SOME APPROACHES TO NONLINEAR MODELING AND PREDICTION

WANG TIANHAO

NATIONAL UNIVERSITY OF SINGAPORE

2013

SOME APPROACHES TO NONLINEAR MODELING AND PREDICTION

WANG TIANHAO

(B.Sc. East China Normal University)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2013

ACKNOWLEDGEMENTS

I would like to give my sincere thanks to my PhD supervisor, Professor Xia Yingcun. It has been an honor to be one of his students. He has taught me, both consciously and unconsciously, how a useful statistical model could be built and applied to the real world. I appreciate all his contributions of time, ideas, and funding to make my PhD experience productive and stimulating. This thesis would not have been possible without his active support and valuable comments.

I would also like to gratefully thank other faculty members and support staffs of the Department of Statistics and Applied Probability for teaching me and helping me in various ways throughout my PhD candidacy.

Last but not the least, I would like to thank my family for all their love and encouragement. For my parents who raised me with a love of science and supported

me in all my pursuits. And most of all for my loving, supportive, encouraging, and patient wife, Chen Jie, whose faithful support during the final stages of this PhD is so appreciated. Thank you.

MANUSCRIPTS

Wang, T. and Xia, Y. (2013) A piecewise single-index model for dimension reduction. *To appear in Technometrics*.

Wang, T. and Xia, Y. (2013) Whittle likelihood estimation of nonlinear autoregressive models with moving average errors. *Submitted to Biometrika*.

CONTENTS

Acknowledgements **iii**

Manuscripts **v**

Summary **xi**

List of Tables **xiii**

List of Figures **xv**

Chapter 1 A Piecewise SIM for Dimension Reduction **1**

1.1 Introduction 1

1.1.1 Effective Dimension Reduction (EDR) Space 2

1.1.2 Single-Index Model (SIM) 5

1.1.3 Piecewise Regression Models 6
1.1.4 Piecewise Single-Index Model (pSIM) 8
1.2 Estimation of pSIM 11
1.2.1 Model Estimation 12
1.2.2 Selection Of Tuning Parameters 16
1.3 Simulations 18
1.4 Real Data Analysis 28
1.5 Asymptotic Analysis 43
1.6 Proofs 48

Chapter 2 WLE of Nonlinear AR Models with MA Errors 71

2.1 Time Series Analysis: A Literature Review 71
2.1.1 Stationarity of Time Series 72
2.1.2 Linear Time Series Models 73
2.1.3 Nonlinear Time Series Models 75
2.1.4 Spectral Analysis and Periodogram 77
2.1.5 Whittle Likelihood Estimation (WLE) 79
2.2 Introduction of the Extended WLE (XWLE) 81
2.3 Estimating Nonlinear Models with XWLE 84
2.4 Model Diagnosis Based on XWLE 87
2.5 Numerical Studies 90
2.6 Asymptotics of XWLE 113

Chapter 3 Conclusion and Future Works 133

Bibliography **137**

SUMMARY

Our work in this thesis consists of two parts. The first part (Chapter 1) deals with dimension reduction in nonparametric regressions. In this Chapter we propose to use different single-index models for observations in different regions of the sample space. This approach inherits the estimation efficiency of the single-index model in each region, and at the same time allows the global model to have multi-dimensionality in the sense of conventional dimension reduction (Li, 1991). On the other hand, the model can be seen as an extension of CART (Breiman et al, 1984) and a piecewise linear model proposed by Li et al (2000). Modeling procedures, including identifying the region for every single-index model and estimation of the single-index models, are developed. Simulation studies and real data analysis are employed to demonstrate the usefulness of the approach.

The second part (Chapter 2) deals with nonlinear time series analysis. In this Chapter, we modify the Whittle likelihood estimation (WLE; Whittle, 1953) such that it is applicable to models in which the theoretical spectral density functions of the models are only partially available. In particular, our modified WLE can be applied to most nonlinear regressive or autoregressive models with residuals following a moving average process. Asymptotic properties of the estimators are established. Its performance is checked by simulated examples and real data examples, and is compared with some existing methods.

List of Tables

Table 1.1 Simulation results of Example 1.3.1: mean of in-sample (IS) and out-of-sample (OS) prediction errors (ASE) from the 100 replications. The percentage numbers in the parenthesis are the proportion of times that the number of regions (m) of the model is identified as three by the proposed BIC method. 23

Table 1.2 Simulation results of Example 1.3.2: mean of in-sample (IS) and out-of-sample (OS) prediction errors (ASE) ($\times 10^{-3}$) from the 100 replications. 25

Table 1.3 Simulation results of Example 1.3.2 (continued): mean of in-sample (IS) and out-of-sample (OS) prediction errors (ASE) ($\times 10^{-3}$) from the 100 replications. 26

Table 1.4 BIC scores for the hitters' salary data (with the outliers removed) 30

Table 1.5 Simulation results of the hitters' salary data: mean of in-sample (IS) and out-of-sample (OS) prediction errors (ASE) from the 100 replications. 33

Table 1.6 BIC scores for the LA Ozone data 35

Table 1.7 Simulation results of the LA ozone data: mean of in-sample (IS) and out-of-sample (OS) prediction errors (ASE) from the 100 replications. 38

Table 1.8 BIC scores for the cars data 39

Table 1.9 Simulation results of the cars data: mean of in-sample (IS) and out-of-sample (OS) prediction errors (ASE) from the 100 replications. 43

Table 2.1 Simulation results for Example 2.5.2. 103

Table 2.2 BIC_W scores for the Niño 3.4 SST anomaly data 111

List of Figures

Figure 1.1 A typical estimation result of Example 1.3.1 with sample size $n = 400$. 21

Figure 1.2 The estimation errors of the three piecewise single-index $D^2(\hat{\boldsymbol{\beta}}_i, \boldsymbol{\beta}_i)$, $i = 1, 2, 3$ in Example 1.3.1. 22

Figure 1.3 Four typical estimation results of Example 1.3.2. 27

Figure 1.4 y plotted against $\boldsymbol{\beta}_0^\top \boldsymbol{x}$ for the hitters' salary data. 29

Figure 1.5 Fitting results for the hitter's salary data. 31

Figure 1.6 The maximum a posteriori (MAP) tree at height 3 estimated by TGP-SIM for the hitters' salary data. 34

Figure 1.7 Fitting results for the LA ozone data. 36

Figure 1.8 The maximum a posteriori (MAP) tree at height 2 estimated by TGP-SIM for the LA ozone data. 37

Figure 1.9 Fitting results for the cars data. 41

Figure 1.10 The tree structures estimated by the TGP-SIM model for the cars data. 42

Figure 2.1 Simulation results for ARMA$(1,1)$ models with $\varepsilon_t \sim N(0,1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '$\square$': MLE, red '$*$': XWLE. 93

Figure 2.2 Simulation results for ARMA$(2,1)$ models with $\varepsilon_t \sim N(0,1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '$\square$': MLE, red '$*$': XWLE. 94

Figure 2.3 Simulation results for ARMA$(5,1)$ models with $\varepsilon_t \sim N(0,1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '$\square$': MLE, red '$*$': XWLE. 95

Figure 2.4 Simulation results for ARMA$(1,1)$ models with $\varepsilon_t \sim t(1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '□': MLE, red '$*$': XWLE. 96

Figure 2.5 Simulation results for ARMA$(2,1)$ models with $\varepsilon_t \sim t(1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '□': MLE, red '$*$': XWLE. 97

Figure 2.6 Simulation results for ARMA$(5,1)$ models with $\varepsilon_t \sim t(1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '□': MLE, red '$*$': XWLE. 98

Figure 2.7 Simulation results for ARMA$(1,1)$ models with $\varepsilon_t \sim U(-1,1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '□': MLE, red '$*$': XWLE. 99

Figure 2.8 Simulation results for ARMA$(2,1)$ models with $\varepsilon_t \sim U(-1,1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '□': MLE, red '$*$': XWLE. 100

Figure 2.9 Simulation results for ARMA$(5,1)$ models with $\varepsilon_t \sim U(-1,1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '□': MLE, red '$*$': XWLE. 101

Figure 2.10 Rate of rejections for the LB(20)-tests and AN(20)-tests in Example 2.5.2. 104

Figure 2.11 Time plots for the transformed sunspot number. 106

Figure 2.12 Root mean squred prediction errors of out-of-sample multi-step forecasts for the original numbers of the sunspots. 109

Figure 2.13 Time plots for the Niño 3.4 anomaly. 110

Figure 2.14 Root mean squred prediction errors of out-of-sample multi-step forecasts for Niño 3.4 SST anomaly data. 113

CHAPTER 1

A Piecewise SIM for Dimension Reduction

1.1 Introduction

Exploring multivariate data under a nonparametric setting is an important and challenging topic in many disciplines of research. Specifically, suppose y is the response variable of interest and $\boldsymbol{x} = (x_1, ..., x_p)^\top$ is the $p-$dimensional covariate. For a nonparametric regression model

$$y = \psi(x_1, ..., x_p) + \varepsilon, \tag{1.1}$$

where ε is the error term with mean 0, the estimation of unknown multivariate function $\psi(x_1, ..., x_p)$ is difficult. There are several different ways to do the nonparametric regression. The two most popular techniques are local polynomial kernel smoothing and spline smoothing. But no matter which technique we use to do the nonparametric regression, as the dimension increases, the estimation efficiency drops dramatically, which is the so-called curse of dimensionality.

1.1.1 Effective Dimension Reduction (EDR) Space

Numerous approaches have been developed to tackle the problem of high dimensionality. One of the most popular approaches is searching for an effective dimension reduction (EDR) space; see for example Li (1991) and Xia, Tong, Li and Zhu (2002). The EDR space was first introduced by Li (1991) who proposed the model

$$y = \tilde{f}(\beta_1^\top \boldsymbol{x}, \cdots, \beta_q^\top \boldsymbol{x}, \varepsilon), \tag{1.2}$$

where $\tilde{f}$ is a real function on $\mathbb{R}^{q+1}$ and ε is the random error independent of $\boldsymbol{x}$. Our primary interest is on the q p-dimensional column vectors β_1, ..., β_q. Of special interest is the additive noise model

$$y = f(\beta_1^\top \boldsymbol{x}, \cdots, \beta_q^\top \boldsymbol{x}) + \varepsilon. \tag{1.3}$$

where f is a real function on $\mathbb{R}^q$. Denote by $B = (\beta_1, \cdots, \beta_q)$ the $p \times q$ matrix pooling all the vectors together. For identification concern, it is usually assumed that $B^\top B = \boldsymbol{I}_q$, where $\boldsymbol{I}_q$ denotes the q by q identity matrix. The space spanned by $B^\top \boldsymbol{x}$ is called the EDR space, and the vectors β_1, ..., β_q are called the EDR directions.

If we know the exact form of $f(\cdot)$, then (1.3) is not much different from a simple neural network model, or a nonlinear regression model. However, (1.3) is special in that $f(\cdot)$ is generally assumed to be unknown and we need to estimate both B and $f(\cdot)$.

There are essentially two approaches to do the estimations. The first is the inverse regression approach first proposed by Li (1991). In his sliced inverse regression (SIR) algorithm, instead of regressing y on $\boldsymbol{x}$, Li (1991) proposed to regress each predictor in $\boldsymbol{x}$ against y. In this way, the original p-dimensional regression problem is reduced to be multiple one-dimensional problems. The SIR method has been proven to be powerful in searching for EDR directions and dimension reduction. However, the SIR method imposes some strong probabilistic structure on $\boldsymbol{x}$. Specifically, this method requires that, for any $\beta \in \mathbb{R}^p$, the conditional expectation

$$E(\beta^\top \boldsymbol{x} | \beta_1^\top \boldsymbol{x}, \cdots, \beta_q^\top \boldsymbol{x})$$

is linear in $\beta_1^\top \boldsymbol{x}, \cdots, \beta_q^\top \boldsymbol{x}$; i.e., there are constants $c_0, \cdots, c_q$ depending on β such

that

$$E(\beta^\top \boldsymbol{x}|\beta_1^\top \boldsymbol{x}, \cdots, \beta_q^\top \boldsymbol{x}) = c_0 + c_1\beta_1^\top \boldsymbol{x} + \cdots + c_q\beta_q^\top \boldsymbol{x}.$$

An important class of random variables that do not satisfy this assumption is the lagged time series variable $\boldsymbol{x} := (y_{t-1}, ..., y_{t-p})$ where $\{y_t\}$ is a time series.

The second approach of searching for the EDR directions is through direct regression of y on $\boldsymbol{x}$. One of the most popular methods in this category is the minimum average variance estimation (MAVE) method introduced by Xia et al (2002). In this method, the EDR directions are found by solving the optimization problem

$$\min_B \{E[y - E(y|B^T\boldsymbol{x})]\},$$

subject to $B^\top B = \boldsymbol{I}_q$, where $E(y|B^T\boldsymbol{x})$ is approximated by a local linear expansion. Through direct regression, the condition on the probability structure of $\boldsymbol{x}$ can be significantly relaxed. So as compared to the inverse-regression based approaches, MAVE method is applicable to a much broadened scope of possible distributions of $\boldsymbol{x}$, including the nonlinear autoregressive modeling aforementioned which violates the basic assumption of the inverse-regression based approaches.

1.1.2 Single-Index Model (SIM)

The single-index model (SIM) is actually a special case of model (1.3) which only has one EDR direction. Specifically, a typical SIM can be written as

$$y = f(\beta_1^\top \boldsymbol{x}) + \varepsilon, \tag{1.4}$$

where ε is independent of $\boldsymbol{x}$. The SIM is singled out here mainly for its popularity in many scientific fields including biostatistics, medicine, economics and financial econometrics. It is in the intersection of both the EDR approaches introduced above and the projection pursuit regression (PPR) approach proposed by Friedman and Stuetzle (1981) which is another popular method in dimension reduction. It is also the non-parametric counterpart of the generalized linear model (GLM) which is one of the prevailing regression models in practice.

In the last two decades a series of papers (Powell, Stock, and Stoker, 1989; Hädle and Stoker, 1989; Ichimura, 1993; Klein and Spady, 1993; Härdle, Hall, and Ichimura, 1993; Sherman, 1994; Horowitz and Härdle, 1996; Hristache, Juditski, and Spokoiny, 2001; Xia et al, 2002; Yu and Ruppert, 2002; Yin and Cook, 2005; Xia, 2006; Cui, Härdle and Zhu, 2011) have investigated the estimation of the parametric index β_1 with focus on root-n estimability and efficiency issues. Among these methods, the most popular ones up to now are the average derivative estimation (ADE) method proposed by Powell, Stock and Stocker (1989) and

Härdle and Stoker (1989), the simultaneous minimization method of Härdle et al (1993) and the MAVE of Xia et al (2002).

As the single-index $\beta_1^\top \boldsymbol{x}$ can be estimated with root-n consistency, the nonparametric estimation of the link function $f(\cdot)$ is able to achieve the best nonparametric efficiency with properly chosen smoothing techniques. However, the flexibility of the SIM in modeling is more or less restricted by involving only one global EDR direction. It has already been observed, e.g., in Xia et al (2002), that some real data sets can have more than one EDR direction for which the SIM does not work well. On the other hand, if we include more EDR directions into the model, we take the risk of losing the optimal estimation efficiency of the link function $f(\cdot)$. There has not been a well-developed method that not only keeps the estimation efficiency of SIM but also allows more than one EDR direction from a global view.

1.1.3 Piecewise Regression Models

Another important approach on approximating the function $\psi(\cdot)$ in (1.1) is through a piecewise regression model, which is also called the tree-structured model. Piecewise models partition the feature space into several disjoint subspaces and fit each subspace with a simple regression model. Specifically, if we assume the subspaces take the shape of rectangles and the function value within each subspace

is a constant, we reach the famous CART model of Breiman, Friedman, Olshen and Stone (1984), i.e., assuming we have M such subspaces $\{R_1, ..., R_M\}$, the function $\psi(\cdot)$ in (1.1) is approximated by

$$\hat{\psi}_c(\boldsymbol{x}) = \sum_{m=1}^{M} c_m \mathbb{I}\{\boldsymbol{x} \in R_m\},$$

where c_m are constants and $\mathbb{I}\{A\}$ is the indicator function of set A. To estimate this model, CART starts from the whole space (the root) and searches for the best cut-point for a univariate split by optimizing a cost function. If we do this recursively on the resulting nodes, we end up with a large initial tree. CART then prune down the size of the tree by a cross-validation procedure. The c_m for region R_m is estimated by the simple average of the response variables within R_m.

Li, Lue and Chen (2000) extended this idea by allowing c_m to be a linear combination of $\boldsymbol{x}$. Their new model is called tree-structured linear regression with the form

$$\hat{\psi}_l(\boldsymbol{x}) = \sum_{m=1}^{M} \beta_m^{\top} \boldsymbol{x}\ \mathbb{I}\{\boldsymbol{x} \in R_m\}.$$

where the regions R_m are partitioned by linear straight lines estimated through the so-called primary PHD directions; see also Li (1992).

In piecewise modeling, to give a reasonable partition of the feature space of $\boldsymbol{x}$ is crucial for building a useful model. Most piecewise methods in current literature rely on some parametric assumptions on the partitioning rules among the regions

$\{R_1, ..., R_M\}$, e.g. rectangle shape as assumed by CART or linear partitions as assumed by tree-structured linear regression. Although by imposing on parametric assumptions we usually improve the stability of the fitted model, we lose the flexibility and capability to model more complicated data structures.

1.1.4 Piecewise Single-Index Model (pSIM)

Following the direction of last subsection and given the efficiency of SIM, it is natural to consider the piecewise SIM defined as

$$\hat{\psi}_s(\boldsymbol{x}) = \sum_{m=1}^{M} f_m(\beta_m^\top \boldsymbol{x})\ \mathbb{I}\{\boldsymbol{x} \in R_m\}. \tag{1.5}$$

Gramacy and Lian (2012) has studied this form of model in the context of Bayesian approaches by restricting that $\{R_1, ..., R_M\}$ are partitioned by binary splits of the coordinates in $\boldsymbol{x}$. In this thesis, model (1.5) is investigated through a frequentist's point of view with weaker restrictions.

Our method will build on the two general categories of approaches to the curse of dimensionality as discussed in subsection 1.1.1 to subsection 1.1.3. First of all, we assume that the link function $\psi(\cdot)$ in model (1.1) satisfies

$$\psi(x_1, ..., x_p) = \phi(\boldsymbol{\eta}_1^\top \boldsymbol{x}, ..., \boldsymbol{\eta}_d^\top \boldsymbol{x})$$

with $d < p$, and thus

$$y = \phi(\boldsymbol{\eta}_1^\top \boldsymbol{x}, ..., \boldsymbol{\eta}_d^\top \boldsymbol{x}) + \varepsilon, \tag{1.6}$$

where ϕ is an unknown link function and $\boldsymbol{\eta}_k,\ k = 1, 2, ..., d$, are constant vectors.

In this Chapter, we consider a piecewise single-index model (pSIM) to perform nonparametric regression in a multidimensional space. Our model can be written as

$$y = \begin{cases} \phi_1(\boldsymbol{\beta}_1^\top \boldsymbol{x}) + \varepsilon_1, & \text{if} \quad \boldsymbol{x} \in R_1, \\ \quad \dots & \quad \dots \\ \phi_m(\boldsymbol{\beta}_m^\top \boldsymbol{x}) + \varepsilon_m, & \text{if} \quad \boldsymbol{x} \in R_m, \end{cases} \tag{1.7}$$

where $\boldsymbol{\beta}_g, g = 1, ..., m$, are $p \times 1$ vectors, $\phi_g, g = 1, ..., m$, are smooth functions on $\mathbb{R}$, $E(\varepsilon_g | \boldsymbol{x},\ R_g) = 0$, $\cup_{g=1}^m R_g = \mathbb{R}^p$ and $R_i \cap R_j = \emptyset$ for any $i \neq j$. The regions $R_i,\ i = 1, ..., m$, need not be contiguous. The error term ε_g is assumed to be independently and identically distributed within region R_g. Heteroscedasticity of the error terms across different regions are allowed. We call $\boldsymbol{\beta}_g$ the piecewise single-index for region R_g. Model (1.7) is an extension of the tree-structured linear regression model proposed by Li et al (2000) that splits the sample space into several regions through linear combinations of $\boldsymbol{x}$. To link model (1.6) with model (1.7), we further assume that the boundaries of $R_1, ..., R_m$ are uniquely determined by $(\boldsymbol{\beta}_1^\top \boldsymbol{x}, ..., \boldsymbol{\beta}_m^\top \boldsymbol{x})$. In other words, the relationship between y and $\boldsymbol{x}$ in model (1.7) is uniquely determined by $(\boldsymbol{\beta}_1^\top \boldsymbol{x}, ..., \boldsymbol{\beta}_m^\top \boldsymbol{x})$, so in this case model (1.7) can also be

written in the form of model (1.6) with $d = m$ and $\boldsymbol{\beta}_k = \boldsymbol{\eta}_k$, for $k = 1, ..., m$. However, model (1.7) enjoys a more specific description of the relationships between y and $\boldsymbol{x}$ with only one effective dimension in each region. Moreover, as compared with the dimension reduction model (1.6), model (1.7) allows more than p regions in the model, i.e., it is possible that $m \geq p$, in which case the dimension can not be reduced by model (1.6).

Similar models have been considered in the literature. Chipman, Geoge and McCulloch (2002) proposed a Bayesian approach to fit the tree models that split the sample space into smaller regions, recursively splitting on a single predictor, applying different linear models on the terminal nodes. Gramacy and Lian (2012) extended this idea to allow single-index link functions in each of the terminal nodes. In fact, the pSIM model can be regarded as a special case of the hierarchical mixture experts (HME) which assign every observation according to a specific rule to different models. HME is more general in its form than the piecewise models, but its estimation is more complicated; see for example Villani, Kohn and Giordani (2009) and Montanari and Viroli (2011) for more details.

In this Chapter, we propose to partition the sample space according to the gradient direction at each sample point. The rationale is the fact that points with the same gradient direction follow the same single-index model and thus should fall into the same region. Many efficient methods are available for the estimation of

gradient directions. See for example Härdle and Stoker (1989), Ruppert and Wand (1994) and Xia et al (2002). In this Chapter, we adopt the estimation method of Xia et al (2002) that uses the first few eigenvectors of the average of outer product of gradients (OPG) as the directions for dimension reduction. A rigorous theoretical justification of the estimation can be found in Xia (2007). This idea will be used in this Chapter to reduce the effect of high dimensionality and to improve the accuracy of estimation.

The rest of the Chapter is organized as follows. Section 1.2 discusses the methodology for model estimation and selection. A method is developed to partition the whole sample space; and local linear smoothing is used to estimate the link functions. A BIC-type criterion is employed to select the number of regions. To check the usefulness of our approach, Section 1.3 gives two simulation examples and Section 1.4 studies three popular real data sets. Section 1.5 and Section 1.6 are devoted to the asymptotic analysis of the estimators.

1.2 Estimation of pSIM

Estimation of model (1.7) consists of two parts. First, we need to partition the whole space into m subsets or regions. Secondly, we need to use semiparametric methods to estimate the single-index model in each region. The selection of m also

needs to be investigated.

1.2.1 Model Estimation

Suppose we have a set of observations $(\boldsymbol{x}_i, y_i), i = 1, ..., n$. To partition the whole sample space, we first estimate the pointwise local gradient direction at each observation, and use them to cluster the observations into m groups. The rationale behind this method is that the estimated local gradient directions for the points in the same single-index model should be close to one another while those in different regions should be apart.

Consider the estimation of the gradient direction at a given point $\boldsymbol{x}_i$. Using local linear approximation, we can get a preliminary estimate for the gradient $\boldsymbol{b}_i$ at $\boldsymbol{x}_i$ through

$$(\hat{a}_i, \hat{\boldsymbol{b}}_i) = \underset{a,\boldsymbol{b}}{\operatorname{argmin}} \sum_{j=1}^{n} \{y_{i_k} - a - \boldsymbol{b}^\top (\boldsymbol{x}_i - \boldsymbol{x}_j)\}^2 w_{i,j}, \tag{1.8}$$

where $w_{i,j}$ is a symmetric weight function of the form $h_i^{-p} \mathrm{K}\{h_i^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)\}$ in which h_i is the bandwidth and $\mathrm{K}(\cdot)$ is the kernel function. If the observations are generated from model (1.7), for any $\boldsymbol{x}_i \in R_{g_i}$, the standardized gradient direction $\tilde{\boldsymbol{b}}_i = \hat{\boldsymbol{b}}_i / \hat{\boldsymbol{b}}_i^\top \hat{\boldsymbol{b}}_i$ is a local estimation for the regional single index $\boldsymbol{\beta}_{g_i}$, where g_i denotes the region index of $\boldsymbol{x}_i$. Suppose conditions (A1) - (A5) in the Appendix hold, a direct application of the Theorem 2 of Lu (1996) gives that $\tilde{\boldsymbol{b}}_i = \boldsymbol{\beta}_{g_i} + o_P(\mathbf{1})$, where

$o_P(\mathbf{1})$ is a infinitesimal item as n approaches to infinity. If $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same region R_g as defined in model (1.7), then we have $\tilde{\boldsymbol{b}}_j = \tilde{\boldsymbol{b}}_i + o_P(\mathbf{1})$. Thus if the observations are generated from model (1.7), the estimated standardized gradient directions $\{\tilde{\boldsymbol{b}}_i : \ i = 1, ..., n\}$ can be separated into m subgroups with centroid directions $\boldsymbol{\beta}_g$ for $g = 1, ..., m$ respectively. Then we can easily identify the regions in model (1.7) by clustering $\{\tilde{\boldsymbol{b}}_i : \ i = 1, ..., n\}$ into m subgroups.

The estimator (1.8) can be improved if the observations are also believed to follow the model (1.6). Based on the idea of the OPG method (Xia et al, 2002), we can estimate the effective dimension reduction directions $\boldsymbol{B} = (\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_q)$ through the first q eigenvectors of the OPG matrix calculated as

$$\hat{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^{n} \hat{\boldsymbol{b}}_i \hat{\boldsymbol{b}}_i^{\top}, \tag{1.9}$$

where the value of q is chosen by a data-driven approach; see Step 2 below for details. Then, the kernel weights $w_{i,j}$ in (1.8) can be refined to work on a lower dimension space $\boldsymbol{B}^{\top}\boldsymbol{x}$ as

$$w_{i,j} = h_i^{-q} \mathrm{K}\{h_i^{-1} \boldsymbol{B}^{\top} (\boldsymbol{x}_i - \boldsymbol{x}_j)\}.$$

The estimated gradients $\{\hat{\boldsymbol{b}}_i : \ i = 1, ..., n\}$ can be updated with the refined kernel weights. In this way, we propose an iterative algorithm to estimate the local direction of gradients as follows.

Step 0. Set $\boldsymbol{B}_0 = \boldsymbol{I}_p$ and $t = 0$, where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. Let

$w_{i,j}^{(0)} = h_i^{-p}\mathrm{K}\{h_i^{-1}\boldsymbol{B}_0^\top(\boldsymbol{x}_i - \boldsymbol{x}_j)\}$ for $i, j = 1, ..., n$.

Step 1. Calculate the solutions to (1.8) for $i = 1, ..., n$,

$$\begin{pmatrix} a_i^{(t)} \\ \boldsymbol{b}_i^{(t)} \end{pmatrix} = \left\{ \sum_{j=1}^{n} w_{i,j}^{(t)} \begin{pmatrix} 1 \\ \boldsymbol{x}_i - \boldsymbol{x}_j \end{pmatrix} \begin{pmatrix} 1 \\ \boldsymbol{x}_i - \boldsymbol{x}_j \end{pmatrix}^\top \right\}^{-1} \sum_{j=1}^{n} w_{i,j}^{(t)} \begin{pmatrix} 1 \\ \boldsymbol{x}_i - \boldsymbol{x}_j \end{pmatrix} y_j.$$

Step 2. Let

$$\hat{\boldsymbol{\Sigma}}^{(t)} = n^{-1} \sum_{i=1}^{n} \boldsymbol{b}_i^{(t)} \boldsymbol{b}_i^{(t)\top},$$

which is the average outer product of gradients (OPG). Make a principal component decomposition of $\hat{\boldsymbol{\Sigma}}^{(t)}$,

$$\hat{\boldsymbol{\Sigma}}^{(t)} = \lambda_1 \boldsymbol{\eta}_1 \boldsymbol{\eta}_1^\top + \cdots + \lambda_p \boldsymbol{\eta}_p \boldsymbol{\eta}_p^\top,$$

where $\lambda_1 > \cdots > \lambda_p \geq 0$. Let $\boldsymbol{B}_t = (\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_{\tilde{q}})$ be the first $\tilde{q}$ eigenvectors of $\hat{\boldsymbol{\Sigma}}^{(t)}$, where $\tilde{q} = \max\{2, \tilde{q}_0\}$ with $\tilde{q}_0$ being determined by

$$\sum_{k \leq (\tilde{q}_0 - 1)} |\lambda_k| \Big/ \sum_{k=1}^{p} |\lambda_k| < \max\{R_0, 1 - 1/\sqrt{n}\},$$

and

$$\sum_{k \leq \tilde{q}_0} |\lambda_k| \Big/ \sum_{k=1}^{p} |\lambda_k| \geq \max\{R_0, 1 - 1/\sqrt{n}\}.$$

To ensure the selected components contain a large proportion of information, we take $R_0 = 0.95$ in our calculation.

Step 3. Set $t = t + 1$. If $\tilde{q} < p$, update $w_{i,j}^{(t)} = h_i^{-\tilde{q}}\mathrm{K}\{h_i^{-1}\boldsymbol{B}_t^\top(\boldsymbol{x}_i - \boldsymbol{x}_j)\}$. Repeat Steps 1 and 2 until convergence. Denote the final value of $\boldsymbol{B}_t$ and $\boldsymbol{b}_i^{(t)}$ by $\boldsymbol{B}$ and $\hat{\boldsymbol{b}}_i$ respectively.

Step 4. Calculate $\tilde{\boldsymbol{b}}_i = \hat{\boldsymbol{b}}_i / \hat{\boldsymbol{b}}_i^\top \hat{\boldsymbol{b}}_i$ for $i = 1, ..., n$.

The above algorithm is inspired by the OPG algorithm of Xia (2007) who proved the convergence of the OPG-related algorithms. In practice, we usually standardize $\boldsymbol{x}_i$ by letting $\boldsymbol{x}_i = \boldsymbol{S}^{-1/2}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$, where $\bar{\boldsymbol{x}} = n^{-1}\sum_{i=1}^n \boldsymbol{x}_i$ and $\boldsymbol{S} = n^{-1}\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$ before applying the above algorithm.

Based only on the Euclidean distances of the estimated gradient directions, we cluster the observations into m groups through the K-means method. Let $\hat{I}_g$ contain all the indices i of observation $(\boldsymbol{x}_i, y_i)$ that are in group $g = 1, ..., m$. After the groups are identified, we estimate the piecewise single-index $\boldsymbol{\beta}_g$ in each group using all the observations in $\hat{I}_g$ through Steps 0 - 3 by fixing $\tilde{q} = 1$ for $t \geq 1$. By doing this, we assume that each cluster group corresponds to a region of model (1.7). Denote the resulting estimate by $\hat{\boldsymbol{\beta}}_g$. Its asymptotic properties are studied in Section 1.5.

As the piecewise single-index model reduces the original p-dimensional predictor to 1-dimensional predictor in each region, the link functions $\phi_g(\cdot)$ for group g can be estimated well by local linear smoothing,

$$(\hat{\phi}_g(x), \hat{\phi}_g'(x)) = \underset{a,b}{\operatorname{argmin}} \sum_{j \in \hat{I}_g} \{y_j - a - b(\hat{\boldsymbol{\beta}}_g^\top \boldsymbol{x}_j - x)\}^2 \mathrm{K}\{(\hat{\boldsymbol{\beta}}_g^\top \boldsymbol{x}_j - x)/H_g\}. \quad (1.10)$$

It is shown in Section 1.5 that $\hat{\phi}_g(x)$ can achieve the same estimation efficiency as if the true indices $\boldsymbol{\beta}_g$, $g = 1, ..., m$ are known.

To make prediction for a newly observed (out of the training sample) predictor $\boldsymbol{x}_{new}$, we need to classify the predictor into the most appropriate region. Based on the partitioning results on the estimated directions $\{\tilde{\boldsymbol{b}}_i : \ i = 1, ..., n\}$, we create a labeled training sample $\{(\boldsymbol{x}_i, g_i), \ i = 1, ..., n\}$, where $g_i \in \{1, ..., m\}$ is the group index of $\boldsymbol{x}_i$. The region identification problem is actually a supervised classification problem. Techniques are available in the literature; see for example Hastie, Tibshirani and Friedman (2009) for a nice review. We propose using k-nearest-neighbor (kNN) based on the distance in the space $\boldsymbol{B}^\top \boldsymbol{x}$. We then apply (1.10) to estimate the response value of $\boldsymbol{x}_{new}$ after its region is identified.

1.2.2 Selection Of Tuning Parameters

Our algorithm involves two sets of tuning parameters: the bandwidth $h_i^{(t)}$ used in gradient direction estimations and the bandwidth H_g used in estimating the link functions.

To ensure convergence of the OPG-related algorithm, Xia (2007) suggested the following sequence of bandwidths

$$h_i^{(t+1)} = \max\{h_i^{(t)} n^{-1/(2(p+6))}, c_0 n^{-1/5}\}$$

for $t \geq 0$ with $h_i^{(0)} = c_0 n^{-1/(p+6)}$, where $c_0 = 2.34$ as suggested by Silverman (1986) for the Epanechnikov kernel. For ease of exposition, we propose to use

$h_i^{(0)} = 2.34n^{-1/(p+6)}$ and then fix h_i for all subsequent iterations, i.e., let $h_i^{(t)} \equiv h_0$, for $t \geq 1$. In later sections of this Chapter, one h_0 is used in the examples.

