=7)} = 1$	
				$f_{(p=8)} = 0$	
2000	$\widehat{\boldsymbol{\Phi}}_m$	.9293	.3463	$f_{(p=1)} = 5$	$f_{(q=1)} = 20$
	$\widehat{\boldsymbol{\Gamma}}_m$	.9244	.3323	$f_{(p=2)} = 2$	$\boldsymbol{f_{(q=2)} = 50}$
				$f_{(p=3)} = 45$	$f_{(q=3)} = 17$
				$f_{(p=4)} = 3$	$f_{(q=4)} = 13$
				$f_{(p=5)} = 0$	
				$\boldsymbol{f_{(p=6)} = 44}$	
				$f_{(p=7)} = 1$	
				$f_{(p=8)} = 0$	

The simultaneous estimation simulation results on Table 6.8 shows that there is no clear improvement on estimates accuracy, or lag selection, compared to the iterative estimation approach. We also observe that the estimates for parameter matrices and selected lags improve on larger sample sizes, corroborating the desired asymptotic convergence.

Table 6.9: Model 5 Simulation Results under Iterative Estimation

n	$\boldsymbol{\Phi}_d/\boldsymbol{\Gamma}_{\widetilde{d}}$	ρ	m^2	$MSBC(p)$	$MSBC(q)$	$CV_s(p)$	$CV_s(q)$
200	$\widehat{\boldsymbol{\Phi}}$	.5077	.8237	$f_{(p=1)} = 32$	$f_{(q=1)} = 36$	$f_{(p=1)} = 53$	$f_{(q=1)} = 67$
	$\widehat{\boldsymbol{\Gamma}}$	.8410	.4768	$f_{(p=2)} = 10$	$\boldsymbol{f_{(q=2)} = 38}$	$f_{(p=2)} = 13$	$\boldsymbol{f_{(q=2)} = 18}$
	$\widehat{\boldsymbol{\Phi}}_s$	.2681	.9294	$f_{(p=3)} = 45$	$f_{(q=3)} = 19$	$f_{(p=3)} = 25$	$f_{(q=3)} = 8$
	$\widehat{\boldsymbol{\Gamma}}_s$	.8292	.5015	$f_{(p=4)} = 4$	$f_{(q=4)} = 07$	$f_{(p=4)} = 4$	$f_{(q=4)} = 7$
				$f_{(p=5)} = 4$		$f_{(p=5)} = 2$	
				$f_{(p=6)} = 2$		$f_{(p=6)} = 1$	
				$f_{(p=7)} = 0$		$f_{(p=7)} = 0$	
				$f_{(p=8)} = 2$		$f_{(p=8)} = 2$	
				$f_{(p=9)} = 1$		$f_{(p=9)} = 0$	
				$f_{(p\geq 10)} = 0$		$f_{(p\geq 10)} = 0$	
600	$\widehat{\boldsymbol{\Phi}}$	.7906	.5871	$f_{(p=1)} = 8$	$f_{(q=1)} = 41$	$f_{(p=1)} = 35$	$f_{(q=1)} = 58$
	$\widehat{\boldsymbol{\Gamma}}$	.8987	.3777	$f_{(p=2)} = 5$	$\boldsymbol{f_{(q=2)} = 34}$	$f_{(p=2)} = 9$	$\boldsymbol{f_{(q=2)} = 25}$
	$\widehat{\boldsymbol{\Phi}}_s$	.6317	.7050	$f_{(p=3)} = 56$	$f_{(q=3)} = 15$	$f_{(p=3)} = 39$	$f_{(q=3)} = 10$
	$\widehat{\boldsymbol{\Gamma}}_s$	.8460	.4752	$f_{(p=4)} = 0$	$f_{(q=4)} = 10$	$f_{(p=4)} = 3$	$f_{(q=4)} = 7$
				$f_{(p=5)} = 0$		$f_{(p=5)} = 1$	
				$f_{(p=6)} = 0$		$f_{(p=6)} = 2$	
				$f_{(p=7)} = 0$		$f_{(p=7)} = 0$	
				$f_{(p=8)} = 26$		$f_{(p=8)} = 7$	
				$f_{(p=9)} = 4$		$f_{(p=9)} = 2$	
				$f_{(p=10)} = 1$		$f_{(p=10} = 0$	
				$f_{(p=11)} = 0$		$f_{(p=11} = 0$	
				$f_{(p=12)} = 0$		$f_{(p=12)} = 1$	
				$f_{(p=13)} = 0$		$f_{(p=13)} = 1$	
				$f_{(p=14)} = 0$		$f_{(p=14)} = 0$	
				$f_{(p\geq 15)} = 0$		$f_{(p\geq 15)} = 0$	
2000	$\widehat{\boldsymbol{\Phi}}$	.9643	.3048	$f_{(p=1)} = 0$	$f_{(q=1)} = 18$	$f_{(p=1)} = 0$	$f_{(q=1)} = 10$
	$\widehat{\boldsymbol{\Gamma}}$	.9286	.2934	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 57}$	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 48}$
	$\widehat{\boldsymbol{\Phi}}_s$	.9759	.1818	$f_{(p=3)} = 0$	$f_{(q=3)} = 15$	$f_{(p=3)} = 8$	$f_{(q=3)} = 25$
	$\widehat{\boldsymbol{\Gamma}}_s$	.9156	.3256	$f_{(p=4)} = 0$	$f_{(q=4)} = 8$	$f_{(p=4)} = 0$	$f_{(q=4)} = 17$
				$f_{(p=5)} = 0$		$f_{(p=5)} = 0$	
				$f_{(p=6)} = 0$		$f_{(p=6)} = 0$	
				$f_{(p=7)} = 0$		$f_{(p=7)} = 0$	
				$f_{(p=8)} = 33$		$f_{(p=8)} = 82$	
				$f_{(p=9)} = 10$		$f_{(p=9)} = 10$	
				$f_{(p=10)} = 0$		$f_{(p=10} = 0$	
				$f_{(p=11)} = 0$		$f_{(p=11} = 0$	
				$f_{(p=12)} = 0$		$f_{(p=12)} = 0$	
				$f_{(p=13)} = 0$		$f_{(p=13)} = 0$	
				$f_{(p=14)} = 0$		$f_{(p=14)} = 0$	
				$f_{(p=15)} = 0$		$f_{(p=15)} = 0$	
				$f_{(p=16)} = 0$		$f_{(p=16)} = 0$	
				$f_{(p=17)} = 0$		$f_{(p=17)} = 0$	
				$f_{(p=18)} = 0$		$f_{(p=18)} = 0$	
				$f_{(p=19)} = 0$		$f_{(p=19)} = 0$	
				$\boldsymbol{f_{(p=20)} = 57}$		$\boldsymbol{f_{(p=20)} = 0}$	
				$f_{(p=21)} = 0$		$f_{(p=21)} = 0$	
				$f_{(p=22)} = 0$		$f_{(p=22)} = 0$	

Model 5 is our last model in which both $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ are vectors. In this simulation scenario $\boldsymbol{\Phi}_d$ has 20 elements, of which 17 are zeros. Therefore, it has a high degree of sparsity and much larger dimension than the previous models. When we compared the results of Models 1 and 2, we observed that larger values of p can have a strong impact over the accuracy of estimates, so it is of our interest to evaluate results under this challenging data simulation model.

According to Table 6.9, the accuracy of both $\widehat{\boldsymbol{\Phi}}$ and $\widehat{\boldsymbol{\Phi}}_s$ are lower for smaller sample sizes than the previous models, reflecting the complexity of Model 5. However, both estimates improve as n increases, yielding large values for ρ when $n = 2000$. An interesting aspect of our simulation results is that the sparse estimator was more accurate on average than the iterative estimators, for the largest sample size case. This indicates that the sparse approach can be useful when dealing with series in which you expect a high degree of sparsity on its dependence structure, if a large sample size is available.

Table 6.9 also shows how precisely our simulation study selected p over a grid from 1 to 22, and chose q from 1 to 4. We observe that selecting p is very challenging for smaller sample sizes, in which neither the $MSBC(p)$ and $CV_s(p)$ approach could identify the correct lag once. However, when $n = 2000$ the MSBC criterion correctly chose the lag in the majority of replicates. Whereas, the sparse approach is still inefficient under the larger sample size scenario, but it shows some signs of convergence. The MSBC criterion also yielded better selection of q for all sample sizes, which accuracy is similar to the one found in previous models. Hence, it appears that this highly complex mean function structure does not have a large impact on the variance parameters estimation.

Table 6.10: Model 5 Simulation Results under Simultaneous Estimation

n	$\boldsymbol{\Phi}_d/\boldsymbol{\Gamma}_{\tilde{d}}$	ρ	m^2	$MSBC(p)$	$MSBC(q)$
200	$\widehat{\boldsymbol{\Phi}}_m$	.4974	.8396	$f_{(p=1)} = 41$	$f_{(q=1)} = 54$
	$\widehat{\boldsymbol{\Gamma}}_m$	.8922	.3773	$f_{(p=2)} = 15$	$\boldsymbol{f_{(q=2)} = 24}$
				$f_{(p=3)} = 30$	$f_{(q=3)} = 11$
				$f_{(p=4)} = 4$	$f_{(q=4)} = 11$
				$f_{(p=5)} = 1$	
				$f_{(p=6)} = 1$	
				$f_{(p=7)} = 1$	
				$f_{(p=8)} = 6$	
				$f_{(p=9)} = 1$	
				$f_{(p\geq 10)} = 0$	
600	$\widehat{\boldsymbol{\Phi}}_m$	.7914	.5933	$f_{(p=1)} = 13$	$f_{(q=1)} = 44$
	$\widehat{\boldsymbol{\Gamma}}_m$	.9060	.3573	$f_{(p=2)} = 6$	$\boldsymbol{f_{(q=2)} = 32}$
				$f_{(p=3)} = 51$	$f_{(q=3)} = 16$
				$f_{(p=4)} = 2$	$f_{(q=4)} = 8$
				$f_{(p=5)} = 1$	
				$f_{(p=6)} = 0$	
				$f_{(p=7)} = 0$	
				$f_{(p=8)} = 25$	
				$f_{(p=9)} = 2$	
				$f_{(p\geq 10)} = 0$	
2000	$\widehat{\boldsymbol{\Phi}}_m$	.9470	.3138	$f_{(p=1)} = 0$	$f_{(q=1)} = 34$
	$\widehat{\boldsymbol{\Gamma}}_m$	.9231	.3331	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 58}$
				$f_{(p=3)} = 0$	$f_{(q=3)} = 0$
				$f_{(p=4)} = 0$	$f_{(q=4)} = 8$
				$f_{(p=5)} = 0$	
				$f_{(p=6)} = 0$	
				$f_{(p=7)} = 0$	
				$f_{(p=8)} = 18$	
				$f_{(p=9)} = 30$	
				$f_{(p=10)} = 0$	
				$f_{(p=11)} = 0$	
				$f_{(p=12)} = 0$	
				$f_{(p=13)} = 0$	
				$f_{(p=14)} = 0$	
				$f_{(p=15)} = 0$	
				$f_{(p=16)} = 0$	
				$f_{(p=17)} = 0$	
				$f_{(p=18)} = 0$	
				$f_{(p=19)} = 0$	
				$\boldsymbol{f_{(p=20)} = 52}$	
				$f_{(p=21)} = 0$	
				$f_{(p=22)} = 0$	

The simultaneous estimation approach results for Model 5 can be found on Table 6.10. The accuracy measures and lag selection behavior are, in general, very similar to those of the iterative estimates. Additionally, no approach has shown to be consistently better than

the other.

In general, the simulation results under the five data generating models allow us to conclude that our proposed iterative estimator approximates the true parameter matrices well, but the rate of convergence will differ depending on the underlying model complexity. The same conclusion can be applied for the sparse and simultaneous estimation approaches. Both the alternative estimation procedures were not consistently more accurate than the iterative approach, so the iterative estimator seems to be a preferred choice. The effectiveness of sparse estimation can be investigated further based on a more comprehensive study in which we can consider more models and different splitting (Training and Test) scenarios of the series. Another subject that requires more investigation is the apparent higher accuracy of $\widehat{\boldsymbol{\Gamma}}_m$, especially on smaller sample sizes. We should note that all five models have the same variance structure, so it is worth investigating what happens when we consider different $g(\cdot)$ functions and parameter matrices.

The overall results of lag selection lead to the conclusion that estimating p is more straightforward than q, and our proposed MSBC tends to perform well on larger sample sizes. Even though estimating the lags is not the main goal of our research, the simulation results have shown us that the proposed criteria can be useful on practical applications when true p and q are usually unknown.

6.2 Models with Parameter Matrices

Now, let us consider the following models where $d > 1$:

Model 6

$$
\begin{aligned}
x_t &= 3 - (1/\sqrt{3})(x_{t-1} + x_{t-3} + x_{t-6}) + cos\{(\pi/2)(1/\sqrt{2})(x_{t-2} - x_{t-4})\} + \varepsilon_t \\
\varepsilon_t &= \left[\sqrt{h_t}\right] e_t \\
h_t &= 1 + (.1)(\varepsilon_{t-1}^2 + \varepsilon_{t-2}^2) \\
e_t &\sim N(0,1) \\
\boldsymbol{\Phi}_d &= \begin{bmatrix} \frac{1}{\sqrt{3}} & 0 \\ 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & 0 \\ 0 & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & 0 \end{bmatrix} \\
\boldsymbol{\Gamma}_{\widetilde{d}} &= \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}
\end{aligned}
\tag{6.9}
$$

Model 7

$$
\begin{aligned}
x_t = -1 + (.4/\sqrt{5})(x_{t-1} + 2x_{t-4}) - cos\{(\pi/2)(1/\sqrt{5})(x_{t-3} + 2x_{t-5})\} + \\
exp\{-[(1/\sqrt{15})(-2x_{t-1} + 2x_{t-2} - 2x_{t-3} + x_{t-4} - x_{t-5} + x_{t-6})]^2\} + \varepsilon_t \\
\varepsilon_t = \left[\sqrt{h_t}\right] e_t \\
h_t = 1 + (.1)(\varepsilon_{t-1}^2 + \varepsilon_{t-2}^2) \\
e_t \sim N(0, 0.2) \\
\boldsymbol{\Phi}_d = \begin{bmatrix} \frac{1}{\sqrt{5}} & 0 & \frac{-2}{\sqrt{15}} \\ 0 & 0 & \frac{2}{\sqrt{15}} \\ 0 & \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{15}} \\ \frac{2}{\sqrt{5}} & 0 & \frac{1}{\sqrt{15}} \\ 0 & 0 & \frac{-1}{\sqrt{15}} \\ 0 & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{15}} \end{bmatrix} \\
\boldsymbol{\Gamma}_{\widetilde{d}} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \\ \frac{1}{\sqrt{2}} \end{bmatrix}
\end{aligned}
\tag{6.10}
$$

Under the parameter matrices setup, we decided to set three levels of the simulation sample sizes: $n = 100$, 300 and 1000. We decided to work with smaller sample sizes than the first five models due to computational time. The numerical optimization of a large number of variables simultaneously is computationally expensive, and using larger sample sizes would require a unreasonable amount of time.

We compute only the estimators based on the iterative estimation and simultaneous estimation approaches for Models 6 and 7. Section 4.4 explains why the sparse estimator is not applicable for the parameter matrices problem. The results reported in the tables for Models 6 and 7 also report the accuracy measures ρ and m^2 based on 100 Monte Carlo

replications. Now, besides identifying p and q, we also use a MSBC criterion for selecting d and $\widetilde{d}$. Each one of these dimension measures are chosen holding all others at their true level.