Then we choose the h_0 and H_g, $g = 1, ..., m$, based on leave-one-out cross validation (LOO-CV). More precisely, for $i \in \hat{I}_g$, let $\hat{\phi}_g^{(-i)}(\boldsymbol{x}_i)$ be the estimator of $\phi_g(\boldsymbol{x}_i)$ obtained by (1.10) with $(\boldsymbol{x}_i, y_i)$ itself being excluded, i.e., $\hat{\phi}_g^{(-i)}(\boldsymbol{x}_i)$ is the LOO prediction of $\phi_g(\boldsymbol{x}_i)$. Note that $\hat{\phi}_g^{(-i)}(\boldsymbol{x}_i)$ is a function of both h_0 and H_g. We thus denote it as $\phi_g^{(-i)}(\boldsymbol{x}_{g_j}; h_0, H_g)$. The CV score of the LOO estimators in $\hat{I}_g$ is defined as

$$\mathrm{CV}_g(h_0, H_g) = \sum_{j \in \hat{I}_g} \left(y_j - \hat{\phi}_g^{(-i)}(\boldsymbol{x}_j; h_0, H_g)\right)^2.$$

The total CV score is then

$$\mathrm{CV}(h_0, H_1, ..., H_m) = \sum_{g=1}^{m} \mathrm{CV}_g(h_0, H_g).$$

It is easy to see that with fixed h_0, each $\mathrm{CV}_g(h_0, H_g)$ is a consistent criterion for choosing the optimal smoothing parameter H_g; see for example Fan and Gijbels (1996). On the other hand, with the optimal H_g, $g = 1, ..., m$, we can find h_0 that minimizes $\mathrm{CV}(h_0, H_1, ..., H_m)$.

There are many viable criteria to select m which determines the complexity of the piecewise single-index model. Because the CV approach is computationally more difficult, we develop a BIC (Schwarz, 1978) approach for the selection. It has been shown that for kernel smoothing, the degree of freedom is of order $1/h$, where

h is the smoothing bandwidth; see Zhang (2003). The BIC score for the model with m regions is calculated as

$$\mathrm{BIC}(m) = \log(\hat{\sigma}^2(m)) + \log(n) \sum_{g=1}^{m} \frac{1}{\hat{n}_g(m) H_g(m)},$$

where $\hat{n}_g(m) = \#\hat{I}_g(m)$ is the number of points in the gth region, $H_g(m)$ is the smoothing bandwidth used in the link function in the gth region, and $\hat{\sigma}^2(m)$ is the estimator of the overall noise variance, i.e.,

$$\hat{\sigma}^2(m) = n^{-1} \sum_{g=1}^{m} \sum_{j \in \hat{I}_g} (y_j - \hat{\phi}_g^{(m)}(\hat{\boldsymbol{\beta}}_g(m)^\top \boldsymbol{x}_j))^2.$$

We choose the number of regions as

$$\hat{m}_{\mathrm{BIC}} = \arg \min_{1 \le m \le M_0} \mathrm{BIC}(m),$$

where M_0 is a predetermined upper bound, usually $M_0 = \lfloor \log(n) \rfloor$. The asymptotic property of the selection is also discussed in Section 5.

1.3 Simulations

To assess the accuracy of model fitting and prediction, we use the average squared error (ASE) defined by

$$\mathrm{ASE} = n^{-1} \sum_{i=1}^{n} \{\phi(\boldsymbol{x}_i) - \hat{\phi}(\boldsymbol{x}_i)\}^2,$$

where $\hat{\phi}$ is the estimate of ϕ. The deviances of the estimated piecewise gradient directions from the true gradient directions are measured by

$$D^2(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) := 1 - (\hat{\boldsymbol{\beta}}^T \boldsymbol{\beta})^2.$$

The noise level is measured by

$$\text{SNR} := \text{corr}(\phi(\boldsymbol{x}), \phi(\boldsymbol{x}) + \varepsilon).$$

The theoretical SNR's of the simulated examples are reported in the corresponding tables below. We study the treed Gaussian process single-index model (TGP-SIM) of Gramacy and Lian (2012) in the simulations for comparison. The TGP-SIM in the simulations studies are all estimated by the "btgp" function in the R package "tgp", see Gramacy (2009) for details. Our method is denoted by "pSIM".

Example 1.3.1. We first study the following piecewise linear model of a triangle pyramid shape used in Li et al (2000).

$$y = \begin{cases} -\boldsymbol{\xi}_1^\top \boldsymbol{x} - \sqrt{3}\boldsymbol{\xi}_2^\top \boldsymbol{x} + 1 + 0.5\varepsilon, & \text{if } \boldsymbol{\xi}_2^\top \boldsymbol{x} \geq 0 \text{ and } \sqrt{3}\boldsymbol{\xi}_1^\top \boldsymbol{x} + \boldsymbol{\xi}_2^\top \boldsymbol{x} \geq 0, \\ -\boldsymbol{\xi}_1^\top \boldsymbol{x} + \sqrt{3}\boldsymbol{\xi}_2^\top \boldsymbol{x} + 1 + 0.5\varepsilon, & \text{if } \boldsymbol{\xi}_2^\top \boldsymbol{x} < 0 \text{ and } \sqrt{3}\boldsymbol{\xi}_1^\top \boldsymbol{x} - \boldsymbol{\xi}_2^\top \boldsymbol{x} \geq 0, \\ 2\boldsymbol{\xi}_1^\top \boldsymbol{x} + 1 + 0.5\varepsilon, & \text{if } \sqrt{3}\boldsymbol{\xi}_1^\top \boldsymbol{x} + \boldsymbol{\xi}_2^\top \boldsymbol{x} < 0 \text{ and } \sqrt{3}\boldsymbol{\xi}_1^\top \boldsymbol{x} - \boldsymbol{\xi}_2^\top \boldsymbol{x} < 0, \end{cases}$$

where $\boldsymbol{\xi}_1 = (1, 1, 1, 1, 1, 0, ..., 0)^\top$, $\boldsymbol{\xi}_2 = (0, ..., 0, 1, 1, 1, 1, 1)^\top$, $\boldsymbol{x} = (x_1, ..., x_{10})^\top$ and $\varepsilon, x_1, ..., x_{10}$ are IID standard normal random variables. After standardization, the gradients in the three regions are respectively

$$\boldsymbol{\beta}_1 = (0.2236,\ ...,\ 0.2236,\ 0.3872,\ ...,\ 0.3872)^\top,$$

$$\boldsymbol{\beta}_2 = (-0.2236,\ ...,\ -0.2236,\ 0.3872,\ ...,\ 0.3872)^\top$$

and

$$\boldsymbol{\beta}_3 = (0.4472,\ ...,\ 0.4472,\ 0\ ,\ ...,\ 0)^\top$$

The sample size is set as $n = 200$ or $n = 400$, and 100 replications are drawn in each case.

An estimation example with size $n = 400$ is shown in Figure 1.1. The panels on the left show the locations of the points on the subspace $(\boldsymbol{\xi}_1^\top \boldsymbol{x}, \boldsymbol{\xi}_2^\top \boldsymbol{x})$: dashed lines represent the true boundaries among the three regions; circles 'o' are the points classified by our proposed pSIM estimator to the respective region; dots '·' are the points classified into the other two regions. We can observe that the circles generally match up with the true regions. The link functions for each group of circles on the left are plotted on the right.

Figure 1.2 shows the boxplots of the gradient estimation errors $D^2(\hat{\boldsymbol{\beta}}_i, \boldsymbol{\beta}_i)$ for $i = 1, 2, 3$. We could see a clear improvement from $n = 200$ to $n = 400$, demonstrating consistency.

To compare the out-of-sample prediction of TGP-SIM and pSIM, we draw an additional test sample of 50 points randomly at each replicate. The in-sample (IS) and out-of-sample (OS) prediction errors are shown in Table 1.1. The percentage numbers in the parenthesis are the proportion of times that the number of regions

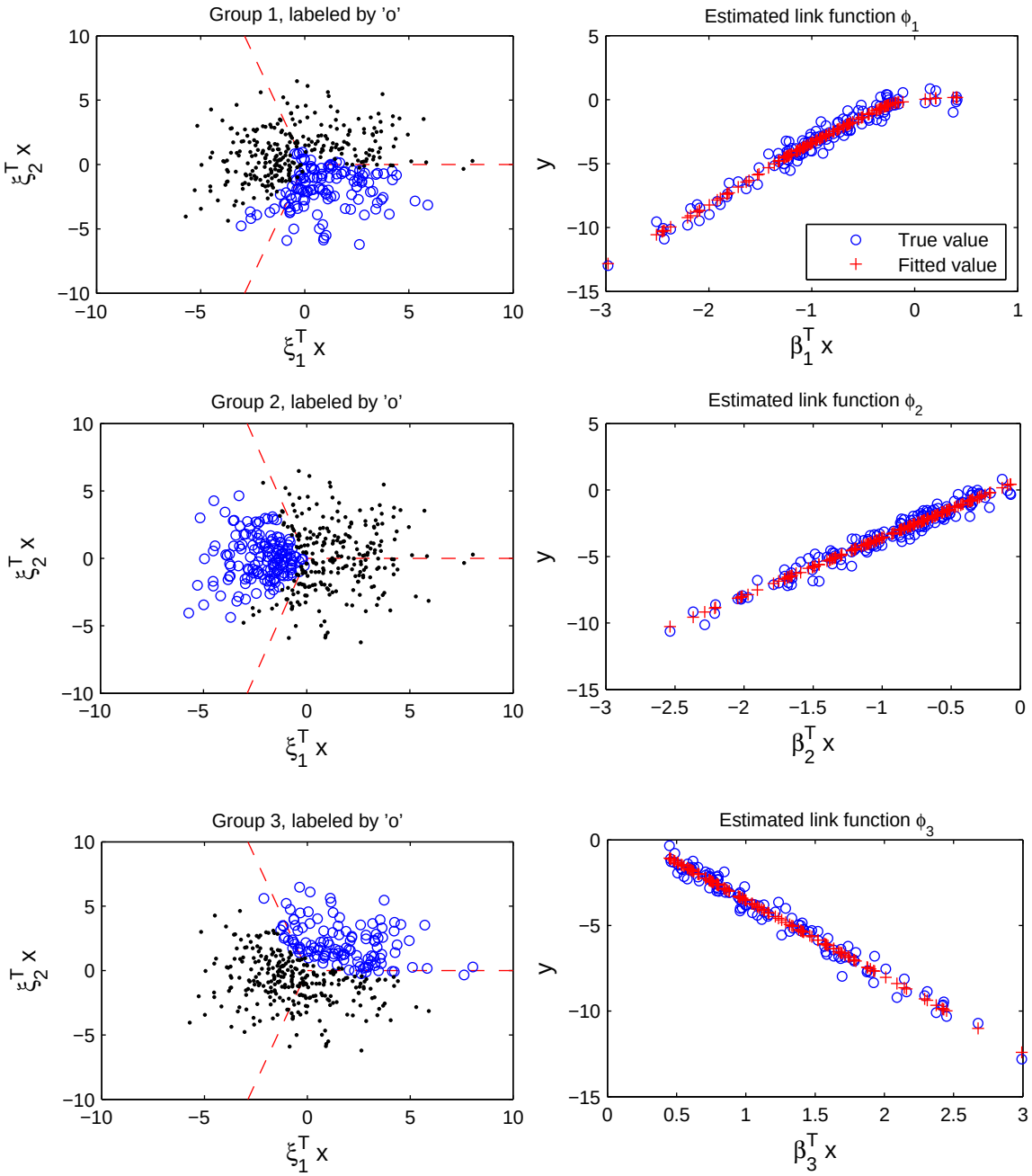

Figure 1.1 A typical estimation result of Example 1.3.1 with sample size $n = 400$.

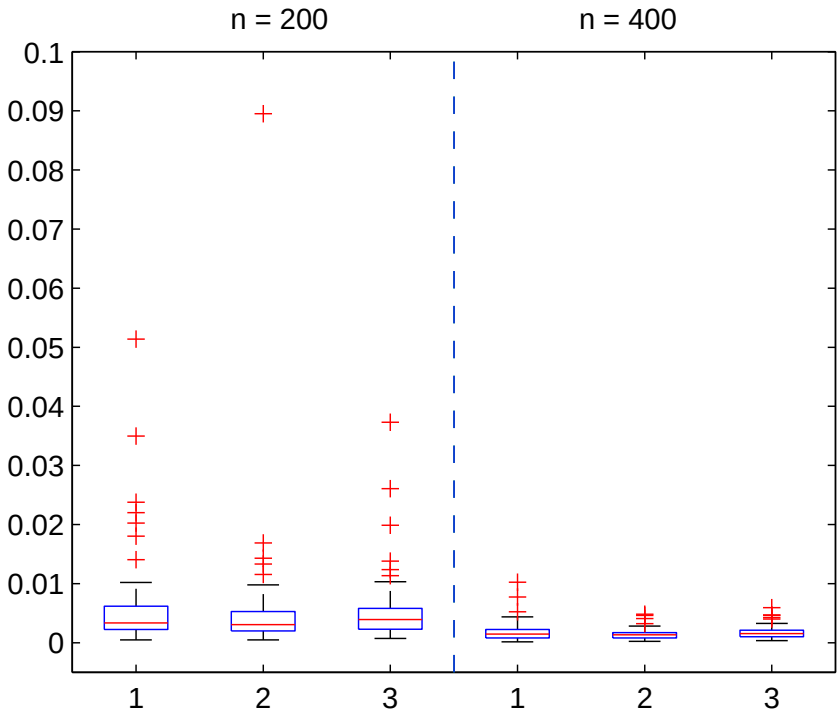

Figure 1.2 The estimation errors of the three piecewise single-index $D^2(\hat{\boldsymbol{\beta}}_i, \boldsymbol{\beta}_i)$, $i = 1, 2, 3$ in Example 1.3.1.

(m) of the model is identified as three by the proposed BIC method. The TGP-SIM method cannot give reliable prediction even though it still fits the data reasonably well. This inferior prediction performance is partially due to the fact that the data generated in this example are not within the tree-SIM class.

Example 1.3.2. This example is inspired by Gramacy and Lee (2008) and Gramacy and Lian (2012). Consider two exponential single-index functions divided by

Table 1.1 Simulation results of Example 1.3.1: mean of in-sample (IS) and out-of-sample (OS) prediction errors (ASE) from the 100 replications. The percentage numbers in the parenthesis are the proportion of times that the number of regions (m) of the model is identified as three by the proposed BIC method.

$SNR = 0.98$			Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$n = 200$	TGP-SIM	IS	0.2144	0.2975	0.3392	0.3429	0.3776	0.5127
		OS	0.7936	1.4867	1.7061	1.8028	2.2058	3.7375
	pSIM (98%)	IS	0.0354	0.0665	0.0897	0.1416	0.1443	1.0956
		OS	0.0517	0.1191	0.1876	0.2749	0.2913	1.8710
$n = 400$	TGP-SIM	IS	0.2477	0.3195	0.3438	0.3454	0.3729	0.4423
		OS	0.8368	1.0666	1.2415	1.2618	1.4221	1.9759
	pSIM (99%)	IS	0.0196	0.0380	0.0471	0.0572	0.0595	0.2039
		OS	0.0202	0.0478	0.0692	0.1011	0.1176	0.6803

a straight line,

$$y = \begin{cases} (\boldsymbol{\xi}_1^\top \boldsymbol{x} + \boldsymbol{\xi}_2^\top \boldsymbol{x} + 1)\exp(-(\boldsymbol{\xi}_1^\top \boldsymbol{x} + \boldsymbol{\xi}_2^\top \boldsymbol{x} + 1)^2) + \varepsilon, & \text{if } \boldsymbol{\xi}_2^\top \boldsymbol{x} \geq 0, \\ (\boldsymbol{\xi}_1^\top \boldsymbol{x} - \boldsymbol{\xi}_2^\top \boldsymbol{x} + 1)\exp(-(\boldsymbol{\xi}_1^\top \boldsymbol{x} - \boldsymbol{\xi}_2^\top \boldsymbol{x} + 1)^2) + \varepsilon, & \text{if } \boldsymbol{\xi}_2^\top \boldsymbol{x} < 0, \end{cases}$$

where $\boldsymbol{x} \sim \text{Unif}([-1,1]^{\otimes p})$, and $\varepsilon \sim N(0, \sigma^2)$. We consider $n = 200$ or 400, $p = 5, 10$ or 20, $\sigma = 0.01,\ 0.05,\ 0.1$ or 0.2, and 100 replications in each case. To

better illustrate how our new method works on high dimension problems, we set

$$\begin{aligned}
\boldsymbol{\xi}_1^{(5)} &= (1,0,0,0,0)^\top,\ \boldsymbol{\xi}_2^{(5)} = (0,1,0,0,0)^\top,\\
\boldsymbol{\xi}_1^{(10)} &= (1,0,0,0,...,0)^\top,\ \boldsymbol{\xi}_2^{(10)} = (0,1,1,0,...,0)^\top,\\
\boldsymbol{\xi}_1^{(20)} &= (1,1,0,0,0,...,0)^\top,\ \boldsymbol{\xi}_2^{(20)} = (0,0,1,1,0,...,0)^\top.
\end{aligned}$$

In this way, as the dimension of $\boldsymbol{x}$ gets larger, more coordinates get involved in the model. Note that when $p = 5$, the model is in the tree-SIM class, but for $p = 10$ and $p = 20$, it is not in tree-SIM. Figure 1.3 shows four typical estimation and classification results for four different settings of sample sizes and noise levels with $p = 20$. Namely, case 1: $n = 200$, $\sigma = 0.01$; case 2: $n = 200$, $\sigma = 0.1$; case 3: $n = 400$, $\sigma = 0.01$; case 4: $n = 400$, $\sigma = 0.1$. Each row belongs to a single case. The panels on the left most column are the classification results on the subspace $(\boldsymbol{\xi}_1^\top \boldsymbol{x}, \boldsymbol{\xi}_2^\top \boldsymbol{x})$. The rest two columns are the true response values ('o') and their fitted values ('+') for the two respective pieces where $\boldsymbol{\beta}_1$'s and $\boldsymbol{\beta}_2$'s are estimated by pSIM. The panels on the left most column are the classification results on the subspace $(\boldsymbol{\xi}_1^\top \boldsymbol{x}, \boldsymbol{\xi}_2^\top \boldsymbol{x})$. The other two columns are the true response values ('o') and their estimated values ('+') for the two respective groups. The nonlinear shape of the link functions are clearly shown in the plots.

The test samples are generated in the same way as in Example 1. The in-sample (IS) and out-of-sample (OS) prediction errors are summarized in Table 1.2 and Table 1.3. The percentage numbers in the parenthesis are the proportion

Table 1.2 Simulation results of Example 1.3.2: mean of in-sample (IS) and out-of-sample (OS) prediction errors (ASE) ($\times 10^{-3}$) from the 100 replications.

			$n = 200$			$n = 400$		
			$p = 5$	$p = 10$	$p = 20$	$p = 5$	$p = 10$	$p = 20$
		SNR	*0.9979*	*0.9978*	*0.9973*	*0.9979*	*0.9978*	*0.9973*
	TGP-SIM	IS	0.0631	0.1365	0.1374	0.0470	0.0981	0.1008
$\sigma = 0.01$		OS	0.2055	1.9287	4.0118	0.1013	0.6005	1.6332
	pSIM	IS	0.2886	0.3573	0.7438	0.2017	0.2946	0.4060
		BIC	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)
		OS	0.4867	0.6579	1.0468	0.2689	0.3731	0.5290
		SNR	*0.9500*	*0.9490*	*0.9380*	*0.9500*	*0.9490*	*0.9380*
	TGP-SIM	IS	0.7397	1.9672	2.130	0.4932	1.3893	2.2794
$\sigma = 0.05$		OS	1.0644	2.8761	5.0652	0.6597	1.6096	2.9425
	pSIM	IS	0.6079	0.9323	1.0790	0.3560	0.4562	0.5868
		BIC	(100%)	(99%)	(100%)	(100%)	(100%)	(100%)
		OS	0.9232	1.5662	1.6663	0.4690	0.5467	0.7379

of times that the number of regions (m) of the model is identified as two by the proposed BIC method. For $n = 200$, both IS and OS predictions of pSIM are better when SNR is high, but its OS prediction performance gets worse more quickly than

Table 1.3 Simulation results of Example 1.3.2 (continued): mean of in-sample (IS) and out-of-sample (OS) prediction errors (ASE) ($\times 10^{-3}$) from the 100 replications.

			$n = 200$			$n = 400$		
			$p = 5$	$p = 10$	$p = 20$	$p = 5$	$p = 10$	$p = 20$
		SNR	*0.8348*	*0.8324*	*0.8020*	*0.8348*	*0.8324*	*0.8020*
	TGP-SIM	IS	1.9671	5.0168	6.1475	1.2672	3.4714	7.5732
$\sigma = 0.1$		OS	2.2142	5.5400	7.4261	1.3149	2.8766	4.0153
	pSIM	IS	1.5008	3.3319	4.1657	0.7420	1.1342	1.1864
		BIC	(98%)	(98%)	(97%)	(100%)	(100%)	(100%)
		OS	1.9050	5.1201	5.4902	0.8757	1.3015	1.4521
		SNR	*0.6050*	*0.5979*	*0.5575*	*0.6050*	*0.5979*	*0.5575*
	TGP-SIM	IS	6.1535	13.854	17.831	3.7130	10.177	18.343
$\sigma = 0.2$		OS	5.7649	10.625	11.931	3.7635	6.7275	8.3592
	pSIM	IS	6.4114	12.1707	15.5733	2.3703	5.9173	6.7964
		BIC	(87%)	(84%)	(80%)	(96%)	(94%)	(89%)
		OS	7.8240	17.6934	23.1272	2.9535	6.7168	8.1093

TGP-SIM as SNR becomes lower. This is the cost of the higher flexibility enjoyed by pSIM with less restrictions on the boundaries. As SNR is low and sample size is

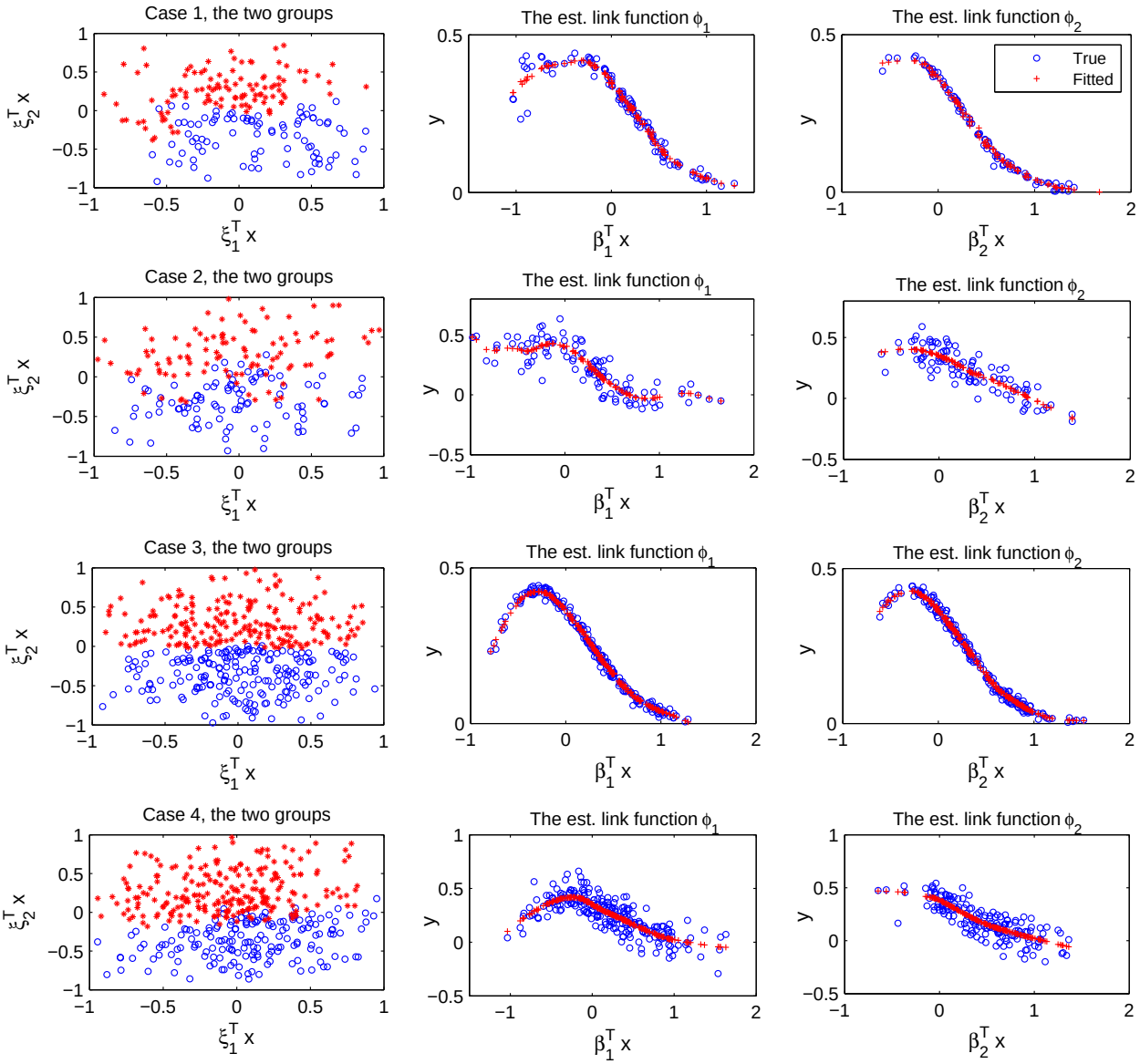

Figure 1.3 Four typical estimation results of Example 1.3.2.

less than sufficient, the models with more specific assumptions such as TGP-SIM are usually able to give more robust predictions. Nevertheless, for larger samples, namely $n = 400$, pSIM seems to perform better in most cases, and incurs similar

errors for both IS prediction and OS prediction. When the SNR is very high, i.e., for $\sigma = 0.01$, the gaps between IS and OS prediction errors from the TGP-SIM are too wide, which is a sign of model over fitting. Keeping σ fixed, as sample size goes bigger, our pSIM approach is more efficient in taking in the additional information provided by more samples to give smaller IS and OS prediction errors.

1.4 Real Data Analysis

In this section, we apply our estimation method to three popular data sets. The first data set concerns the salary of 263 baseball players; it was originally given at 1988 ASA Graphics Poster Session (Chaudhuri, Huang, Loh and Yao, 1994). The second data set studies the atmospheric ozone concentration in Los Angeles basin (Breiman and Friedman, 1985). The last data set considered in this section is the cars data set which studies the fuel efficiency for automobiles; it is obtained from the ASA Data Exposition dataset (1983) collected by Professor Ernesto Ramos and Professor David Donoho.

Hitters' salary data. The hitters' salary dataset consists of 16 covariates: times at bat (x_1), hits (x_2), home runs (x_3), runs (x_4), runs batted in (x_5) and walks (x_6) in 1986, years in major leagues (x_7), times at bat (x_8), hits (x_9), home runs (x_{10}), runs (x_{11}), runs batted in (x_{12}) and walks (x_{13}) during their entire career up

to 1986, put-outs (x_{14}), assistances (x_{15}), errors (x_{16}), and a dependent variable: annual salary (y) in 1987. In our modeling, all covariates are standardized. The response (y) is logarithmically transformed (to natural base). It is well known that there is "aging effect" that makes the dependence of y on $\boldsymbol{x}$ nonlinear.

To begin with, we first fit a one-piece single index model for the data. The estimated single-index is denoted by $\boldsymbol{\beta}_0^\top \boldsymbol{x}$. Figure 1.4 plots y against $\boldsymbol{\beta}_0^\top \boldsymbol{x}$, suggesting that there are five outliers, all of which were also detected by Li et al (2000) and Xia et al (2002). After removing the outliers from the data set, in total we have 258 observations in our analysis. Denote by $B = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ the estimated effective dimension reduction (EDR) directions which is a by-product of the algorithm Step 1 - 4.

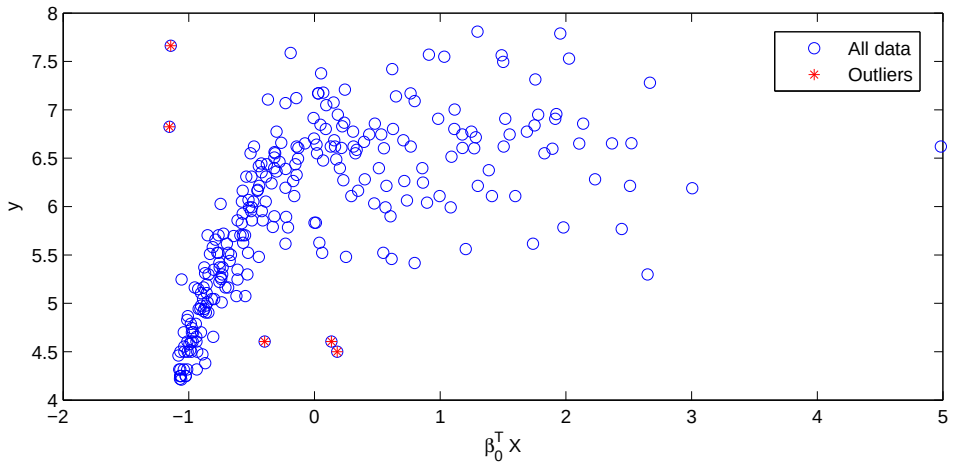

Figure 1.4 y plotted against $\boldsymbol{\beta}_0^\top \boldsymbol{x}$ for the hitters' salary data.

Table 1.4 BIC scores for the hitters' salary data (with the outliers removed)

No. of Regions	ASE	BIC score
1	0.1935	-1.5711
2	0.0960	-2.0634
3	0.1372	-1.3559
4	0.0674	1.4370
5	0.0433	2.8004

Applying the BIC with scores shown on Table 1.4, we select the numbers of regions as two, leading to the following model,

$$y = \begin{cases} g_1(\boldsymbol{\beta}_1^\top \boldsymbol{x}) + \varepsilon_1, & \text{for} \quad \boldsymbol{x} \in R_1, \\ g_2(\boldsymbol{\beta}_2^\top \boldsymbol{x}) + \varepsilon_2, & \text{for} \quad \boldsymbol{x} \in R_2, \end{cases}$$

where the estimated piecewise single-indices for the two regions are respectively

$$\begin{aligned} \hat{\boldsymbol{\beta}}_1 \;=\; & (\, -0.20,\ 0.20,\ 0.03,\ -0.05,\ 0.01,\ 0.04,\ 0.14(x_7),\ 0.39(x_8),\ 0.70(x_9), \\ & 0.19,\ -0.31,\ -0.23,\ 0.27,\ -0.02,\ -0.01,\ 0.04\ )^\top, \end{aligned}$$

$$\begin{aligned} \hat{\boldsymbol{\beta}}_2 \;=\; & (\, -0.07,\ 0.06,\ 0.05,\ -0.13,\ -0.04,\ 0.01,\ 0.26(x_7),\ 0.69(x_8),\ -0.63(x_9), \\ & 0.00,\ -0.02,\ -0.10,\ -0.14,\ -0.10,\ 0.02,\ -0.01\ )^\top. \end{aligned}$$

The fitting results are shown in Figure 1.5. The upper panels (a) and (b) plot y against the two estimated piecewise single indices $\hat{\boldsymbol{\beta}}_1^\top \boldsymbol{x}$ and $\hat{\boldsymbol{\beta}}_2^\top \boldsymbol{x}$, for the points

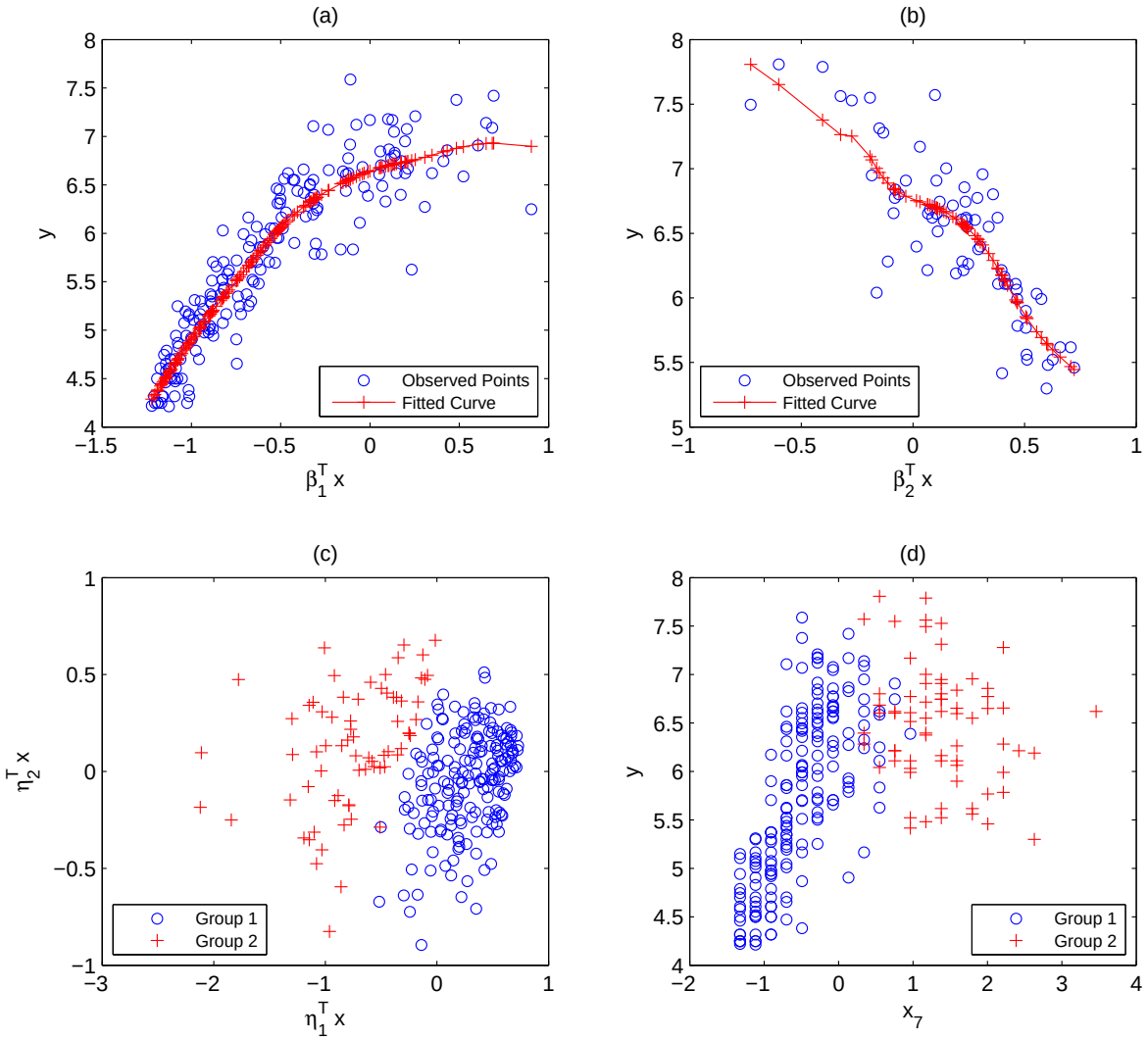

Figure 1.5 Fitting results for the hitter's salary data.

clustered in the respective groups. The left lower panel (c) plots the points of the two clustered groups on the effective dimension reduction space $(\boldsymbol{\eta}_1^\top \boldsymbol{x}, \boldsymbol{\eta}_2^\top \boldsymbol{x})$. Based on the panel (c), R_1 roughly corresponds to $\boldsymbol{\eta}_1^\top \boldsymbol{x} > 0$, and R_2 roughly corresponds to $\boldsymbol{\eta}_1^\top \boldsymbol{x} \leq 0$. An alternative perspective, looking at the partition of the whole space,

is provided by Figure 1.5(d) which plots y against x_7 (years in major league). The sign of x_7 is also a good indicator of which region an observation belongs to. Note that y increases as $\hat{\boldsymbol{\beta}}_1^\top \boldsymbol{x}$ increases in region R_1, while y is a decreasing function of $\hat{\boldsymbol{\beta}}_2^\top \boldsymbol{x}$ in the second region R_2. The same sign of coefficients on x_7 for the two piecewise single-index actually show the "aging effect" for hitters' salary. Namely, for small x_7 (junior hitters), i.e., in the first region, y increases as x_7 increases; for large x_7 (senior hitters), i.e., in the second region, x_7 is a negative factor for y. This aging effect was first noticed by Li et al (2000). If we judge the importance of a variable by the magnitude of its corresponding coefficient, we also observe that within each age group, x_7 is not the most influencing factor. Instead, x_8 and x_9 seem to have the greatest influences on players' salaries if we look at the two age groups separately. Specifically, the salaries of the junior group are positively correlated with the sum of x_8 and x_9, which can be viewed as a measure of a player's experience on the field; for the senior group, the salary increases as the difference $(x_9 - x_8)$ increases, which actually measures their hitting efficiencies on the field.

We also applied TGP-SIM to this data. Interestingly, TGP-SIM also splits the space based on the value of x_7. The maximum a posteriori (MAP) 3-node tree estimated by TGP-SIM is shown in Figure 1.6. The MAP 2-node tree is the same, but with lower branch pruned and $\log(p) = 261.343$. In addition, the data

are randomly partitioned 100 times into training/test sets of size 208/50. The in-sample and out-of-sample fitting errors are reported in Table 1.5. Although TGP-SIM gives slightly smaller error in the median, it suffers more from some extreme cases. As a result, the mean out-of-sample fitting error of our pSIM is lower than that of TGP-SIM by 17%.

Table 1.5 Simulation results of the hitters' salary data: mean of in-sample (IS) and out-of-sample (OS) prediction errors (ASE) from the 100 replications.

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
TGP-SIM	IS	0.0065	0.0163	0.0191	0.0189	0.0215	0.0356
	OS	0.0641	0.1084	0.1291	0.1664	0.1703	1.3847
pSIM	IS	0.0693	0.0831	0.0873	0.0887	0.0945	0.1242
	OS	0.0824	0.1176	0.1328	0.1417	0.1631	0.2532

LA Ozone data. The LA Ozone data consists of 330 observations on 10 variables: daily maximum 1-hour average ozone reading at Upland (y), 500mb pressure height (m) measured at Vandenberg AFB (x_1), wind speed (mph) at Los Angles International Airport(LAX) (x_2), humidity (%) at LAX (x_3), temperature measured at Sandburg (x_4), inversion base height (feet) at LAX (x_5), pressure gradient (mm Hg) from LAX to Daggett (x_6), inversion base temperature (°F) at LAX (x_7), visibility (miles) measured at LAX (x_8), day of the year (x_9). The goal

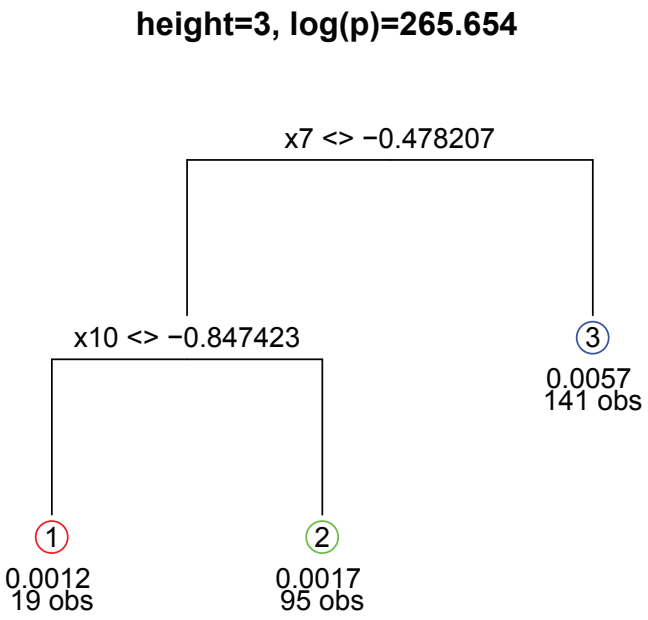

Figure 1.6 The maximum a posteriori (MAP) tree at height 3 estimated by TGP-SIM for the hitters' salary data.

is to explore the relationship between response value y and the covariates $X = (x_1, ..., x_9)$. To make the coefficients of each variable comparable, we standardize all covariates separately.

The BIC scores for $m = 1, ..., 5$ are shown in Table 1.6, suggesting $m = 2$. So similar to the previous example, we fit the data with the following model

$$y = \begin{cases} g_1(\boldsymbol{\beta}_1^\top \boldsymbol{x}) + \varepsilon_1, & \text{for} \quad \boldsymbol{x} \in R_1, \\ g_2(\boldsymbol{\beta}_2^\top \boldsymbol{x}) + \varepsilon_2, & \text{for} \quad \boldsymbol{x} \in R_2, \end{cases} \tag{1.11}$$

Table 1.6 BIC scores for the LA Ozone data

No. of Regions	ASE	BIC score
1	15.5506	2.7982
2	11.9941	2.6810
3	12.9890	3.7057
4	12.6943	4.8083
5	12.1982	6.0141

where the estimated piecewise single-indices for the two regions are respectively

$$\hat{\boldsymbol{\beta}}_1 = (-0.14,\ 0.06,\ -0.02,\ 0.20,\ -0.04,\ -0.27(x_6),\ 0.91,\ -0.09,\ -0.17)^\top,$$

$$\hat{\boldsymbol{\beta}}_2 = (0.27,\ -0.28,\ 0.35,\ 0.23,\ 0.26,\ 0.62(x_6),\ 0.35,\ -0.15,\ -0.28)^\top.$$

The estimated single-index link functions are shown in the upper two panels (a) and (b) in Figure 1.7, which plot y against the two estimated piecewise single indices $\hat{\boldsymbol{\beta}}_1^\top \boldsymbol{x}$ and $\hat{\boldsymbol{\beta}}_2^\top \boldsymbol{x}$.

Denote the estimated dimension reduction directions by $\boldsymbol{B} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$. The left lower panel (c) of Figure 1.7 plots the points of the two clustered groups on the effective dimension reduction space $(\boldsymbol{\eta}_1^\top \boldsymbol{x}, \boldsymbol{\eta}_2^\top \boldsymbol{x})$. The panel (c) suggests that $\boldsymbol{\eta}_2^\top \boldsymbol{x}$ is a good indicator for the two regions. As a comparison, TGP-SIM selects x_7 to split the space into two regions as shown in Figure 1.8. In fact, the regions of pSIM

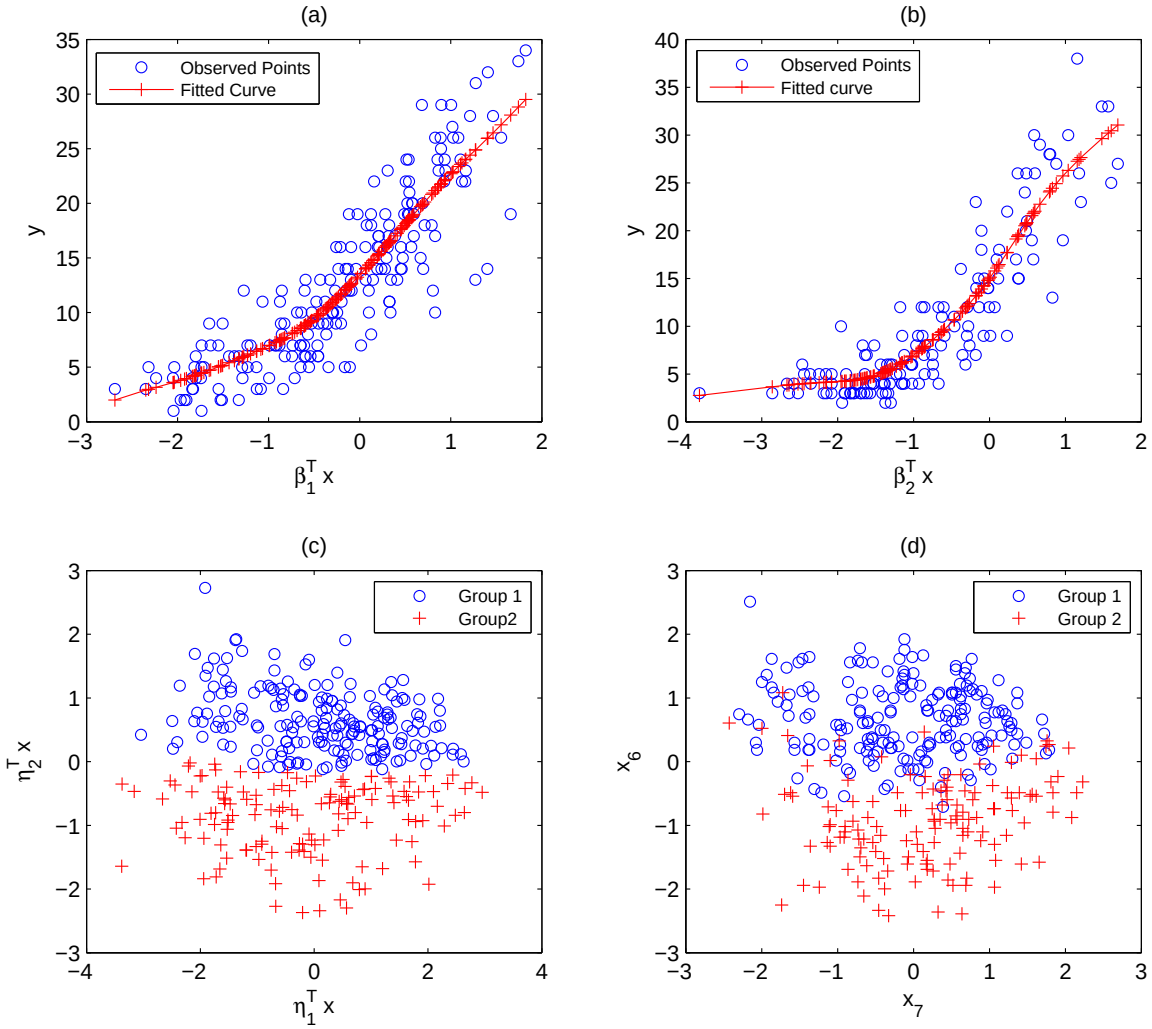

Figure 1.7 Fitting results for the LA ozone data.

can roughly be separated by the sign of x_6 as shown in Figure 1.7(d).

In addition, we notice an interesting "pressure gradient effect" based on x_6. Namely, as both link functions are increasing monotonic with their respective single-index, the sign of x_6 plays different roles in the two regions. For negative

x_6, y increases as x_6 increases; while for positive x_6, y decreases as x_6 increases. In other words, keeping other factors fixed, the ozone level y will attain its maximum value when the standardized x_6 is around 0.

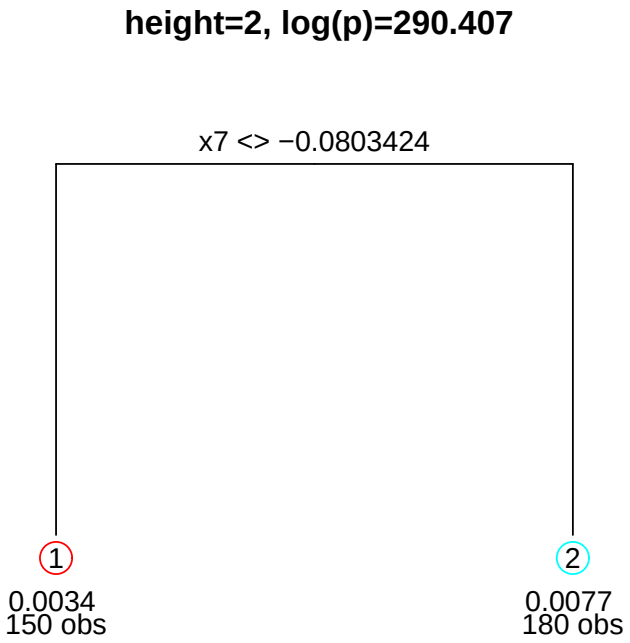

Figure 1.8 The maximum a posteriori (MAP) tree at height 2 estimated by TGP-SIM for the LA ozone data.

Similar to the previous example, we randomly partition the data 100 times into training/test sets of size 280/50 with the fitting results shown in Table 1.7. The mean out-of-sample fitting error of pSIM is lower than that of TGP-SIM by 12%.

Cars data. This real data analysis gives an example that TGP-SIM model

Table 1.7 Simulation results of the LA ozone data: mean of in-sample (IS) and out-of-sample (OS) prediction errors (ASE) from the 100 replications.

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
TGP-SIM	IS	1.849	5.619	6.819	6.962	8.530	11.390
	OS	8.246	13.184	15.739	17.391	19.887	68.109
pSIM	IS	9.650	11.389	11.894	11.807	12.277	14.428
	OS	7.686	12.764	14.458	15.507	18.434	29.330

can give better out-of-sample prediction performances. We think this is common in real data applications since no method can dominate the others in all data sets collected from the real world.

The original Cars data consists of 406 observations on 7 variables: miles per gallon (y), number of cylinders (x_1), engine displacement (x_2), horsepower(x_3), vehicle weight(x_4), time to accelerate from 0 to 60 miles per hour (x_5), model year(x_6), and origin of a car (1 for American, 2 for European and 3 for Japanese). There are 14 subjects having missing values in at least one variable, so we exclude them in our analysis leaving 392 observations. Li et al. (2000) has studied its piecewise property.

Since the last variable is a categorical variable, we define two dummy variables

x_7 and x_8 to account for the 3 scenarios of the origin of a car. Namely, let $x_7 = 1$ if a car is from America and 0 otherwise; $x_8 = 1$ if a car is from Europe and 0 otherwise. In this way, we have $(x_7, x_8) = (1, 0),\ (0, 1),\ (0, 0)$ corresponding to American cars, European cars and Japanese cars respectively. The main goal of our analysis is to explore the relationship between response value y and the covariates $\boldsymbol{x} = (x_1, ..., x_8)$. To make the coefficients of each variable comparable, we standardize all covariates separately.

Table 1.8 BIC scores for the cars data

No. of Regions	ASE	BIC score
1	7.7225	2.0944
2	6.5257	2.0510
3	6.2863	2.2392
4	6.2341	2.5531
5	6.1332	2.5868

The BIC scores for $m = 1, ..., 5$ are shown in Table 1.8, suggesting $m = 2$. So similar to the previous example, we fit the data with the following model

$$y = \begin{cases} g_1(\beta_1 \boldsymbol{x}) + \varepsilon_1, & \text{for} \quad \boldsymbol{x} \in R_1, \\ g_2(\beta_2 \boldsymbol{x}) + \varepsilon_2, & \text{for} \quad \boldsymbol{x} \in R_2, \end{cases} \tag{1.12}$$

where the estimated piecewise single-indices for the two regions are respectively

$$\hat{\beta}_1 = (\,0.90,\ -0.20,\ -0.10,\ -0.31,\ 0.02,\ 0.21,\ -0.05,\ 0.02)^\top,$$

$$\hat{\beta}_2 = (\,-0.02,\ -0.27,\ 0.39,\ 0.71,\ 0.04,\ -0.37,\ 0.35,\ -0.05)^\top.$$

In-sample fitting results are shown in the upper two panels (a) and (b) in Figure 1.9. The upper panels (a) and (b) plot y against the two estimated piecewise single indices $\hat{\beta}_1^\top \boldsymbol{x}$ and $\hat{\beta}_1^\top \boldsymbol{x}$. The lower panels (c) and (d) plot the points of the two clustered groups on the effective dimension reduction space $B^\top \boldsymbol{x} = (\eta_1^\top \boldsymbol{x}, \eta_2^\top \boldsymbol{x})$.

Denote the effective dimension reduction directions estimated in the algorithm by $B = (\eta_1, \eta_2)$. Panel (c) of Figure 1.9 plots the spread of each region on the effective dimension reduction space $B^\top \boldsymbol{x}$. It shows three isolated "clusters". Panel (d) further suggests that the three clusters correspond to three sets of different values of x_1 (number of cylinders). Namely, from bottom to top the first group consists of cases with $x_1 = 3, 4$; the second group cases with $x_1 = 5, 6$; and the third group cases with $x_1 = 8$. Note that the local gradients for the upper two clusters corresponding to $x_1 = 5, 6$ and 8 do not differ too much from each other. So the pSIM model puts them together into one group. As a result, R_1 in the pSIM model (1.12) corresponds to $x_1 = 3, 4$ and the R_2 corresponds to $x_1 = 5, 6, 8$. For group one, i.e., for cars with small number of cylinders (≤ 4), we do not see significant differences among the three origins of cars since the last two coefficients of the piecewise single index are very small. But for the other region R_2, we have

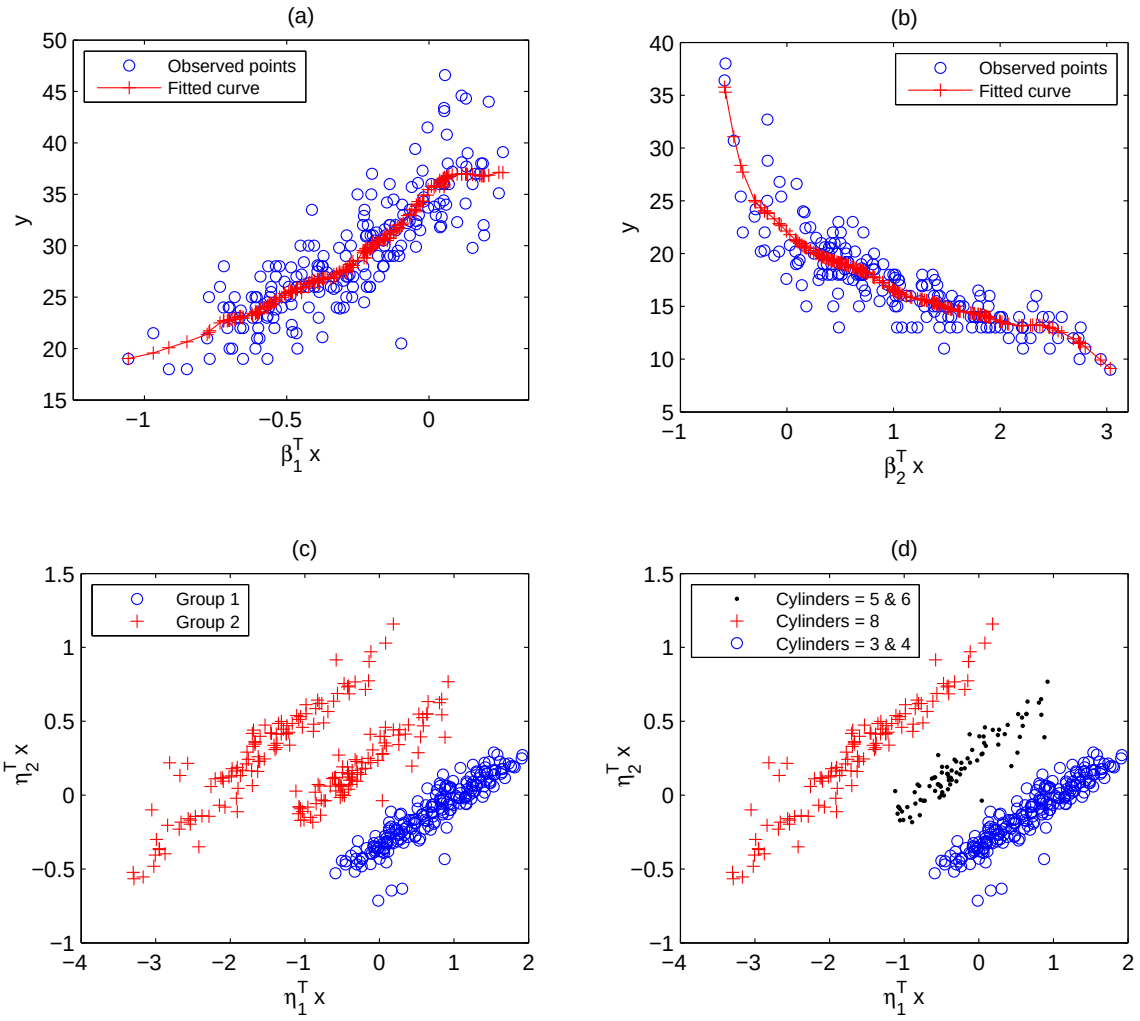

Figure 1.9 Fitting results for the cars data.

a significant positive coefficient on the dummy variable for American cars. Since in R_2 the response value y (miles per gallon) is a decreasing function of $\beta_2^\top \boldsymbol{x}$, we conclude that American cars with more cylinders (≥ 5) have lower values of miles per gallon as compared to the European and Japanese cars. Considering the data

was collected in the US, the conclusion is quite reasonable in that the foreign cars have to be more fuel-efficient to be competitive to the local cars.

It is interesting to point out that Li et al (2000) noticed the similar fact about the three cylinder groups by looking at the fitted residuals of a linear regression model based on a different estimating procedure. The theory of Li et al (2000) also suggested to partitioning the space into two regions. Coincidentally, the TGP-SIM model also identifies two classes based on the value of x_3 (horse power); see Figure 1.10.

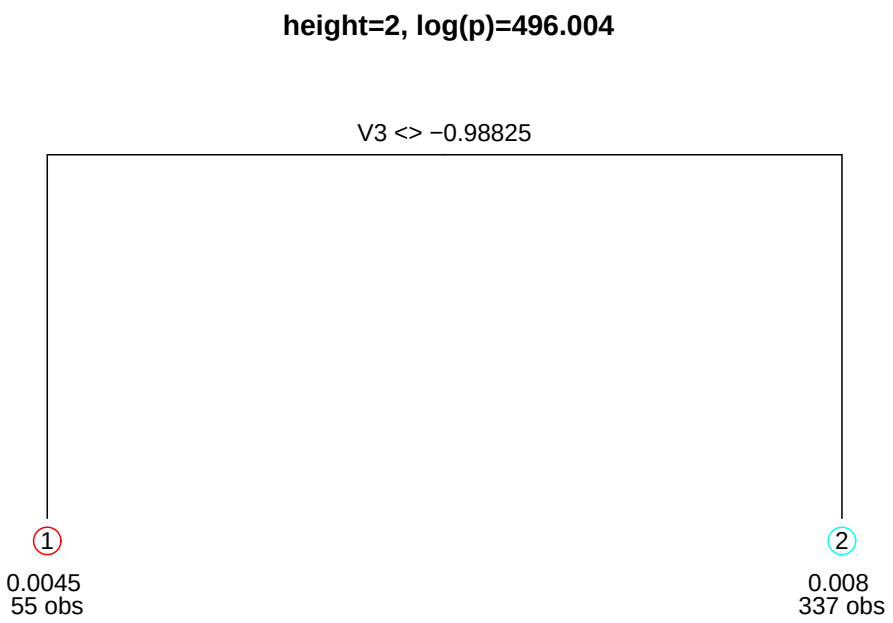

Figure 1.10 The tree structures estimated by the TGP-SIM model for the cars data.

In addition, the data are randomly partitioned 100 times into training/test sets of size 342/50. The in-sample and out-of-sample fitting errors are reported in Table 1.9. The mean out-of-sample fitting error of TGP-SIM is lower than that of pSIM by 8.4%.

Table 1.9 Simulation results of the cars data: mean of in-sample (IS) and out-of-sample (OS) prediction errors (ASE) from the 100 replications.

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
TGP-SIM	IS	1.422	2.740	3.384	3.390	4.041	6.163
	OS	3.015	6.360	8.320	8.923	10.697	29.742
pSIM	IS	5.146	6.334	7.487	7.521	8.488	9.035
	OS	4.346	7.138	9.082	9.669	11.571	22.614

1.5 Asymptotic Analysis

In this section we consider the statistical theory of our proposed method in Sections 1.2. Some of the proofs are given in the appendix. Suppose the sample $\{(\boldsymbol{x}_i, y_i),\ i = 1, ..., n\}$ is generated by model (1.7) and let I_g be the index set for the observations in R_g. For any matrix $\boldsymbol{A}$, let $\|\boldsymbol{A}\|$ denote its largest singular value, which is same as the Euclidean norm if $\boldsymbol{A}$ is a vector.

Similar to Lu (1996) and Xia (2007), we need the following assumptions for (1.7) to prove our theoretical results. Let $\mu_{\boldsymbol{\beta}_i}(u) = E(\boldsymbol{x}|\boldsymbol{\beta}_i^\top \boldsymbol{x} = u,\ \boldsymbol{x} \in R_i)$ and $w_{\boldsymbol{\beta}_i}(u) = E(\boldsymbol{x}\boldsymbol{x}^\top|\boldsymbol{\beta}_i^\top \boldsymbol{x} = u,\ \boldsymbol{x} \in R_i)$. We write $B(\boldsymbol{x}; h) = \{\boldsymbol{x}' \in \mathbb{R}^p :\ \|\boldsymbol{x}' - \boldsymbol{x}\| \le h\}$ and Vol(h) as the volume of $B(\boldsymbol{x}; h)$.

(A1) [Design of $\boldsymbol{x}$] The density function $f(\boldsymbol{x})$ of $\boldsymbol{x}$ has a compact support and bounded second order derivatives on $\mathbb{R}^p$, and there are positive constants $0 < c_f \le C_f$ such that $c_f/\text{Vol}(1) \le f(\boldsymbol{x}) \le C_f/\text{Vol}(1)$; $E|\boldsymbol{x}|^r < \infty$ for some $r > 8$; functions $\mu_{\boldsymbol{\beta}_i}(u)$ and $w_{\boldsymbol{\beta}_i}(u)$ have bounded derivatives with respect to u and $\hat{\boldsymbol{\beta}}_i$ for $\hat{\boldsymbol{\beta}}_i \in \{\hat{\boldsymbol{\beta}}_i; \|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\| \le \delta\}$ for some $\delta > 0$.

(A2) [Density function] The conditional density functions $f_{y|\boldsymbol{x}}(y|\boldsymbol{x})$ and $f_{y|\{\hat{\boldsymbol{\beta}}_i^\top \boldsymbol{x},\ \boldsymbol{x} \in R_i\}}(y|u)$ have bounded fourth order derivatives with respect to x, u and $\hat{\boldsymbol{\beta}}_i \in \{\hat{\boldsymbol{\beta}}_i; \|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\| \le \delta\}$ for some $\delta > 0$.

(A3) [Boundaries between regions] For any region R_g considered in model (1.7), its boundary ∂R_g is a continuously derivative function of $(\boldsymbol{\beta}_1^\top \boldsymbol{x}, ..., \boldsymbol{\beta}_m^\top \boldsymbol{x})$ a.s. and has a measure 0 in space $\mathbb{R}^p$.

(A4) [Kernel function] The kernel $\text{K}(\cdot)$ is a spherically symmetric density function, i.e., there exists a univariate function $k(\cdot)$ such that $\text{K}(\boldsymbol{z}) = k(\|\boldsymbol{z}\|)$ for all $\boldsymbol{z} \in \mathbb{R}^d$, where d is the effective dimension for $\text{K}(\cdot)$.

(A5) [Regression functions] The regression functions $\phi_i(\boldsymbol{\beta}_i^\top \boldsymbol{x})$ have bounded second order derivatives within its own region R_i and $\phi_i'(\boldsymbol{\beta}_i^\top \boldsymbol{x}) \neq 0$ almost

surely in R_i.

The accuracy of the estimated gradient direction $\tilde{\boldsymbol{b}}_i$ is summarized in the following lemma.

Lemma 1.5.1. Under model (1.7) and assumptions (A1) - (A5), we have

$$\max_{i\in I_g} \|\tilde{\boldsymbol{b}}_i\tilde{\boldsymbol{b}}_i^\top - \boldsymbol{\beta}_g\boldsymbol{\beta}_g^\top\| = O_P\Big\{h_0^2 + \Big(\frac{\log(n)}{nh_0^{p+2}}\Big)^{1/2}\Big\}, \tag{1.13}$$

for $g = 1, ..., m$.

Lemma 1.5.1 is a direct application of the Theorem 2 of Lu (1996) and large deviation theory (Chapter 8, De la Pena, Lai and Shao, 2009). Let $\tilde{k} = \min_{g_1,g_2}(\|\boldsymbol{\beta}_{g_1}\boldsymbol{\beta}_{g_1}^\top - \boldsymbol{\beta}_{g_2}\boldsymbol{\beta}_{g_2}^\top\|)/6$, we have

$$Pr\left(\max_{i\in I_g} \|\tilde{\boldsymbol{b}}_i\tilde{\boldsymbol{b}}_i^\top - \boldsymbol{\beta}_g\boldsymbol{\beta}_g^\top\| \geq \frac{1}{6}\min_{g_1,g_2}(\|\boldsymbol{\beta}_{g_1}\boldsymbol{\beta}_{g_1}^\top - \boldsymbol{\beta}_{g_2}\boldsymbol{\beta}_{g_2}^\top\|)\right) \leq O\Big(\exp\Big\{-\frac{\tilde{k}^2}{2}\Big(\frac{n}{\log(n)}\Big)^{4/(p+6)}\Big\}\Big), \tag{1.14}$$

where we implicitly assume that the h_0 used in Steps 1 - 3 is the asymptotically optimal bandwidth in the sense of minimizing mean squared error. Let $\boldsymbol{\beta}_g^{(i)}$ be the true regional single index corresponding to the ith observation and

$$M_i = \mathbf{1}\Big\{\|\tilde{\boldsymbol{b}}_i\tilde{\boldsymbol{b}}_i^\top - \boldsymbol{\beta}_g^{(i)}\boldsymbol{\beta}_g^{(i)\top}\| \geq \tilde{k}\Big\}.$$

Define $n_e = \sum_{i=1}^n M_i$ which is the number of observations that are not estimated well.

Lemma 1.5.2. Under the same conditions as Lemma 1.5.1, we have

$$\frac{n_e}{n} \leq O_P\Big(\exp\Big\{ -\frac{\tilde{k}^2}{2}\Big(\frac{n}{\log(n)}\Big)^{4/(p+6)}\Big\}\Big), \tag{1.15}$$

where $\tilde{k} = \min_{g_1,g_2}(\|\boldsymbol{\beta}_{g_1}\boldsymbol{\beta}_{g_1}^\top - \boldsymbol{\beta}_{g_2}\boldsymbol{\beta}_{g_2}^\top\|)/6$ is a nonzero constant.

Lemma 1.5.2 implies that with probability tending to 1, the estimated gradient directions will gather around their true values with a "safe" distance from those in different regions and that the proportion of badly estimated gradient directions decreases exponentially as n increases. Equation (1.15) gives an upper bound for the proportion of the points that are mis-clustered in a single region, say R_g. More precisely, for any cluster group whose main part is in R_g, the group would only contain an exponentially dampening proportion of points that do not belong to R_g. We have the following result for the OPG estimator $\hat{\boldsymbol{\beta}}_g$.

Theorem 1.5.3. Under model (1.7) and assumptions (A1) - (A5), if $h_0 \to 0$ and $n_g \to \infty$, then

$$\|\hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\beta}}_g^\top - \boldsymbol{\beta}_g\boldsymbol{\beta}_g^\top\| = O_P\big(h_0^2 + {n_g}^{-1/2}\big),\ g = 1, ..., m, \tag{1.16}$$

where n_g is the sample size in group g, i.e., $n_g = \#I_g$.

Theorem 1.5.3 states that with the refined weights based on the 1-dimensional space $\hat{\boldsymbol{\beta}}_g^\top \boldsymbol{x}$, we can achieve optimal parametric convergence in gradients estimations

with the OPG method. Next, we give a result on the estimation efficiency of the local linear smoother in (1.10).

Corollary 1.5.4. Suppose model (1.7) and assumptions (A1) - (A5) hold. If $h_0 \to 0$, $n_g/n > c_g > 0$, $H_g \to 0$ and $n_g H_g \to \infty$ as $n \to \infty$, and $\left(h_0^2 + \sqrt{n_g}^{-1}\right)/H_g \to 0$, we have

$$\hat{\phi}_g(\hat{\boldsymbol{\beta}}_g^\top \boldsymbol{x}_i) - \phi_g(\boldsymbol{\beta}_g^\top \boldsymbol{x}_i) = O_P\{H_g^2 + (n_g H_g)^{-1/2}\}, \tag{1.17}$$

where n_g is the sample size for group g and H_g is the bandwidth used in (1.10).

The convergence rate implied by (1.17) is the typical rate in nonparametric regression analysis. Finally we present a theorem concerning the consistency of the BIC proposed in subsection 1.2.2.

Theorem 1.5.5. Under the same conditions as Corollary 1.5.4 and assuming that $H_g = O(n^{-1/5})$ for all g, we have

$$\hat{m}_{\mathrm{BIC}} \to m_0 \qquad \text{in probability.}$$

The $\hat{\phi}_{m,g}(\hat{\boldsymbol{\beta}}_{m,g}^\top \boldsymbol{x}_i)$ involved in the estimation of BIC can be estimated either by the Nadaraya-Watson estimator or the local linear kernel estimator, both of which lead to a consistent estimator $\hat{m}_{\mathrm{BIC}}$.

1.6 Proofs

In this chapter, we consider a piecewise single-index model (pSIM) to perform nonparametric regression in a multidimensional space. Our model can be written as

$$y = \begin{cases} \phi_1(\boldsymbol{\beta}_1^\top \boldsymbol{x}) + \varepsilon_1, & \text{if} \quad \boldsymbol{x} \in R_1, \\ \quad \dots & \quad \dots \\ \phi_m(\boldsymbol{\beta}_m^\top \boldsymbol{x}) + \varepsilon_m, & \text{if} \quad \boldsymbol{x} \in R_m, \end{cases} \tag{1.18}$$

where $\boldsymbol{\beta}_g, g = 1, ..., m$, are $p \times 1$ vectors, $\phi_g, g = 1, ..., m$, are smooth functions on $\mathbb{R}$, $E(\varepsilon_g|\boldsymbol{x}) = 0$, $\cup_{g=1}^m R_g = \mathbb{R}^p$ and $R_i \cap R_j = \emptyset$ for any $i \neq j$.