Table 6.11: Model 6 Simulation Results under Iterative Estimation

n	$\boldsymbol{\Phi}_d/\boldsymbol{\Gamma}_{\widetilde{d}}$	ρ	m^2	$MSBC(p)$	$MSBC(q)$	$MSBC(d)$	$MSBC(\widetilde{d})$
100	$\widehat{\boldsymbol{\Phi}}$	.3537	[.291; .847]	$f_{(p=1)} = 0$	$f_{(q=1)} = 57$	$f_{(d=1)} = 98$	$\boldsymbol{f_{(\widetilde{d}=1)} = 99}$
	$\widehat{\boldsymbol{\Gamma}}$	.8377	.483	$f_{(p=2)} = 1$	$\boldsymbol{f_{(q=2)} = 1}$	$\boldsymbol{f_{(d=2)} = 0}$	$f_{(\widetilde{d}=2)} = 1$
				$f_{(p=3)}1$	$f_{(q=3)} = 04$	$f_{(d=3)} = 2$	$f_{(\widetilde{d}=3)} = 0$
				$f_{(p=4)} = 7$	$f_{(q=4)} = 14$		
				$f_{(p=5)} = 10$			
				$\boldsymbol{f_{(p=6)} = 79}$			
				$f_{(p=7)} = 2$			
				$f_{(p=8)} = 0$			
300	$\widehat{\boldsymbol{\Phi}}$	.4501	[.165; .736]	$f_{(p=1)} = 0$	$f_{(q=1)} = 48$	$f_{(d=1)} = 87$	$\boldsymbol{f_{(\widetilde{d}=1)} = 92}$
	$\widehat{\boldsymbol{\Gamma}}$	.8560	.453	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 3}$	$\boldsymbol{f_{(d=2)} = 0}$	$f_{(\widetilde{d}=2)} = 8$
				$f_{(p=3)} = 0$	$f_{(q=3)} = 14$	$f_{(d=3)} = 13$	$f_{(\widetilde{d}=3)} = 0$
				$f_{(p=4)} = 0$	$f_{(q=4)} = 15$		
				$f_{(p=5)} = 0$			
				$\boldsymbol{f_{(p=6)} = 90}$			
				$f_{(p=7)} = 8$			
				$f_{(p=8)} = 2$			
1000	$\widehat{\boldsymbol{\Phi}}$	.6948	[.088; .297]	$f_{(p=1)} = 0$	$f_{(q=1)} = 25$	$f_{(d=1)} = 0$	$\boldsymbol{f_{(\widetilde{d}=1)} = 55}$
	$\widehat{\boldsymbol{\Gamma}}$	.9067	.351	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 33}$	$\boldsymbol{f_{(d=2)} = 1}$	$f_{(\widetilde{d}=2)} = 45$
				$f_{(p=3)} = 0$	$f_{(q=3)} = 20$	$f_{(d=3)} = 99$	$f_{(\widetilde{d}=3)} = 0$
				$f_{(p=4)} = 0$	$f_{(q=4)} = 22$		
				$f_{(p=5)} = 0$			
				$\boldsymbol{f_{(p=6)} = 91}$			
				$f_{(p=7)} = 8$			
				$f_{(p=8)} = 1$			

Models 6 and 7 illustrate the case when $\boldsymbol{\Phi}$ has more than one column. Table 6.11 summarizes the results of the iterative estimation for data simulated through Model 6, in which the mean parameter matrix has six rows and two columns. We observe that the

values of ρ are relatively low on smaller sample sizes. The m^2 measure is more informative to understand why estimating $\boldsymbol{\Phi}_d$ is challenging, since we can clearly notice that the first column[3] is estimated much more precisely than the second, which affects the overall ρ. This behavior could be explained by the fact that the first column has a linear effect on the mean function, whereas the second column affects $E(x_t|\boldsymbol{X}_{t-1})$ through a cosine function, which may make its detection more complex. Figure 8.12 in Appendix illustrates the Nadaraya-Watson fit for one of the 1000 observations replicate, and it is possible to observe that the monotonic linear effect can be easily seen, whereas the wave-shaped behavior is more nuanced. Nevertheless, $\widehat{\boldsymbol{\Phi}}$ and $\widehat{\boldsymbol{\Gamma}}$ improve as the sample size increases, indicating consistency.

The accuracy measure results show that, under Model 6, we might need larger sample sizes in order to achieve reasonable accuracy. Therefore, lag selection can also be challenging under the scenarios of our simulation experiment. We observe on Table 6.11 that the selection of p is correct for all three scenarios, and there is an apparent convergence of $\widehat{q}$ to its true value. However, the choosing process of d and $\widetilde{d}$ was ineffective, which may be caused by the low accuracy of the parameter matrices estimates.

[3]Defined on Equation (6.9).

Table 6.12: Model 6 Simulation Results under Simultaneous Estimation

n	$\boldsymbol{\Phi}_d/\boldsymbol{\Gamma}_{\widetilde{d}}$	ρ	m^2	$MSBC(p)$	$MSBC(q)$	$MSBC(d)$	$MSBC(\widetilde{d})$
100	$\widehat{\boldsymbol{\Phi}}_m$	.3256	[.280; .867]	$f_{(p=1)} = 0$	$f_{(q=1)} = 58$	$f_{(d=1)} = 99$	$\boldsymbol{f_{(\widetilde{d}=1)} = 97}$
	$\widehat{\boldsymbol{\Gamma}}_m$	.9016	.357	$f_{(p=2)} = 2$	$\boldsymbol{f_{(q=2)} = 17}$	$\boldsymbol{f_{(d=2)} = 0}$	$f_{(\widetilde{d}=2)} = 3$
				$f_{(p=3)} = 0$	$f_{(q=3)} = 12$	$f_{(d=3)} = 1$	$f_{(\widetilde{d}=3)} = 0$
				$f_{(p=4)} = 4$	$f_{(q=4)} = 13$		
				$f_{(p=5)} = 4$			
				$\boldsymbol{f_{(p=6)} = 89}$			
				$f_{(p=7)} = 1$			
				$f_{(p=8)} = 0$			
300	$\widehat{\boldsymbol{\Phi}}_m$	.4634	[.163; .709]	$f_{(p=1)} = 0$	$f_{(q=1)} = 50$	$f_{(d=1)} = 75$	$\boldsymbol{f_{(\widetilde{d}=1)} = 84}$
	$\widehat{\boldsymbol{\Gamma}}_m$	.8853	.405	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 28}$	$\boldsymbol{f_{(d=2)} = 0}$	$f_{(\widetilde{d}=2)} = 16$
				$f_{(p=3)} = 0$	$f_{(q=3)} = 14$	$f_{(d=3)} = 25$	$f_{(\widetilde{d}=3)} = 0$
				$f_{(p=4)} = 0$	$f_{(q=4)} = 08$		
				$f_{(p=5)} = 0$			
				$\boldsymbol{f_{(p=6)} = 99}$			
				$f_{(p=7)} = 1$			
				$f_{(p=8)} = 0$			
1000	$\widehat{\boldsymbol{\Phi}}_m$	.7998	[.099; .301]	$f_{(p=1)} = 0$	$f_{(q=1)} = 41$	$f_{(d=1)} = 0$	$\boldsymbol{f_{(\widetilde{d}=1)} = 43}$
	$\widehat{\boldsymbol{\Gamma}}_m$	.8565	.459	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 29}$	$\boldsymbol{f_{(d=2)} = 1}$	$f_{(\widetilde{d}=2)} = 57$
				$f_{(p=3)} = 0$	$f_{(q=3)} = 15$	$f_{(d=3)} = 99$	$f_{(\widetilde{d}=3)} = 0$
				$f_{(p=4)} = 0$	$f_{(q=4)} = 15$		
				$f_{(p=5)} = 0$			
				$\boldsymbol{f_{(p=6)} = 84}$			
				$f_{(p=7)} = 14$			
				$f_{(p=8)} = 2$			

Under simultaneous estimation, the simulation results on Table 6.12 are less precise than those from the iterative estimation approach. In fact, the estimation of $\boldsymbol{\Gamma}_{\widetilde{d}}$ seems to become less accurate as the sample size increases. Additionally the selection of p, q, d and $\widetilde{d}$ are, in general, also less efficient than in the iterative estimation approach.

Table 6.13: Model 7 Simulation Results under Iterative Estimation

n	$\boldsymbol{\Phi}_d/\boldsymbol{\Gamma}_{\widetilde{d}}$	ρ	m^2	$MSBC(p)$	$MSBC(q)$	$MSBC(d)$	$MSBC(\widetilde{d})$
100	$\widehat{\boldsymbol{\Phi}}$	.7250	[.353; .199; .401]	$f_{(p=1)} = 0$	$f_{(q=1)} = 95$	$f_{(d=1)} = 95$	$\boldsymbol{f_{(\widetilde{d}=1)} = 100}$
	$\widehat{\boldsymbol{\Gamma}}$	.8042	.534	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 03}$	$f_{(d=2)} = 5$	$f_{(\widetilde{d}=2)} = 0$
				$f_{(p=3)} = 21$	$f_{(q=3)} = 02$	$\boldsymbol{f_{(d=3)} = 0}$	$f_{(\widetilde{d}=3)} = 0$
				$f_{(p=4)} = 47$	$f_{(q=4)} = 0$	$f_{(d=4)} = 0$	
				$f_{(p=5)} = 0$			
				$\boldsymbol{f_{(p=6)} = 32}$			
				$f_{(p=7)} = 2$			
				$f_{(p=8)} = 0$			
300	$\widehat{\boldsymbol{\Phi}}$	.9708	[.152; .104; .128]	$f_{(p=1)} = 0$	$f_{(q=1)} = 92$	$f_{(d=1)} = 1$	$\boldsymbol{f_{(\widetilde{d}=1)} = 100}$
	$\widehat{\boldsymbol{\Gamma}}$	.7951	.557	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 6}$	$f_{(d=2)} = 99$	$f_{(\widetilde{d}=2)} = 0$
				$f_{(p=3)} = 0$	$f_{(q=3)} = 2$	$\boldsymbol{f_{(d=3)} = 0}$	$f_{(\widetilde{d}=3)} = 0$
				$f_{(p=4)} = 0$	$f_{(q=4)} = 0$	$f_{(d=4)} = 0$	
				$f_{(p=5)} = 0$			
				$\boldsymbol{f_{(p=6)} = 100}$			
				$f_{(p=7)} = 0$			
				$f_{(p=8)} = 0$			
1000	$\widehat{\boldsymbol{\Phi}}$	.9907	[.087; .071; .066]	$f_{(p=1)} = 0$	$f_{(q=1)} = 89$	$f_{(d=1)} = 0$	$\boldsymbol{f_{(\widetilde{d}=1)} = 100}$
	$\widehat{\boldsymbol{\Gamma}}$	.8236	.508	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 7}$	$f_{(d=2)} = 48$	$f_{(\widetilde{d}=2)} = 0$
				$f_{(p=3)} = 0$	$f_{(q=3)} = 2$	$\boldsymbol{f_{(d=3)} = 52}$	$f_{(\widetilde{d}=3)} = 0$
				$f_{(p=4)} = 0$	$f_{(q=4)} = 2$	$f_{(d=4)} = 0$	
				$f_{(p=5)} = 0$			
				$\boldsymbol{f_{(p=6)} = 100}$			
				$f_{(p=7)} = 0$			
				$f_{(p=8)} = 0$			

Table 6.13 summarizes results of the simulation experiment under Model 7. Our last data generating model assumes that $\boldsymbol{\Phi}_d$ has three columns and six rows, totaling 18 parameters only on the mean function. This model assumes the same mean structure and white noise term ($e_t \sim N(0, 0.2)$) as in Park et al. (2009), but also adding a conditional variance structure to the problem.

The accuracy measures of $\boldsymbol{\Phi}_d$ are much higher than those from Model 6, indicating

that adopting a weaker white noise term, allows us to use smaller sample sizes in order to achieve precise estimates. Additionally, when $n = 100$ and $n = 300$, our average ρ measure is similar to those found in (Park et al. 2009, p. 723), in which the error term was homoscedastic. Interestingly, the accuracy of $\boldsymbol{\Gamma}_{\widetilde{d}}$ estimates are smaller than those in Table 6.11, so we have an indication that reducing the white noise amplitude might also make variance terms estimation less efficient.

Under the Model 7 scenario, the selection of p is overwhelmingly precise, but q appears to be majorly underestimated. Again, this behavior could be a reflection of our choice of distribution of e_t. The selection of d is better than in the previous model, as it improves for larger sample sizes. Lastly, the selection of $\widetilde{d}$ is highly accurate for all three scenarios.

Table 6.14: Model 7 Simulation Results under Simultaneous Estimation

n	$\boldsymbol{\Phi}_d/\boldsymbol{\Gamma}_{\widetilde{d}}$	ρ	m^2	$MSBC(p)$	$MSBC(q)$	$MSBC(d)$	$MSBC(\widetilde{d})$
100	$\widehat{\boldsymbol{\Phi}}_m$	.7812	[.320; .176; .358]	$f_{(p=1)} = 0$	$f_{(q=1)} = 95$	$f_{(d=1)} = 95$	$\boldsymbol{f_{(\widetilde{d}=1)} = 100}$
	$\widehat{\boldsymbol{\Gamma}}_m$	.8026	.540	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 03}$	$f_{(d=2)} = 05$	$f_{(\widetilde{d}=2)} = 0$
				$f_{(p=3)} = 0$	$f_{(q=3)} = 2$	$\boldsymbol{f_{(d=3)} = 0}$	$f_{(\widetilde{d}=3)} = 0$
				$f_{(p=4)} = 1$	$f_{(q=4)} = 0$	$f_{(d=4)} = 0$	
				$f_{(p=5)} = 0$			
				$\boldsymbol{f_{(p=6)} = 99}$			
				$f_{(p=7)} = 0$			
				$f_{(p=8)} = 0$			
300	$\widehat{\boldsymbol{\Phi}}_m$	.9693	[.149; .105; .130]	$f_{(p=1)} = 0$	$f_{(q=1)} = 97$	$f_{(d=1)} = 0$	$\boldsymbol{f_{(\widetilde{d}=1)} = 100}$
	$\widehat{\boldsymbol{\Gamma}}_m$	.8514	.452	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 03}$	$f_{(d=2)} = 100$	$f_{(\widetilde{d}=2)} = 0$
				$f_{(p=3)} = 0$	$f_{(q=3)} = 00$	$\boldsymbol{f_{(d=3)} = 0}$	$f_{(\widetilde{d}=3)} = 0$
				$f_{(p=4)} = 0$	$f_{(q=4)} = 00$	$f_{(d=4)} = 0$	
				$f_{(p=5)} = 0$			
				$\boldsymbol{f_{(p=6)} = 100}$			
				$f_{(p=7)} = 0$			
				$f_{(p=8)} = 0$			
1000	$\widehat{\boldsymbol{\Phi}}_m$	.9911	[.088; .066; .064]	$f_{(p=1)} = 0$	$f_{(q=1)} = 92$	$f_{(d=1)} = 0$	$\boldsymbol{f_{(\widetilde{d}=1)} = 99}$
	$\widehat{\boldsymbol{\Gamma}}_m$	.9029	.356	$f_{(p=2)} = 0$	$\boldsymbol{f_{(q=2)} = 04}$	$f_{(d=2)} = 45$	$f_{(\widetilde{d}=2)} = 1$
				$f_{(p=3)} = 0$	$f_{(q=3)} = 02$	$\boldsymbol{f_{(d=3)} = 55}$	$f_{(\widetilde{d}=3)} = 0$
				$f_{(p=4)} = 0$	$f_{(q=4)} = 02$	$f_{(d=4)} = 0$	
				$f_{(p=5)} = 0$			
				$\boldsymbol{f_{(p=6)} = 100}$			
				$f_{(p=7)} = 0$			
				$f_{(p=8)} = 0$			

Lastly, we present the simulation results under simultaneous estimation for Model 7. In general, we can conclude the results of simultaneously estimating $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ in Table 6.14 are much similar to those found by the iterative estimation approach, so we cannot conclude if one is more accurate than the other.

Overall, our simulation results when $\boldsymbol{\Phi}_d$ has multiple columns are consistent with those from Section 6.1. Our main estimators $\widehat{\boldsymbol{\Phi}}$ and $\widehat{\boldsymbol{\Gamma}}$ demonstrate the expected consistency,

and the selection of p is more accurate than choosing q. Our proposed selection approach for d and $\widetilde{d}$ seems to work well when we can obtain accurate estimates for the parameter matrices under the number of columns true values. Nevertheless, we should stress that the main purpose of this research is to establish estimators for the parameter matrices, for which consistency was shown theoretically and further validated through simulations. The task of selecting the dimensions of $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ deserves an entire study dedicated to it, which could not be largely investigated here because of the high computational cost of each simulation scenario.

Chapter 7

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, thus major investment decisions cannot be done. 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.

Over the second half of the twentieth century, Brazil has experienced an accelerating inflationary process. During this time, several economic measures were attempted [4] to control the general price increase. When we observe the General Price Index - Overall Supply (IGP-OG) inflation index provided by the Institute for Applied Economic Research (IPEA) on Figure 7.1, it is clear that it was only after adopting the Real Currency (BRL) that inflation withdrew to reasonable standards.

[4] Including the adoption of new currencies.

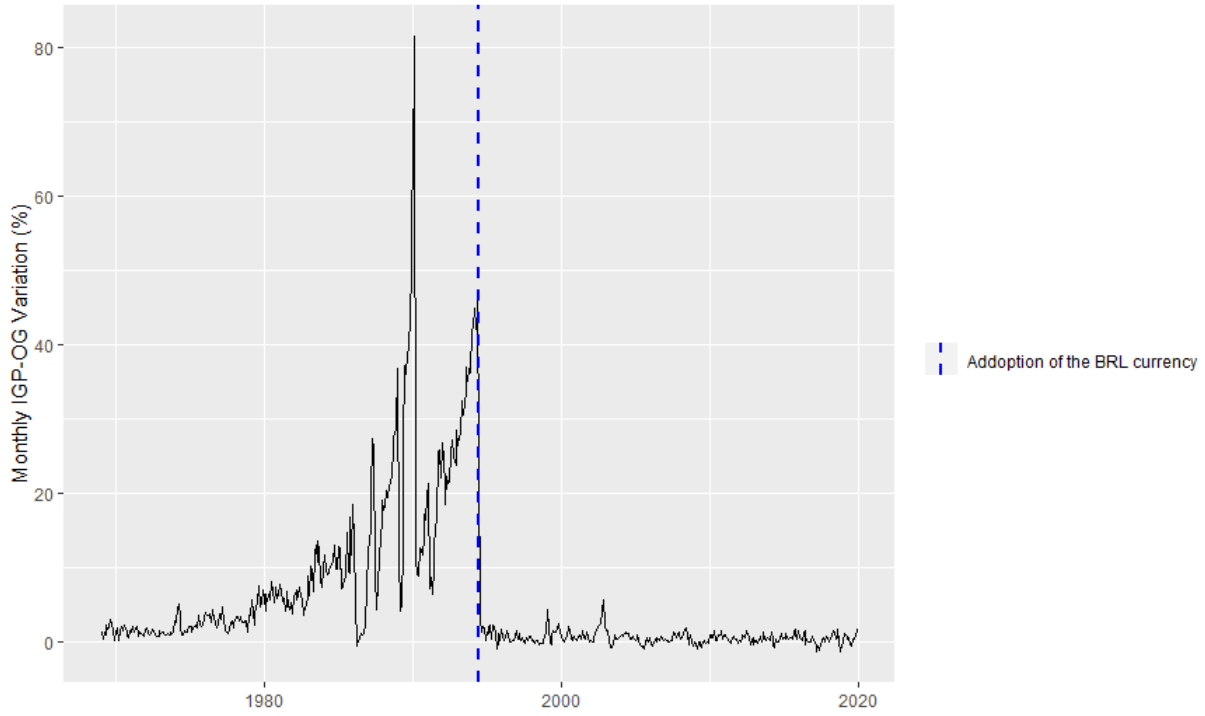

Figure 7.1: Monthly Brazilian Inflation (%)

Figure 7.1 illustrates how important the Real currency is to the Brazilian recent economic history. Moreover, it is still relevant to evaluate how strong the newest currency is not only compared to domestic products, but also in relation to other currencies as well. 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 thesis, we analyze the monthly BRL/USD. Foreign Exchange Rate series from January 1999 to December 2019; see Figure 7.2 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, in which the currency price fluctuates mostly according to market forces, instead of government fixation. From the adoption of the BRL as the official currency on July/1994 until early January/1999 the fixed exchange regime was in place, so for many months there would not be any variation on the series, and it is sensible to not include this period into the analysis.

We split the time series into two parts. The training dataset ranges from 01/1999 to 10/2015 (202 observations) whereas the test set starts at 11/2015 and goes until 12/2019 (50 observations). First, we consider only the training dataset in order to build our time series models. Our goal is to evaluate how well our fitted models are capable of generating accurate out-of-sample forecasts of the exchange rate series.

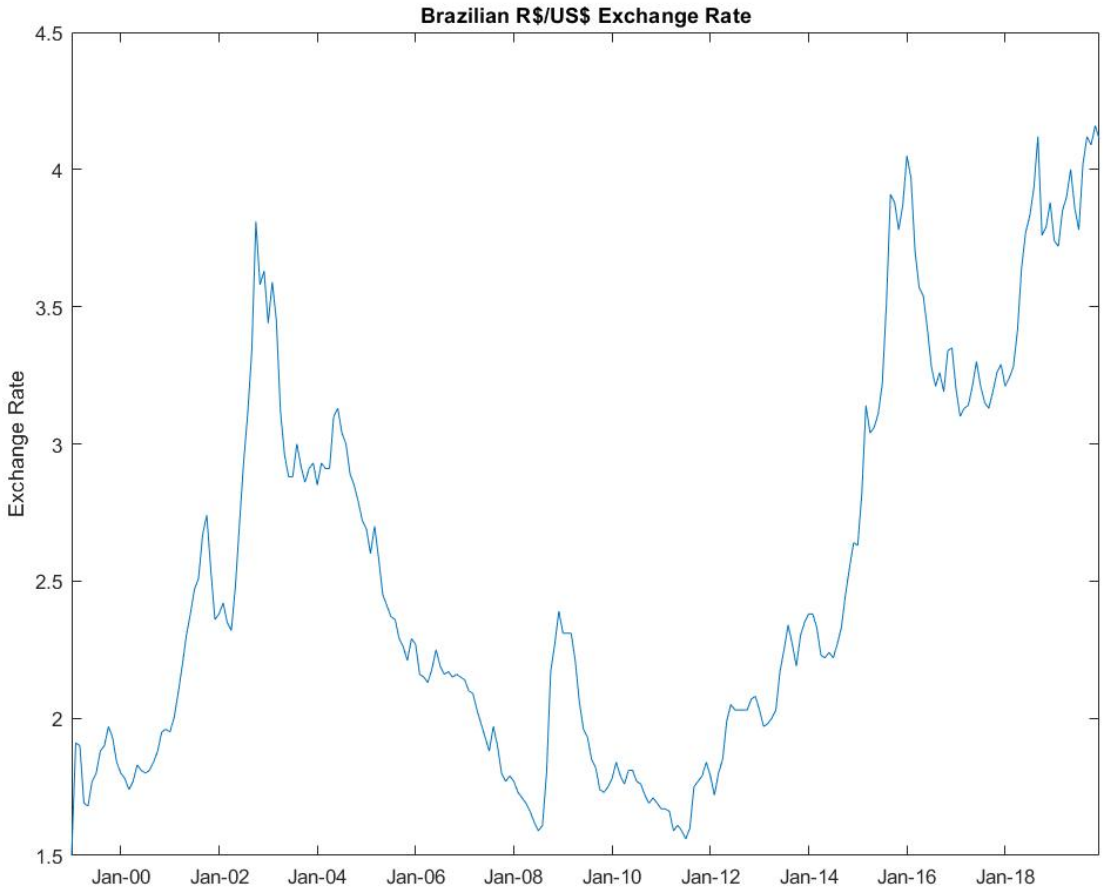

Figure 7.2: BRL/USD Exchange Rate monthly series.

7.1 AR-ARCH Model

A benchmark in conditional mean and variance parametric modeling is the family of AR(p)-ARCH(q) models. Hence, we first find a suitable AR-ARCH model that fits the monthly

BRL/USD Exchange Rate series in the training dataset (01/1999 to 10/2015). Our goal is to compare the performance of the fitted AR-ARCH model to other models fitted here based on respective out-of-sample forecasts.

The first step is to find appropriate lags p and q for the mean and variance functions, respectively, using the SBC criterion. Thus, we fit all possible combinations within a grid of values for p and q in MATLAB through the *estimate* function. Table 7.1 gives the SBC values for various choices of p and q, respectively.

Table 7.1: AR(p)-ARCH(q) lag selection by SBC

SBC	$q = 1$	$q = 2$
$p = 1$	-371.0644	-378.9687
$p = 2$	-409.0871	-409.1751
$p = 3$	-406.3224	-408.8852
$p = 4$	-413.3367	-410.5773
$p = 5$	–408.0576	-406.6763
$p = 6$	-407.0810	-403.6766

We can observe from Table 7.1 that the SBC criteria is minimized when $p = 4$ and $q = 1$. Then, we obtained following AR(4)-ARCH(1) fitted model using Matlab:

$$\begin{aligned} \widehat{x}_t &= -0.023455 + 1.6059x_{t-1} - 0.90397x_{t-2} + 0.55746x_{t-3} - 0.26837x_{t-4} + \widehat{\varepsilon}_t \\ \widehat{\varepsilon}_t &= \sqrt{\widehat{h}_t}\,\widehat{e}_t \qquad (7.1) \\ \widehat{h}_t &= 0.0029 + 0.7695\widehat{\varepsilon}^2_{t-1} \end{aligned}$$

where $\widehat{e}_t$ are residuals that are approximately normally distributed with mean zero and variance 1. All estimated coefficients are significant at the 1% level, except for the intercept in the mean function.

7.2 Model fitting based on iterative estimation with one linear combination.

Now, we will use our proposed iterative estimation approach to fit a model for the training dataset. First, we will consider the case when the mean and variance functions depend on only one linear combination, i.e., we assume for now that $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ are vectors. The first step of the analysis is the selection of lag p using the quantity defined in (4.19). However, since q is also unknown, we compute the Modified SBC criterion [in (4.19)] using $\widehat{\boldsymbol{\Phi}}_0$ [instead of $\widehat{\boldsymbol{\Phi}}$], which is defined in (4.6).

Table 7.2: Selection of p based on Modified SBC

p	MSBC
1	-3.0462
2	-3.2063
3	-3.2015
4	-3.1635
5	-3.1198
6	-3.0743

From Table 7.2, the MSBC criteria is smallest when $p = 2$, so we define $\widehat{p} = 2$. In order to find q, we compute the MSBC value for $q \in [1, 4]$, and it is based on the final iterative estimates $\widehat{\boldsymbol{\Phi}}$ and $\widehat{\boldsymbol{\Gamma}}$ defined in (4.10) and (4.9), respectively, for fixed $(\widehat{p}, q)$ pairs.

Table 7.3: Selection of q based on Modified SBC

q	MSBC
1	-5.2754
2	-5.2320
3	-5.1808
4	-5.1305

According to Table 7.3, the selection criterion is minimized when $\widehat{q} = 1$. After choosing

the dimensions of $\boldsymbol{\Phi}$ and $\boldsymbol{\Gamma}$ we computed the iterative estimates: $\widehat{\boldsymbol{\Phi}} = (0.8273, 0.5617)^\intercal$ and $\widehat{\boldsymbol{\Gamma}} = 1$.

Let $l_{\boldsymbol{\Phi},t} = \widehat{\boldsymbol{\Phi}}^T \mathbf{X}_{t-1}$. We first look at the time series plot of x_t and $l_{\boldsymbol{\Phi},t}$ to check how the fitted linear combination of past observations relates to x_t. Figure 7.3 shows us how the fitted linear combination series, $l_{\boldsymbol{\Phi},t}$, itself seems to capture the dynamics of the observed series even before we have modeled the series.

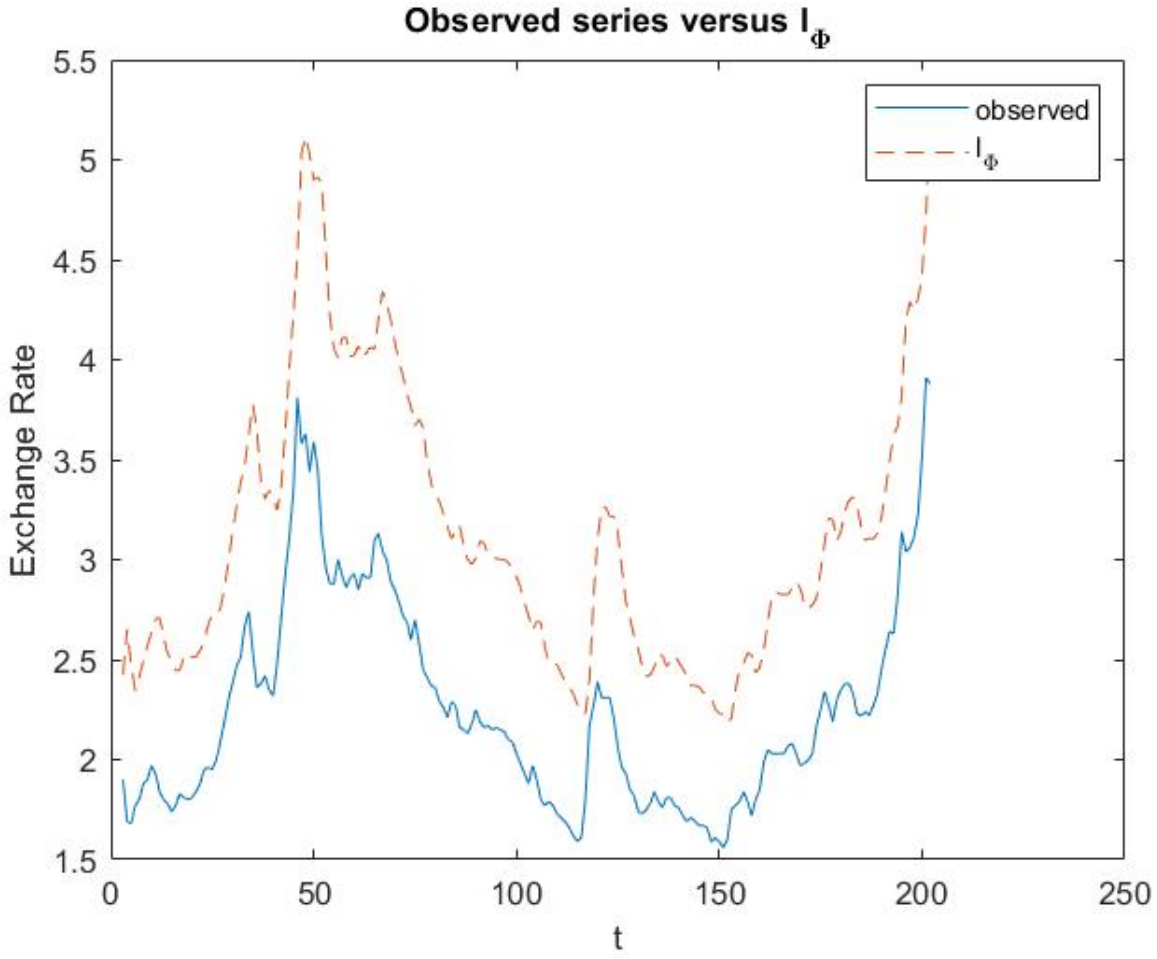

Figure 7.3: x_t and $l_{\Phi t}$ over time.

Our goal now is to build a scatter plot of x_t (the original series) vs $l_{\boldsymbol{\Phi},t}$ to determine a reasonable choice for $f(\cdot)$ to model the mean function based on $l_{\boldsymbol{\Phi},t}$. The scatter plot given in Figure 7.4 presents a clear increasing relationship between x_t and $l_{\boldsymbol{\Phi},t}$. Additionally, a visual inspection suggests that a linear function seems to be an appropriate choice in this

case.