Similar to Lu (1996) and Xia (2007), we need the following assumptions for (1.18) to prove our theoretical results. Let $\mu_{\boldsymbol{\beta}_i}(u) = E(\boldsymbol{x}|\boldsymbol{\beta}_i^\top \boldsymbol{x} = u,\ \boldsymbol{x} \in R_i)$ and $w_{\boldsymbol{\beta}_i}(u) = E(\boldsymbol{x}\boldsymbol{x}^\top|\boldsymbol{\beta}_i^\top \boldsymbol{x} = u,\ \boldsymbol{x} \in R_i)$. We write $B(\boldsymbol{x}; h) = \{\boldsymbol{x}' \in \mathbb{R}^p :\ \|\boldsymbol{x}' - \boldsymbol{x}\| \leq h\}$ and $\text{Vol}(h)$ as the volume of $B(\boldsymbol{x}; h)$.

(A1) [Design of $\boldsymbol{x}$] The density function $f(\boldsymbol{x})$ of $\boldsymbol{x}$ has a compact support and bounded second order derivatives on $\mathbb{R}^p$, and there are positive constants $0 < c_f \leq C_f$ such that $c_f/\text{Vol}(1) \leq f(\boldsymbol{x}) \leq C_f/\text{Vol}(1)$; $E|\boldsymbol{x}|^r < \infty$ for some $r > 8$; functions $\mu_{\boldsymbol{\beta}_i}(u)$ and $w_{\boldsymbol{\beta}_i}(u)$ have bounded derivatives with respect to u and $\hat{\boldsymbol{\beta}}_i$ for $\hat{\boldsymbol{\beta}}_i \in \{\hat{\boldsymbol{\beta}}_i; \|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\| \leq \delta\}$ for some $\delta > 0$.

(A2) [Density function] The conditional density functions $f_{y|\boldsymbol{x}}(y|\boldsymbol{x})$ and $f_{y|\{\hat{\boldsymbol{\beta}}_i^\top \boldsymbol{x},\ \boldsymbol{x}\in R_i\}}(y|u)$ have bounded fourth order derivatives with respect to x, u and $\hat{\boldsymbol{\beta}}_i \in \{\hat{\boldsymbol{\beta}}_i; \|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\| \leq \delta\}$ for some $\delta > 0$.

(A3) [Boundaries between regions] For any region R_g considered in model (1.7), its boundary ∂R_g is a continuously derivative function of $(\boldsymbol{\beta}_1^\top \boldsymbol{x}, ..., \boldsymbol{\beta}_m^\top \boldsymbol{x})$ a.s. and has a measure 0 in space $\mathbb{R}^p$.

(A4) [Kernel function] The kernel $\mathrm{K}(\cdot)$ is a spherically symmetric density function, i.e., there exists a univariate function $k(\cdot)$ such that $\mathrm{K}(\boldsymbol{z}) = k(\|\boldsymbol{z}\|)$ for all $\boldsymbol{z} \in \mathbb{R}^d$, where d is the effective dimension for $\mathrm{K}(\cdot)$.

(A5) [Regression functions] The regression functions $\phi_i(\boldsymbol{\beta}_i^\top \boldsymbol{x})$ have bounded second order derivatives within its own region R_i and $\phi_i'(\boldsymbol{\beta}_i^\top \boldsymbol{x}) \neq 0$ almost surely in R_i.

Proof of Theorem 1.5.3: By Lemma 5.2, we have that the probability of point $\boldsymbol{x}_i$ being misclassified diminishes exponentially to zero, so the misclassifications are negligible in the asymptotic sense as compared to the parametric convergence rate to be shown in this lemma. For ease of exposition, we assume no misclassification exists. Consider the gth region with piecewise single index $\boldsymbol{\beta}_g$. Let

$$\boldsymbol{\varepsilon}_i = \hat{\boldsymbol{b}}_i - \phi_g'(\boldsymbol{\beta}_g^\top \boldsymbol{x}_i)\boldsymbol{\beta}_g,$$

where $\phi_g'(\boldsymbol{\beta}_g^\top \boldsymbol{x})$ is the first derivative of $\phi_g(\boldsymbol{\beta}_g^\top \boldsymbol{x})$ and $\boldsymbol{\varepsilon}_i$ is the estimation error studied extensively in nonparametric literatures; see Fan and Gijbels(1996). We

have the OPG matrix for the gth matrix

$$\hat{\boldsymbol{\Sigma}}_g = \Big(n_g^{-1}\sum_{i\in I_g}\{\phi_g'(\boldsymbol{\beta}_g^\top \boldsymbol{x}_i)\}^2\Big)\boldsymbol{\beta}_g\boldsymbol{\beta}_g^\top + \mathscr{E}_g,$$

where

$$\mathscr{E}_g = \Big(n_g^{-1}\sum_{i\in I_g}\phi_g'(\boldsymbol{\beta}_g^\top \boldsymbol{x}_i)\boldsymbol{\varepsilon}_i\Big)\boldsymbol{\beta}_g^\top + \boldsymbol{\beta}_g\Big(n_g^{-1}\sum_{i\in I_g}\phi_g'(\boldsymbol{\beta}_g^\top \boldsymbol{x}_i)\boldsymbol{\varepsilon}_i\Big)^\top + n_g^{-1}\sum_{i\in I_g}\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\top.$$

Since eigenvector $\hat{\boldsymbol{\beta}}_g$ corresponds to the largest eigenvalue of $\hat{\boldsymbol{\Sigma}}_g$, it follows from spectral analysis of random matrix that

$$\hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\beta}}_g^\top - \boldsymbol{\beta}_g\boldsymbol{\beta}_g^\top = O(\|\mathscr{E}_g\|), \tag{1.19}$$

which implies that it is sufficient to study the asymptotic behavior of $\mathscr{E}_g$, or equivalently, the asymptotic behavior of $n_g^{-1}\sum_{i\in I_g}\phi_g'(\boldsymbol{\beta}_g^\top \boldsymbol{x}_i)\boldsymbol{\varepsilon}_i$. So it is equivalent to prove the following lemma:

Lemma 1.6.1. Under model (1.7) and (A1) - (A5), if $n_g \to \infty$ and $h_{g0} \to 0$ we have

$$\|\mathscr{E}_i\| = O_P\left(h_0^2 + \frac{1}{\sqrt{n_0}}\right).$$

Suppose $\hat{\boldsymbol{a}}_i$ and $\hat{\boldsymbol{b}}_i$ are the solution to

$$(\hat{a}_i, \hat{\boldsymbol{b}}_i) = \underset{a,\boldsymbol{b}}{\operatorname{argmin}}\sum_{j=1}^{n}\{y_{i_k} - a - \boldsymbol{b}^\top(\boldsymbol{x}_i - \boldsymbol{x}_j)\}^2 w_{i,j}, \tag{1.20}$$

where $w_{i,j}$ is a symmetric weight function of the form $h_i^{-p}\mathrm{K}\{h_i^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j)\}$ in which h_i is the bandwidth and $\mathrm{K}(\cdot)$ is the kernel function.

Define

$$\begin{aligned}\boldsymbol{\psi}_i &= (\phi(\boldsymbol{\beta}^\top\boldsymbol{x}_i), \phi'(\boldsymbol{\beta}^\top\boldsymbol{x}_i)\boldsymbol{\beta}_g^\top)^\top,\\ \hat{\boldsymbol{\psi}}_i &= (\hat{\boldsymbol{a}}_i, \hat{\boldsymbol{b}}_i^\top)^\top.\end{aligned}$$

We denote $\boldsymbol{Y}_g = (y_{g,1}, ..., y_{g,n_g})^\top$ which is the vector of the response values of the gth region, and the corresponding sub-sample is denoted as ${}_g = \{\boldsymbol{x}_{g,1}, ..., \boldsymbol{x}_{g,n_g}\}$.

$$\boldsymbol{W}_g^{(i)} = diag\Big\{w_{1,i}^{(g)}, ..., w_{n_g,i}^{(g)}\Big\},$$

and

$$\mathbf{x}_g^{(i)} = \begin{pmatrix} 1 & (\boldsymbol{x}_{g,1}-\boldsymbol{x}_i)^\top \\ \vdots & \vdots \\ 1 & (\boldsymbol{x}_{g,n_g}-\boldsymbol{x}_i)^\top \end{pmatrix}.$$

If there are at least $(p+1)$ points with positive weights, $\mathbf{x}_g^{(i)\top}\boldsymbol{W}_g^{(i)}\mathbf{x}_g^{(i)}$ is invertible with probability one, and

$$\hat{\boldsymbol{\psi}}_i = (\mathbf{x}_g^{(i)\top}\boldsymbol{W}_g^{(i)}\mathbf{x}_g^{(i)})^{-1}\mathbf{x}_g^{(i)\top}\boldsymbol{W}_g^{(i)}\boldsymbol{Y}_g. \tag{1.21}$$

Note that

$$\boldsymbol{\varepsilon}_i = \hat{\boldsymbol{b}}_i - \phi'(\boldsymbol{\beta}_g^\top\boldsymbol{x}_i)\boldsymbol{\beta}_g$$

is a part of $\hat{\boldsymbol{\psi}}_i - \boldsymbol{\psi}_i$. we will derive the property of $\boldsymbol{\varepsilon}_i$ through that of $\hat{\boldsymbol{\psi}}_i - \boldsymbol{\psi}_i$.

It is suffice to show that

$$E[\mathscr{E}_g|\boldsymbol{x}_1,...,\boldsymbol{x}_n] = O_P(h_0^2), \tag{1.22}$$

$$Var[\mathscr{E}_g|\boldsymbol{x}_1,...,\boldsymbol{x}_n] = O_P\left(\frac{1}{n_g}\right). \tag{1.23}$$

The key idea in Steps 1 to 3 is the refinement of the kernel functions (from p-dim to 1-dim), which divides the proof into two parts:

(1) Asymptotic properties of using p-dim kernel;

(2) Asymptotic properties of using 1-dim kernel.

The first part follows directly from the Theorem 3 of Xia et al (2002), which claims that under some regularity assumptions and if $nh^p/\log(n) \to \infty$ and $h \to 0$ as $n \to \infty$, then

$$\|\hat{\boldsymbol{\beta}}_g^{(0)}\hat{\boldsymbol{\beta}}_g^{(0)T} - \boldsymbol{\beta}_g\boldsymbol{\beta}_g^\top\| = O_P(h_0^2 + \log(n_g)/(n_g(h_0)^{p+1})), \tag{1.24}$$

where h_0 is the bandwidth. For ease of notation, we omit the suffix g hereafter, e.g., replace $\hat{\boldsymbol{\beta}}_g^{(0)}$ by $\hat{\boldsymbol{\beta}}^{(0)}$, $\boldsymbol{\beta}_g$ by $\boldsymbol{\beta}$, $\mathbf{x}_g^{(i)}$ by $\mathbf{x}_i$ and $\boldsymbol{W}_g^{(i)}$ by $\boldsymbol{W}_i$. The $\hat{\boldsymbol{\beta}}^{(0)}$ serves as an initial estimation of $\boldsymbol{\beta}$ for the following iterations.

For $t \geq 1$, we have the updated weight functions

$$w_i^{(t)} = \frac{1}{h^{(t)}}\mathrm{K}(\hat{\boldsymbol{\beta}}^{(t-1)\top}(\boldsymbol{x}_i - \boldsymbol{x}_j)/h^{(t)}),$$

where $h^{(t)}$ is chosen by

$$h^{(t)} = \max\{h^{(t-1)}n^{-1/(2(p+6))}, c_0 n^{-1/4}\},$$

which ensures that $(\hat{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{\beta})/h^{(t)} = O_P(1)$ as will be shown later.

With the updated weight functions, the conditional bias and conditional covariance matrix of $\hat{\boldsymbol{\beta}}_i^{(t)}$ are given respectively by

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_i^{(t)}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n) - \boldsymbol{\beta}_i &= (\mathbf{x}_i^\top \boldsymbol{W}_i \mathbf{x}_i)^{-1} \mathbf{x}_i^\top \boldsymbol{W}_i (\Phi - \mathbf{x}_i \boldsymbol{\beta}) \\ &= \boldsymbol{S}_i^{-1} \boldsymbol{R}_i, \\ Cov(\hat{\boldsymbol{\beta}}_i^{(t)}, \hat{\boldsymbol{\beta}}_j^{(t)}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n) &= (\mathbf{x}_i^\top \boldsymbol{W}_i \mathbf{x}_i)^{-1} \mathbf{x}_i^\top \boldsymbol{W}_i \boldsymbol{V} \boldsymbol{W}_j \mathbf{x}_j (\mathbf{x}_j^\top \boldsymbol{W}_j \mathbf{x}_j)^{-1} \\ &= n^{-1} \boldsymbol{S}_i^{-1} \boldsymbol{C}_{ij} \boldsymbol{S}_j^{-1}, \end{aligned}$$

where $\Phi = (\phi(\boldsymbol{\beta}^\top \boldsymbol{x}_1), ..., \phi(\boldsymbol{\beta}^\top \boldsymbol{x}_n))^\top$, $\boldsymbol{V} = (v(\boldsymbol{\beta}^\top \boldsymbol{x}_1), ..., v(\boldsymbol{\beta}^\top \boldsymbol{x}_n))$,

$$\begin{aligned} \boldsymbol{S}_i &= n^{-1} \sum_{l=1}^{n} \tilde{\mathbf{x}}_{li} \tilde{\mathbf{x}}_{li}^\top \mathrm{K}_h(\hat{\boldsymbol{\beta}}^{(t-1)\top}(\boldsymbol{x}_i - \boldsymbol{x}_l)), \\ \boldsymbol{C}_{ij} &= n^{-1} \sum_{l=1}^{n} \tilde{\mathbf{x}}_{li} \tilde{\mathbf{x}}_{lj}^\top \mathrm{K}_h(\hat{\boldsymbol{\beta}}^{(t-1)\top}(\boldsymbol{x}_i - \boldsymbol{x}_l)) \mathrm{K}_h(\hat{\boldsymbol{\beta}}^{(t-1)\top}(\boldsymbol{x} y_j - \boldsymbol{x}_l)) v(\boldsymbol{\beta}^\top \boldsymbol{x}_l), \\ \boldsymbol{R}_i &= n^{-1} \sum_{l=1}^{n} \tilde{\mathbf{x}}_{li} (\phi(\boldsymbol{\beta}^\top \boldsymbol{x}_l) - \phi(\boldsymbol{\beta}^\top \boldsymbol{x}_i) - \phi'(\boldsymbol{\beta}^\top \boldsymbol{x}_i) \boldsymbol{\beta}^\top (\boldsymbol{x}_l - \boldsymbol{x}_i)) \mathrm{K}_h(\hat{\boldsymbol{\beta}}^{(t-1)\top}(\boldsymbol{x}_i - \boldsymbol{x}_l)), \end{aligned}$$

where

$$\tilde{\mathbf{x}}_{li} = \begin{pmatrix} 1 \\ \\ \boldsymbol{x}_l - \boldsymbol{x}_i \end{pmatrix}$$

Let $\boldsymbol{\Sigma}_{ij}^{(t)} = Cov(\hat{\boldsymbol{\beta}}_i^{(t)}, \hat{\boldsymbol{\beta}}_j^{(t)}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$, which can be written into 4 blocks:

$$\boldsymbol{\Sigma}_{ij}^{(t)} = \begin{pmatrix} \boldsymbol{\Sigma}_{ij,11}^{(t)} & \boldsymbol{\Sigma}_{ij,12}^{(t)} \\ \\ \boldsymbol{\Sigma}_{ij,21}^{(t)} & \boldsymbol{\Sigma}_{ij,22}^{(t)} \end{pmatrix}$$

where $\boldsymbol{\Sigma}_{ij,22}^{(t)}$ is a $p \times p$ matrix that is the covariance matrix of $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\varepsilon}_j$. Actually, the asymptotic behavior of $n^{-1} \sum_{j=1}^{n} y_j \boldsymbol{\varepsilon}_j$ is included in that of $n^{-1} \sum_{j=1}^{n} y_j \boldsymbol{\beta}_j$. So

we first study the asymptotic properties of $n^{-1}\sum_{j=1}^{n} y_j\boldsymbol{\beta}_j$ and then extract the information about $n^{-1}\sum_{j=1}^{n} y_j\boldsymbol{\varepsilon}_j$ from them.

To simplify and calculate $\boldsymbol{S}_i$, $\boldsymbol{C}_{ij}$ and $\boldsymbol{R}_i$, we first note that they all have forms ready to apply the law of large numbers. But it should be pointed out that the $\hat{\boldsymbol{\beta}}^{(t-1)}$ used in the expressions is estimated from the $\boldsymbol{x}_i$ and thus is correlated to $\tilde{\boldsymbol{x}}_{li}$, so the LLN is not directly applicable. However, we can evade this problem by viewing $\hat{\boldsymbol{\beta}}^{(t-1)}$ as a point in the neighborhood of $\boldsymbol{\beta}$ denoted as $\Omega(\boldsymbol{\beta}; h, t)$ which is determined by two deterministic parameters: h and t. If we can prove an uniform property on the neighborhood, then the case of $\hat{\boldsymbol{\beta}}^{(t-1)}$ will follow accordingly. To this end, we apply the tricks commonly used in the nonparametric proofs that we first pretend that the $\hat{\boldsymbol{\beta}}^{(t-1)}$ is estimated from some another set of observations independent with the one in hand and has the same distribution and then prove the required result is valid uniformly for any such $\hat{\boldsymbol{\beta}}^{(t-1)}$. In this way, by LLN, it can be easily seen that

$$
\begin{aligned}
\boldsymbol{S}_i &= \int \begin{pmatrix} 1 \\ \boldsymbol{x}-\boldsymbol{x}_i \end{pmatrix} (1, (\boldsymbol{x}-\boldsymbol{x}_i)^\top) \frac{1}{h}\mathrm{K}\left(\frac{1}{h}\hat{\boldsymbol{\beta}}^{(t-1)\top}(\boldsymbol{x}_i-\boldsymbol{x})\right) f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + O_P\left(\frac{1}{\sqrt{nh}}\right), \\
\boldsymbol{C}_{ij} &= \int \begin{pmatrix} 1 \\ \boldsymbol{x}-\boldsymbol{x}_i \end{pmatrix} (1, (\boldsymbol{x}-\boldsymbol{x}_j)^\top) \frac{1}{h}\mathrm{K}\left(\frac{1}{h}\hat{\boldsymbol{\beta}}^{(t-1)\top}(\boldsymbol{x}_i-\boldsymbol{x})\right) \frac{1}{h}\mathrm{K}\left(\frac{1}{h}\hat{\boldsymbol{\beta}}^{(t-1)\top}(\boldsymbol{x}_j-\boldsymbol{x})\right) \\
&\qquad \times f(\boldsymbol{x}) v(\boldsymbol{\beta}^\top\boldsymbol{x})\mathrm{d}\boldsymbol{x} \quad + \quad O_P\left(\frac{1}{\sqrt{nh}}\right),
\end{aligned}
$$

$$\begin{aligned}\boldsymbol{R}_i &= \int \begin{pmatrix} 1 \\ \boldsymbol{x}-\boldsymbol{x}_i \end{pmatrix} (\phi(\boldsymbol{\beta}^\top \boldsymbol{x}) - \phi(\boldsymbol{\beta}^\top \boldsymbol{x}_i) - \phi'(\boldsymbol{\beta}^\top \boldsymbol{x}_i)\boldsymbol{\beta}^\top(\boldsymbol{x}-\boldsymbol{x}_i)) \\ &\qquad \times \frac{1}{h}\mathrm{K}\left(\frac{1}{h}\hat{\boldsymbol{\beta}}^{(t-1)\top}(\boldsymbol{x}_i-\boldsymbol{x})\right) f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + O_P\left(\frac{1}{\sqrt{nh}}\right).\end{aligned}$$

Moreover,

$$\begin{aligned}E(n^{-1}\sum_{j=1}^n y_j(\hat{\boldsymbol{\beta}}_i^{(t)} - \boldsymbol{\beta})|\boldsymbol{x}_1,...,\boldsymbol{x}_n) &= n^{-1}\sum_{j=1}^n y_j \boldsymbol{S}_j^{-1}\boldsymbol{R}_j, \\ Cov(n^{-1}\sum_{j=1}^n y_j\hat{\boldsymbol{\beta}}_j^{(t)}|\boldsymbol{x}_1,...,\boldsymbol{x}_n) &= n^{-2}\sum_{i=1}^n\sum_{j=1}^n y_i y_j n^{-1}\boldsymbol{S}_i^{-1}\boldsymbol{C}_{ij}\boldsymbol{S}_j^{-1}.\end{aligned}$$

To prove (1.22) and (1.23), we first prove

$$E(n^{-1}\sum_{j=1}^n y_j(\hat{\boldsymbol{\beta}}_j^{(t)} - \boldsymbol{\beta}_j)|\boldsymbol{x}_1,...,\boldsymbol{x}_n) = O(h^2), \tag{1.25}$$

and

$$Cov(n^{-1}\sum_{j=1}^n y_j\hat{\boldsymbol{\beta}}_j^{(t)}|\boldsymbol{x}_1,...,\boldsymbol{x}_n) = O\left(\frac{1}{n}\right), \tag{1.26}$$

i.e.,

$$\boldsymbol{\Gamma} \stackrel{\text{def}}{=} n^{-2}\sum_{i=1}^n\sum_{j=1}^n y_i y_j \boldsymbol{S}_i^{-1}\boldsymbol{C}_{ij}\boldsymbol{S}_j^{-1} = O(1). \tag{1.27}$$

Consider $\boldsymbol{S}_i$ first. Without loss of generality, we assume $\hat{\boldsymbol{\beta}}_1^{(t-1)} \neq 0$. Let

$$A(\boldsymbol{x}_i) = \int \begin{pmatrix} 1 \\ \boldsymbol{x}-\boldsymbol{x}_i \end{pmatrix} (1, (\boldsymbol{x}-\boldsymbol{x}_i)^\top)\frac{1}{h}\mathrm{K}\left(\frac{1}{h}\hat{\boldsymbol{\beta}}^{(t-1)\top}(\boldsymbol{x}_i-\boldsymbol{x})\right) f(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

and

$$\boldsymbol{U} = \begin{pmatrix} \hat{\beta}_1^{(t-1)} & \hat{\beta}_2^{(t-1)} & \cdots & \hat{\beta}_p^{(t-1)} \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \boldsymbol{x} = \boldsymbol{J}_u \boldsymbol{x},$$

so $u_1 = \hat{\boldsymbol{\beta}}^{(t-1)\top}\boldsymbol{x}$, $\boldsymbol{x} = \boldsymbol{J}_b^{-1}\boldsymbol{U}$ and $\mathrm{d}\boldsymbol{x} = \mathrm{d}\boldsymbol{U}/\hat{\beta}_1^{(t-1)}$.

By changing the integral variable, we have

$$A(\boldsymbol{x}_i) = (\hat{\beta}_1^{(t-1)})^{-1} \int \begin{pmatrix} 1 \\ \boldsymbol{J}_b^{-1}\boldsymbol{U} - \boldsymbol{x}_i \end{pmatrix} (1, (\boldsymbol{J}_b^{-1}\boldsymbol{U} - \boldsymbol{x}_i)^\top) \frac{1}{h} \mathrm{K}\left(\frac{1}{h}(u_1 - \hat{\boldsymbol{\beta}}^{(t-1)\top}\boldsymbol{x}_i)\right) f(\boldsymbol{J}_b^{-1}\boldsymbol{U}) \mathrm{d}\boldsymbol{U}.$$

Let

$$\boldsymbol{W} = diag(h^{-1}, 1, ..., 1)\boldsymbol{U} - (h^{-1}\hat{\boldsymbol{\beta}}^{(t-1)\top}\boldsymbol{x}_i, 0, ..., 0)^\top = \boldsymbol{J}_h \boldsymbol{U} - \boldsymbol{\eta}_{bh},$$

i.e., $w_1 = h^{-1}(u_1 - \hat{\boldsymbol{\beta}}^{(t-1)\top}\boldsymbol{x}_i))$, $\boldsymbol{U} = \boldsymbol{J}_h^{-1}(\boldsymbol{W} + \boldsymbol{\eta}_{bh})$ and $\mathrm{d}\boldsymbol{U} = h\mathrm{d}\boldsymbol{W}$.

By changing the integral variable, we have

$$\begin{aligned} A(\boldsymbol{x}_i) \;=\; & (\hat{\beta}_1^{(t-1)})^{-1} \int \begin{pmatrix} 1 \\ \boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W} + \boldsymbol{\eta}_{bh}) - \boldsymbol{x}_i \end{pmatrix} \times \\ & (1, (\boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W} + \boldsymbol{\eta}_{bh}) - \boldsymbol{x}_i)^\top) \mathrm{K}(w_1) f(\boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W} + \boldsymbol{\eta}_{bh})) \mathrm{d}\boldsymbol{W}, \end{aligned}$$

where

$$\boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W} + \boldsymbol{\eta}_{bh}) = \left(\frac{\hat{\boldsymbol{\beta}}^{(t-1)\top}\boldsymbol{x}_i + hw_1}{\hat{\beta}_1^{(t-1)}} - \frac{\hat{\beta}_2^{(t-1)}}{\hat{\beta}_1^{(t-1)}} w_2 - \cdots - \frac{\hat{\beta}_p^{(t-1)}}{\hat{\beta}_1^{(t-1)}} w_p, w_2, ..., w_p\right)^\top.$$

Write $A(\boldsymbol{x}_i)$ as

$$A(\boldsymbol{x}_i) = \begin{pmatrix} A_{11}(\boldsymbol{x}_i) & A_{12}(\boldsymbol{x}_i) & A_{13}(\boldsymbol{x}_i) \\ A_{21}(\boldsymbol{x}_i) & A_{22}(\boldsymbol{x}_i) & A_{23}(\boldsymbol{x}_i) \\ A_{31}(\boldsymbol{x}_i) & A_{32}(\boldsymbol{x}_i) & A_{33}(\boldsymbol{x}_i) \end{pmatrix},$$

where

$$\begin{aligned} A_{11}(\boldsymbol{x}_i) &= (\hat{\beta}_1^{(t-1)})^{-1} \int \mathrm{K}(w_1) f(\boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W}+\boldsymbol{\eta}_{bh}))\mathrm{d}\boldsymbol{W}, \\ A_{22}(\boldsymbol{x}_i) &= (\hat{\beta}_1^{(t-1)})^{-1} \int (\boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W}+\boldsymbol{\eta}_{bh}) - x_{i1})^2 \mathrm{K}(w_1) f(\boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W}+\boldsymbol{\eta}_{bh}))\mathrm{d}\boldsymbol{W}, \\ A_{33}(\boldsymbol{x}_i) &= (\hat{\beta}_1^{(t-1)})^{-1} \int (\boldsymbol{W}-\boldsymbol{x}_i)_{-1}(\boldsymbol{W}-\boldsymbol{x}_i)_{-1}^{\top} \mathrm{K}(w_1) f(\boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W}+\boldsymbol{\eta}_{bh}))\mathrm{d}\boldsymbol{W}, \end{aligned}$$

with $(\boldsymbol{W}-\boldsymbol{x}_i)_{-1} = (w_2 - x_{2i}, w_3 - x_{3i}, ..., w_q - x_{qi})^{\top}$ and other items of $A(\boldsymbol{x}_i)$ can be defined accordingly.

Let $\hat{\boldsymbol{\beta}}_{-1}^{(t-1)} \stackrel{\text{def}}{=} (\hat{\beta}_2^{(t-1)}, ..., \hat{\beta}_p^{(t-1)})^{\top}$, we have

$$A_{22}(\boldsymbol{x}_i) = (\hat{\beta}_1^{(t-1)})^{-1} \int \left(\frac{hw_1 + \hat{\boldsymbol{\beta}}_{-1}^{(t-1)\top}(\boldsymbol{W}-\boldsymbol{x}_i)_{-1}}{\hat{\boldsymbol{\beta}}_1^{(t-1)}} \right)^2 \mathrm{K}(w_1) f(\boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W}+\boldsymbol{\eta}_{bh}))\mathrm{d}\boldsymbol{W}.$$

Based on the values of $\hat{\boldsymbol{\beta}}_{-1}^{(t-1)}$, we have two possible scenarios:

(1) $\hat{\boldsymbol{\beta}}_{-1}^{(t-1)} \neq \mathbf{0}_{(p-1)\times 1}$.

(2) $\hat{\boldsymbol{\beta}}_{-1}^{(t-1)} = \mathbf{0}_{(p-1)\times 1}$,

For scenario 1, since the diagonal items of A_{11}, A_{22}, A_{33} are all integrals of positive functions, there exists a positive number $s(\boldsymbol{x}_i)$, which is a function of $\boldsymbol{x}_i$, such that

$$\min\{A_{11}, A_{22}, \mathrm{diag}(A_{33})\} > s(\boldsymbol{x}_i) > s_0 > 0. \tag{1.28}$$

It is known that $A(\boldsymbol{x}_i)$ is invertible with probability 1, which together with (1.28) imply that $\|A(\boldsymbol{x}_i)\|$ is $O(1)$ bounded below and so all entries in $A(\boldsymbol{x}_i)^{-1}$ is $O(1)$ bounded above. Since $A(\boldsymbol{x}_i) \in C^2(\boldsymbol{x}_i)$, it follows that $A(\boldsymbol{x}_i)^{-1} \in C^2(\boldsymbol{x}_i)$ too. Then we have

$$\begin{aligned}\boldsymbol{\Gamma} &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j (A(\boldsymbol{x}_i)^{-1})^{-1}\boldsymbol{C}_{ij}(A(\boldsymbol{x}_j)^{-1}) + O_P\left(\frac{1}{\sqrt{nh}}\right) \\ &= \int \boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_2 f(\boldsymbol{x})v(\boldsymbol{\beta}^\top\boldsymbol{x})\mathrm{d}\boldsymbol{x} + O_P\left(\frac{1}{\sqrt{nh}}\right),\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{\Gamma}_1 &= \int A(\boldsymbol{x}_i)^{-1}(1,(\boldsymbol{x}-\boldsymbol{x}_i)^\top)^\top \frac{1}{h}\mathrm{K}\left(\frac{1}{h}\hat{\boldsymbol{\beta}}^{(t-1)\top}(\boldsymbol{x}_i-\boldsymbol{x})\right) f(\boldsymbol{x}_i)\phi(\boldsymbol{\beta}^T\boldsymbol{x}_i)\mathrm{d}\boldsymbol{x}_i, \\ \boldsymbol{\Gamma}_2 &= \int (1,(\boldsymbol{x}-\boldsymbol{x}_j)^\top)A(\boldsymbol{x}_j)^{-1}\frac{1}{h}\mathrm{K}\left(\frac{1}{h}\hat{\boldsymbol{\beta}}^{(t-1)\top}(\boldsymbol{x}_j-\boldsymbol{x})\right) f(\boldsymbol{x}_j)\phi(\boldsymbol{\beta}^T\boldsymbol{x}_j)\mathrm{d}\boldsymbol{x}_j.\end{aligned}$$

By changing the integral variable as before, it can be easily shown that $\boldsymbol{\Gamma}_1 = O(1)$ and $\boldsymbol{\Gamma}_2 = O(1)$, which implies that $\boldsymbol{\Gamma} = O(1)$ and as such (1.26) is valid.

As to the bias term (1.25), we can simply prove for each i, $\boldsymbol{S}_i^{-1}\boldsymbol{R}_i = O(h^2)$. Note that $(\hat{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{\beta})/h^{(t)} = O_P(1)$ and

$$\begin{aligned}\mathrm{K}\left(\frac{1}{h}\hat{\boldsymbol{\beta}}^{(t-1)\top}(\boldsymbol{x}_i-\boldsymbol{x})\right) &= \mathrm{K}\left(\frac{1}{h}\boldsymbol{\beta}^T(\boldsymbol{x}_i-\boldsymbol{x})\right) \\ &\quad +\mathrm{K}'\left(\frac{1}{h}\boldsymbol{\beta}^T(\boldsymbol{x}_i-\boldsymbol{x})\right)\frac{1}{h}(\hat{\boldsymbol{\beta}}^{(t-1)}-\boldsymbol{\beta})^T(\boldsymbol{x}_i-\boldsymbol{x}) \\ &\quad +O\left(\frac{1}{h}(\hat{\boldsymbol{\beta}}^{(t-1)}-\boldsymbol{\beta})^T(\boldsymbol{x}_i-\boldsymbol{x})\right)^2,\end{aligned}$$

It can be shown that

$$\boldsymbol{S}_i^{-1}\boldsymbol{R}_i = O(h^2) + O(h(\hat{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{\beta})) = O(h^2),$$

and so

$$E(n^{-1}\sum_{j=1}^{n} y_j(\hat{\boldsymbol{\beta}}_i^{(t)} - \boldsymbol{\beta})|\boldsymbol{x}_1, ..., \boldsymbol{x}_n) = n^{-1}\sum_{j=1}^{n} y_j \boldsymbol{S}_j^{-1}\boldsymbol{R}_j = O(h^2).$$

For scenario 2, A_{11} and A_{33} are similar to scenario 1, but A_{22} is now simplified to

$$A_{22}(\boldsymbol{x}_i) = (\hat{\beta}_1^{(t-1)})^{-1}\int\left(\frac{hw_1}{\hat{\beta}_1^{(t-1)}}\right)^2 \mathrm{K}(w_1) f(\boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W}+\boldsymbol{\eta}_{bh}))\mathrm{d}\boldsymbol{W} = O(h^2)a_{22}(\boldsymbol{x}_i),$$

where $a_{22}(\boldsymbol{x}_i)$ is a positive function in $C^2(\boldsymbol{x}_i)$. Note that

$$\int \mathrm{K}(w_1) w_1 dw_1 = 0,$$

and

$$\begin{aligned} f(\boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W}+\boldsymbol{\eta}_{bh})) &= f\left(\frac{w_i h}{\hat{\beta}_1^{(t-1)}} + \frac{\hat{\boldsymbol{\beta}}_{-1}^{(t-1)\top}(\boldsymbol{W}-\boldsymbol{x}_i)_{-1}}{\hat{\beta}_1^{(t-1)}} + x_{i1}, w_2, ..., w_p\right) \\ &= f\left(\frac{\hat{\boldsymbol{\beta}}_{-1}^{(t-1)\top}(\boldsymbol{W}-\boldsymbol{x}_i)_{-1}}{\hat{\beta}_1^{(t-1)}} + x_{i1}, w_2, ..., w_p\right) + \frac{\partial f}{\partial w_1}(\cdot)\frac{w_i h}{\hat{\beta}_1^{(t-1)}} + O(h^2), \end{aligned}$$

where the first item on the right hand side of the second equation dose not include w_1. It can be easily shown that

$$A(\boldsymbol{x}_i) = \begin{pmatrix} a_{11}(\boldsymbol{x}_i) & a_{12}(\boldsymbol{x}_i)O(h^2) & a_{13}(\boldsymbol{x}_i)O(h^2) \\ a_{21}(\boldsymbol{x}_i)O(h^2) & a_{22}(\boldsymbol{x}_i)O(h^2) & a_{23}(\boldsymbol{x}_i)O(h^2) \\ a_{31}(\boldsymbol{x}_i)O(h^2) & a_{32}(\boldsymbol{x}_i)O(h^2) & a_{33}(\boldsymbol{x}_i) \end{pmatrix},$$

where $a_{kl}(\boldsymbol{x}_i) = O(1) \in C^2(\boldsymbol{x}_i)$ for $k, l = 1, 2, 3$, and $a_{11}(\boldsymbol{x}_i)$, $a_{22}(\boldsymbol{x}_i)$ and $a_{33}(\boldsymbol{x}_i)$ are bounded below by some positive constant. Then we have

$$A(\boldsymbol{x}_i)^{-1} = \begin{pmatrix} \tilde{a}_{11}(\boldsymbol{x}_i) & \tilde{a}_{12}(\boldsymbol{x}_i) & \tilde{a}_{13}(\boldsymbol{x}_i) \\ \tilde{a}_{21}(\boldsymbol{x}_i) & \tilde{a}_{22}(\boldsymbol{x}_i)O(h^{-2}) & \tilde{a}_{23}(\boldsymbol{x}_i) \\ \tilde{a}_{31}(\boldsymbol{x}_i) & \tilde{a}_{32}(\boldsymbol{x}_i) & \tilde{a}_{33}(\boldsymbol{x}_i) \end{pmatrix} \stackrel{\text{def}}{=} \tilde{A}(\boldsymbol{x}_i),$$

where $\tilde{a}_{kl}(\boldsymbol{x}_i) = O(1) \in C^2(\boldsymbol{x}_i)$ for $k, l = 1, 2, 3$. By changing integral variables, we have

$$\begin{aligned} \boldsymbol{\Gamma}_1 &= \int \tilde{A}(\boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W} + \boldsymbol{\eta}_{bh})) \left(1, \frac{hw_1}{\hat{\boldsymbol{\beta}}_1^{(t-1)}}, \boldsymbol{W}_{-1}\right)^\top \\ &\quad \times \mathrm{K}(w_1) f(\boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W} + \boldsymbol{\eta}_{bh})) \phi(\boldsymbol{\beta}^T \boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W} + \boldsymbol{\eta}_{bh})) \mathrm{d}\boldsymbol{W}, \end{aligned}$$

where

$$\boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W} + \boldsymbol{\eta}_{bh}) = (hw_1/\hat{\beta}_1^{(t-1)} + x_1, w_2, ..., w_3)^\top \stackrel{\text{def}}{=} (hw_1/\hat{\beta}_1^{(t-1)} + x_1, \boldsymbol{W}_{-1})^\top.$$

By Taylor's expansion,

$$\begin{aligned} \tilde{A}(\boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W} + \boldsymbol{\eta}_{bh})) &= \tilde{A}(x_1, \boldsymbol{W}_{-1}) + \frac{\partial \tilde{A}}{\partial w_1}(x_1, \boldsymbol{W}_{-1}) \frac{hw_1}{\hat{\beta}_1^{(t-1)}} + O(h^2) \frac{\partial^2 \tilde{A}}{\partial^2 w_1}(x_1, \boldsymbol{W}_{-1}), \\ f(\boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W} + \boldsymbol{\eta}_{bh})) &= f(x_{i1}, \boldsymbol{W}_{-1}) + \frac{\partial f}{\partial w_1}(x_1, \boldsymbol{W}_{-1}) \frac{hw_1}{\hat{\beta}_1^{(t-1)}} + O(h^2), \\ \phi(\boldsymbol{\beta}^T \boldsymbol{J}_b^{-1}\boldsymbol{J}_h^{-1}(\boldsymbol{W} + \boldsymbol{\eta}_{bh})) &= \phi(\boldsymbol{\beta}^{(T}(x_1, \boldsymbol{W}_{-1})) + \phi'(\boldsymbol{\beta}^T(x_1, \boldsymbol{W}_{-1}))\boldsymbol{\beta}_1 \frac{hw_1}{\hat{\beta}_1^{(t-1)}} + O(h^2) \end{aligned}$$

then we have

$$\begin{aligned} \boldsymbol{\Gamma}_1 &= \int \tilde{A}(x_1, \boldsymbol{W}_{-1}) f(x_1, \boldsymbol{W}_{-1}) \phi(\boldsymbol{\beta}^{(T}(x_1, \boldsymbol{W}_{-1})) \left(1, \frac{hw_1}{\hat{\beta}_1^{(t-1)}}, \boldsymbol{W}_{-1}\right)^\top \mathrm{K}(w_1) \mathrm{d}\boldsymbol{W} \\ &\quad + \int \left(\frac{\partial \tilde{A}}{\partial w_1}(x_1, \boldsymbol{W}_{-1}) + \frac{\partial f}{\partial w_1}(x_1, \boldsymbol{W}_{-1}) + \phi'(\boldsymbol{\beta}^{(T}(x_1, \boldsymbol{W}_{-1}))\boldsymbol{\beta}_1 \right) \frac{hw_1}{\hat{\beta}_1^{(t-1)}} \end{aligned}$$

$$\times \left(1, \frac{hw_1}{\hat{\beta}_1^{(t-1)}}, \boldsymbol{W}_{-1}\right)^\top \mathrm{K}(w_1)\mathrm{d}\boldsymbol{W} + O(h^2)\left(I + \frac{\partial^2 \tilde{A}}{\partial^2 w_1}(x_1, \boldsymbol{W}_{-1})\right)$$

$$= \quad (\gamma_{11}(x_1), \gamma_{21}(x_1), \gamma_{31}(x_1))^\top$$

where $\gamma_{k1} = O(1)$ for $k = 1, 2, 3$.

Similarly, we can prove that

$$\boldsymbol{\Gamma}_2 = (\gamma_{12}(x_1), \gamma_{22}(x_1), \gamma_{32}(x_1)),$$

where $\gamma_{k2} = O(1)$ for $k = 1, 2, 3$. As such it is easy to show that

$$\boldsymbol{\Gamma} = \int \boldsymbol{\Gamma}_1 \boldsymbol{\Gamma}_2 f(\boldsymbol{x}) v(\boldsymbol{\beta}^\top \boldsymbol{x}) \mathrm{d}\boldsymbol{x} + O_P\left(\frac{1}{\sqrt{nh}}\right) = O(1).$$

The bias term in scenario 2 can be studied quite similarly as the scenario 1; the only difference here is to take care of the second entry of $\boldsymbol{R}_i$ which is corresponding to A_{22}. The details are not included here.

Combining the two scenarios discussed before, we have proved that (1.25) and (1.26) are valid for all possible values of $\boldsymbol{\beta}$. □

Proof of Theorem 1.5.5: To simplify the notations in the proof, let us assume that the sample is from a model with two regions, recorded as

$$\mathscr{S}_1 = \Big\{(\boldsymbol{x}_{1,1}, y_{1,1}), ..., (\boldsymbol{x}_{1,n_1}, y_{1,n_1})\Big\},$$

and

$$\mathscr{S}_2 = \Big\{(\boldsymbol{x}_{2,1}, y_{2,1}), ..., (\boldsymbol{x}_{2,n_2}, y_{2,n_2})\Big\},$$

where $n_1 + n_2 = n$. The true gradients at two regions are $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Let the corresponding estimations of local gradients at each points be

$$\mathscr{S}_1(\boldsymbol{\beta}) = \left\{\hat{\boldsymbol{\beta}}_{1,1}, ..., \hat{\boldsymbol{\beta}}_{1,n_1}\right\},$$

and

$$\mathscr{S}_2(\boldsymbol{\beta}) = \left\{\hat{\boldsymbol{\beta}}_{2,1}, ..., \hat{\boldsymbol{\beta}}_{2,n_2}\right\}.$$

In the following proof, each point is labeled by its estimated local gradient. Let

$$\tilde{\mathscr{S}_1}(\boldsymbol{\beta}) = \left\{\hat{\boldsymbol{\beta}}_{1,k}; \|\hat{\boldsymbol{\beta}}_{1,k}\hat{\boldsymbol{\beta}}_{1,k}^\top - \boldsymbol{\beta}_1\boldsymbol{\beta}_1^\top\| \leq \frac{1}{6}(\|\boldsymbol{\beta}_1\boldsymbol{\beta}_1^\top - \boldsymbol{\beta}_2\boldsymbol{\beta}_2^\top\|)\right\} \quad \text{with} \quad \#\tilde{\mathscr{S}_1}(\boldsymbol{\beta}) = \tilde{n}_1,$$

and

$$\tilde{\mathscr{S}_2}(\boldsymbol{\beta}) = \left\{\hat{\boldsymbol{\beta}}_{2,k}; \|\hat{\boldsymbol{\beta}}_{2,k}\hat{\boldsymbol{\beta}}_{2,k}^\top - \boldsymbol{\beta}_2\boldsymbol{\beta}_2^\top\| \leq \frac{1}{6}(\|\boldsymbol{\beta}_1\boldsymbol{\beta}_1^\top - \boldsymbol{\beta}_2\boldsymbol{\beta}_2^\top\|)\right\} \quad \text{with} \quad \#\tilde{\mathscr{S}_2}(\boldsymbol{\beta}) = \tilde{n}_2.$$

It is easy to see that points in $\tilde{\mathscr{S}_1}(\boldsymbol{\beta})$ will never share the same group with those in $\tilde{\mathscr{S}_2}(\boldsymbol{\beta})$.

If we choose $m = 2$, denote the two estimated clustering groups as $\hat{\mathscr{S}}_{2,1}$ and $\hat{\mathscr{S}}_{2,2}$, then with probability exponentially going to 1, we have

$$\hat{\mathscr{S}}_{2,1} \subseteq \tilde{\mathscr{S}_1}(\boldsymbol{\beta}) \quad \#\hat{\mathscr{S}}_1 = \hat{n}_1 \quad \text{and} \quad \hat{\mathscr{S}}_{2,2} \subseteq \tilde{\mathscr{S}_2}(\boldsymbol{\beta}) \quad \#\hat{\mathscr{S}}_2 = \hat{n}_2.$$

In light of this fact, similar to the proof of Lemma 1.5.1, we can assume there is no misclassifications in $\hat{\mathscr{S}}_{2,1}$ and $\hat{\mathscr{S}}_{2,2}$. Let $(\boldsymbol{x}_{1,k}, y_{1,k})$ be the points labeled in $\hat{\mathscr{S}}_{2,1}$ and $\#\hat{\mathscr{S}}_{2,1} = n_{2,1}$, by definition,

$$\hat{\sigma}_{2,1}^2 \quad = \quad \frac{1}{n_{2,1}}\sum_{k=1}^{n_{2,1}}(y_{1,k} - \hat{\phi}_1(\hat{\boldsymbol{\beta}}_{2,1}^\top\boldsymbol{x}_{1,k}))^2,$$

$$
\begin{aligned}
&= \frac{1}{n_{2,1}}\sum_{k=1}^{n_{2,1}}\varepsilon_{1,k}'^2 + \frac{1}{n_{2,1}}\sum_{k=1}^{n_{2,1}}(\phi_1(\boldsymbol{\beta}_1^\top \boldsymbol{x}_{1,k}) - \hat{\phi}_1(\hat{\boldsymbol{\beta}}_{2,1}^\top \boldsymbol{x}_{1,k}))^2 \\
&\quad -\frac{2}{n_{2,1}}\sum_{k=1}^{n_{2,1}}(\phi_1(\boldsymbol{\beta}_1^\top \boldsymbol{x}_{1,k}) - \hat{\phi}_1(\hat{\boldsymbol{\beta}}_{2,1}^\top \boldsymbol{x}_{1,k}))\varepsilon_{1,k},
\end{aligned}
$$

where $\hat{\phi}_1(\cdot)$ is the NW estimator on the 1st piece. By exploiting the microstructure of $\hat{\phi}_1(\cdot)$ as a weighted summation based on a symmetric kernel function, we have

$$
\hat{\phi}_1(\hat{\boldsymbol{\beta}}_{2,1}^\top \boldsymbol{x}_{1,k}) = \hat{\phi}_1(\boldsymbol{\beta}_1^\top \boldsymbol{x}_{1,k}) + O_P\left(\frac{h_{2,1}}{\sqrt{n_1}}\right),
$$

so it follows that

$$
\begin{aligned}
\hat{\sigma}_{2,1}^2 &= \frac{1}{n_{2,1}}\sum_{k=1}^{n_{2,1}}(y_{1,k} - \hat{\phi}_1(\hat{\boldsymbol{\beta}}_{2,1}^\top \boldsymbol{x}_{1,k}))^2, \\
&= \frac{1}{n_{2,1}}\sum_{k=1}^{n_{2,1}}\varepsilon_{1,k}'^2 + \frac{1}{n_{2,1}}\sum_{k=1}^{n_{2,1}}(\phi_1(\boldsymbol{\beta}_1^\top \boldsymbol{x}_{1,k}) - \hat{\phi}_1(\boldsymbol{\beta}_1^\top \boldsymbol{x}_{1,k}))^2 \\
&\quad +\frac{1}{n_{2,1}}\sum_{k=1}^{n_{2,1}}(\phi_1(\boldsymbol{\beta}_1^\top \boldsymbol{x}_{1,k}) - \hat{\phi}_1(\boldsymbol{\beta}_1^\top \boldsymbol{x}_{1,k}))O_P\left(\frac{h_1}{\sqrt{n_1}}\right) \\
&\quad -\frac{2}{n_{2,1}}\sum_{k=1}^{n_{2,1}}(\phi_1(\boldsymbol{\beta}_1^\top \boldsymbol{x}_{1,k}) - \hat{\phi}_1(\boldsymbol{\beta}_1^\top \boldsymbol{x} y_{1,k}))\varepsilon_{1,k} + \frac{2}{n_{2,1}}\sum_{k=1}^{n_{2,1}}\varepsilon_{1,k}O_P\left(\frac{h_1}{\sqrt{n_1}}\right) \quad (1.29)
\end{aligned}
$$

We have

$$
\begin{aligned}
\hat{\sigma}^2(2) &= n^{-1}(n_{2,1}\hat{\sigma}_{2,1}^2 + n_{2,2}\hat{\sigma}_{2,2}^2) \\
&= \frac{1}{n}\sum_{k=1}^{n_1}(y_{1,k} - \hat{\phi}_1(\hat{\boldsymbol{\beta}}_{2,1}^\top \boldsymbol{x}_{1,k}))^2 + \frac{1}{n}\sum_{k=1}^{n_2}(y_{2,k} - \hat{\phi}_2(\hat{\boldsymbol{\beta}}_{2,2}^\top \boldsymbol{x}_{2,k}))^2 \\
&= \frac{1}{n}\sum_{k=1}^{n_1}\varepsilon_{1,k}^2 + \frac{1}{n}\sum_{k=1}^{n_2}\varepsilon_{2,k}^2 + O_P\left(h_{2,1}^4 + \frac{1}{n_{2,1}h_{2,1}} + h_{2,2}^4 + \frac{1}{n_{2,2}h_{2,2}}\right). \quad (1.30)
\end{aligned}
$$

As discussed in Section 5, $\{h_{2,g},\ g = 1, 2\}$ are chosen to be optimal minimizing $O_P\left(h_{2,g}^4 + \frac{1}{n_{2,g}h_{2,g}}\right)$ and as such $h_{2,g}^4 = O_P\left(\frac{1}{n_{2,g}h_{2,g}}\right)$. So (1.30) can be further

simplified to be

$$\hat{\sigma}^2(2) = \frac{1}{n}\sum_{k=1}^{n_1}\varepsilon_{1,k}^2 + \frac{1}{n}\sum_{k=1}^{n_2}\varepsilon_{2,k}^2 + O_P\left(\frac{1}{n_{2,1}h_{2,1}} + \frac{1}{n_{2,2}h_{2,2}}\right). \tag{1.31}$$

For a given sample, the first two summation parts of (1.31) is a constant for all possible numbers of regions.

We are to prove that

$$Pr\Big(\inf_{m\neq m_0} \mathrm{BIC}(m) - \mathrm{BIC}(m_0) > 0\Big) \to 1, \quad \text{as} \quad n \to \infty.$$

By definition, for any $m \neq m_0$,

$$\mathrm{BIC}(m) - \mathrm{BIC}(m_0) = \log\left(\frac{\hat{\sigma}^2(m)}{\hat{\sigma}^2(m_0)}\right) + k(m) - k(m_0),$$

where $k(m)$ is the penalty term.

The proof is divided into two parts:

(1) $m = 1$;

(2) $m \geq 3$;

Case $m = 1$**.** For $m = 1$, let the estimated single-index be $\hat{\boldsymbol{\beta}}_{1,1}$, by definition,

$$\begin{aligned}\hat{\sigma}^2(1) &= \frac{1}{n}\sum_{k=1}^{n_1}(y_{1,k} - \hat{\phi}_{1,1}(\hat{\boldsymbol{\beta}}_{1,1}^\top \boldsymbol{x}_{1,k}))^2 + \frac{1}{n}\sum_{k=1}^{n_2}(y_{2,k} - \hat{\phi}_{1,1}(\hat{\boldsymbol{\beta}}_{1,1}^\top \boldsymbol{x}_{2,k}))^2, \\ &= I_1 + I_2 - I_3, \end{aligned} \tag{1.32}$$

where

$$I_1 = \frac{1}{n}\sum_{k=1}^{n_1}\varepsilon_{1,k}^2 + \frac{1}{n}\sum_{k=1}^{n_2}\varepsilon_{2,k}^2,$$

$$I_2 = \frac{1}{n}\sum_{k=1}^{n_1}(\phi_1(\boldsymbol{\beta}_1^\top \boldsymbol{x}_{1,k}) - \hat{\phi}_{1,1}(\hat{\boldsymbol{\beta}}_{1,1}^\top \boldsymbol{x}_{1,k}))^2 + \frac{1}{n}\sum_{k=1}^{n_2}(\phi_2(\boldsymbol{\beta}_2^\top \boldsymbol{x} y_{2,k}) - \hat{\phi}_{1,1}(\hat{\boldsymbol{\beta}}_{1,1}^\top \boldsymbol{x}_{2,k}))^2,$$

$$I_3 = \frac{2}{n}\sum_{k=1}^{n_1}(\phi_1(\boldsymbol{\beta}_1^\top \boldsymbol{x}_{1,k}) - \hat{\phi}_{1,1}(\hat{\boldsymbol{\beta}}_{1,1}^\top \boldsymbol{x}_{1,k}))\varepsilon_{1,k} + \frac{2}{n}\sum_{k=1}^{n_2}(\phi_2(\boldsymbol{\beta}_2^\top \boldsymbol{x}_{2,k}) - \hat{\phi}_{1,1}(\hat{\boldsymbol{\beta}}_{1,1}^\top \boldsymbol{x} y_{2,k}))\varepsilon_{2,k}.$$

To compare (1.32) to (1.31), we need to prove that

$$Pr(I_2 - I_3 > c_0 > 0) \to 1 \quad \text{as} \quad n \to \infty, \tag{1.33}$$

and

$$k(2) - k(1) \to 0 \quad \text{in probability} \quad \text{as} \quad n \to \infty, \tag{1.34}$$

which ensure that

$$Pr(\text{BIC}(1) - \text{BIC}(2) > 0) \to 1, \quad \text{as} \quad n \to \infty.$$

It is easy to see that (1.34) follows directly from the definition of $k(m)$. We can prove (1.33) by showing the following two results:

$$I_2 = |O_P(1)|, \tag{1.35}$$

and

$$I_3 \leq O_P\left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}}\right). \tag{1.36}$$

Without loss of generality, we assume that $\boldsymbol{\beta}_1 \neq \hat{\boldsymbol{\beta}}_{1,1}$ and $\boldsymbol{\beta}_1 = \lambda\hat{\boldsymbol{\beta}}_{1,1} + \tilde{\lambda}\tilde{\boldsymbol{\beta}}_1$ with $\hat{\boldsymbol{\beta}}_{1,1} \perp \tilde{\boldsymbol{\beta}}_1$. By definition,

$$\hat{\phi}_{1,1}(\hat{\boldsymbol{\beta}}_{1,1}^\top \boldsymbol{x}_{1,k}) = \frac{n_1^{-1}\sum_{i=1}^{n_1} \mathrm{K}_h(\hat{\boldsymbol{\beta}}_{1,1}^\top(\boldsymbol{x}_{1,i} - \boldsymbol{x}_{1,k}))y_{1,i}}{n_1^{-1}\sum_{i=1}^{n_1} \mathrm{K}_h(\hat{\boldsymbol{\beta}}_{1,1}^\top(\boldsymbol{x}_{1,i} - \boldsymbol{x}_{1,k}))}$$

$$= \frac{n_1^{-1}\sum_{i=1}^{n_1} \mathrm{K}_h(\hat{\boldsymbol{\beta}}_{1,1}^\top(\boldsymbol{x}_{1,i}-\boldsymbol{x}_{1,k}))(\phi_1(\boldsymbol{\beta}_1^\top\boldsymbol{x}_{1,i})+\varepsilon_{1,i})}{n_1^{-1}\sum_{i=1}^{n_1} \mathrm{K}_h(\hat{\boldsymbol{\beta}}_{1,1}^\top(\boldsymbol{x}_{1,i}-\boldsymbol{x}_{1,k}))},$$

we have

$$\begin{aligned} E(\hat{\phi}_{1,1}(\hat{\boldsymbol{\beta}}_{1,1}^\top\boldsymbol{x}_{1,k})-\phi(\boldsymbol{\beta}_1^\top\boldsymbol{x}_{1,k})) &= \frac{n_1^{-1}\sum_{i=1}^{n_1} \mathrm{K}_h(\hat{\boldsymbol{\beta}}_{1,1}^\top(\boldsymbol{x}_{1,i}-\boldsymbol{x}_{1,k}))(\phi_1(\boldsymbol{\beta}_1^\top\boldsymbol{x}_{1,i})-\phi(\boldsymbol{\beta}_1^\top\boldsymbol{x}_{1,k}))}{n_1^{-1}\sum_{i=1}^{n_1} \mathrm{K}_h(\hat{\boldsymbol{\beta}}_{1,1}^\top(\boldsymbol{x}_{1,i}-\boldsymbol{x}_{1,k}))}, \\ &= \frac{\int \mathrm{K}_h(\hat{\boldsymbol{\beta}}_{1,1}^\top(\boldsymbol{x}-\boldsymbol{x}_{1,k}))(\phi_1(\boldsymbol{\beta}_1^\top\boldsymbol{x})-\phi(\boldsymbol{\beta}_1^\top\boldsymbol{x}_{1,k})))f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{f_{\hat{\boldsymbol{\beta}}_{1,1}}(\hat{\boldsymbol{\beta}}_{1,1}^\top\boldsymbol{x}_{1,k})} \\ &\quad +O_P(n_1^{-1/2}). \end{aligned}$$

By changing of variables and Taylor's expansion similar to the proofs of Lemma 1.5.1, we have

$$E(\hat{\phi}_{1,1}(\hat{\boldsymbol{\beta}}_{1,1}^\top\boldsymbol{x}_{1,k})-\phi(\boldsymbol{\beta}_1^\top\boldsymbol{x}_{1,k})) = \frac{\int[\phi_1(\lambda u_0^{(1,1)}+\tilde{\lambda}w)-\phi_1(\lambda u_0^{(1,1)}+\tilde{\lambda}\tilde{u}_0)]\tilde{f}(w)\mathrm{d}w}{f_{\hat{\boldsymbol{\beta}}_{1,1}}(\hat{\boldsymbol{\beta}}_{1,1}^\top\boldsymbol{x}_{1,k})}+O_P(n_1^{-1/2}), \tag{1.37}$$

where $u_0^{(1,1)} = \hat{\boldsymbol{\beta}}_{1,1}^\top\boldsymbol{x}_{1,k}$ and $\tilde{u}_0 = \tilde{\boldsymbol{\beta}}_1^\top\boldsymbol{x}_{1,k}$. Without loss of generality, we assume that $P(u_0^{(1,1)}=0)>0$, otherwise, we may consider $\boldsymbol{\beta}_2$ in the first place. Then the nominator of (1.37) is $\int[\phi_1(\tilde{\lambda}w)-\phi_1(\tilde{\lambda}\tilde{u}_0)]\tilde{f}(w)\mathrm{d}w$ which is non-zero almost surely by assumption (A5), and it follows that $I_2 = |O_P(1)|$.

To prove (1.36), first if we let $\hat{\boldsymbol{\beta}}_{1,1}^{(-k)}$ be the gradient estimated without $\boldsymbol{x}_{1,k}$, then we have

$$\hat{\boldsymbol{\beta}}_{1,1}^{(-k)} - \hat{\boldsymbol{\beta}}_{1,1} = O_P(n_1^{-1}),$$

which follows from the structure of (1.21). Noting that $\hat{\phi}_{1,1}(\hat{\boldsymbol{\beta}}_{1,1}^\top\boldsymbol{x}_{1,k}))$ is calculated

with $\boldsymbol{x}_{1,k}$ itself left out, we have

$$E(\hat{\phi}_{1,1}(\hat{\boldsymbol{\beta}}_{1,1}^{\top}\boldsymbol{x}_{1,k})\varepsilon_{1,k}) = O(n_1^{-1}),$$

and $E(I_3) = O(n_1^{-1} + n_2^{-1})$. With similar reasoning, we can show that $Var(I_3) = O(n_1^{-1} + n_2^{-1})$, which leads to (1.36) directly. Then we complete the proof for case $m = 1$.

Case $m \geq 3$**.** For $m \geq 3$, our main task is to show that

$$\frac{\hat{\sigma}^2(m)}{\hat{\sigma}^2(m_0)} = 1 + \sum_{g=1}^{m} \alpha_{m,g}(n_{m,g}) - \sum_{g=1}^{m_0} \alpha_{m_0,g}(n_{m_0,g}),$$

where $m_0 = 2$ and

$$\left\{ \sum_{g=1}^{m} \alpha_g(n_{m,g}) - \sum_{g=1}^{m_0} \alpha_{m_0,g}(n_{m_0,g}) \right\} \Big/ \{k(m) - k(m_0)\} \to 0. \tag{1.38}$$

Then as long as $Pr(k(m) - k(m_0) > 0) = 1$, we have

$$Pr(\mathrm{BIC}(m) - \mathrm{BIC}(m_0) > 0) \to 1, \quad \text{as} \quad n \to \infty.$$

In our proof, we only give the discussion about the case of $m = 3$, which can be easily extended to the cases of $m > 3$. For $m = 3$, let the three groups resulted from Algorithm 2 be $\hat{\mathscr{S}}_{3,1}$, $\hat{\mathscr{S}}_{3,2}$ and $\hat{\mathscr{S}}_{3,3}$, corresponding to the three "core groups": $\hat{\mathscr{S}}_{3,1}$, $\hat{\mathscr{S}}_{3,2}$ and $\hat{\mathscr{S}}_{3,3}$, with

$$\#\hat{\mathscr{S}}_{3,1} \geq \#\hat{\mathscr{S}}_{3,2} \geq \#\hat{\mathscr{S}}_{3,3}.$$

Without loss of generality, we assume that

$$\hat{\mathscr{S}}_{3,1} \cap \hat{\mathscr{S}}_{2,1} \neq \emptyset, \quad \text{and} \quad \hat{\mathscr{S}}_{3,2} \cap \hat{\mathscr{S}}_{2,2} \neq \emptyset.$$

By the characteristics of K-means clustering we have, with probability 1, two scenarios:

(1) $\#\hat{\mathscr{S}}_{3,3} \leq np(n) \quad \Rightarrow \quad \hat{\mathscr{S}}_{3,1} = \hat{\mathscr{S}}_{2,1},\ \hat{\mathscr{S}}_{3,2} = \hat{\mathscr{S}}_{2,2};$

(2) $\#\hat{\mathscr{S}}_{3,3} > np(n) \quad \Rightarrow \quad \hat{\mathscr{S}}_{3,1} \subset \hat{\mathscr{S}}_{2,1},\ \hat{\mathscr{S}}_{3,2} \subset \hat{\mathscr{S}}_{2,2}.$

Scenario 1 is easy to handle since it only attaches an additional group to the original two groups.

For scenario 2, since $\#\hat{\mathscr{S}}_{3,3} > np(n)$, it can not be "ignored" by both $\hat{\mathscr{S}}_{2,1}$ and $\hat{\mathscr{S}}_{2,2}$, otherwise, we would have $\#\hat{\mathscr{S}}_{2,1} + \#\hat{\mathscr{S}}_{2,2} < n(1 - p(n))$, which is not sufficient to terminate the clustering Step III. By hierarchial clustering, $\hat{\mathscr{S}}_{3,3}$ must be aggregated by either $\hat{\mathscr{S}}_{2,1}$ or $\hat{\mathscr{S}}_{2,2}$, and let's assume that in scenario 2

$$\hat{\mathscr{S}}_{3,3} \subset \hat{\mathscr{S}}_{2,1}.$$

Consequently, we have $\hat{\mathscr{S}}_{3,1} \cup \hat{\mathscr{S}}_{3,3} \subseteq \hat{\mathscr{S}}_{2,1}$. The only difference between $\hat{\mathscr{S}}_{3,1} \cup \hat{\mathscr{S}}_{3,3}$ and $\hat{\mathscr{S}}_{2,1}$ is a diminishing proportion of points $O(p_0(n))$. So again we have, with probability going to 1,

$$\hat{\mathscr{S}}_{3,1} \cup \hat{\mathscr{S}}_{3,3} \subseteq \hat{\mathscr{S}}_{2,1} \quad \text{and} \quad \hat{\mathscr{S}}_{3,2} \subseteq \hat{\mathscr{S}}_{2,2}.$$

Then equalities

$$\hat{\mathscr{S}}_{3,1} \cup \hat{\mathscr{S}}_{3,3} = \hat{\mathscr{S}}_{2,1} \quad \text{and} \quad \hat{\mathscr{S}}_{3,2} = \hat{\mathscr{S}}_{2,2} \tag{1.39}$$

are followed with probability tending to 1 by the fact that

$$\hat{\mathscr{S}}_{3,1} \cup \hat{\mathscr{S}}_{3,2} \cup \hat{\mathscr{S}}_{3,3} = \hat{\mathscr{S}}_{2,1} \cup \hat{\mathscr{S}}_{2,2}.$$

The equalities (1.39) show that in scenario 2, we create one more group by splitting one of the original groups into two which should have been together as one. So intuitively, we can not expect such action can improve the accuracy of model fitting significantly. Following the same reasoning that leads to (1.31), we can easily show that

$$\hat{\sigma}^2(3) = \frac{1}{n}\sum_{k=1}^{n_1} \varepsilon_{1,k}^2 + \frac{1}{n}\sum_{k=1}^{n_2} \varepsilon_{2,k}^2 + O_P\left(\frac{1}{n_{3,1}h_{3,1}} + \frac{1}{n_{3,2}h_{3,2}} + \frac{1}{n_{3,3}h_{3,3}}\right),$$

where the first two items are unchanged since the noise terms for a given sample are fixed. Finally we have

$$\frac{\hat{\sigma}^2(3)}{\hat{\sigma}^2(2)} = 1 + O_P\left(\sum_{g=1}^{3} \frac{1}{n_{3,g}h_{3,g}}\right) + O_P\left(\sum_{l=1}^{2} \frac{1}{n_{2,l}h_{2,l}}\right),$$

which implies that $k(m) = \log(n)\sum_{g=1}^{m}(n_{m,g}h_{m,g})^{-1}$ is sufficient to ensure (1.38) to be satisfied. □

CHAPTER 2

WLE of Nonlinear AR Models with MA Errors

2.1 Time Series Analysis: A Literature Review

Time series data typically refer to the observations collected sequentially over time, in which the data in the future depend on the observations in the past. A fundamental task of time series analysis is to discover the stochastic law that governs the observed time series which helps us to understand the underlying dynamics and forecast future events. To this end, time series analysis typically rests

on proper statistical modeling. In this section, we introduce several popular time series models and some related analytical techniques that we will use later on.

2.1.1 Stationarity of Time Series

In time series analysis, statistical inference is useful only when the observed underlying dynamics are sustained over a time period of interest. This leads to the definition of stationarity which requires that time series exhibit certain time-invariant property. Here we present the definitions of both (weak) stationarity and strict stationarity.

Definition 2.1. A time series $\{y_t\}$ is stationary if $E(y_t^2) < \infty$ for each t, and

(1) $E(y_t)$ is a constant, independent of t, and

(2) $\text{Cov}(y_t, y_{t+k})$ is independent of t for each k.

Definition 2.2. A time series $\{y_t\}$ is strictly stationary if (y_1, ..., y_n) and (y_{1+k}, ..., y_{n+k}) have the same joint distribution for any integer $n \geq 1$ and any integer k.

Obviously stationarity is generally weaker than strict stationarity if the process has finite second moments. As will be introduced later, the assumption of weak stationarity is usually sufficient in analyzing linear time series models. In contrast, if we are to investigate nonlinear relationships, restrictions on only the first two

moments are sometimes inadequate to yield the desired asymptotic properties. This is why strict stationarity is introduced here.

2.1.2 Linear Time Series Models

It has been a very long history of linear time series modeling in statistics society dating back to Yule's autoregressive (AR) models (1927). Specifically, the class of AR models can be represented as

$$y_t = a_0 + a_1 y_{t-1} + \cdots + a_p y_{t-p} + \varepsilon_t, \tag{2.1}$$

where the a_j are real constants, p is a finite positive integer referred to as the order of the AR model, and the ε_t are zero-mean uncorrelated random variables, called white noise, with a finite common variance σ_ε^2. If $\{y_t\}$ follows model (2.1), we denote $y_t \sim \text{AR}(p)$. Model (2.1) represents the current state y_t through its immediate p past values y_{t-1}, ... , y_{t-p} in a linear regressive manner.