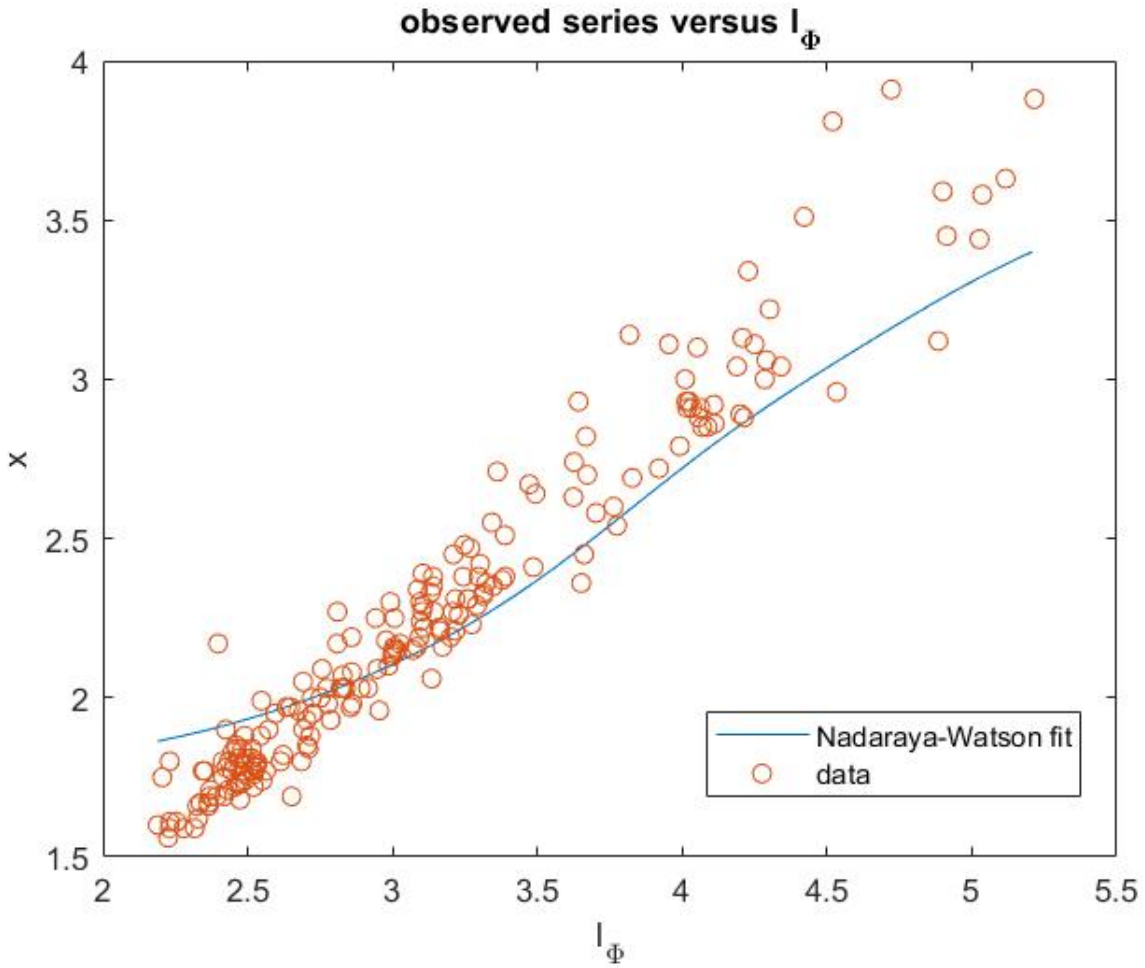

Figure 7.4: x_t vs. $l_{\boldsymbol{\Phi},t}$ Scatter plot

The Figure 7.4 also overlays a plot of the fitted Nadaraya-Watson (N-W) curve over the scatter plot to further help us visualize what function $\widehat{f}$ was estimated based on the data and $\widehat{\boldsymbol{\Phi}}$. This N-W curve visualization is a better way determine the functional relationships than the ones proposed in Park and Samadi (2019), where $f(\cdot)$ and $g(\cdot)$ are not directly estimated[5]. Therefore, the fitted curve can aid us into further choosing similar parametric functions when the scatter plots are not entirely informative. In fact, this visualization feature is even more important when we have two linear combinations, since interpreting a 3-D scatter plot is usually much more challenging. For instance, Figure 8.12 on Appendix

[5]The functions $f(\cdot)$ and $g(\cdot)$ in our notation are represented by $g_1(\cdot)$ and $g_2(\cdot)$ in Park and Samadi (2019).

B shows us how the fitted curve can detect the original patterns on a simulation based on Model 6 from Section 6.

The overall behavior in Figure 7.4 looks roughly linear, so we decide to fit the following linear model for x_t as a function of $d_{1,t}$:

$$x_t = \beta_0 + \beta_1 l_{\boldsymbol{\Phi},t} + \varepsilon_t \tag{7.2}$$

The estimates of the parameters are given by $\widehat{\beta}_0 = -0.0065$ and $\widehat{\beta}_1 = 0.7250$, so the estimated mean function is already defined and residuals can be computed. In order to model the variance function, we analyze case of the squared fitted residuals $\widehat{\varepsilon}_t^2$. Let $l_{\boldsymbol{\Gamma},t} = \widehat{\boldsymbol{\Gamma}}^T \widehat{\boldsymbol{\varepsilon}}_{t-1}^2$. First, we inspect the time series plot of $\widehat{\varepsilon}_t^2$ and $l_{\boldsymbol{\Gamma},t}$ of Figure 7.5. From this plot, we conclude that $l_{\boldsymbol{\Gamma},t}$ can contain much information about the mean value of the squared residuals.

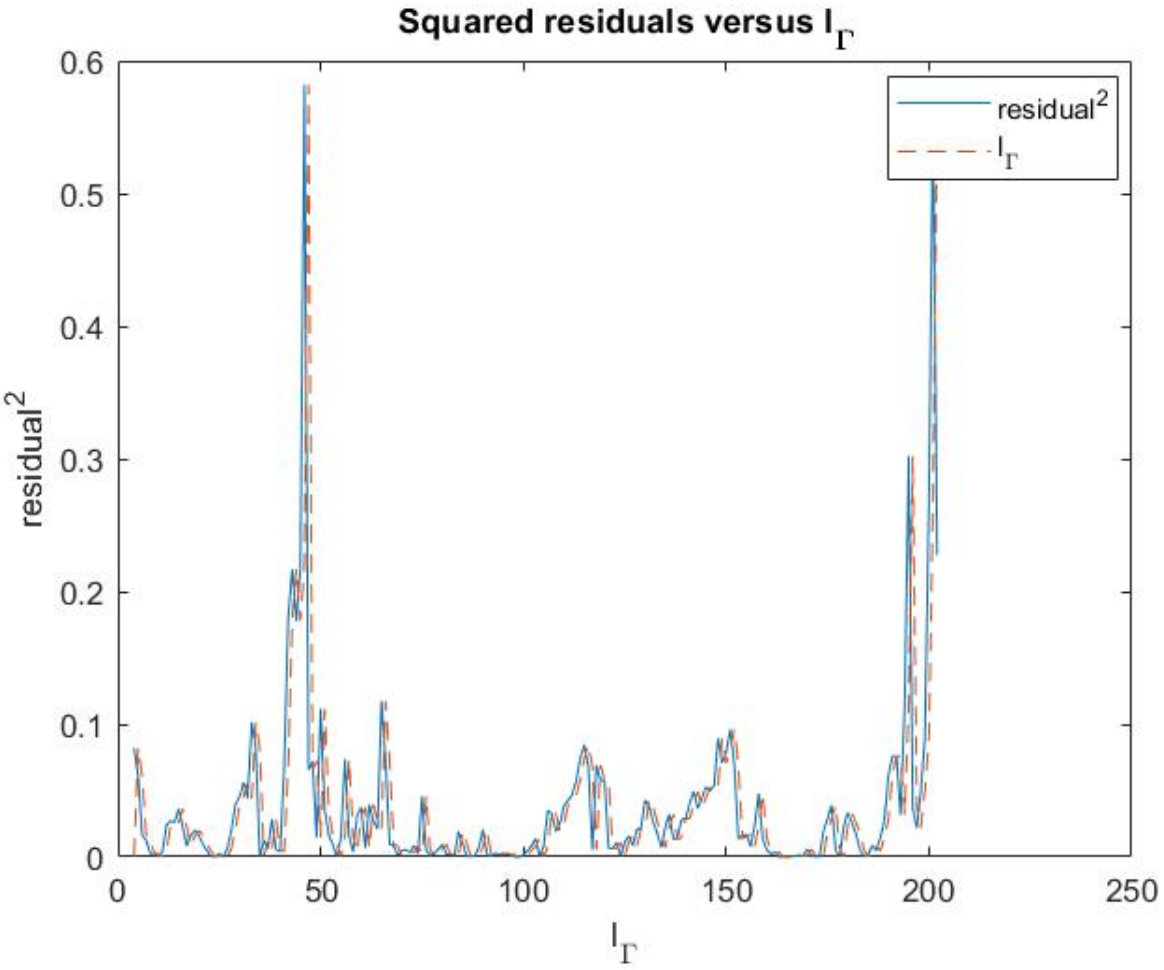

Figure 7.5: $\widehat{\varepsilon}_t^2$ and $l_{\boldsymbol{\Gamma},t}$ over time.

The last task of modeling the variance function is to find a parametric relation that describes the relationship between $\widehat{\varepsilon}_t^2$ and $l_{\boldsymbol{\Gamma},t}$. We produce a scatter plot of these two variables and try to identify a function that could be a good approximation of $g(\cdot)$.

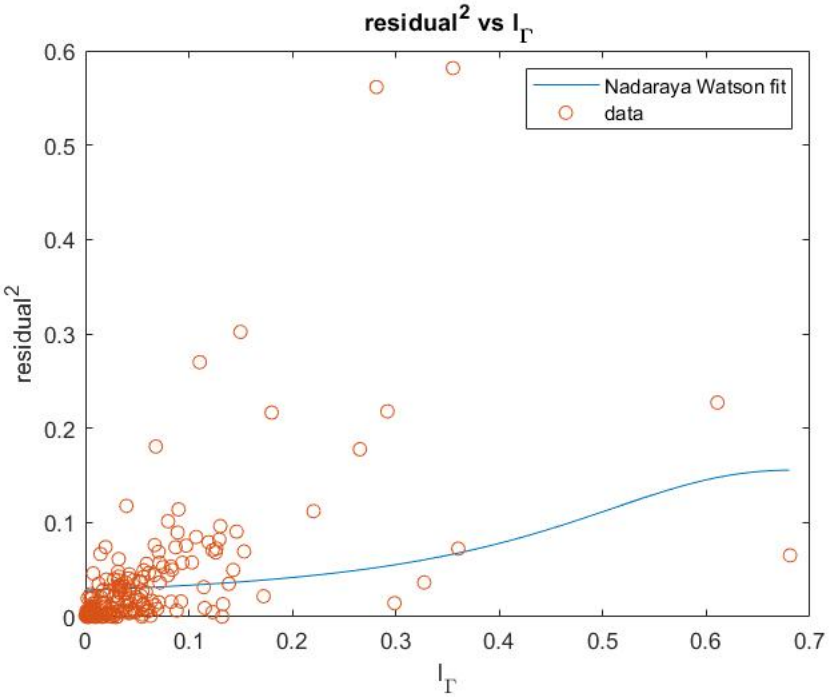

Figure 7.6: $\widehat{\varepsilon}_t^2$ vs. $l_{\boldsymbol{\Gamma},t}$ Scatter plot.

Interpreting the scatter plot of the conditional variance is usually less straightforward since it has multiplicative errors instead of additive ones, so the plot will not be as clear as in mean function modeling. Figure 7.6 shows the sctatter plot of $\widehat{\varepsilon}_t^2$ and $l_{\boldsymbol{\Gamma},t}$. By analyzing the plot we verify that it is harder to visualize what an appropriate choice of $g(\cdot)$ might be. In this scenario, the Nadaraya-Watson fitted curve can help us find a suitable choice, as previously discussed. Again, there is no strong evidence of a nonlinear relation, so we take $g(\cdot)$ to be a linear function. The final fitted model (including the variance function) is:

$$\begin{aligned} \widehat{x}_t &= -0.0065 + 0.7250 l_{\boldsymbol{\Phi},t} + \widehat{\varepsilon}_t \\ \widehat{\varepsilon}_t &= \sqrt{\widehat{h}_t}\widehat{e}_t \\ \widehat{h}_t &= 0.0167 + 0.5713 l_{\boldsymbol{\Gamma},t}. \end{aligned} \tag{7.3}$$

7.3 Model fitting based on iterative estimates with multiple linear combinations

Lastly, let us consider our iterative estimation approach for the case where $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Gamma}_{\widetilde{d}}$ have more than one column. We want to check if adding/detecting more linear combinations improves the modeling of conditional mean and variance functions.

To this end, first we estimate the lag parameters (p, q) and the number of linear combinations ($d, \widetilde{d}$) for the mean and variance parameter matrices. Our strategy is to first consider p and d . We select them through the MSBC criteria (see (4.19)), where we use the initial estimator $\widehat{\boldsymbol{\Phi}}_0$ in (4.6) in the place of $\widehat{\boldsymbol{\Phi}}$.

Table 7.4: Selection of p and d based on Modified SBC

MSBC	$d=1$	$d=2$	$d=3$	$d=4$
$p=1$	-3.0462			
$p=2$	-3.2063	-3.2295		
$p=3$	-3.2015	-3.1440	-2.8237	
$p=4$	-3.1635	-3.0379	-2.5887	-1.8993

According to Table 7.4, the combination that minimizes the selection criterion is $\widehat{p}=2$ and $\widehat{d}=2$. Now that the dimensions of $\boldsymbol{\Phi}$ are chosen, we search for q and $\widetilde{d}$ based again on minimizing the MSBC value.

Table 7.5: Selection of q and $\widetilde{d}$ based on Modified SBC

MSBC	$\widetilde{d}=1$	$\widetilde{d}=2$	$\widetilde{d}=3$
$q=1$	-5.5443		
$q=2$	-5.4998	-5.3462	
$q=3$	-5.4452	-5.2158	-4.8220
$q=4$	-5.3953	-5.0908	-4.5691

The combination that minimizes the selection criterion for the variance parameter matrix

is $\widehat{q} = 1$ and $\widehat{\widetilde{d}} = 1$ according to Table 7.5. Now that all dimensions are determined, we us use the iterative estimation approach to obtain our parameter matrices estimates:

$$\widehat{\boldsymbol{\Phi}} = \begin{bmatrix} 0.9783 & 0.2071 \\ -0.2071 & 0.9783 \end{bmatrix},$$

$$\widehat{\boldsymbol{\Gamma}} = 1.$$

After estimating the parameter matrices, our final task is to build a time series model from them. Let us define the estimated linear combinations of the past values of x_t as $l_{\boldsymbol{\Phi}_{1,t}} = [1\ 0]\widehat{\boldsymbol{\Phi}}^T\mathbf{X}_{t-1}$ and $l_{\boldsymbol{\Phi}_{2,t}} = [0\ 1]\widehat{\boldsymbol{\Phi}}^T\mathbf{X}_{t-1}$.

Note that the true parameter matrices are not identifiable. In the context of this application, the numerical optimization of equation (4.10) could lead to similar matrices, where columns of $\widehat{\boldsymbol{\Phi}}$ may be switched and/or re-scaled by -1. For instance, the following matrices are also equivalent estimates of $\boldsymbol{\Phi}$:

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

The reason why the lack of identifiability is not a problem is that any alternative version of $\widehat{\boldsymbol{\Phi}}$ will lead us to the same linear combinations, but with a possible sign change. Therefore, all versions are equally useful in the next step of building a predictive model from them.

Next, we look at the time series plot of x_t, $l_{\boldsymbol{\Phi}_{1,t}}$ and $l_{\boldsymbol{\Phi}_{2,t}}$ to verify how well the linear combinations can capture the dynamics of the original series.