A more general class of linear models is obtained by replacing ε_t by a moving average process $\xi_t := \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$, i.e.,

$$y_t = a_0 + a_1 y_{t-1} + \cdots + a_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \tag{2.2}$$

where the θ_j are real constants. Model (2.2) is referred as ARMA(p, q) model. If

we define the backshift operator B as

$$B^k y_t = y_{t-k},\ k = \pm 1, \pm 2, ...,$$

then the model (2.2) can be written as

$$a(B)y_t = \theta(B)\varepsilon_t, \tag{2.3}$$

where $a(\cdot)$ and $\theta(\cdot)$ are polynomials defined as

$$a(s) = 1 - a_1 s - \cdots - a_p s^p,$$

and

$$\theta(s) = 1 - \theta_1 s - \cdots - \theta_q s^q.$$

For ARMA models as defined in (2.3), it is always assumed that polynomials $a(s)$ and $\theta(s)$ do not have common factors, i.e., the p and q involved in the model are assumed to be the smallest respectively among all possible choices. The following theorem gives a sufficient condition for the stationary of the ARMA models (pp. 31, Chapter 2, Fan and Yao, 2003).

Theorem 2.1. *The process $\{y_t\}$ given by (2.3) is stationary if $a(s) \neq 0$ for all complex numbers s such that $|s| \leq 1$.*

The condition imposed in this theorem has become a standard assumption for most linear time series analysis.

2.1.3 Nonlinear Time Series Models

Linear models have a reasonable flexibility in approximating many stationary processes. Nonetheless, the linear models do not approximate well the nonlinear phenomena we observe in many real time series data, such as sunspot data and Canadian lynx data. Those nonlinear phenomena include, for example, nonnormality, asymmetric cycles, bimodality, nonlinear relationship between lagged variables, variation of prediction performance over the state-space, time irreversibility, sensitivity to initial conditions and others. Modeling the nonlinearity in time series is beyond the scope of traditional linear models.

We have seen fruitful developments on various nonlinear parametric time series models. The successful examples include, among others, the ARCH-modeling of volatility of financial data (Engle, 1982; Bollerslev, 1986) and the (smooth) threshold autoregressive modeling of biological and economic data (Tong, 1990; Terasvirta, 1994). The focus of this thesis is on the latter class of nonlinear models, i.e., the nonlinear autoregressive models. Specifically, $\{y_t\}$ is said to follow a nonlinear autoregressive model of order p if there exists a function $\tilde{\phi}$ such that

$$y_t = \tilde{\phi}(y_{t-1}, y_{t-2}, ..., y_{t-p}, \varepsilon_t),\ t = \pm 1, \pm 2, ..., \tag{2.4}$$

where ε_t is a sequence of stationary process with $E\{\varepsilon_t \varepsilon_s\} = 0$ for $t \neq s$. It is of

special interest to study the additive noise model defined as

$$y_t = \phi(y_{t-1}, y_{t-2}, ..., y_{t-p}) + \varepsilon_t, \ t = \pm 1, \pm 2, ..., \tag{2.5}$$

for some real function ϕ. A typical example of model (2.5) is the threshold autoregressive model (TAR; Tong, 1990). The TAR model is of the form

$$y_t = a_0^{(i)} + a_1^{(i)} y_{t-1} + \cdots + a_p^{(i)} y_{t-p} + \varepsilon_t, \text{ if } y_{t-d} \in R_i, \tag{2.6}$$

for $i = 1, ..., k$, where $\{R_i\}$ forms a nonoverlapping partition of the real line. There are also many successful smoothing extensions of the TAR model. Most of them can be included in the class of function-coefficient autoregressive (FAR; Chen and Tsay, 1994) model which has the form

$$y_t = \phi_0(y_{t-d}) + \phi_1(y_{t-d}) y_{t-1} + \cdots + \phi_p(y_{t-d}) y_{t-p} + \varepsilon_t, \tag{2.7}$$

where $\phi_j(\cdot)$ are unknown coefficient functions.

Similar to the generalization from AR models to ARMA models, we can define a more general class of nonlinear models by replacing the ε_t in model (2.5) with a moving average (MA) process $\xi_t := \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$, i.e.,

$$y_t = \phi(y_{t-1}, y_{t-2}, ..., y_{t-p}) + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \ t = \pm 1, \pm 2, ..., \tag{2.8}$$

We call model (2.8) nonlinear autoregressive/moving average model. The TAR and FAR model can be generalized in a similar way. The estimation method of model (2.8) is currently not well developed. A main contribution of this thesis is

having established an efficient method for the estimation of the parametric models that take the form of (2.8). The necessity of adding an MA part to the original model is also supported by real data examples. Detailed discussions appear in the subsequent sections of the Chapter.

2.1.4 Spectral Analysis and Periodogram

For a stationary time series $\{y_t\}$, it follows that $\text{Cov}(y_t, y_{t+n})$ is simply a function of n. This function is called the autocovariance function of $\{y_t\}$ at lag n and is denoted by $\gamma(n)$. The ratio $\rho(n) = \gamma(n)/\gamma(0)$ is called the autocorrelation function (ACF) of $\{y_t\}$ of lag n. The following theorem states that the ACF can be denoted by a Fourier transform of a certain distribution function G (pp. 51, Chapter 2, Fan and Yao, 2003).

Theorem 2.2. *A real function defined by* $\{\rho_n : \ n = 0, \pm 1, \pm 2, ...\}$ *is the ACF of a stationary time series if and only if there exists a symmetric probability distribution on* $[-\pi, \pi]$ *with distribution function* F *for which*

$$\rho(n) = \int_{-\pi}^{\pi} e^{inw} dG(w),$$

where $i = \sqrt{-1}$ *stands for the imaginary unit.*

The function G is called the normalized spectral distribution function of the

time series $\{y_t\}$. If G has a density function g, then

$$\rho(n) = \int_{-\pi}^{\pi} e^{inw} g(w) dw,$$

and g is called the normalized spectral density function. Moreover, If $\rho(n)$ is absolutely summable in the sense that $\sum_{n=1}^{\infty} |\rho(n)| < \infty$, then g exists and is given by

$$g(w) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \rho(n) e^{-inw}. \tag{2.9}$$

In many applications such as engineering, spectral decomposition of the total power, i.e., the variance, is of main interest. To this end, we define the (non-normalized) spectral distribution function as

$$F(w) = \gamma(0) G(w),$$

and the (non-normalized) spectral density function as

$$f(w) = \gamma(0) g(w) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \gamma(n) e^{-inw}.$$

Given an observed time series $\{y_t,\ t = 1, ...T\}$, a nature estimation of the spectral density function $f(w)$ is obtained by replacing the $\gamma(n)$ with the sample autocovariance

$$\hat{\gamma}(n) = \frac{1}{T} \sum_{t=1}^{T-|n|} (y_{t+|n|} - \bar{y})(y_t - \bar{y}),$$

for $-T < n < T$, where

$$\bar{y} = \frac{1}{T} \sum_{t=1}^{T} y_t.$$

Then we have

$$\hat{f}(w) = \frac{1}{2\pi} \sum_{n=-T+1}^{T-1} \hat{\gamma}(n) e^{-inw}.$$

For a time series $\{y_t\}$, define the discrete Fourier transform (DFT) as $(Y(w_1), Y(w_2),$ ..., $Y(w_{T-1}))$, where

$$Y(w_k) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} y_t e^{-itw_k}, \tag{2.10}$$

and $w_k = 2\pi k/T$ are called the Fourier frequencies. The periodogram of $\{y_t\}$ is defined as

$$I(w_k) = |Y(w_k)|^2 = \frac{1}{T} \left| \sum_{t=1}^{T} y_t e^{-itw_k} \right|^2,$$

where w_k is the Fourier frequency. The theorem below establishes the link between periodogram and spectral density function (pp. 62, Chapter 2, Fan and Yao, 2003).

Theorem 2.3. *For $k = 1, ..., T-1$,*

$$I(w_k) = \frac{1}{T} \left| \sum_{t=1}^{T} y_t e^{-itw_k} \right|^2 \equiv \sum_{n=-T+1}^{T-1} \hat{\gamma}(n) e^{-inw_k} = 2\pi \hat{f}(w_k),$$

where $\hat{f}(\cdot)$ and $\hat{\gamma}(\cdot)$ are as defined above.

2.1.5 Whittle Likelihood Estimation (WLE)

One of the most successful applications of spectral analysis and periodogram is the Whittle's approximation to the Gaussian likelihood function. A time series $\{z_t\}$ is said to be Gaussian if all its finite-dimensional distributions are normal. If

ε_t are i.i.d $N(0, \sigma^2)$ and $a(s) \neq 0$ for all $|s| \leq 1$, $\{y_t\}$ defined by (2.3) is a stationary Gaussian process.

Consider a set of observations $Z_T = (z_1, ..., z_T)^\top$ generated by a univariate stationary Gaussian process. Then we have the Gaussian -2log-likelihood function

$$L(\beta, \theta, \sigma^2) = \log |\sigma^2 G_T| + \sigma^{-2} Z_T^\top G_T^{-1} Z_T, \tag{2.11}$$

where G_T is the $T \times T$ covariance matrix of Z_T. The maximum likelihood estimator (MLE), $(\hat{\beta}, \hat{\theta}, \hat{\sigma}^2)$, is obtained by maximizing (2.11) over the parameter space. However, the direct involvement of $|G_T|$ and G_T^{-1} in the evaluations of $L(\beta, \theta, \sigma^2)$ intensifies the computation burden to a daunting scale for moderately large samples. Moreover, since the dimension of G_T goes to infinity at the same rate as the sample size T, the asymptotic properties of the estimator are not so straightforward.

To avoid such a problem, Whittle (1953) used several ingenious matrix calculus and approximated the quadratic form in (2.11) by a summation of the ratios of the periodogram of the observations and the corresponding spectral density function of the model taking value at the Fourier frequencies $\lambda_j = 2\pi j/T$, $j = 1, ..., T-1$. Suppose the spectral density function of z_t is $f(\lambda; \beta, \theta) = \sigma^2 k(\lambda; \beta, \theta)/(2\pi)$. The Whittle's approximation to the likelihood function (2.11) is

$$L_W(\beta, \theta, \sigma^2) = \sum_{j=1}^{T-1} \Big[\log(\sigma^2 k(\lambda_j; \beta, \theta)) + \frac{I(\lambda_j)}{\sigma^2 k(\lambda_j; \beta, \theta)} \Big], \tag{2.12}$$

where $I(\lambda_j)$ is the periodogram of z_t. Since the periodogram can be calculated easily via the fast Fourier transform, the Whittle likelihood estimation can be implemented easily as long as the spectral density function $f(\lambda;\beta,\theta)$ has an explicit form. The estimator based on (2.12) is called Whittle likelihood estimation (WLE).

The traditional WLE has played a fundamental role in the theoretical development of linear and nonlinear time series analysis. Most notably, the asymptotic theory of MLE of ARMA models was first derived by Hannan (1973) based on the equivalence of the Whittle likelihood function $L_W(\beta,\theta,\sigma^2)$ and the usual likelihood function $L(\beta,\theta,\sigma^2)$. Without this equivalence, the asymptotic theory is extremely difficult, and thus was derived many years later by Yao and Brockwell (2006).

2.2 Introduction of the Extended WLE (XWLE)

For linear or nonlinear autoregressive (AR) time series models, it is known that the regression errors are usually not linearly independent. There are two possible approaches to accommodate the dependence. The first approach is by increasing the order of the autoregressive models, and the second by introducing moving average (MA) residuals. The latter is usually more efficient in the sense that it needs less parameters. As an example, an ARMA model is more efficient than an AR model even though any ARMA model can be approximated by a higher order AR

model. For nonlinear time series models, to use a higher order nonlinear AR model to approximate a nonlinear ARMA model is even more intractable because the resulted model might have a very complicated functional form. Therefore, investigating nonlinear AR models with a moving average error is very important in time series modeling. In this Chapter we consider the following nonlinear autoregressive model with MA errors

$$y_t = \phi(X_t, \beta) + \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q}, \tag{2.13}$$

where ϕ is a twice continuously differentiable function with unknown parameters $\beta = (\beta_1, ..., \beta_p)$, X_t is a vector variable that can contain either lags of y_t or a collection of exogenous variables, or both, $E(\varepsilon_t) = 0$, $E(\varepsilon_t\varepsilon_s) = 0$ if $t \neq s$ and σ_0^2 otherwise. For ease of exposition, let

$$\xi_t(\theta) = \theta(B)\varepsilon_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q}, \tag{2.14}$$

where B is the backshift operator on t and

$$\theta(s) = 1 + \theta_1 s + \cdots + \theta_q s^q.$$

The linear ARMA model is included in model (2.13). Another special case of model (2.13) is the smooth threshold AR model (STAR, Chan and Tong, 1986; Terasvirta, 1994) with MA regression errors

$$y_t = \beta_1^\top X_t + \beta_2^\top X_t \times I_{t-d} + \xi_t,$$

where I_{t-d} is a smooth function of y_{t-d} with $d \geq 1$.

The estimation of model (2.13) is not trivial. First, note that the least square method might not get a consistent estimator because

$$E(\xi_t(\theta)|X_t) \neq 0.$$

The maximum likelihood estimation is also not easily tractable as the nonlinearity of the model complicates the marginal distribution of y_t. On the other hand, direct application of the traditional WLE to model (2.13) faces at least two problems. Firstly, the likelihood function (2.11) is based on Gaussian distribution of y_t, which is usually not correct if $\phi(X_t, \beta)$ is not linear in X_t. Secondly, a time series y_t following model (2.13) usually has no theoretical spectral density function of the parameters, and thus the Whittle's approximation (2.12) is not available.

In this Chapter, we extend the Whittle likelihood estimation to handle these problems by exploiting the periodogram of residuals which are assumed to follow an MA process. So we convert a nonlinear and non-Gaussian problem to be a linear Gaussian problem. The idea of transforming a nonlinear problem to be a linear problem is also seen in the Whittle estimation of ARCH models (Giraitis and Robinson, 2001). With respect to investigating the periodogram of residuals, Shimotsu and Phillips (2005) employed a similar idea to give a semiparametric

estimation of the memory parameter in fractionally integrated time series. However, the estimation method of Shimotsu and Phillips (2005) relies on an explicit representation of y_t by a linear combination of ε_t, which is usually not attainable in nonlinear models.

The rest of this Chapter is organized as follows. In Section 2.3, we describe in details our estimation method, called the extended Whittle likelihood estimation (XWLE). Section 2.4 discuses the model diagnostics based on XWLE. In Section 2.5, some numerical studies are employed to check the performance of the estimation method, especially as compared with the existing estimation methods if they are applicable; two real data sets are used to illustrate the application of the methods. Theoretical justification of the proposed methods is given in Section 2.6.

2.3 Estimating Nonlinear Models with XWLE

Suppose we have observations $\{y_t,\ t = 1, 2, ..., T\}$ and $\{X_t,\ t = 1, 2, ..., T\}$ satisfying model (2.13), i.e.,

$$y_t = \phi(X_t, \beta) + \xi_t(\theta),$$

where X_t is a vector variable that can contain either lags of y_t or a collection of exogenous variables, or both, and $\xi_t(\theta)$ is a moving average (MA) process defined

in (2.14). For a nonlinear autoregressive function $\phi(\cdot)$, the theoretical spectral density function of y_t is generally not available. Instead, we know that $\xi_t(\theta)$ is an MA(q) process with spectral density function

$$f_0(\lambda, \theta, \sigma^2) = \frac{\sigma^2}{2\pi} k_0(\lambda, \theta) = \frac{\sigma^2}{2\pi} \Big| \sum_{j=0}^{q} \theta_j e^{ij\lambda} \Big|^2,$$

where $\theta_0 = 1$ and i stands for the imaginary unit. The calculation of $k_0(\lambda, \theta)$ is very easy and so are its derivatives. Let $z_t(\beta) = y_t - \phi(X_t, \beta), \quad t = 1, ..., T$. When β and θ are both correctly specified, we have that $z_t(\beta) = \xi_t(\theta)$, i.e., $z_t(\beta)$ is an MA process whose theoretical spectral density function is known. In order to estimate β and θ, instead of considering the periodogram of $\{y_t : t = 1, ..., n\}$ directly, we consider the periodogram of $z_t(\beta)$ which would coincide with $f_0(\lambda, \theta, \sigma^2)$ if both β and θ approach to their true values as $T \to \infty$.

We assume that $z_t(\beta)$ is attainable from $t = 1$ to $t = T$, for simplicity of notations. We also write $z_t(\beta)$ as z_t in the following context. The periodogram of z_t is defined as

$$I_z(\lambda; \beta) = (2\pi T)^{-1} \Big| \sum_{t=1}^{T} z_t e^{it\lambda} \Big|^2.$$

Let $c_z(n; \beta) = T^{-1} \sum_{t=1}^{T-n} z_{t+n} z_t$, then

$$I_z(\lambda; \beta) = (2\pi)^{-1} \sum_{n=-T+1}^{T-1} c_z(n; \beta) e^{-in\lambda}. \tag{2.15}$$

We define the extended Whittle likelihood function for z_t as

$$L_T(\beta, \theta, \sigma^2) = \frac{1}{T} \sum_{j=1}^{T-1} \Big[\log(\sigma^2 k_0(\lambda_j; \theta)) + \frac{I_z(\lambda_j; \beta)}{\sigma^2 k_0(\lambda_j; \theta)} \Big], \tag{2.16}$$

where $\lambda_j = 2\pi j/T,\ j = 1, ..., T-1$. Following Hannan (1973) and assumption (A5) in Section 2.6, we have that

$$\frac{1}{T}\sum_{j=1}^{T-1}\log(k_0(\lambda_j;\theta)) = O(T^{1/2-\alpha})$$

for some $\alpha > 1/2$. So similar to Giraitis and Robinson (2001), we estimate parameters $(\beta, \theta, \sigma^2)$ by minimizing

$$W_T(\beta,\theta,\sigma^2) = \log(\sigma^2) + \frac{1}{T\sigma^2}\sum_{j=1}^{T-1}\frac{I_z(\lambda_j;\beta)}{k_0(\lambda_j,\theta)}$$

with respect to β, θ and σ^2. It is easy to see that given (β, θ) the solution for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{j=1}^{T-1}\frac{I_z(\lambda_j;\beta)}{k_0(\lambda_j,\theta)} \stackrel{def}{=} Q_T(\beta,\theta). \tag{2.17}$$

Then minimizing $W_T(\beta,\theta,\sigma^2)$ is equivalent to first minimizing

$$\tilde{W}_T(\beta,\theta) = \log(Q_T(\beta,\theta)) + 1$$

and then solving $\hat{\sigma}^2$ by (2.17), which further induces us to estimate (β, θ) by solving

$$(\hat{\beta}_T, \hat{\theta}_T) = \arg\min Q_T(\beta,\theta) = \arg\min \frac{1}{T}\sum_{j=1}^{T-1}\frac{I_z(\lambda_j;\beta)}{k_0(\lambda_j,\theta)}. \tag{2.18}$$

We call the above estimation method the extended Whittle likelihood estimation (XWLE).

Comparing our extended Whittle likelihood function (2.16) with the classic one (2.12), the main difference between them is that we "move" the unknown parameters β from the denominator to the numerator to avoid direct involvement

of the spectral density function of the original time series y_t. This difference makes XWLE applicable to a much more general class of time series models. It is also applicable to the case in which exogenous variables are involved in the model, for which the classical WLE cannot be used. In theory, however, XWLE is much more complicated than the conventional WLE. By moving β into $I_z(\lambda;\beta)$ which is a random variable with non-negligible noise (see e.g. Theorem 10.3.2 of Brockwell and Davis, 1991), to investigate the asymptotic properties of XWLE is not an easy job at all. The details are given in Section 2.6. Although the asymptotic variance matrix of the XWLE is less explicit than the classical WLE, our intensive simulation studies in Section 4 suggest that XWLE is sometimes more stable and more efficient than the classic WLE when both methods are applicable.

2.4 Model Diagnosis Based on XWLE

For linear ARMA models with normal innovations, it is proved that the Whittle likelihood function W_T is asymptotically equivalent to the maximum likelihood function; see Hannan (1973) for details. This fact induces us to apply the traditional idea of model diagnostics to our new estimator for its corresponding diagnostics.

The extensions of the classic BIC (Schwarz, 1978) to the Whittle likelihood function is defined as,

$$\text{BIC}_W = \log(Q_T(\hat{\beta}_T, \hat{\theta}_T)) + k \log(T)/T,$$

where k is the number of parameters involved in the model. The model with the smallest BIC_W score is the model preferred. The consistency of BIC_W in selecting the number of parameters follows directly from the asymptotic equivalence of the Whittle likelihood function and the maximum likelihood function and that $z_t(\beta)$ is an MA process when β is correctly specified.

To validate the estimated model, it is also common to carry out a white noise test for the fitted residuals $\hat{\varepsilon}_t$. The $\hat{\varepsilon}_t$ here are calculated in a similar way as we commonly do to the ARMA model. Namely, we first define $\hat{\varepsilon}_{-j} = 0$, for $j = 0, 1, ..., q-1$. Then the $\hat{\varepsilon}_t$ are calculated as

$$\hat{\varepsilon}_t = y_t - \phi(X_t, \beta) - \theta_1 \hat{\varepsilon}_{t-1} - ... - \theta_q \hat{\varepsilon}_{t-q}$$

for $t = 1, ..., T$. The most popular white noise test is probably the χ^2 portmanteau test (Box and Pierce, 1970, and Ljung and Box, 1978), which depends on the following statistic:

$$R_T^{(m)} = T(T+2) \sum_{k=1}^{m} \frac{\hat{\rho}^2(k)}{T-k},$$

where m is the so-called lag truncation number and is fixed. The empirical autocorrelation, $\rho^2(k)$, is defined as

$$\hat{\rho}^2(k) = \frac{\sum_{t=k+1}^{T}(\hat{\varepsilon}_t - \bar{\varepsilon})(\hat{\varepsilon}_{t-k} - \bar{\varepsilon})}{\sum_{t=1}^{T}(\hat{\varepsilon}_t - \bar{\varepsilon})^2},$$

where $\bar{\varepsilon} = T^{-1}\sum_{t=1}^{T}\hat{\varepsilon}_t$. Under the assumption that ε_t are independent and identically distributed (i.i.d.), it can be shown that $R_T(m) \xrightarrow{D} \chi^2(m)$, where $\xrightarrow{D}$ stands for convergence in distribution.

Different values of m will result in different test statistics $R_T^{(m)}$. One way to overcome this problem is to use the adaptive Neyman test (Fan, 1996)

$$R_T^{(AN)} = \max_{1\le m\le a_T} \frac{R_T^{(m)} - m}{\sqrt{2m}},$$

where a_T is some upper limit. Fan (1996) showed that under the null hypothesis

$$P(\tilde{R}_T^{(AN)} < x) \to \exp(-\exp(-x)) \quad \text{as } T \to \infty,$$

where

$$\tilde{R}_T^{(AN)} = \sqrt{2\log\log a_T}R_T^{(AN)} - \{2\log\log a_T + 0.5\log\log\log a_T - 0.5\log(4\pi)\}.$$

Although we still have a parameter a_T to choose, the adaptive Neyman test is less sensitive to a_T than $R_T(m)$ to m. We call $R_T^{(m)}$ LB(m)-test and $\tilde{R}_T^{(AN)}$ AN(a_T)-test. The empirical performances of both tests are examined in the next section by a simulated example.

2.5 Numerical Studies

In this section, we illustrate the proposed modeling procedure of Section 2.3 and Section 2.4 by applying it to some simulated and real examples. We study the performance of our estimation method by (1) comparing the estimation efficiency of XWLE with the original WLE for the ARMA model to which both estimation methods are applicable, and (2) checking the estimation efficiency of XWLE and the model selection for a nonlinear time series model to which WLE is not applicable.

Example 2.5.1. We first consider the following ARMA(p,1) model,

$$y_t = \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \varepsilon_t,$$

where $\varepsilon_t's$ are i.i.d with mean 0. Three lag-values of p are considered in this example, namely, we study the three models:

- ARMA$(1, 1)$,
- ARMA$(2, 1)$,
- ARMA$(5, 1)$.

Moreover, we take three distribution assumptions on the innovation ε_t respectively for each value of p. The three distributions are

- $N(0,1)$: Standard normal distribution,
- $t(1)$: Student's t distribution with one degree of freedom,
- $U(-1,1)$: Uniform distribution between -1 and 1.

So in combination we have 9 different settings of models.

To make a fair comparison, we consider 5 values of β_1: 0.1, 0.3, 0.5, 0.7 and 0.9, and let θ_1 go through all its invertible region $(-1,1)$. For $p > 1$, The rest parameters β_2, ..., β_p are randomly sampled from the $(p-1)$ dimension uniform distribution $U(-\frac{1}{2},\frac{1}{2})^{\otimes(p-1)}$ in which only the stationary choices are used for further simulation studies, otherwise, we resample them till stationarity is satisfied. For each parameter setting, we draw a time series with length n and estimate the parameters using different methods, including the Whittle likelihood estimation (WLE), the extended WLE (XWLE), and the maximum likelihood estimation (MLE). The estimation error is defined as

$$\mathrm{Err}(\hat{\boldsymbol{\beta}},\hat{\theta}_1)=\sum_{k=1}^{p}(\hat{\beta}_k-\beta_k)^2+(\hat{\theta}_1-\theta_1)^2,$$

where $\hat{\boldsymbol{\beta}}=(\beta_1,\cdots,\beta_p)$.

Based on 100 replications for each setting of parameters, innovation distribution and sample size n, the logarithms of the average estimation errors are shown in Figure 2.1 to 2.9. In each panel, the blue line with 'o', the green line with '$\square$' and the red line with '$*$' represent the estimation error of WLE, MLE and XWLE

respectively. The y-axis is for log(Err) and the x-axis is for θ_1.

Although WLE was proved to be asymptotically equivalent to MLE under normality assumption of ε_t, the former is commonly found not so stable as the latter in some situations, which is also observed in our simulations as shown in many panels of figures when θ_1 approaches to 1. It seems, however, that XWLE is more stable than WLE in most cases. The choice of the innovation distribution seems to be not a crucial influencing factor for the estimation accuracies of any of the three methods.

From $p = 1$ to $p = 5$ and $n = 100$ to $n = 400$ we observe that MLE become less attractive as compared to WLE and XWLE when both p and n become larger. Moreover, for some of the settings, especially for $\theta_1 >= -0.5$, the errors from XWLE method always stick to the smaller values of MLE and WLE, or attain the minima of the three methods by themselves. This phenomenon is clearly shown in the rows $\beta_1 = 0.3$ and $\beta_1 = 0.9$ of most Figures.

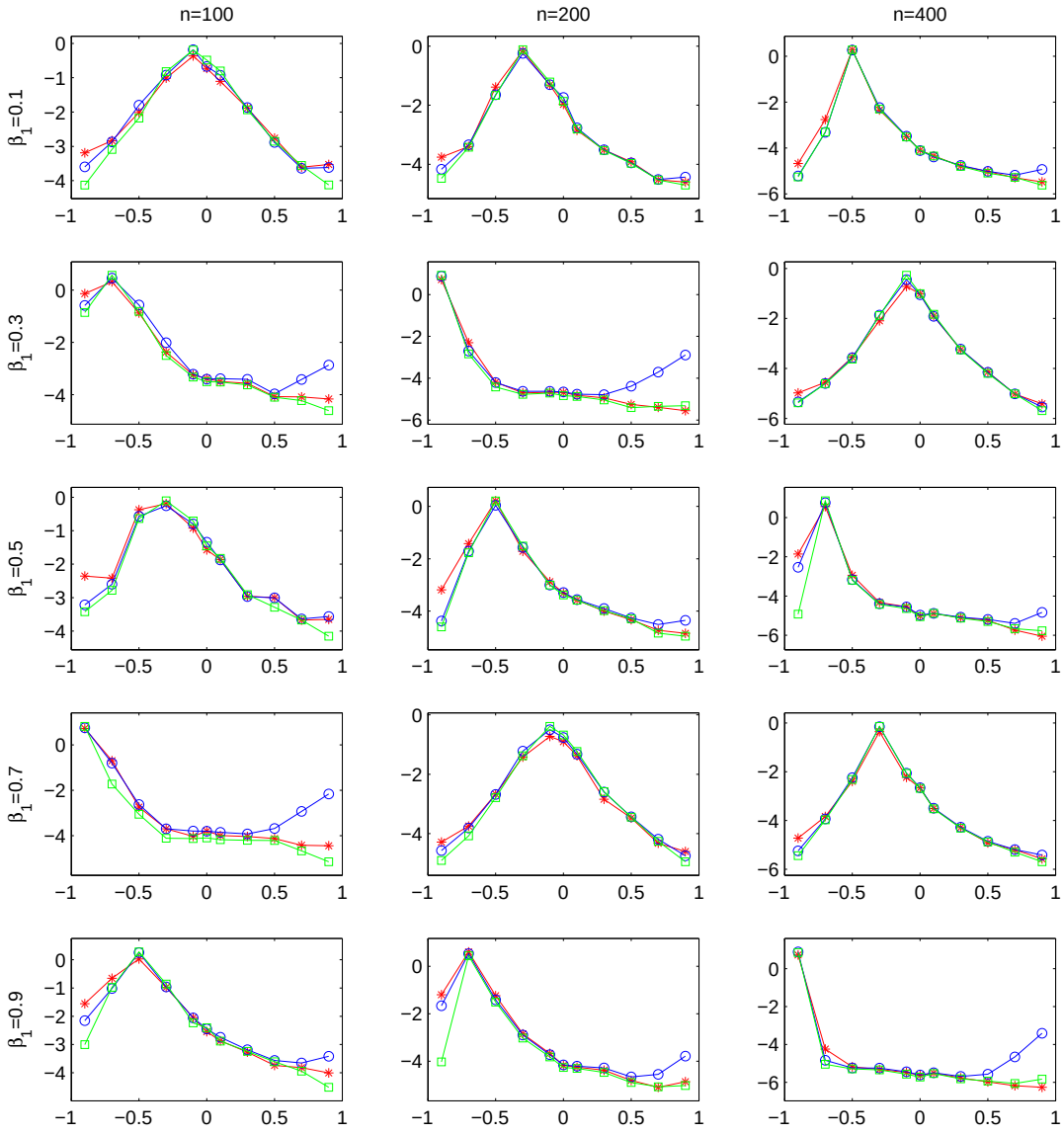

Figure 2.1 Simulation results for ARMA$(1,1)$ models with $\varepsilon_t \sim N(0,1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '□': MLE, red '$*$': XWLE.

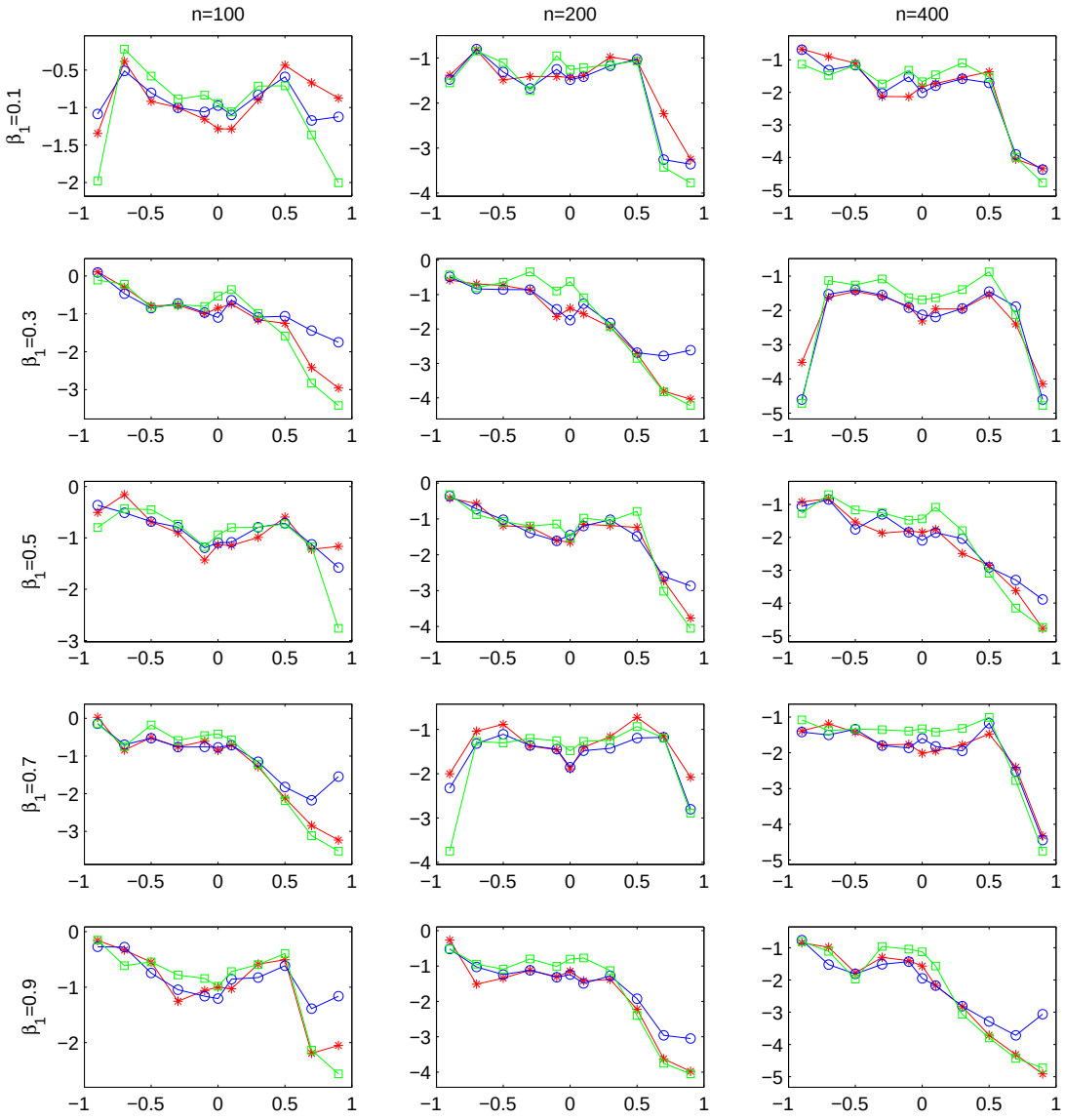

Figure 2.2 Simulation results for ARMA$(2,1)$ models with $\varepsilon_t \sim N(0,1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '$\square$': MLE, red '$*$': XWLE.

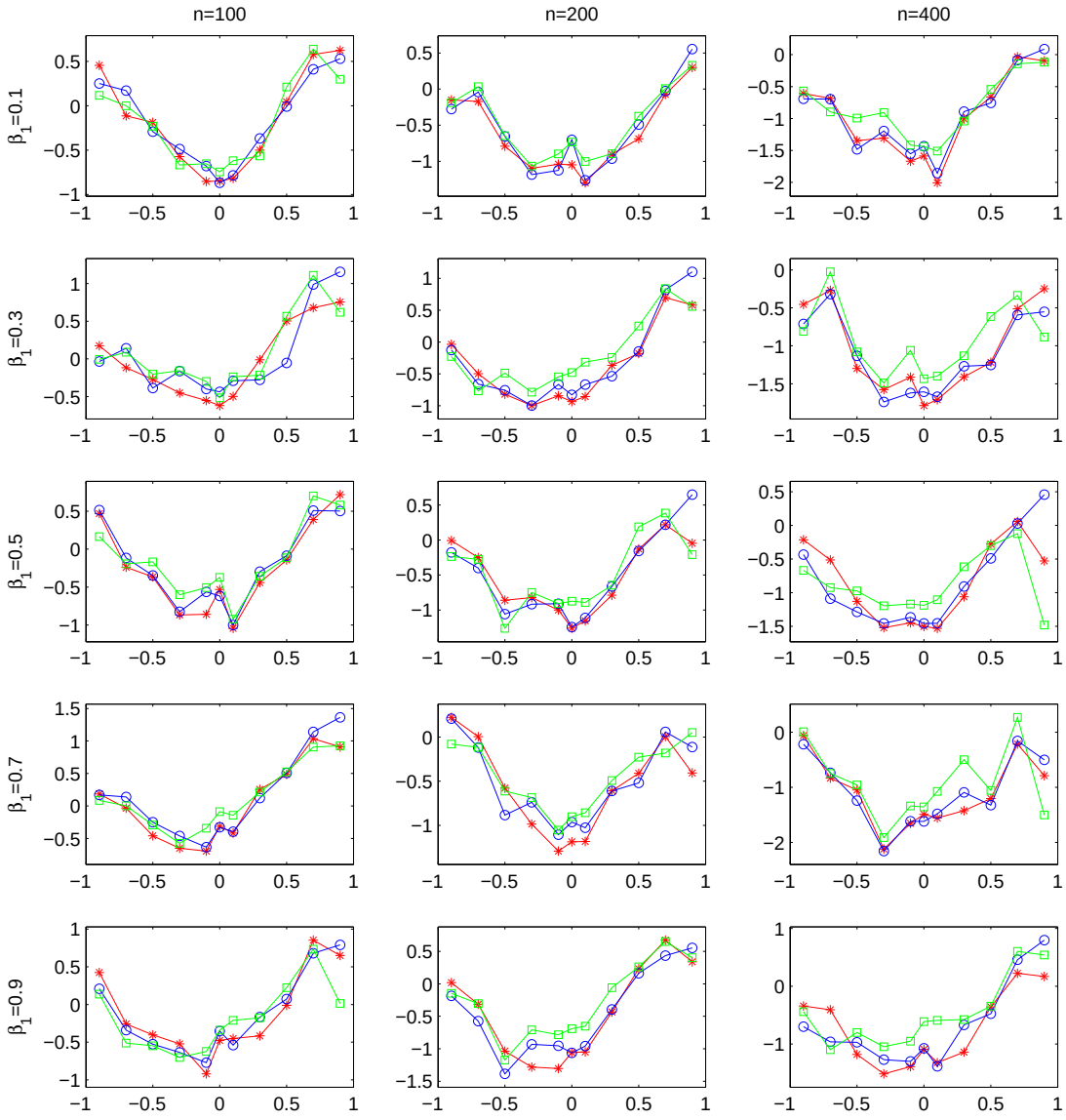

Figure 2.3 Simulation results for ARMA$(5,1)$ models with $\varepsilon_t \sim N(0,1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '□': MLE, red '$*$': XWLE.

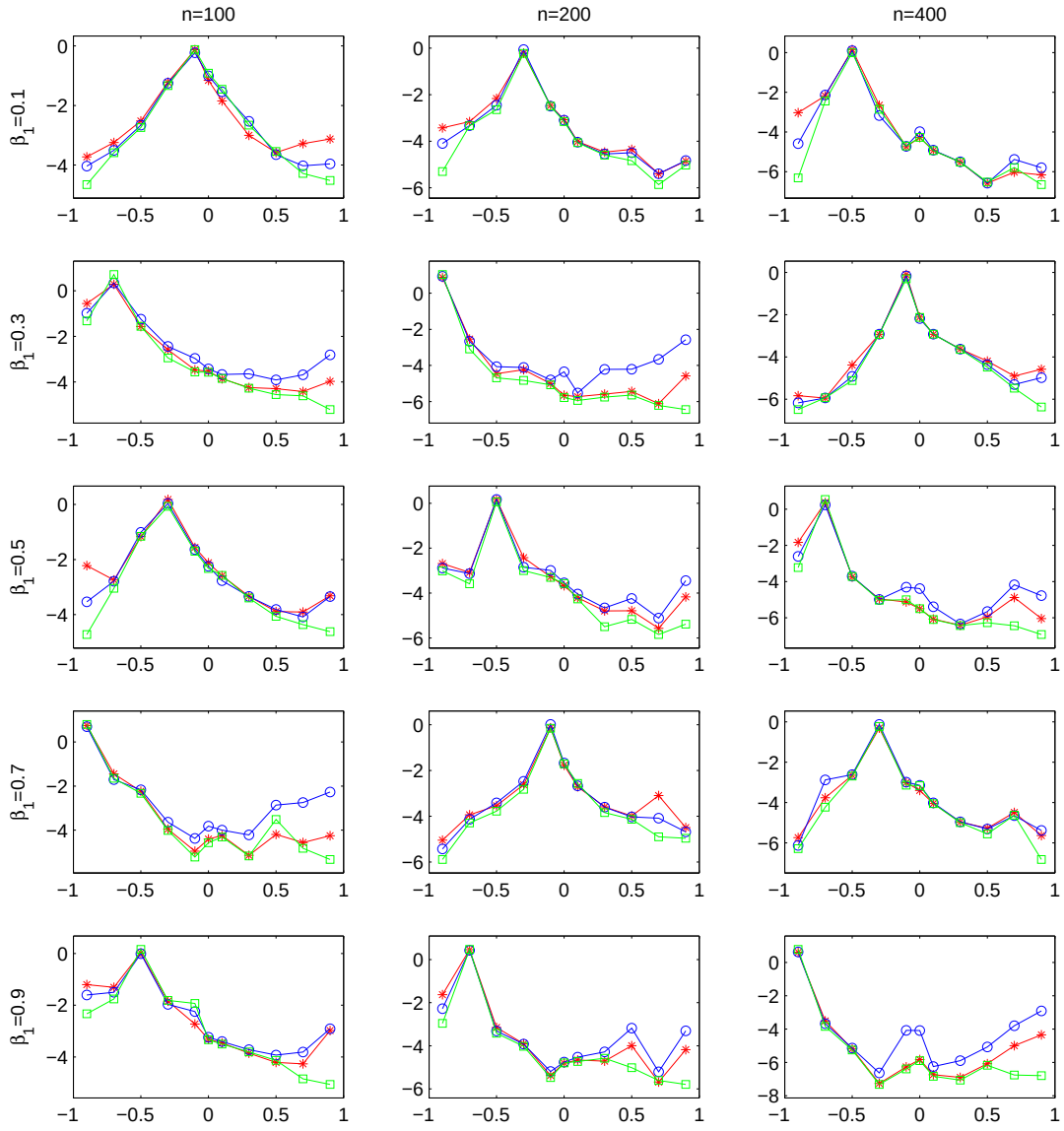

Figure 2.4 Simulation results for ARMA$(1,1)$ models with $\varepsilon_t \sim t(1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '□': MLE, red '$*$': XWLE.

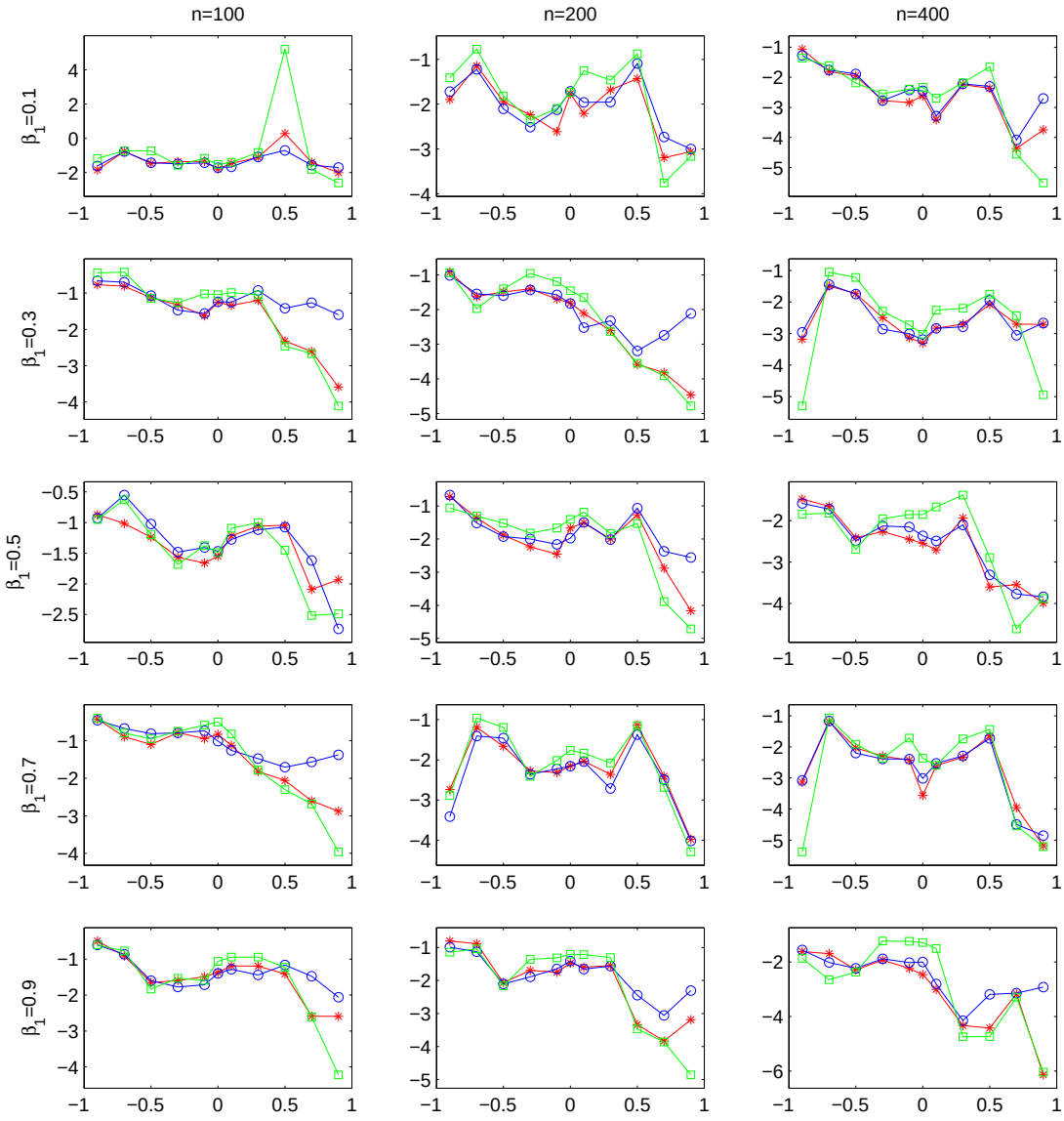

Figure 2.5 Simulation results for ARMA$(2,1)$ models with $\varepsilon_t \sim t(1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '□': MLE, red '$*$': XWLE.

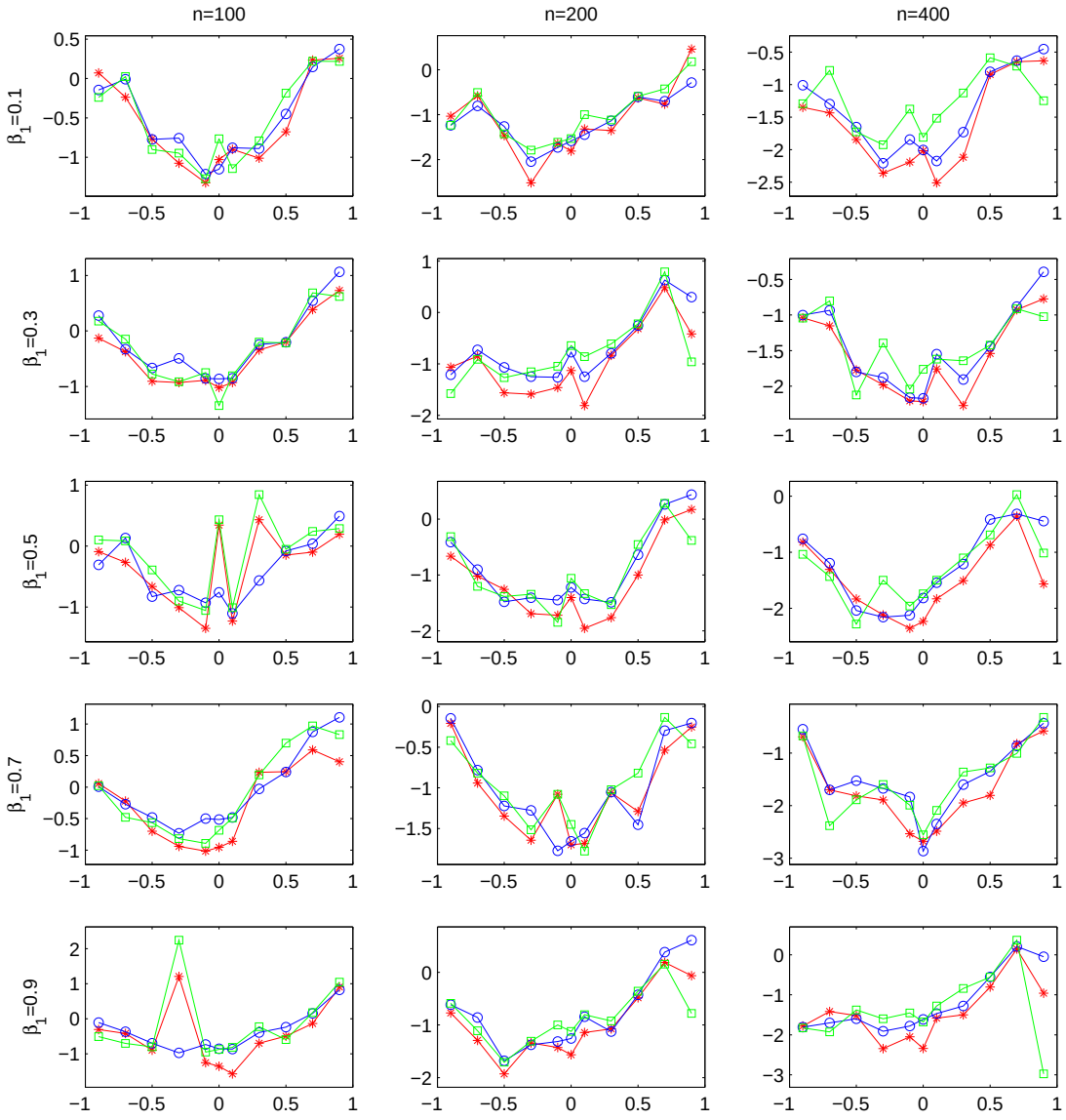

Figure 2.6 Simulation results for ARMA$(5,1)$ models with $\varepsilon_t \sim t(1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '□': MLE, red '$*$': XWLE.

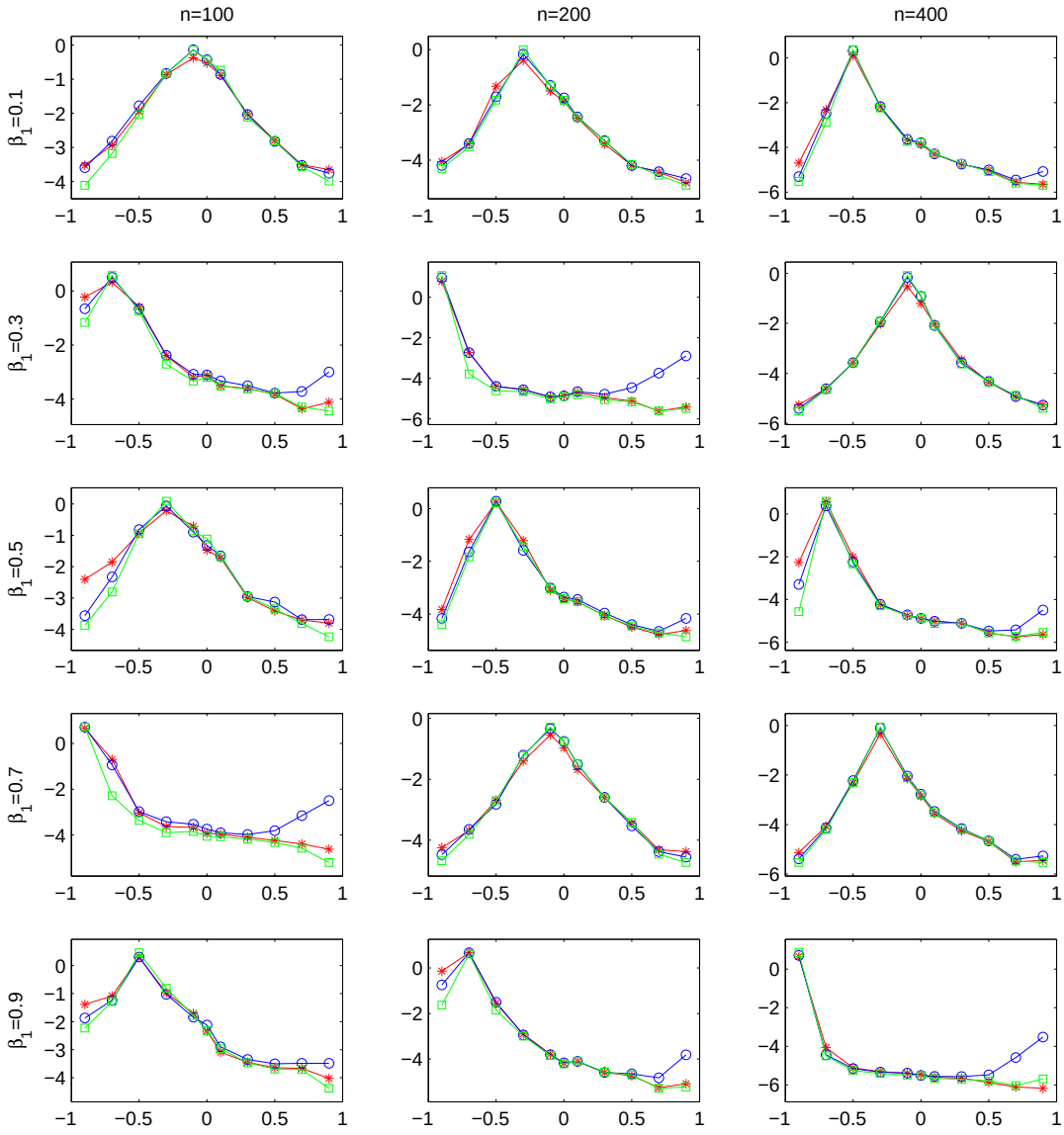

Figure 2.7 Simulation results for ARMA$(1,1)$ models with $\varepsilon_t \sim U(-1,1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '□': MLE, red '$*$': XWLE.

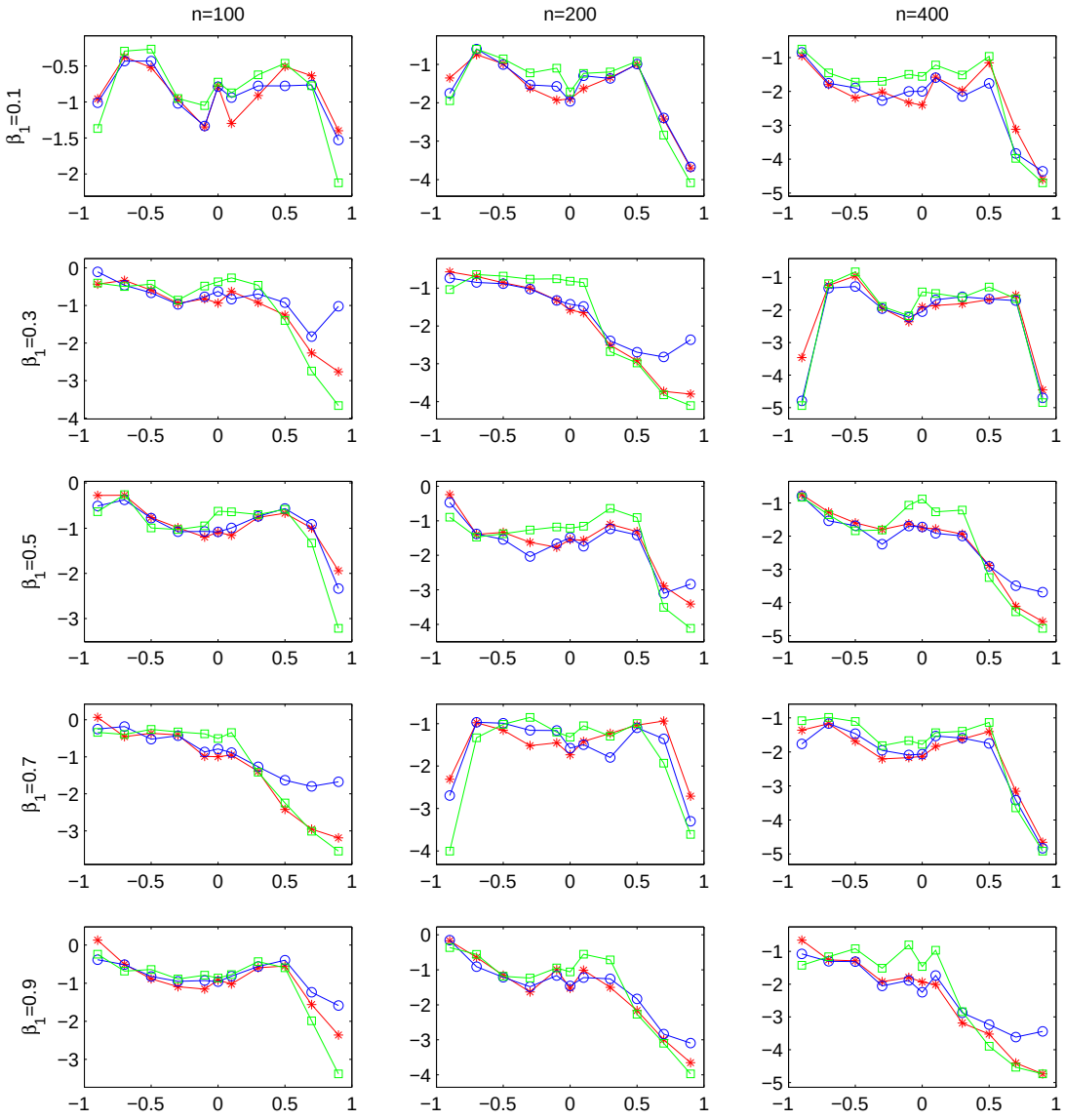

Figure 2.8 Simulation results for ARMA$(2,1)$ models with $\varepsilon_t \sim U(-1,1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '$\square$': MLE, red '$*$': XWLE.

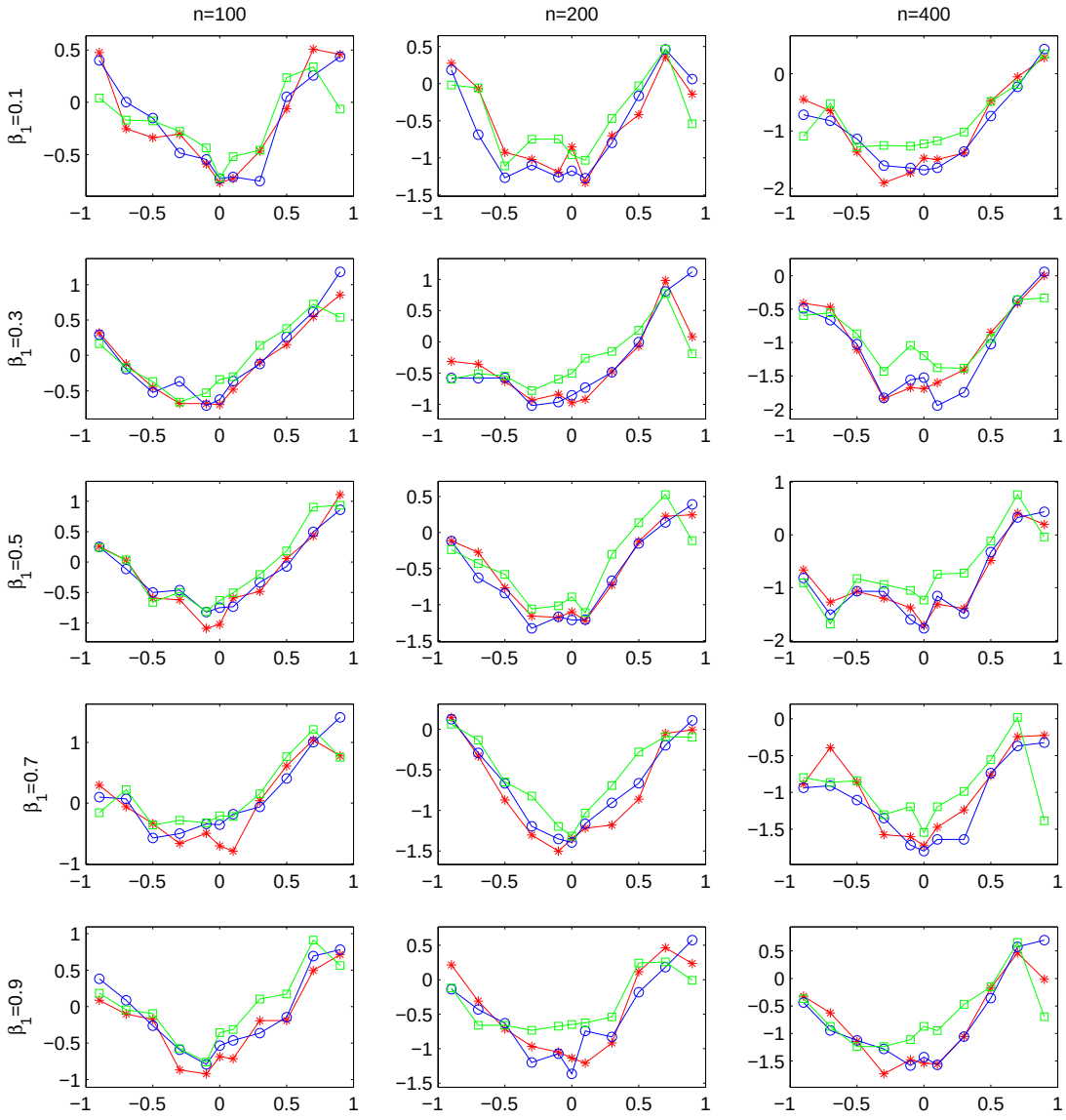

Figure 2.9 Simulation results for ARMA$(5,1)$ models with $\varepsilon_t \sim U(-1,1)$, where y-axes represent log(Err) and x-axes represent θ_1; blue 'o': WLE, green '□': MLE, red '$*$': XWLE.

Example 2.5.2. In this example we study the effects of MA errors on the following logistic smooth threshold AR model (LSTAR) with an MA(1) error, denoted by LSTAR(p)-MA(1),

$$y_t = \alpha_0+\alpha_1 y_{t-1}+...+\alpha_p y_{t-p}+(\beta_0+\beta_1 y_{t-1}+...+\beta_p y_{t-p})\times I_{t-1}+\theta\varepsilon_{t-1}+\varepsilon_t, \quad (2.19)$$

where $\varepsilon_t's$ are i.i.d $N(0,\sigma_\varepsilon^2)$ and $I_{t-1} = 1/(1+\exp\{-\gamma(y_{t-1}-c)\})$. Set $p = 2$, $\alpha_0 = 0$, $\alpha_1 = 1.8$, $\alpha_2 = -1.06$, $\beta_0 = 0.02$, $\beta_1 = -0.9$, $\beta_2 = 0.8$, $c = 0.02$ and $\gamma = 100$, such that the time series is explosive in the lower regime and stationary in the higher regime, but the time series generated is stationary. We consider three values for σ_ε: 0.25, 0.5, and 1, and θ is chosen from the range -0.9 to 0.9.

Based on 200 replications, the average of estimation errors defined in the previous example are summarised in Table 2.1. We can see a clear improvement in parameter estimations as sample size increases, demonstrating the estimation consistency. Denote by LSTAR(p) the classic LSTAR model (Terasvirta, 1994) of order p. In each replication, we fitted the data to 20 models {LSTAR(p)-MA(1), $p = 1,...,10$} and {LSTAR(p), $p = 1,...,10$} respectively, and calculate the BIC_W scores. Table 2.1 reports the proportion of replications that BIC_W attained its minimum at the true model LSTAR(2)-MA(1) among the 20 candidate models. The powers of the LB-test and AN-test in detecting the existence of MA errors are displayed in Figure 2.10 where the y-axis of each panel is the percent of replications of rejecting the null hypothesis that the residuals from the LSTAR(2) model are

Table 2.1 Simulation results for Example 2.5.2.

		$n = 100$			$n = 200$			$n = 400$		
θ		$\sigma_\varepsilon = 0.25$	$\sigma_\varepsilon = 0.5$	$\sigma_\varepsilon = 1$	$\sigma_\varepsilon = 0.25$	$\sigma_\varepsilon = 0.5$	$\sigma_\varepsilon = 1$	$\sigma_\varepsilon = 0.25$	$\sigma_\varepsilon = 0.5$	$\sigma_\varepsilon = 1$
-0.9	Err	0.0767	0.1384	0.1715	0.0378	0.0404	0.0589	0.0134	0.0151	0.0229
	BIC_W	0.71	0.72	0.85	0.73	0.7	0.77	0.8	0.81	0.75
-0.7	Err	0.1352	0.1692	0.2112	0.0447	0.0446	0.0808	0.0165	0.0223	0.0362
	BIC_W	0.8	0.73	0.77	0.92	0.93	0.92	0.98	0.93	1
-0.5	Err	0.1568	0.1845	0.2247	0.0584	0.0630	0.1010	0.0248	0.0386	0.0536
	BIC_W	0.7	0.73	0.74	0.98	0.93	0.95	0.99	0.98	0.97
-0.3	Err	0.1495	0.1708	0.2823	0.0571	0.0698	0.1187	0.0225	0.0328	0.0531
	BIC_W	0.44	0.46	0.57	0.73	0.67	0.68	0.96	0.91	0.98
-0.1	Err	0.1316	0.1570	0.2720	0.0452	0.0716	0.1193	0.0266	0.0294	0.0750
	BIC_W	0.08	0.1	0.1	0.13	0.14	0.17	0.2	0.15	0.21
0	Err	0.0699	0.1049	0.2521	0.0418	0.0534	0.1369	0.0153	0.0253	0.0513
	BIC_W	0.03	0.07	0.04	0.04	0.02	0.04	0.04	0	0.02
0.1	Err	0.1111	0.1589	0.2690	0.0497	0.0544	0.1267	0.0221	0.0306	0.0598
	BIC_W	0.09	0.06	0.06	0.09	0.08	0.06	0.15	0.15	0.19
0.3	Err	0.0817	0.1331	0.2841	0.0404	0.0582	0.1314	0.0205	0.0284	0.0615
	BIC_W	0.42	0.51	0.46	0.82	0.8	0.87	0.96	1	0.96
0.5	Err	0.0765	0.1244	0.2878	0.0341	0.0514	0.1274	0.0173	0.0230	0.0631
	BIC_W	0.89	0.88	0.87	0.97	0.97	0.94	0.99	1	1
0.7	Err	0.0618	0.1108	0.2900	0.0269	0.0479	0.0968	0.0144	0.0224	0.0605
	BIC_W	0.91	0.88	0.84	0.93	0.94	0.97	0.98	0.93	0.96
0.9	Err	0.0558	0.0975	0.2317	0.0219	0.0433	0.0953	0.0111	0.0167	0.0402
	BIC_W	0.73	0.75	0.75	0.79	0.73	0.85	0.8	0.82	0.87

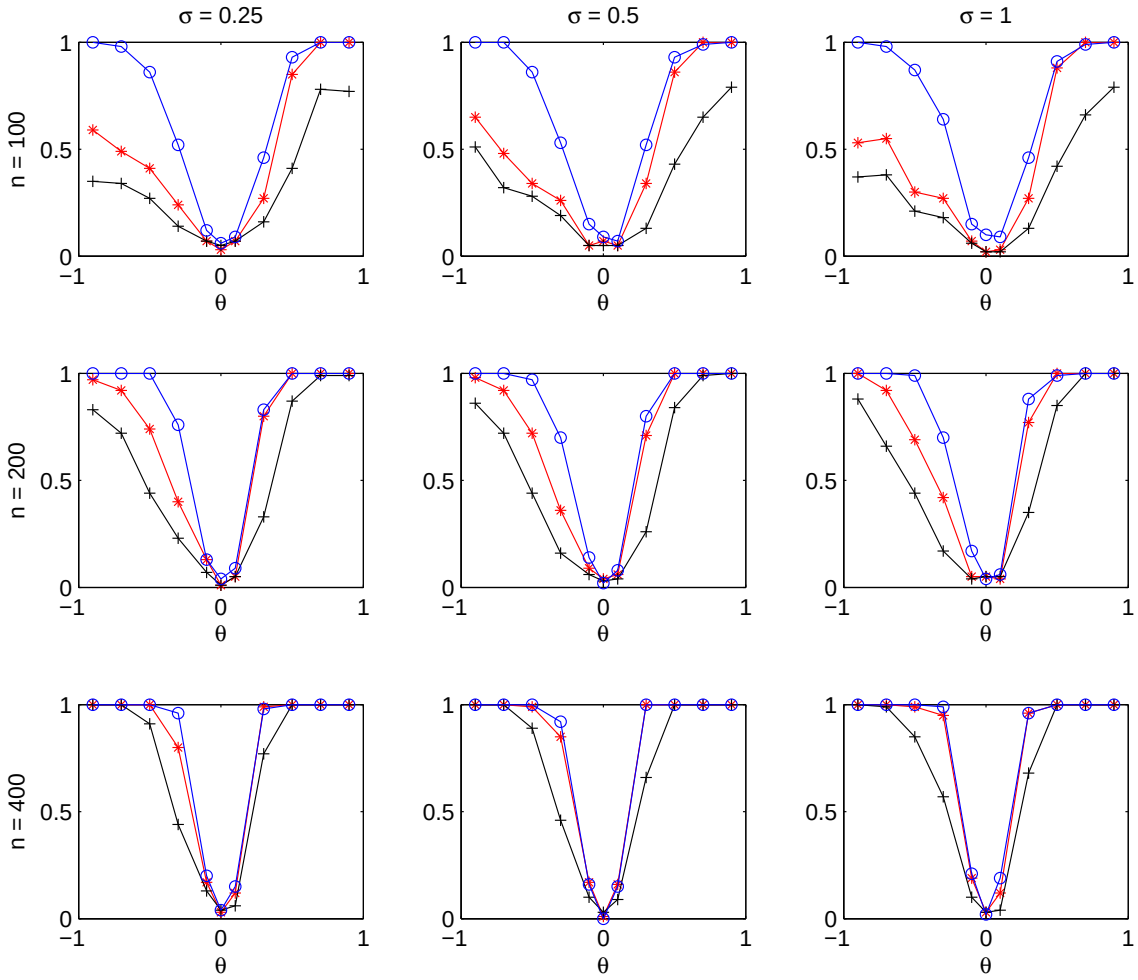

Figure 2.10 Rate of rejections for the LB(20)-tests and AN(20)-tests in Example 2.5.2.

from a white noise process, at level $p_0 = 0.05$. The value of $\mathrm{BIC}_W^{(0)}$ is the percents of replications that the best model chosen by BIC_W is within LSTAR(p)-MA(1) for $p = 1, ..., 10$, i.e., it is the frequency that BIC_W favors the existence of an MA part. In each panel, lines with 'o', '$*$' and '$+$' denote $\mathrm{BIC}_W^{(0)}$, AN(20)-test and

LB(20)-test, respectively. Figure 2.10 shows that AN-test generally has a higher or equivalent power as compared to LB-test when the null hypothesis is false. When the null is true, AN-test also preserves the size well. All in all, BIC_W seems to be the most trust-worthy criterion in detecting the existence of MA errors.