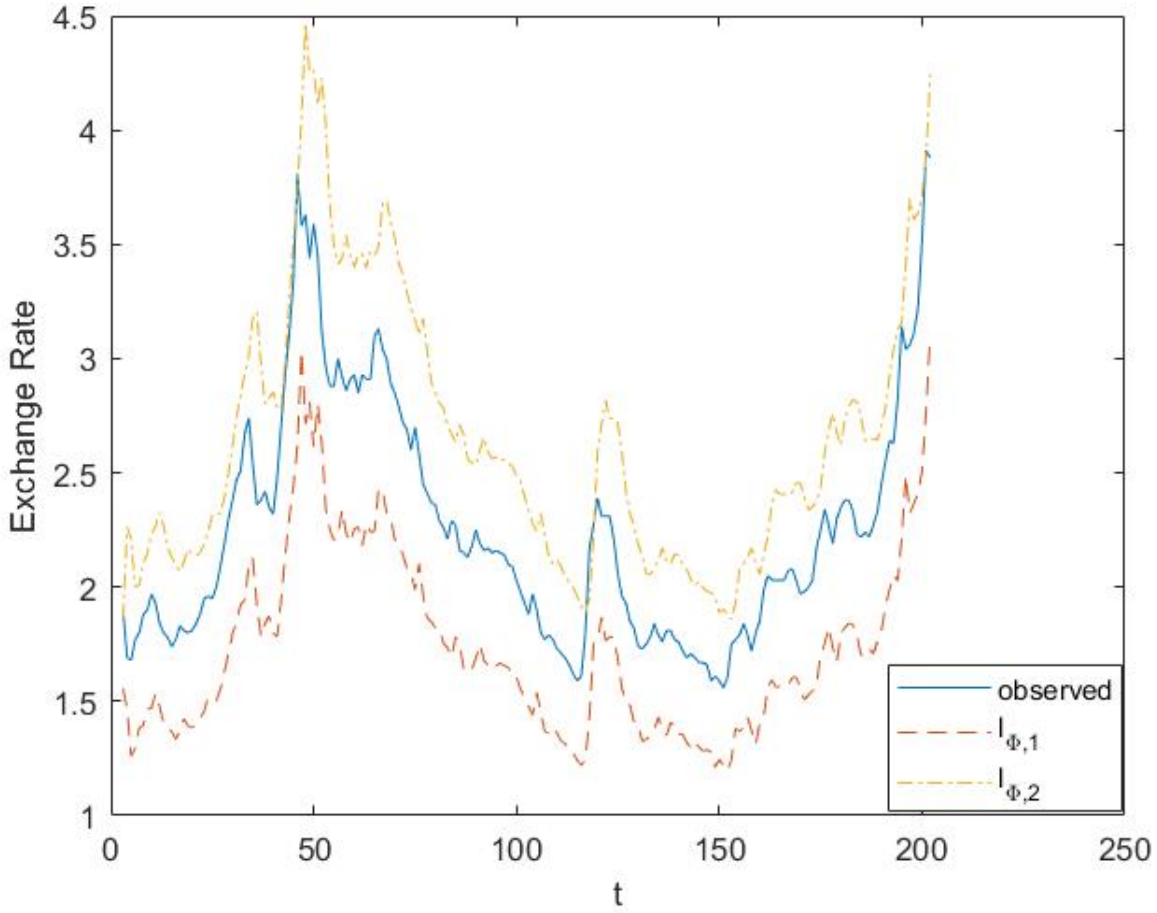

Figure 7.7: x_t , $l_{\Phi_{1,t}}$ and $l_{\Phi_{2,t}}$ over time

Figure 7.7 show us that in general, x_t has a directly proportional relation to $l_{\mathbf{\Phi}_{1,t}}$ and $l_{\mathbf{\Phi}_{2,t}}$. Thus, it seems plausible that we can build a more accurate predictive model based on them. The next step of the graphical analysis is to build a 3-D scatter plot with its Nadaraya-Watson fitted surface to visualize what choice of $f()$ might be appropriate for further modeling specifications.

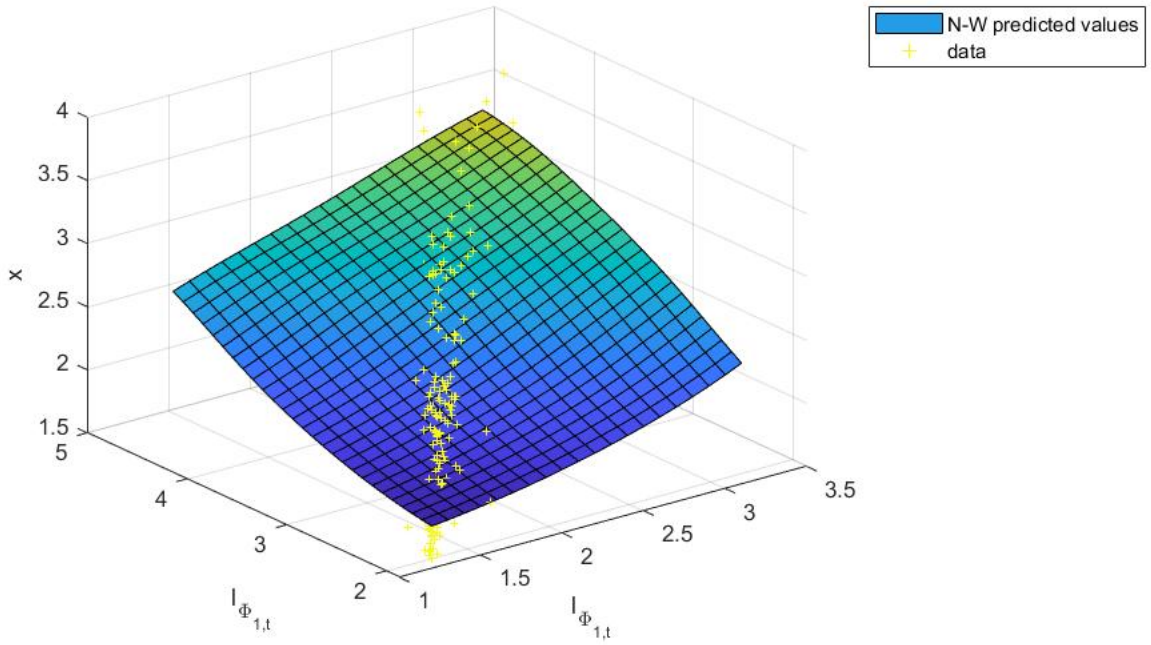

Figure 7.8: x_t , $l_{\mathbf{\Phi}_{1,t}}$ and $l_{\mathbf{\Phi}_{2,t}}$ Scatter Plot

From Figure 7.8, we see that both linear combinations appear to have an increasing effect over the response. It is hard to identify a clear nonlinear behavior, thus it is a sensible choice to fit the mean function as a linear model. The fitted equation is:

$$\widehat{x}_t = 0.0274 + 1.3569 l_{\mathbf{\Phi}_{1,t}} - 0.0473 l_{\mathbf{\Phi}_{2,t}} + \varepsilon_t \tag{7.4}$$

After determining the fitted values for the mean function, we compute the residuals following equation (7.4). We can work now with the squared residuals and the estimate $\widehat{\boldsymbol{\Gamma}}$ in order to model the variance function. Again, let us define $l_{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\Gamma}} \widehat{\boldsymbol{\varepsilon}}^2_{t-1}$.

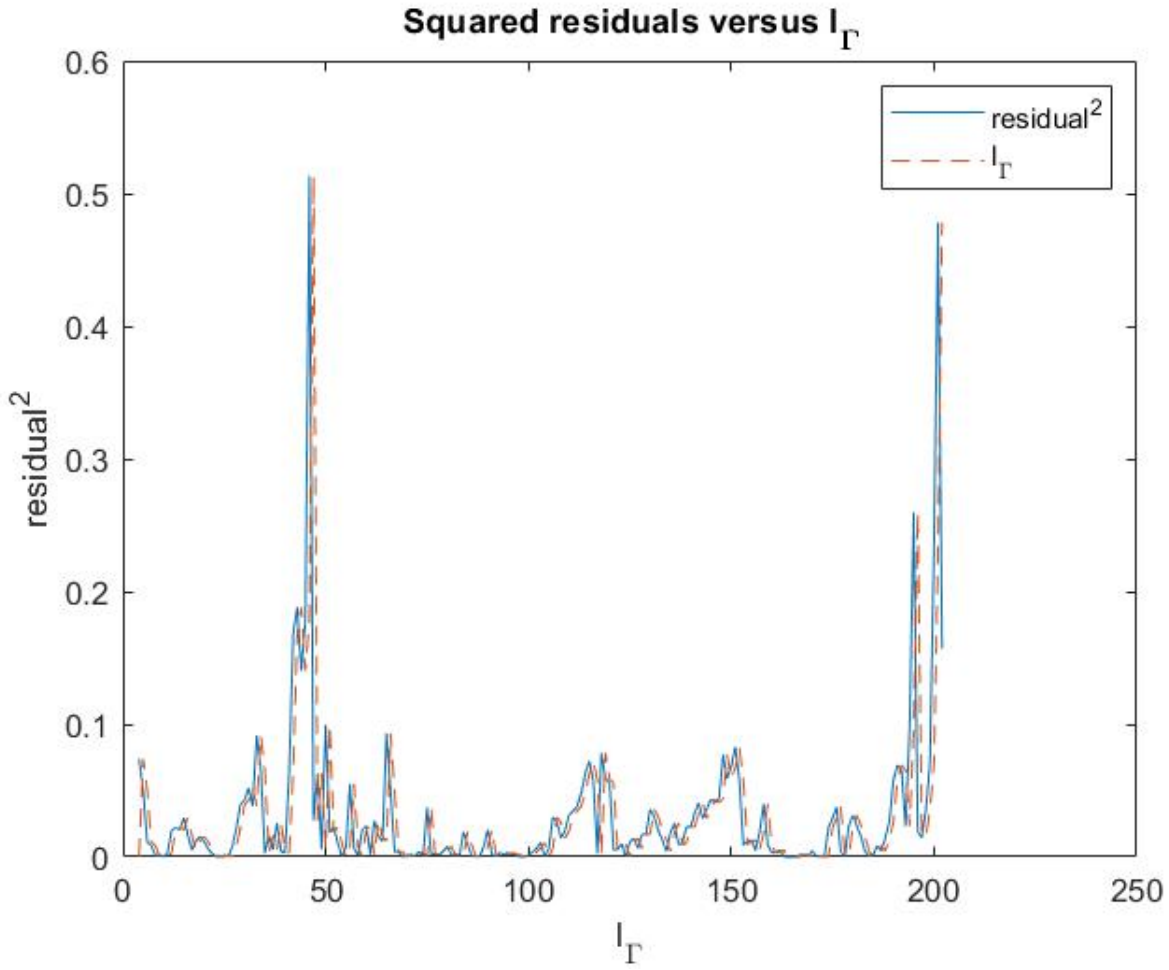

Figure 7.9: ε_t^2 and $l_{\Gamma t}$ time series plot for multidimensional analysis

The linear combination of past squared residuals seems to be a good potential predictor for $\widehat{\varepsilon}_t^2$, according to Figure 7.9. Hence, we investigate the scatterplot between them to choose an appropriate $g(\cdot)$ function.

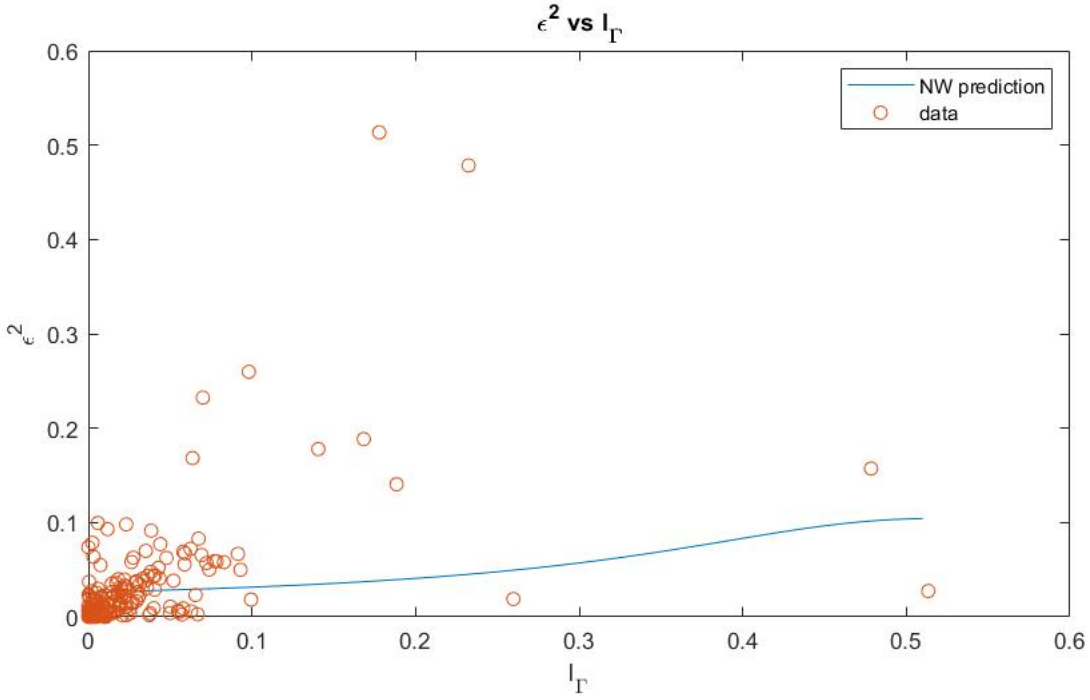

Figure 7.10: ε_t^2 vs. $l_{\Gamma t}$ for multidimensional analysis

The scatterplot in Figure 7.10 shows no clear nonlinear pattern, so we estimate $g(\cdot)$ through a linear model as well. The overall fitted model is then:

$$\begin{aligned} \widehat{x}_t &= 0.0274 + 1.3569 l_{\boldsymbol{\Phi}_{1,t}} - 0.0473 l_{\boldsymbol{\Phi}_{2,t}} + \widehat{\varepsilon}_t \\ \widehat{\varepsilon}_t &= \sqrt{\widehat{h}_t}\widehat{e}_t \\ \widehat{h}_t &= 0.0158 + 0.5056 l_{\boldsymbol{\Gamma}_t} \end{aligned} \tag{7.5}$$

7.4 Comparison of Out-of-Sample Forecasts

In the final step of our analysis, we use the test set of the BRL/USD Foreign Exchange Rate data from November 2015 to December 2019 in order to compute out-of-sample forecasts based on the three final models described by equations (7.1), (7.3) and (7.5). Our goal is to compare the three different forecasts based on the MSPE (Mean Squared Prediction Error):

$$MSPE = (1/n)\sum_{t=1}^{n}[(x_t - \widehat{x}_t)^2] \tag{7.6}$$

Table 7.6: MSPE values for three models

Model	$MSPE$
(7.1) AR-ARCH	0.0157
(7.3) Iterative Estimation with one linear combination	0.0197
(7.5) Iterative Estimation with two linear combinations	0.0151

According to Table 7.6, it is clear that all the approaches provide similar MSPE values, whereas the smallest one was obtained by our iterative estimation method with matrices. This demonstrates how our proposed approach can build a competitive predictive model by using the fitted linear combinations as explanatory variables after all the nonparametric estimation steps.

In this application, we had a moderate sample size, which may not be enough for complex dependence structures, as demonstrated in chapter 6. Nevertheless, for the Brazilian Exchange rate data, we are still able to yield a good predictive power, demonstrating that our approach is still applicable for moderate sample sizes, and is expected to be even more accurate for larger sample sizes.

Chapter 8

Concluding Remarks

This thesis considers a univariate time series x_t where the conditional mean is assumed to be an unknown function of linear combinations of past observations and the conditional variance is assumed to be an unknown function of linear combinations of past squared residuals. Without assuming a time series model, we have developed an iterative, nonparametric estimation approach for dimension reduction in time series which estimates these linear combinations.

We have shown theoretically that the estimators are consistent as the sample size tends to infinity. Our theoretical results are further validated through extensive simulation studies and a data analysis. Additionally, we have proposed a data-driven approach similar to the SBC criterion to select the dimension of the parameter matrices in applications where it is not known. The simulation and the data analysis results indicate that our selection criteria provide a useful tool in real data applications, especially when both the parameter matrices of interest have only one column.

This research also investigates, through simulations, the accuracy of possible alternatives to the iterative estimation approach. For example, a sparse estimation approach is proposed in order to potentially produce more precise estimators when our parameter vectors contain a large proportion of zero elements. Considering the scope of our simulation scenarios, we observed that the sparse estimators do not outperform the estimators obtained by the iterative approach. Thus, our original estimator does not seem to be sensitive to a high degree of sparsity. Nonetheless, we still plan to investigate the accuracy of the sparse estimators by considering new data generating models and different ratios of training and test data. The proposed simultaneous estimation approach, however, yielded better results than the

iterative approach on many occasions, especially when estimating parameters associated with the conditional variance function. Hence, we will establish the theoretical convergence results for estimators obtained through the simultaneous estimation approach.

In our research as well as in similar research on dimension reduction in time series, the estimators do not have a closed form solution. Therefore, we have to rely on computational algorithms to minimize the objective functions. In addition, the parameter matrices have an orthonormality restriction which reduces the numerical search to a highly complex space. In order to overcome some computational challenges, we have developed a new reparametrization based on angles, where every element can freely vary inside $(-\frac{\pi}{2}, \frac{\pi}{2})$. Under this reparametrization framework, all of our estimates from simulation replicates were able to numerically converge, while ensuring that the original restriction is fully met and reducing the computational cost of the problem. In addition, this reparametrization approach allows us to randomly sample from the space of orthonormal matrices to find multiple appropriate initial points, and is general enough to be applicable for any numerical optimization problem with respect to a matrix of fixed dimension under a orthonormality restriction.

Lastly, we have applied our proposed iterative estimation approach to the task of forecasting the Brazilian Real/ U.S. Dollar Exchange Rate. This financial series is well known for its importance to the Brazilian economy. Its modeling is still an ongoing task, which involves both understanding how its expected value depends on past information and accounting for a possible heteroscedastic variance[6]. We have demonstrated that the estimated linear combination of past values can be used to generate competitive forecasts when compared to an AR-ARCH model. Here, our approach is shown to be useful even for moderate sample sizes.

This research thoroughly investigated the theoretical and application aspects of the iterative estimator for parameter matrices of both mean and variance functions. Due to computational time and cost, we did not perform simulations on larger sample sizes. Our future research will study the convergence behavior of our estimators for larger sample sizes. Other aspects that can be further investigated include the study of asymptotic properties of the sparse and simultaneous estimators and if the proposed dimension selection criterion can be improved.

[6]A common feature on financial assets series.

Bibliography

Agrawal, R., Faloutsos, C. and Swami, A. (1993), Efficient similarity search in sequence databases, *in* D. B. Lomet, ed., 'Foundations of Data Organization and Algorithms', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 69–84.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998), 'Robust and efficient estimation by minimising a density power divergence', *Biometrika* **85**(3), 549–559.
URL: *http://www.jstor.org/stable/2337385*

Becker, C. and Fried, R. (2003), Sliced inverse regression for high-dimensional time series, *in* M. Schwaiger and O. Opitz, eds, 'Exploratory Data Analysis in Empirical Research', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 3–11.

Bollerslev, T. (1986), 'Generalized autoregressive conditional heteroskedasticity', *Journal of Econometrics* **31**(3), 307 – 327.
URL: *http://www.sciencedirect.com/science/article/pii/0304407686900631*

Broman, K. W. and Speed, T. P. (2002), 'A model selection approach for the identification of quantitative trait loci in experimental crosses', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4), 641–656.
URL: *https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00354*

Chakrabarti, K., Keogh, E., Mehrotra, S. and Pazzani, M. (2002), 'Locally adaptive dimensionality reduction for indexing large time series databases', *ACM Transactions on Database Systems (TODS)* .
URL: *https://www.microsoft.com/en-us/research/publication/locally-adaptive-dimensionality-reduction-for-indexing-large-time-series-databases/*

Chan, K.-P. and Fu, A. W.-C. (1999), Efficient time series matching by wavelets, *in* 'Proceedings 15th International Conference on Data Engineering (Cat. No.99CB36337)', pp. 126–133.

Cook, R. and Li, B. (2002), 'Dimension reduction for conditional mean in regression', *Ann. Statist.* **30**(2), 455–474.
URL: *https://doi.org/10.1214/aos/1021379861*

Engle, R. F. (1982), 'Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation', *Econometrica* **50**(4), 987–1007.
URL: *http://www.jstor.org/stable/1912773*

Fan, J. and Yao, Q. (2003), *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer.

Hall, P. and Yao, Q. (2005), 'Approximating conditional distribution functions using dimension reduction', *Ann. Statist.* **33**(3), 1404–1421.
URL: *https://doi.org/10.1214/009053604000001282*

Hannan, E. J. and Quinn, B. G. (1979), 'The determination of the order of an autoregression', *Journal of the Royal Statistical Society. Series B (Methodological)* **41**(2), 190–195.
URL: *http://www.jstor.org/stable/2985032*

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

Hong, S. Y. and Linton, O. (2020), 'Nonparametric estimation of infinite order regression and its application to the risk-return tradeoff.', *Journal of Econometrics* .

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

Iaci, R. and Sriram, T. (2013), 'Robust multivariate association and dimension reduction using density divergences', *Journal of Multivariate Analysis* **117**, 281 – 295.
URL: *http://www.sciencedirect.com/science/article/pii/S0047259X13000316*

IPEA, IPEADATA database (2020), `http://www.ipeadata.gov.br`. Accessed: 2020-05-28.

Keogh, E., Chakrabarti, K., Pazzani, M. and Mehrotra, S. (2002), 'Dimensionality reduction for fast similarity search in large time series databases', *Knowledge and Information Systems* **3**.

Lee, C. E. and Shao, X. (2018), 'Martingale difference divergence matrix and its application to dimension reduction for stationary multivariate time series', *Journal of the American Statistical Association* **113**(521), 216–229.
URL: *https://doi.org/10.1080/01621459.2016.1240083*

Li, K.-C. and Shedden, K. (2002), 'Identification of shared components in large ensembles of time series using dimension reduction', *Journal of the American Statistical Association* **97**(459), 759–765.
URL: *https://doi.org/10.1198/016214502388618573*

Lin, J., Keogh, E., Lonardi, S. and Chiu, B. (2003), A symbolic representation of time series, with implications for streaming algorithms, *in* 'Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, DMKD '03', pp. 2–11.

Luo, W., Li, B. and Yin, X. (2014), 'On efficient dimension reduction with respect to a statistical functional of interest.', *The annals of statistics* **42**(1), 384–412.

Mack, Y. and Müller, H.-G. (1989), 'Derivative estimation in nonparametric regression with random predictor variable', *Sankhyā: The Indian Journal of Statistics, Series A* pp. 59–72.

Nadaraya, E. (1964), 'On estimating regression', *Theory of Probability and its Applications.* **9**, 141–142.

Ng, S. and Perron, P. (2005), 'A note on the selection of time series models', *Oxford Bulletin of Economics and Statistics* **67**(1), 115–134.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0084.2005.00113.x*

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

Park, J.-H. (2011), 'Dimension reduction transfer function model', *Journal of Statistical Computation and Simulation* **81**, 2131–2140.

Park, J.-H. and Samadi, S. Y. (2014), 'Heteroscedastic modelling via the autoregressive conditional variance subspace', *Canadian Journal of Statistics* **42**(3), 423–435.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/cjs.11222*

Park, J.-H. and Sriram, T. N. (2017), 'Robust estimation of conditional variance of time series using density power divergences', *Journal of Forecasting* **36**(6), 703–717.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2465*

Park, J.-H., Sriram, T. N. and Yin, X. (2009), 'Central mean subspace in time series', *Journal of Computational and Graphical Statistics* **18**(3), 717–730.
URL: *https://doi.org/10.1198/jcgs.2009.08076*

Park, J.-H., Sriram, T. N. and Yin, X. (2010), 'Dimension reduction in time series', *Statistica Sinica* **20**(2), 747–770.
URL: *http://www.jstor.org/stable/24309020*

Park, J. and Samadi, S. (2019), 'Dimension reduction for the conditional mean and variance functions in time series', *Scandinavian Journal of Statistics* .

Schwarz, G. (1978), 'Estimating the dimension of a model', *Ann. Statist.* **6**(2), 461–464.
URL: *https://doi.org/10.1214/aos/1176344136*

Scott, D. W. (1992), *Multivariate densisty 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', *The annals of statistics* pp. 1040–1053.

Tiao, G. C. and Tsay, R. S. (1994), 'Some advances in non-linear and adaptive modelling in time-series', *Journal of Forecasting* **13**(2), 109–131.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/for.3980130206*

Tjøstheim, D. (1994), 'Non-linear time series: A selective review', *Scandinavian Journal of Statistics* **21**(2), 97–130.
URL: *http://www.jstor.org/stable/4616304*

Watson, G. (1964), 'Smooth regression analysis', *Sankhyā: The Indian Journal of Statistics, Series A* **26**(4), 359–372.

Wu, D., Singh, A., Agrawal, D., El Abbadi, A. and Smith, T. R. (1996), Efficient retrieval for browsing large image databases, *in* 'Proceedings of the Fifth International Conference on Information and Knowledge Management', CIKM '96, Association for Computing Machinery, New York, NY, USA, p. 11–18.
URL: *https://doi.org/10.1145/238355.238365*

Xia, Y. and Li, W. K. (1999), 'On the estimation and testing of functional-coefficient linear models', *Statistica Sinica* **9**(3), 735–757.
URL: *http://www.jstor.org/stable/24306613*

Xia, Y., Tong, H. and Li, W. K. (1999), 'On extended partially linear single-index models', *Biometrika* **86**(4), 831–842.
URL: *http://www.jstor.org/stable/2673588*

Xia, Y., Tong, H., Li, W. K. and Zhu, L.-X. (2002), 'An adaptive estimation of dimension reduction space', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64**(3), 363–410.
URL: *http://www.jstor.org/stable/3088779*

Ye, Z. and Weiss, R. E. (2003), 'Using the bootstrap to select one of a new class of dimension reduction methods', *Journal of the American Statistical Association* **98**(464), 968–979.
URL: *https://doi.org/10.1198/016214503000000927*

Zhu, L., Miao, B. and Peng, H. (2006), 'On sliced inverse regression with high-dimensional covariates', *Journal of the American Statistical Association* **101**(474), 630–643.
URL: *https://doi.org/10.1198/016214505000001285*

Appendix

In Appendix A, we give a detailed proof of the three theorems stated in Section 4.3.2. In Appendix B, we give 13 figures that are referred in various sections of the dissertation.

A: Proof of the Three Theorems in Section 4.3.2

A.1: Proof of Theorem 1

Let $\delta_n = \left(\frac{\ln(n-p)}{n-p}\right)^{2/5}$ and η_n be a sequence that 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 is suffice 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

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

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

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

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 (4.14).

According to Lemma 4.3.2 and the optimally of $\boldsymbol{\Phi}_0$, we have

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

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

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

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

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

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

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

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

With Lemma 4.3.1 and assumption (A4), we can derive

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

where C_2 is a large enough positive constant.

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

$$
I_2 \le C_2(n-p)\big(\delta_n^2 + \|\boldsymbol{r} - \boldsymbol{\Phi}_0\|^2\big)^{1/2}(1 + o_p(1)). \qquad \text{(A.5)}
$$

Similarly, we can upper bound I_3 by

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

where C_3 is a large enough positive constant and the last inequality follows Lemma 4.3.1 and (A.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. $\square$

A.2: Proof of Theorem 2

Let $\delta_n = \left(\frac{\ln(n-p)}{n-p}\right)^{2/5}$, $\delta_n' = \left(\frac{\ln(n-p-q)}{n-p-q}\right)^{2/5}$ and η_n be a sequence that 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 is suffice to show that

$$\inf_{\|\boldsymbol{s}-\boldsymbol{\Gamma}\|=\eta_n\delta_n} \sum_{t=p+q+1}^{n} \left(\widehat{\varepsilon}_t^2 - \widehat{g}_n(\boldsymbol{s}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right)^2 - \sum_{t=p+q+1}^{n} \left(\widehat{\varepsilon}_t^2 - \widehat{g}_n(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right)^2 > 0$$

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

Similar to the proof of (A.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

$$J_2 \leq \left\{ \sum_{t=p+q+1}^{n} \left[\widehat{\varepsilon}_t^2 - \varepsilon_t^2\right]^2 \sum_{t=p+q+1}^{n} [\widehat{g}_n(\boldsymbol{s}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\big]^2 \right\}^{1/2},$$

$$J_3 \leq \left\{ \sum_{t=p+q+1}^{n} \left[\varepsilon_t^2 - g(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]^2 \sum_{t=p+q+1}^{n} [\widehat{g}_n(\boldsymbol{s}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\big]^2 \right\}^{1/2},$$

and $$J_4 \leq \left\{ \sum_{t=p+q+1}^{n} \left[g(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\right]^2 \sum_{t=p+q+1}^{n} [\widehat{g}_n(\boldsymbol{s}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\big]^2 \right\}^{1/2}.$$

Similar to the derivation of (A.4), Lemma 4.3.2 and assumption (A4) yields

$$\begin{aligned}
&[\widehat{g}_n(\boldsymbol{s}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - \widehat{g}_n(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\big]^2 \\
&\leq 4\Big\{ \widehat{g}_n(\boldsymbol{s}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - g(\boldsymbol{s}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\big]^2 + \big[\widehat{g}_n(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - g(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\big]^2 \\
&\quad + \big[g(\boldsymbol{s}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2) - g(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)\big]^2 \Big\} \\
&\leq C_5(\delta_n'^2 + \|\boldsymbol{s} - \boldsymbol{\Gamma}\|^2), \qquad \text{(A.7)}
\end{aligned}$$

where C_5 is a large enough positive constant.

Follow Lemma 4.3.1 and Theorem 1, we can show

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

Notice the fact that $E(g(\boldsymbol{\Gamma}^\intercal \widehat{\boldsymbol{\varepsilon}}_{t-1}^2)) = \varepsilon_t^2$, strong law of large numbers and Lemma 4.3.1 suggest

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

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

The results in (A.7) – (A.10) implies

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

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. □

A.1: Proof of Theorem 3

Let $\delta_n = \left(\frac{\ln(n-p)}{n-p}\right)^{2/5}$, $\delta_n' = \left(\frac{\ln(n-p-q)}{n-p-q}\right)^{2/5}$ and η_n be a sequence that 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 is suffice to show that

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

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

To keep the presentation neat, we introduce the following notations

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

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

With some calculations, one can show

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

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

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

Follow assumption (A1), Lemma 4.3.1, Theorem 1 and Theorem 2, K_{11} and K_{12} can be upper bounded by

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

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

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

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

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

By plugging (A.14), (A.15) and (A.16) back to (A.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. □

B: Simulation Experiment Figures

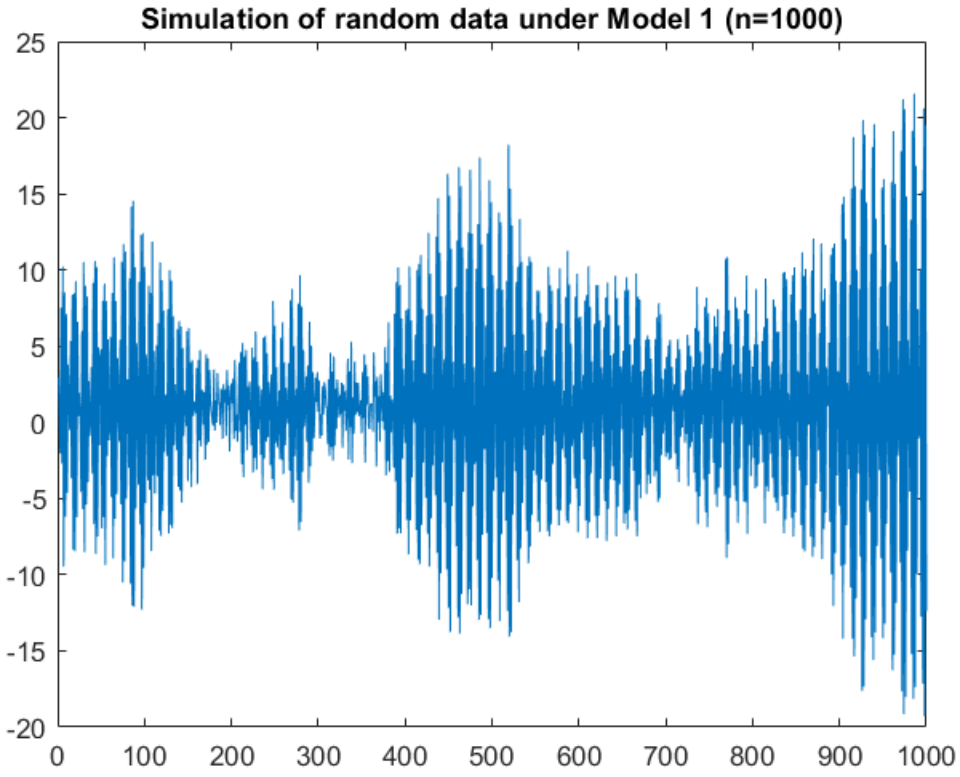

Figure 8.1: Example of simulated data from Model 1

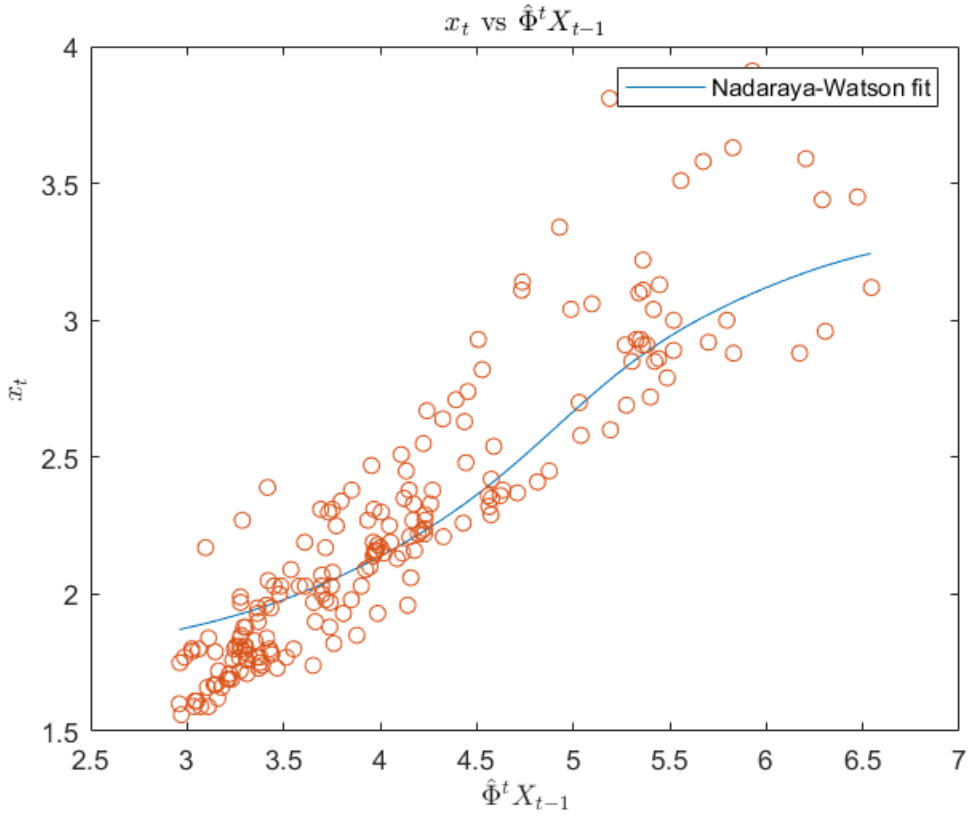

Figure 8.2: Example of fitted curve using simulated data from Model 1

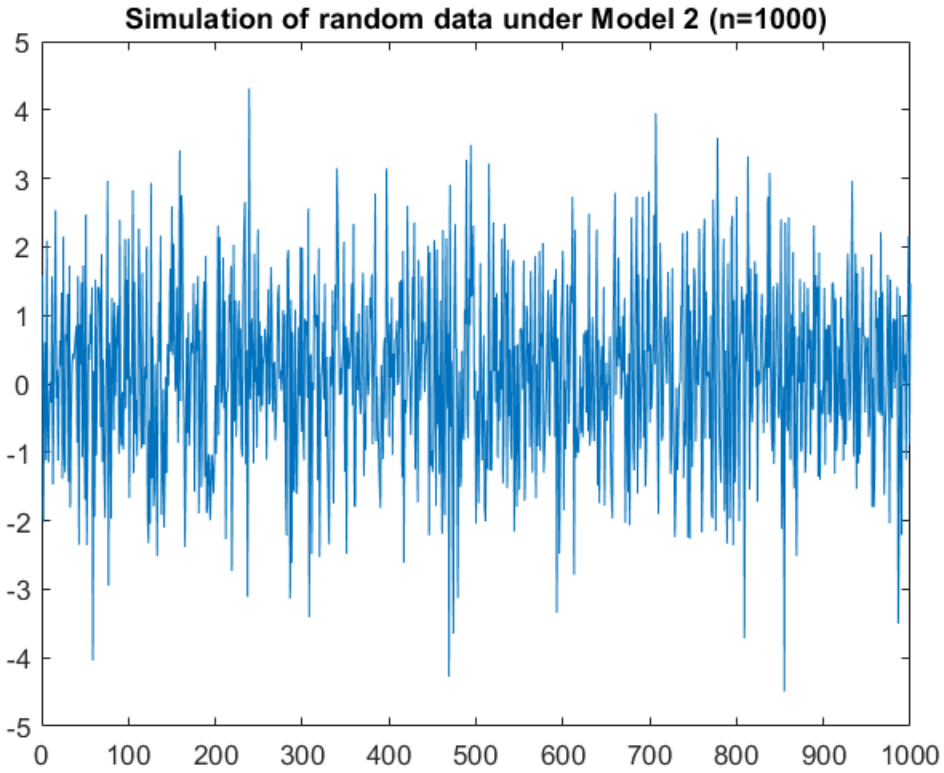

Figure 8.3: Example of simulated data from Model 2

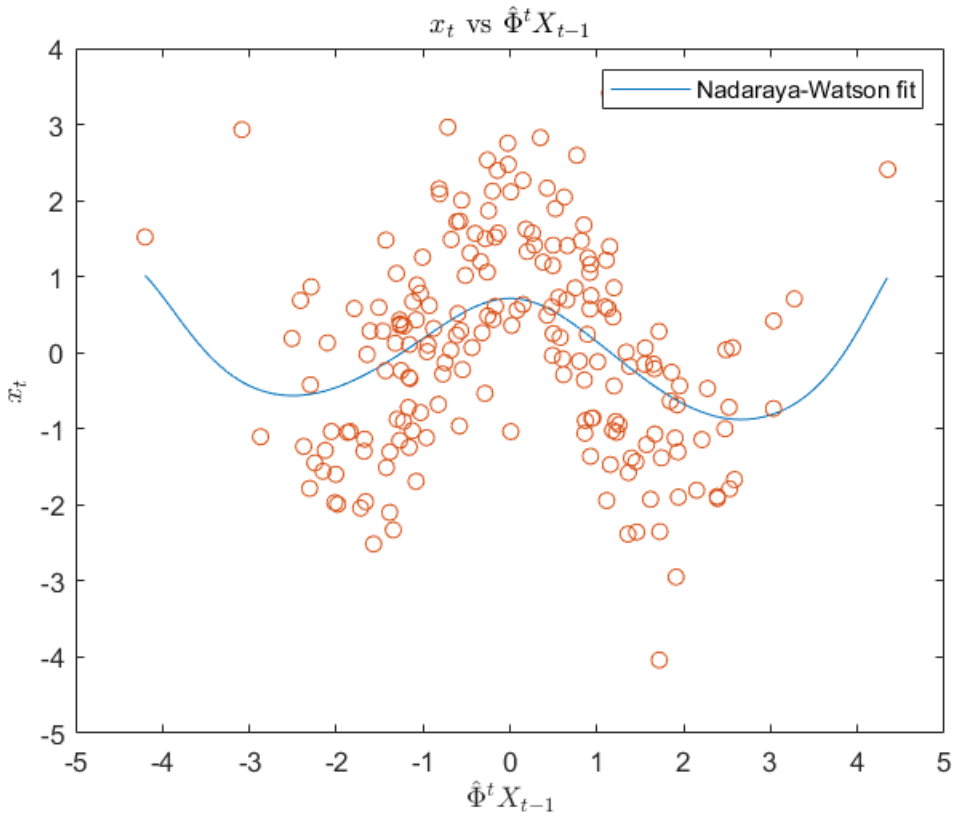

Figure 8.4: Example of fitted curve using simulated data from Model 2

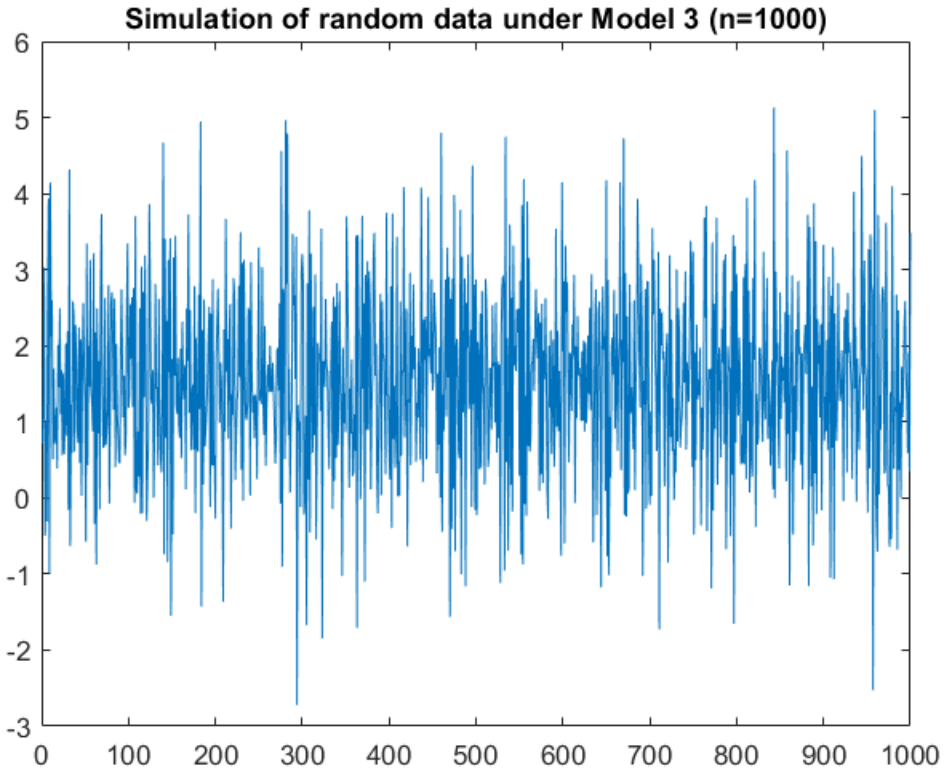

Figure 8.5: Example of simulated data from Model 3

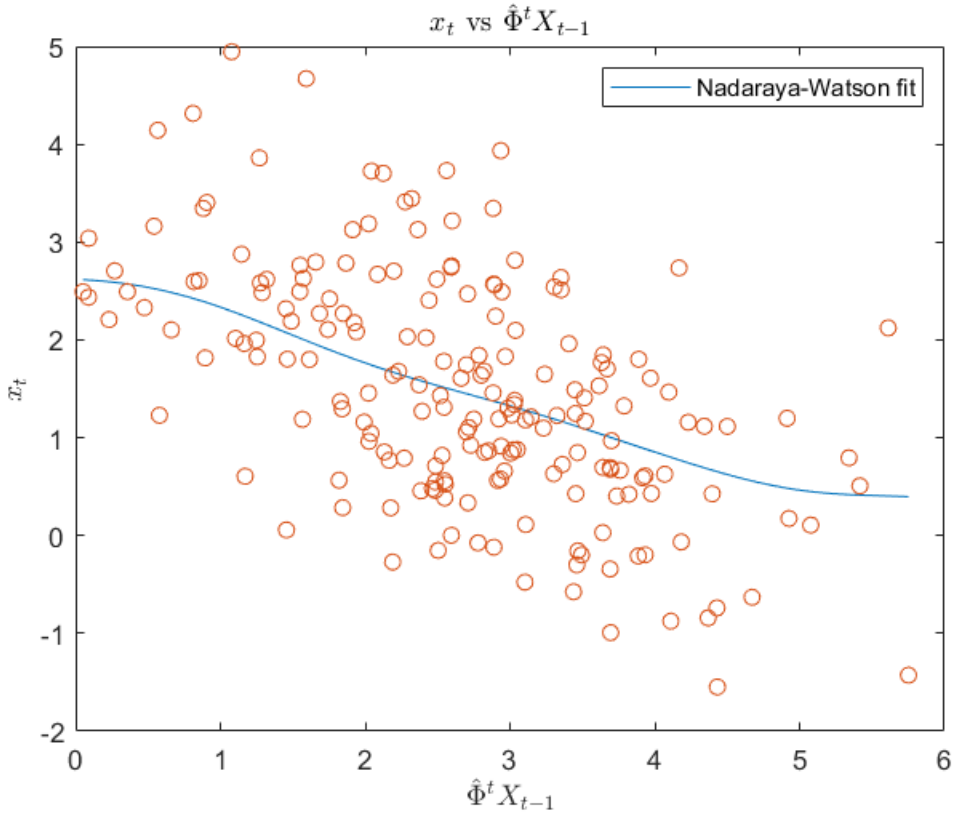

Figure 8.6: Example of fitted curve using simulated data from Model 3

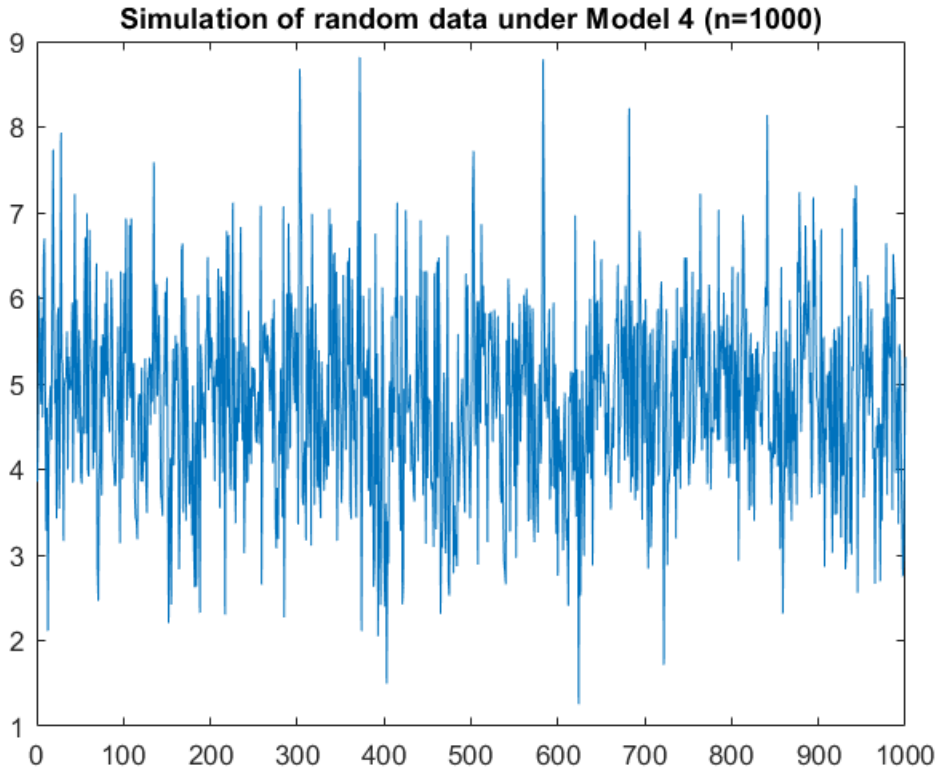

Figure 8.7: Example of simulated data from Model 4

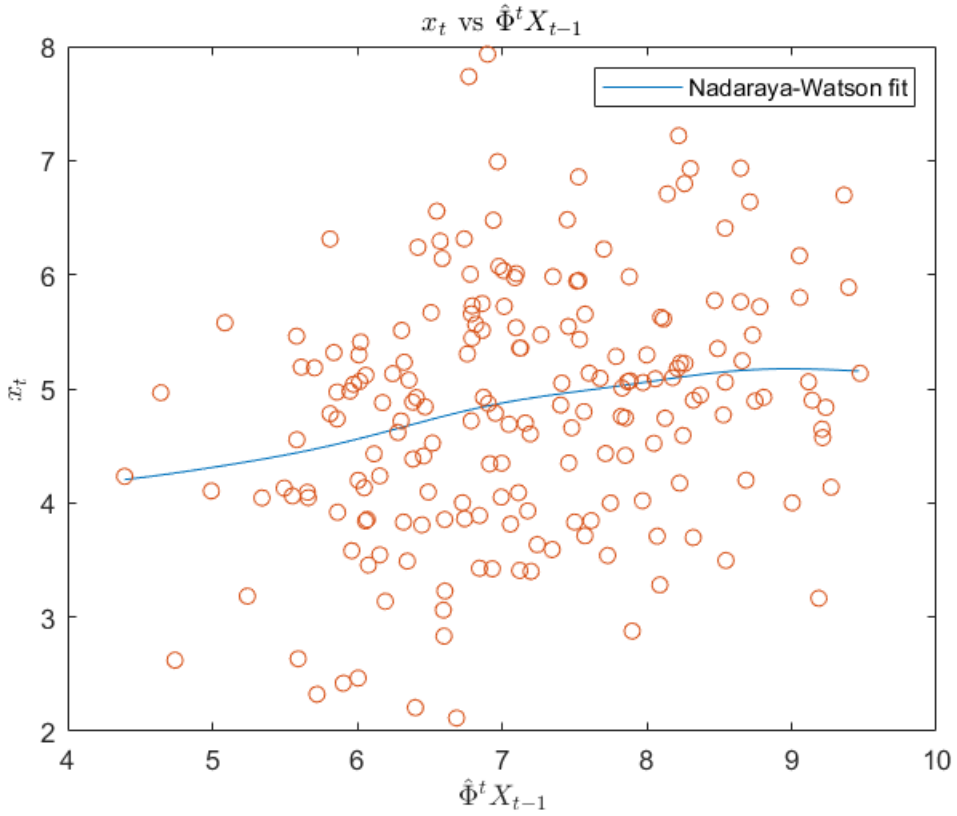

Figure 8.8: Example of fitted curve using simulated data from Model 4

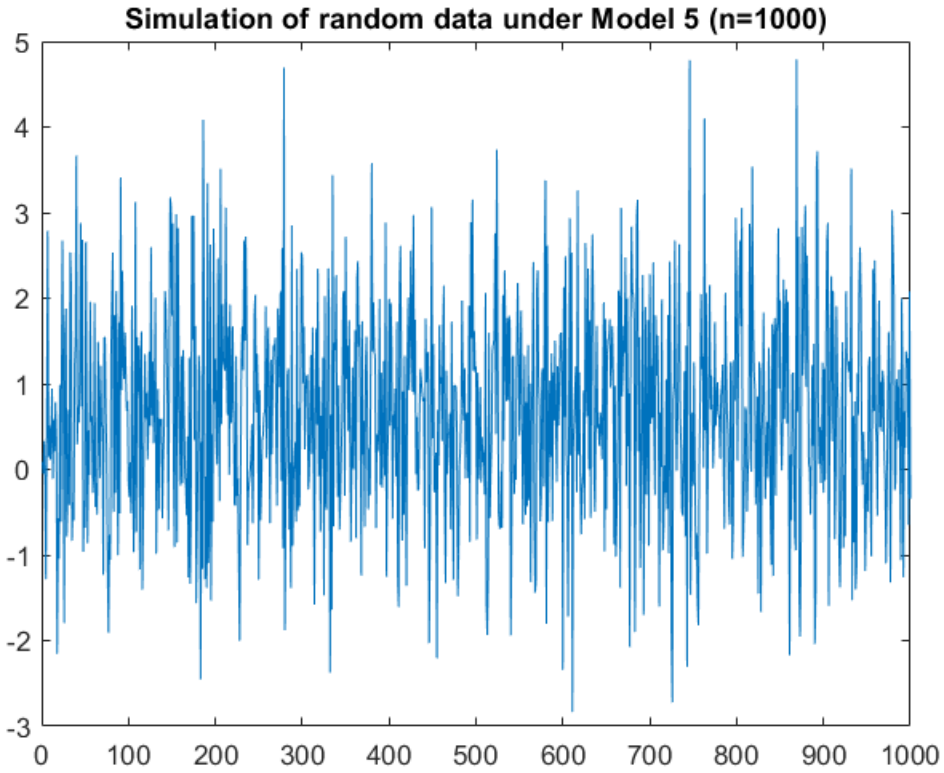

Figure 8.9: Example of simulated data from Model 5

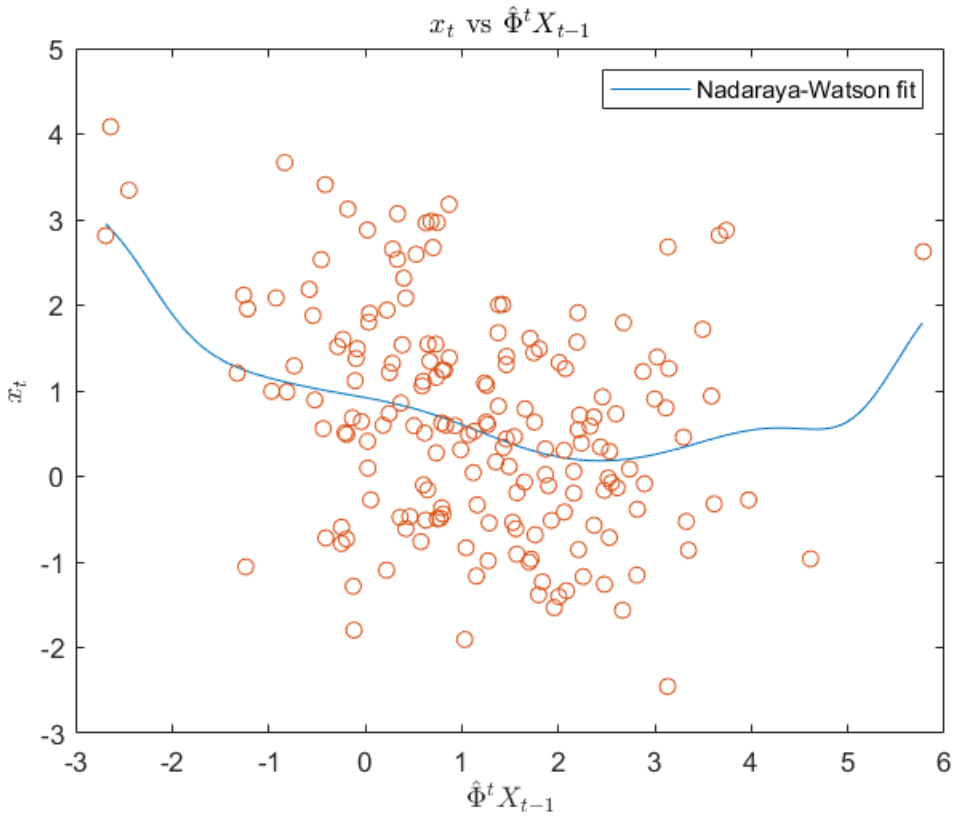

Figure 8.10: Example of fitted curve using simulated data from Model 5

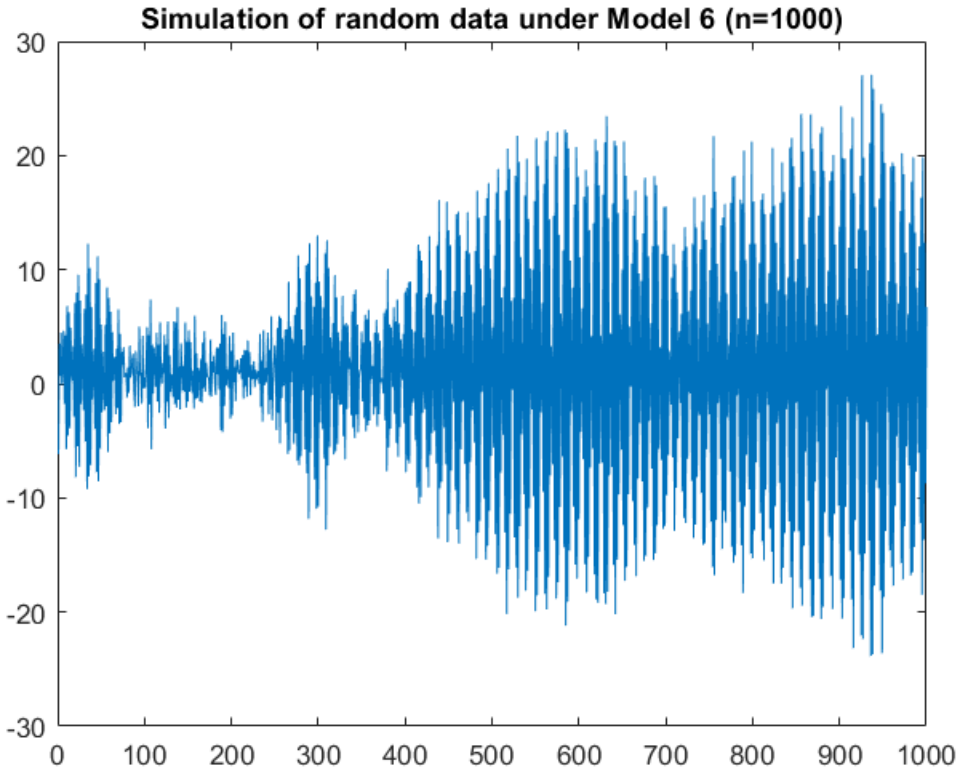

Figure 8.11: Example of simulated data from Model 6

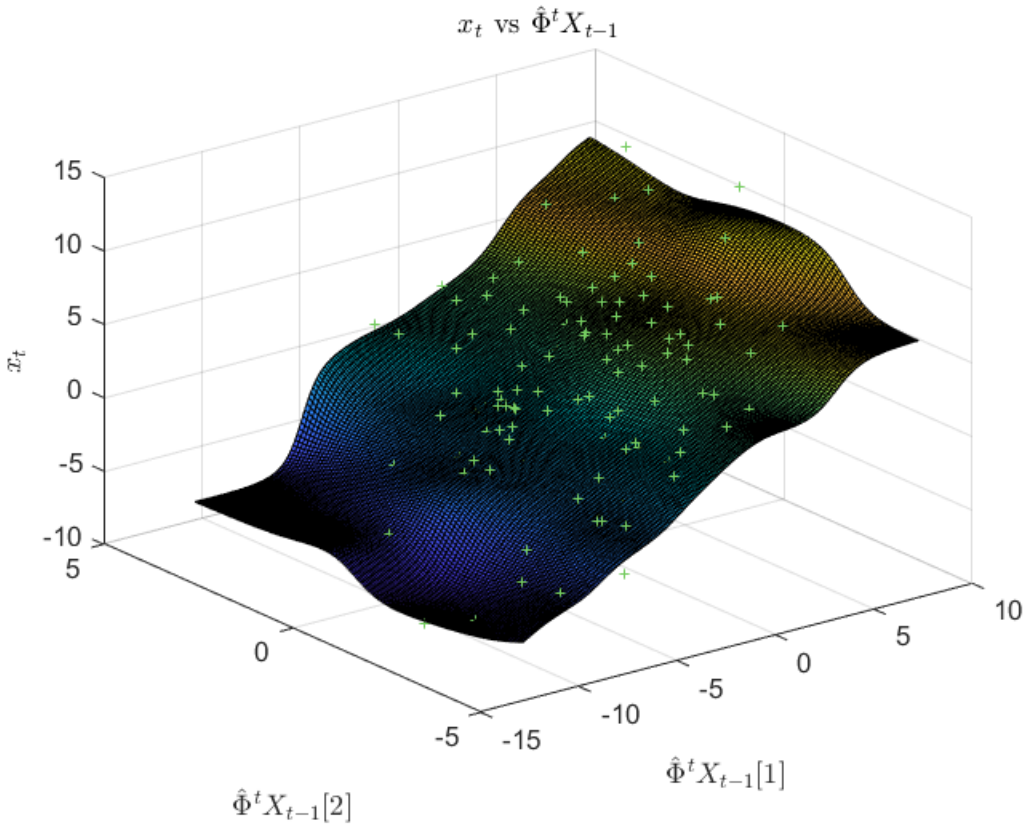

Figure 8.12: Example of fitted curve using simulated data from Model 6

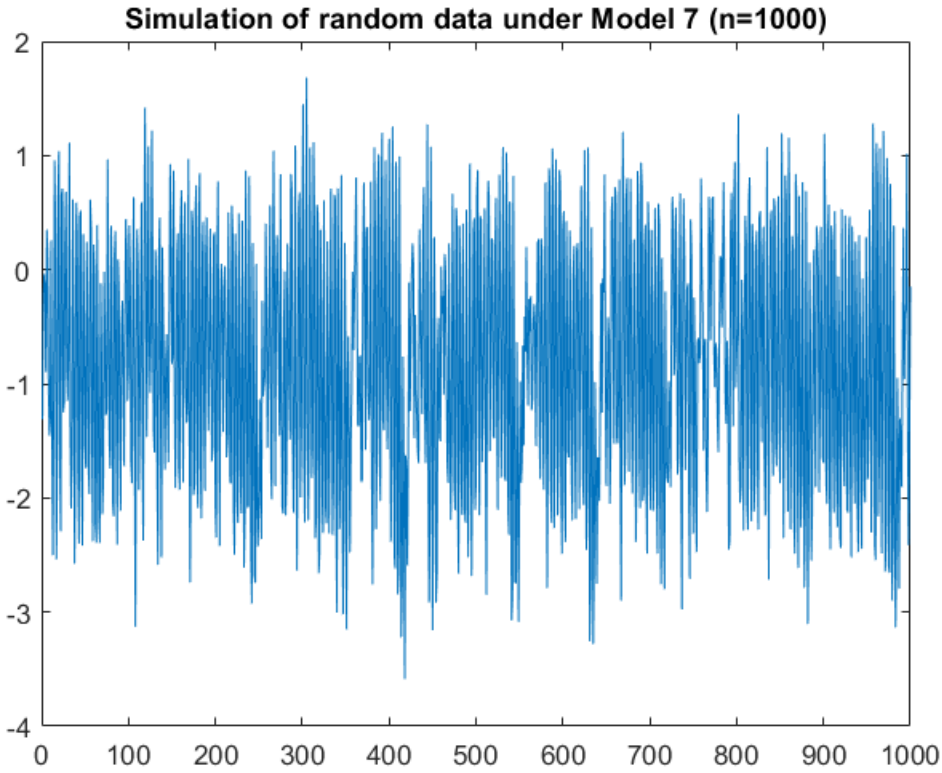

Figure 8.13: Example of simulated data from Model 7