Example 2.5.3. Next we analyze the square root transformed series $y_t = 2(\sqrt{1+x_t} - 1)$ of annual sunspot numbers x_t for the period $1700 - 2012$. The raw data were downloaded from the official website of the Solar Influences Data Analysis Center (SIDC), Brussels, Belgium. The transformed data for the period $1700 - 1979$ have been analyzed in details by Tong (1990) and other researchers. It is believed that this data can be better fitted with a nonlinear time series model; see also Chen and Tsay (1993) for the nonlinearity tests on this data. In this example we try to improve the threshold autoregressive (TAR) model of Tong (1990) by adding an MA term, and estimate the model by XWLE. We shall check the necessity of adding the MA term by the prediction ability of the models.

The TAR model fitted by Tong (1990) has two regimes with lag-11 in one regime and lag-3 in the other. We believe the long AR lags is abundant if a lag-1 MA term is employed. By also taking advantage of the correlation analysis reported in

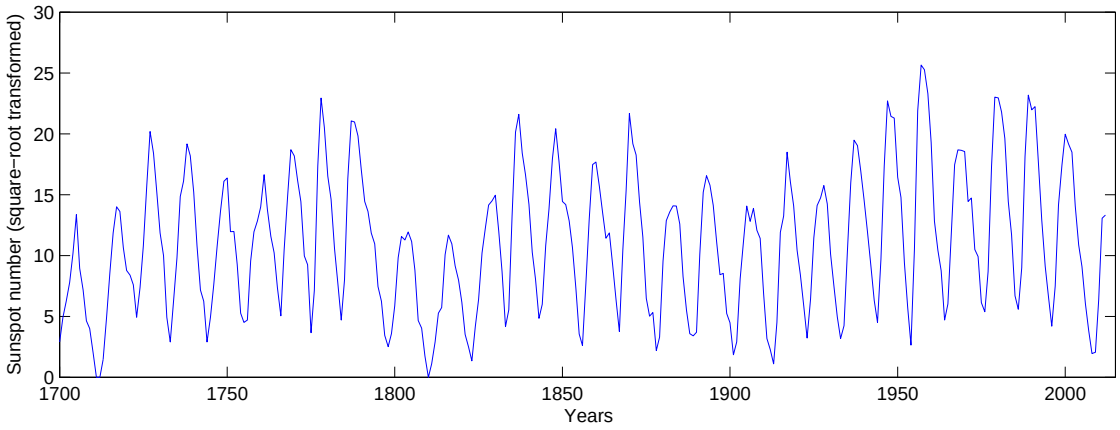

Figure 2.11 Time plots for the transformed sunspot number.

Chen and Tsay (1993), we propose the following TAR-MA model,

$$y_t = \begin{cases} \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_3 y_{t-3} + \alpha_8 y_{t-8} + \varepsilon_t + \theta_1 \varepsilon_{t-1}, & \text{if} \quad y_{t-8} \leq 11.93, \\ \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \beta_8 y_{t-8} + \varepsilon_t + \theta_1 \varepsilon_{t-1}, & \text{if} \quad y_{t-8} > 11.93, \end{cases} \tag{2.20}$$

where we use the same structure parameter as Tong (1990) to facilitate the comparison. We use the same period $1700 - 1979$ for in-sample fitting and reserve the period $1980 - 2012$ for out-of-sample predictions. The XWLE estimates of the parameters are

$$\hat{\theta}_1 = -0.5162,$$

$$\hat{\alpha}_0 = 0.8436,\ \hat{\alpha}_1 = 1.4011,\ \hat{\alpha}_2 = -0.4446,\ \hat{\alpha}_3 = -0.1341,\ \hat{\alpha}_8 = 0.0581,$$

$$\hat{\beta}_0 = 1.4264,\ \hat{\beta}_1 = 1.8603,\ \hat{\beta}_2 = -1.4067,\ \hat{\beta}_3 = -0.4116,\ \hat{\beta}_8 = 0.0437.$$

This model has 10 coefficient parameters plus one structure parameter. The overall

residual variance is 3.817 which is slightly higher than Tong's TAR model (3.734) that has 16 parameters and one structure parameter. Nevertheless, as will be shown later, this model outperforms Tong's TAR model by a significant margin in out-of-sample predictions. The p-values of the LB(20)-test and AN(20)-test on the fitted residuals are respectively 0.0573 and 0.1618, both of which do not reject the white noise hypothesis at level 0.05. The BIC_W of model (2.20) is 1.5599.

For comparison, we also fit a TAR(8) model to the data using XWLE as follows,

$$y_t = \begin{cases} \alpha_0^{(0)} + \alpha_1^{(0)} y_{t-1} + \alpha_2^{(0)} y_{t-2} + \alpha_3^{(0)} y_{t-3} + \alpha_8^{(0)} y_{t-8} + \varepsilon_t & \text{if} \quad y_{t-8} \le 11.93, \\ \\ \beta_0^{(0)} + \beta_1^{(0)} y_{t-1} + \beta_2^{(0)} y_{t-2} + \beta_3^{(0)} y_{t-3} + \beta_8^{(0)} y_{t-8} + \varepsilon_t & \text{if} \quad y_{t-8} > 11.93, \end{cases} \tag{2.21}$$

where

$$\hat{\alpha}_0^{(0)} = 1.4635,\ \hat{\alpha}_1^{(0)} = 0.9950,\ \hat{\alpha}_2^{(0)} = 0.1469,\ \hat{\alpha}_3^{(0)} = -0.3942,\ \hat{\alpha}_8^{(0)} = 0.0643,$$

$$\hat{\beta}_0^{(0)} = 2.5939,\ \hat{\beta}_1^{(0)} = 1.4246,\ \hat{\beta}_2^{(0)} = -0.7932,\ \hat{\beta}_3^{(0)} = 0.0189,\ \hat{\beta}_8^{(0)} = 0.1074.$$

The overall residual variance of model (2.21) is 3.963. The p-values of the LB(20)-test and AN(20)-test on the fitted residuals are respectively 0.0128 and 0.0095, both of which reject the white noise hypothesis at level 0.05. The BIC_W of model (2.21) is 1.5783, which is higher than that of model (2.20). These p-values and BIC_W suggest the necessity of adding the MA term.

Next, we compare the multi-step ahead forecasts of four models: TAR(8)-MA(1) in (2.20), the TAR(8) in (2.21), the TAR(11) of Tong(1990) and the FAR(8) of Chen and Tsay (1993). Based on the estimated models with data from 1700-1979, the h-step ahead prediction for 1980-2012 is made in a rolling approach with $h = 1, 2, ..., 20$. More specifically, for any year t we first predict y_{t+1}, denoted by $\hat{y}_{t+1}$, using previous values $\{y_s, s \leq t\}$, and then predict y_{t+2} using $\{\hat{y}_{t+1}, y_s, s \leq t\}$. The procedure is repeated until the last value y_{t+h} is predicted. We calculate the prediction error for the original numbers of the sunspots by taking the inverse transformation. The results are shown on Figure 2.12. Our TAR-MA model almost dominates the other three models in all the steps h except for $h = 11$ and $h = 12$. It is interesting to note that the TAR(8) model is better than or comparable to Tong's TAR(11) model up to $t = 13$. The usefulness of the higher lags of TAR(11) model starts to appear from the lead-time 14, but it is still less efficient as compared to TAR(8)-MA(1) in (2.20).

Example 2.5.4. El Niño Southern Oscillation (ENSO) is a large-scale medium-frequency event in the equatorial Pacific Ocean that is manifested in an abnormal increase (El Niño) or decrease (La Niña) of the Sea Surface Temperatures (SST). The time series variable representing the ESNO anomaly, *Niño 3.4*, is derived from the index tabulated by the Climate Prediction Center at the National Oceanic and Atmospheric Administration. This index measures the difference in SST in the area of the Pacific Ocean between $5°N - 5°S$ and $170°W - 120°W$ (Trenberth and

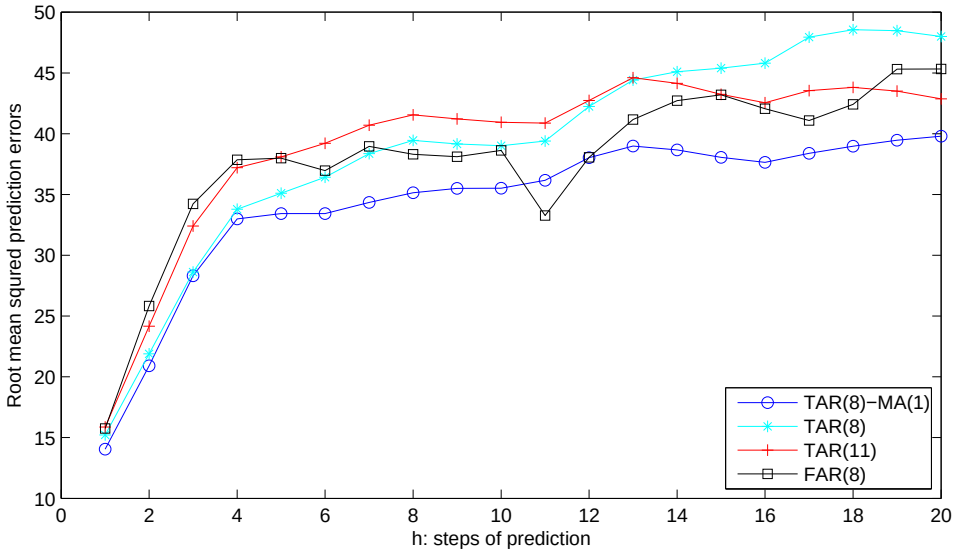

Figure 2.12 Root mean squred prediction errors of out-of-sample multistep forecasts for the original numbers of the sunspots.

Stepanyak, 2001). The SST anomaly is the deviation of the *Niño 3.4* monthly measure from the average historic measure for that particular month from the period 1971-2000.

In this study, we consider the monthly SST anomaly between January 1950 and December 2012. The nonlinearity of the this time series has been tested and validated by Ubilava and Helmers (2013) who propose to fit the data with an LSTAR model. Using data between January 1950 and December 2007, they estimated the optimal lag as 6 based on the classic BIC (Schwarz, 1978). It is interesting to study whether the lag can be shortened by an LSTAR-MA model as

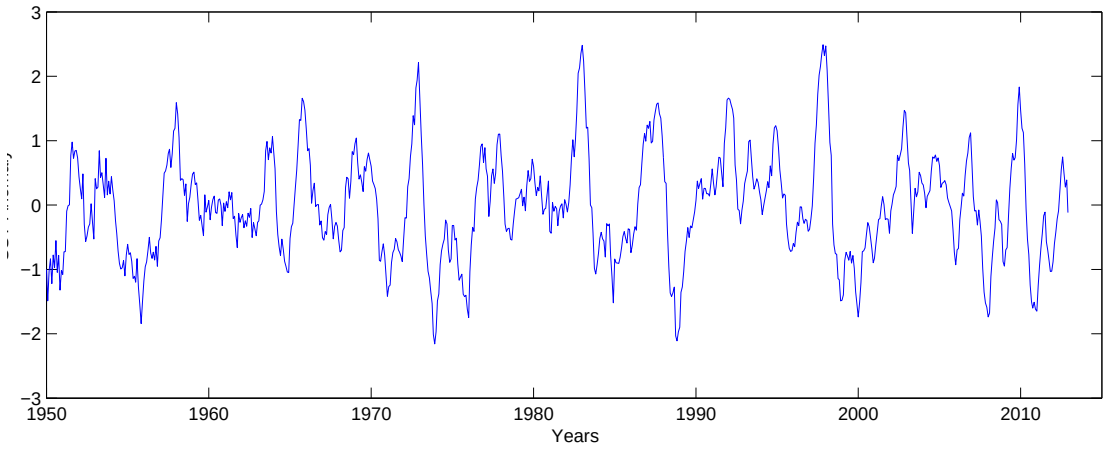

Figure 2.13 Time plots for the Niño 3.4 anomaly.

follows

$$\begin{aligned} y_t &= \beta_{1,0} + \beta_{1,1} y_{t-1} + ... + \beta_{1,p} y_{t-p} + \delta_1^\top D_t \\ &\quad + (\beta_{2,0} + \beta_{2,1} y_{t-1} + ... + \beta_{2,p} y_{t-p} + \delta_2^\top D_t) \times I_t \\ &\quad + \varepsilon_t + \theta_1 \varepsilon_{t-1} \end{aligned} \tag{2.22}$$

where $I_t = (1 + \exp(-\gamma(y_{t-d} - c)))^{-1}$ and $D_t = (D_{t,1}, ..., D_{t,11})^\top$ is a vector of dummy variables for the month. To facilitate the comparison, we use the same structure parameters as used by Ubilava and Helmers (2013), and let

$$I_t = (1 + \exp(-1.196/0.835(y_{t-1} + 0.447)))^{-1},$$

and

$$\delta_1 = (0.114, 0.354, 0.340, 0.177, 0.040, 0.097, -0.036, -0.177, -0.166, -0.370, -0.183)^\top,$$

$$\delta_2 = (-0.159, -0.569, -0.535, -0.269, -0.037, -0.078, 0.061, 0.218, 0.162, 0.651, 0.252)^\top.$$

The BIC_W scores for each choice of p in model (2.22) from 1 to 6 are reported in table 2.2 in which the lag-2 has the smallest value. So with the same data between

Table 2.2 BIC_W scores for the Niño 3.4 SST anomaly data

Lag length	$Q_T(\hat{\sigma}^2)$	BIC_W
1	0.0614	-2.5358
2	0.0552	-2.6242
3	0.0550	-2.6090
4	0.0543	-2.6031
5	0.0536	-2.5966
6	0.0531	-2.5884

January 1950 and December 2007, we obtained the following LSTAR(2)-MA(1) model to fit the data:

$$\begin{aligned} y_t &= -0.0421 + 1.2019y_{t-1} - 0.2207y_{t-2} + \delta_1^\top D_t \\ &\quad + (0.0702 + 0.5692y_{t-1} - 0.6140y_{t-2} + \delta_2^\top D_t) \times I_t \\ &\quad + \varepsilon_t - 0.5530\varepsilon_{t-1}, \end{aligned} \tag{2.23}$$

which significantly reduces the number of parameters as compared to the LSTAR(6) model used by Ubilava and Helmers (2013). The in-sample root mean squared fitting error (0.2350) from model (2.23) is only slightly larger than that obtained from

their LSTAR(6) model (0.2316). The p-values of the LB(20)-test and AN(20)-test on the fitted residuals are respectively 0.1159 and 0.3155, both accept the white noise hypothesis at level 0.05.

For comparison, we also fit an LSTAR(2) model using XWLE as follows

$$\begin{aligned} y_t &= 0.0865 + 1.0629y_{t-1} - 0.0455y_{t-2} + \delta_1^\top D_t \\ &\quad + (-0.1286 + 0.1543y_{t-1} - 0.2082y_{t-2} + \delta_2^\top D_t) \times I_t \\ &\quad + \varepsilon_t. \end{aligned} \tag{2.24}$$

The in-sample root mean squared fitting error of model (2.24) is 0.2432. The p-values of the LB(20)-test and AN(20)-test on the fitted residuals are respectively 4.31×10^{-5} and 5.01×10^{-4}, suggesting strong evidence that model (2.24) is inadequate. Its BIC_W is -2.5647 which is also larger than that of model (2.23).

We use the data from January 2008 to December 2012 to asses the out-of-sample prediction accuracies of three models: the LSTAR(6) of Ubilava and Helmers (2013), model (2.23) (LSTAR(2)-MA(1)) and model (2.24) (LSTAR(2)). The h-step ahead predictions are made in a similar way as the previous example with $h = 1, ..., 36$. The prediction error in Figure 2.14 shows that the LSTAR(2)-MA(1) model is the best among the three up to lead-time 16. After that, there is no much difference between LSTAR(2)-MA(1) and LSTAR(6). The LSTAR(2) model is generally the worst predictor up to lead-time 27 (more than 2 years), which

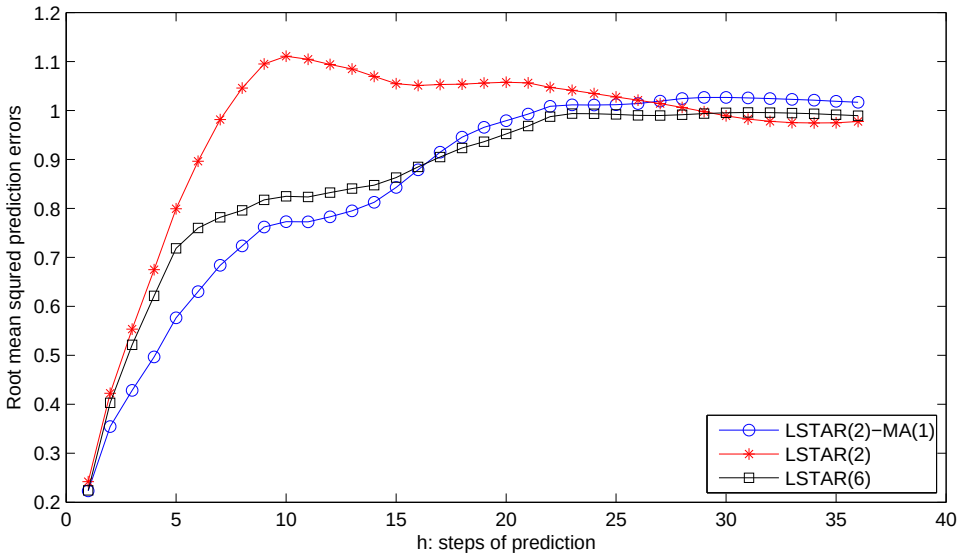

Figure 2.14 Root mean squred prediction errors of out-of-sample multistep forecasts for Niño 3.4 SST anomaly data.

provides a strong proof that the MA part in model (2.23) plays a crucial role in improving the out-of-sample prediction accuracies.

2.6 Asymptotics of XWLE

Let y_t and X_t be two time series satisfying model (2.13), i.e.,

$$y_t = \phi(X_t, \beta) + \xi_t(\theta),$$

where X_t is a vector variable that can be either lags of y_t or a collection of exogenous variables, or both, and $\xi_t(\theta)$ is a moving average (MA) process defined in (2.14).

In the following discussions we shall denote the true parameters by $(\beta_0, \theta_0, \sigma_0^2)$. Define $z_t(\tilde{\beta}) = y_t - \phi(X_t, \tilde{\beta})$. Let $\gamma_z(j;\tilde{\beta}) = cov(z_t(\tilde{\beta}), z_{t-j}(\tilde{\beta}))$. Define the spectral density function of $z_t(\tilde{\beta})$ as

$$k_z(\lambda;\tilde{\beta}) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\gamma_z(j;\tilde{\beta})e^{-ij\lambda}.$$

We need the following assumptions in our theoretical justification of the proposed methods.

(A1) Time series $\{y_t\}$ is stationary with autocovariance function $\gamma_y(k), k = 0, \pm1, \pm2, ...$

(A2) There is a compact parameter space for $\tilde{\beta}$, denoted by $\mathscr{B}$, such that the time series $z_t(\tilde{\beta})$ is stationary and $\sup_{\tilde{\beta}\in\mathscr{B}}\sum_{j=1}^{\infty}|\gamma_z(j;\tilde{\beta})| < \infty$. Moreover, the second order derivatives of $\phi(X_t, \tilde{\beta})$ with respect to $\tilde{\beta}$ exists for $\tilde{\beta} \in \mathscr{B}$.

(A3) The MA part is invertible, i.e., $1 + \sum_j \theta_{0,j}x^j$ has no zeros inside the unit circle, and the parameter space for $\theta_0 = (\theta_{0,1}, ..., \theta_{0,q})$ is Θ.

(A4) Assume the time series can be written as

$$y_t - \psi(\beta, \theta, y_{t-1}, y_{t-2}, ...) = \varepsilon_t.$$

In this form, we further assume y_t has a unique set of parameter values $\beta = \beta_0$ and $\theta = \theta_0$ such that $E(\varepsilon_t\varepsilon_s) = 0$ for $t \neq s$ and σ_0^2 for $t = s$.

(A5) The spectral densities $k_z(\lambda;\beta)$, $k_0(\lambda;\theta)$ and their first order differentials $\partial k_z(\lambda;\beta_0)/\partial\beta$, $\partial k_0(\lambda;\theta_0)/\partial\theta$ belong to the Lipshitz class Λ_α, $\alpha > 1/2$, i.e.,

for example for $k_z(\lambda;\theta_0)$

$$\sup_{\lambda} |k_z(\lambda;\theta_0) - k_z(\lambda+\delta;\theta_0)| = O(\delta^{\alpha}).$$

(A6) The stationary process $\{y_t, \xi_t, \partial\phi(X_t;\beta_0)/\beta, \partial^2\phi(X_t;\beta_0)/\beta^\top\beta\}$ is α-mixing with mixing coefficients α_j satisfying $\sum_{j=1}^{\infty}\alpha_j^{\delta/(2+\delta)} < \infty$ for some $\delta > 0$. Also, $E(|y_t|^{2+\delta}) < \infty$, $E(|\xi_t|^{2+\delta}) < \infty$, $E(\|\partial\phi(X_t;\beta_0)/\beta\|^{2+\delta}) < \infty$, $E(\|\partial^2\phi(X_t;\beta_0)/\beta^\top\beta\}\|^{2+\delta}) < \infty$ and $E(\{|\xi_{t+n}| \cdot \|\partial\phi(X_{t+m};\beta_0)/\beta\|\}^{2+\delta}) \leq K < \infty$ for $m, n \geq 1$.

Assumptions (A1)-(A3) are standard assumptions for time series models. (A4) is equivalent to assuming that the Whittle likelihood below has only one global minimum point. To find the limiting distribution of $(\hat{\beta}_T, \hat{\theta}_T)$, we need an additional condition restricting the smoothness of $k_z(\lambda;\beta)$ and $k_0(\lambda;\theta)$. (A5) is similar to the "Condition B" of Hannan (1973) in requiring higher order smoothness of spectral density functions. (A6) is a common assumption to obtain the limit theorems for α-mixing processes; see for example Fan and Yao (2003, pp. 74). The assumption on $\{\partial^2\phi(X_t;\beta_0)/\beta^\top\beta\}$ is new in nonlinear time series analysis as compared to its linear counterpart in which case $\{\partial^2\phi(X_t;\beta_0)/\beta^\top\beta\}$ is just a zero matrix.

Let X and Y be two real random variables. Define

$$\alpha = \sup_{A\in\sigma(X), B\in\sigma(Y)} |P(A)P(B) - P(AB)|,$$

where $\sigma(X)$ and $\sigma(Y)$ are respectively the σ-algebra for X and Y. The proposition below presents the bound for $\mathrm{Cov}(X, Y)$ in terms of the dependence measure α. Its proof can be found in §1.2.2 of Doukhan (1994).

Proposition 2.6.1. If $E(|X|^p + |Y|^q) < \infty$ for some $p, q > 1$ and $1/p + 1/q < 1$, it holds that

$$|\mathrm{Cov}(X, Y)| \leq 8\alpha^{1/r}\{E|X|^p\}^{1/p}\{E|Y|^q\}^{1/q},$$

where $r = (1 - 1/p - 1/q)^{-1}$.

In this Proposition, the smallest choices for p and q is $(2 + \delta)$ if we let $p = q$, which explains how the parameters are selected in (A6).

Lemma 2.6.2. Under assumptions (A1)-(A3), we have

$$\lim_{T\to\infty} \frac{2\pi}{T} \sum_{j=1}^{T-1} \frac{I_z(\lambda_j; \beta)}{k_0(\lambda_j, \theta)} = \frac{\sigma_0^2}{2\pi} \int_{-\pi}^{\pi} \frac{k_z(\lambda; \beta)}{k_0(\lambda, \theta)} d\lambda,$$

and the convergence is uniformly on $\mathscr{B} \otimes \Theta_\delta$, where $\Theta_\delta = \{\theta : \theta \in \Theta, k_0(\lambda, \theta) \geq \delta > 0, \lambda \in [-\pi, \pi]\}$.

Note that by (A2) and the property of fast Fourier transformation, we have

$$2\pi \sum_{j=1}^{T-1} I_z(\lambda_j, \beta)/T = \sum_{j=1}^{T} z_j(\beta)^2/T$$

which converges almost surely to $E[z_t(\beta)^2](< \infty)$ uniformly for $\beta \in \mathscr{B}$. The proof of Lemma 2.6.2 is similar to that of Lemma 1 of Hannan (1973).

Lemma 2.6.3. Under (A1)-(A4), we have

$$\frac{\sigma_0^2}{2\pi}\int_{-\pi}^{\pi}\frac{k_z(\lambda;\beta)}{k_0(\lambda,\theta)}d\lambda \geq \frac{\sigma_0^2}{2\pi}\int_{-\pi}^{\pi}\frac{k_z(\lambda;\beta_0)}{k_0(\lambda,\theta_0)}d\lambda = \sigma_0^2, \quad (\beta,\theta)\in\mathscr{B}\otimes\Theta,$$

and the equality holds only when $\beta=\beta_0$ and $\theta=\theta_0$.

PROOF: If the integral on the left hand side diverges to $+\infty$, the equality holds. We only consider the case that the integral is finite. By Corollary 7.5.3 of Anderson (1971, pp. 412), $\{\sigma_0^2 k_z(\lambda;\beta)\}/\{2\pi k_0(\lambda,\theta)\}$ can be taken as the spectral density function of the stationary process

$$\varepsilon_t(\beta,\theta) = \{\theta(B)\}^{-1}\{y_t-\phi(X_t,\beta)\} = y_t - \psi(\beta,\theta,y_{t-1},y_{t-2},...)$$

which is the prediction error of model y_t using parameters β and θ. Its prediction variance is not less than the prediction variance under true parameters, i.e.,

$$\sigma^2(\beta,\theta) = \frac{\sigma_0^2}{2\pi}\int_{-\pi}^{\pi}\frac{k_z(\lambda;\beta)}{k_0(\lambda,\theta)}d\lambda \geq \sigma_0^2.$$

Furthermore, assumption (A4) guarantees that the equality holds only at $\beta=\beta_0$ and $\theta=\theta_0$. □

Theorem 2.6.1. Suppose assumptions (A1)-(A6) hold. For the estimator $(\hat{\beta}_T,\hat{\theta}_T)$ in (2.18), we have

$$\lim_{T\to\infty}(\hat{\beta}_T,\hat{\theta}_T) = (\beta_0,\theta_0) \text{ a.s.},$$

$$\lim_{T\to\infty}\hat{\sigma}_T^2 = \lim_{T\to\infty}\frac{1}{T}\sum_{j=1}^{T-1}\frac{I_z(\lambda_j;\hat{\beta}_T)}{k_0(\lambda_j,\hat{\theta}_T)} = \sigma_0^2 \text{ a.s.}$$

PROOF: If $(\hat{\beta}_T, \hat{\theta}_T) \not\to (\beta_0, \theta_0)$, then we can find a subsequence of $\{(\hat{\beta}_T, \hat{\theta}_T)\}$, denoted by $\{(\hat{\beta}_m, \hat{\theta}_m)\}$ such that $(\hat{\beta}_m, \hat{\theta}_m) \to (\beta', \theta') \neq (\beta_0, \theta_0)$ and

$$\varliminf_{m\to\infty} \frac{2\pi}{m} \sum_{j=1}^{m-1} \frac{I_z(\lambda_j; \hat{\beta}_m)}{k_0(\lambda_j, \hat{\theta}_m)} = \frac{\sigma_0^2}{2\pi} \int_{-\pi}^{\pi} \frac{k_z(\lambda; \beta')}{k_0(\lambda, \theta')} d\lambda > \sigma_0^2,$$

where the last inequality follows from Lemma 2.6.3. On the other hand, since $(\hat{\beta}_T, \hat{\theta}_T)$ minimizes $Q_T(\beta, \theta)$, it follows that $Q_T(\hat{\beta}_T, \hat{\theta}_T) \leq Q_T(\beta, \theta)$ for any $(\beta, \theta) \in \mathscr{B} \otimes \Theta$, which implies

$$\varlimsup_{m\to\infty} Q_m(\hat{\beta}_m, \hat{\theta}_m) \leq \varlimsup_{m\to\infty} Q_m(\beta, \theta) = \frac{\sigma_0^2}{2\pi} \int_{-\pi}^{\pi} \frac{k_z(\lambda; \beta)}{k_0(\lambda, \theta)} d\lambda,$$

i.e.

$$\varlimsup_{m\to\infty} Q_m(\hat{\beta}_m, \hat{\theta}_m) \leq \inf_{(\beta,\theta)\in\mathscr{B}\otimes\Theta} \frac{\sigma_0^2}{2\pi} \int_{-\pi}^{\pi} \frac{k_z(\lambda; \beta)}{k_0(\lambda, \theta)} d\lambda = \sigma_0^2.$$

Thus we arrive at a contradiction and $(\beta', \theta') = (\beta_0, \theta_0)$, i.e., any subsequence of $\{(\hat{\beta}_T, \hat{\theta}_T)\}$ must converge to (β_0, θ_0), hence $(\hat{\beta}_T, \hat{\theta}_T)$ converges to (β_0, θ_0) almost surely. As a by product, we have also proved that $\hat{\sigma}_T^2 = Q_T(\hat{\beta}_T, \hat{\theta}_T)$ converges to σ_0^2 almost surely. □

Theorem 2.6.2. Suppose assumptions (A1)-(A6) hold. For the estimator $(\hat{\beta}_T, \hat{\theta}_T)$ in (2.18), we have

$$T^{1/2}\big\{(\hat{\beta}_T, \hat{\theta}_T) - (\beta_0, \theta_0)\big\} \xrightarrow{D} N(\mathbf{0}, \Omega^{-1}\Phi\Omega^{-1}),$$

where

$$\Omega = \frac{\sigma_0^2}{2\pi} \int_{-\pi}^{\pi} f(\lambda; \beta_0, \theta_0) \Big\{ \frac{\partial^2 [f(\lambda; \beta_0, \theta_0)^{-1}]}{\partial(\beta, \theta)^\top \partial(\beta, \theta)} \Big\} d\lambda$$

and

$$\Phi = \begin{pmatrix} \sum_{m,n=-\infty}^{\infty} \varrho_0(n;\theta_0)\varrho_0(m;\theta_0)\Phi_\infty^{(11)}(m,n), & \sum_{m,n=-\infty}^{\infty} 2\varrho_0(n;\theta_0)\Phi_\infty^{(12)}(m,n)\varrho_0'(m;\theta_0) \\ \sum_{m,n=-\infty}^{\infty} 2\varrho_0(n;\theta_0)\{\varrho_0'(m;\theta_0)\}^\top\{\Phi_\infty^{(12)}(m,n)\}^\top, & \Psi \end{pmatrix},$$

with

$$\begin{aligned} \Phi_\infty^{(11)}(m,n) &= \sum_{r=-\infty}^{\infty} \Big\{ E\Big[\frac{\partial\phi_{r+n}}{\partial\beta^\top}\frac{\partial\phi_m}{\partial\beta}\xi_r\xi_0\Big] - E\Big[\frac{\partial\phi_{r+n}}{\partial\beta}\xi_r\Big]^\top E\Big[\frac{\partial\phi_m}{\partial\beta}\xi_0\Big] \\ &\quad + E\Big[\frac{\partial\phi_{r+n}}{\partial\beta^\top}\frac{\partial\phi_0}{\partial\beta}\xi_r\xi_m\Big] - E\Big[\frac{\partial\phi_{r+n}}{\partial\beta}\xi_r\Big]^\top E\Big[\frac{\partial\phi_0}{\partial\beta}\xi_m\Big] \\ &\quad + E\Big[\frac{\partial\phi_r}{\partial\beta^\top}\frac{\partial\phi_m}{\partial\beta}\xi_{r+n}\xi_0\Big] - E\Big[\frac{\partial\phi_r}{\partial\beta}\xi_{r+n}\Big]^\top E\Big[\frac{\partial\phi_m}{\partial\beta}\xi_0\Big] \\ &\quad + E\Big[\frac{\partial\phi_r}{\partial\beta^\top}\frac{\partial\phi_0}{\partial\beta}\xi_{r+n}\xi_m\Big] - E\Big[\frac{\partial\phi_r}{\partial\beta}\xi_{r+n}\Big]^\top E\Big[\frac{\partial\phi_0}{\partial\beta}\xi_m\Big]\Big\}, \end{aligned}$$

$$\begin{aligned} \Phi_\infty^{(12)}(m,n) &= 2\sum_{r=-\infty}^{\infty} \Big\{ E\Big[\frac{\partial\phi_{r+n}}{\partial\beta^\top}\xi_m\xi_r\xi_0\Big] - E\Big[\frac{\partial\phi_{r+n}}{\partial\beta}\xi_r\Big]^\top E\Big[\xi_m\xi_0\Big] \\ &\quad + E\Big[\frac{\partial\phi_r}{\partial\beta^\top}\xi_0\xi_{r+n}\xi_m\Big] - E\Big[\frac{\partial\phi_r}{\partial\beta}\xi_{r+n}\Big]^\top E\Big[\xi_0\xi_m\Big]\Big\}, \end{aligned}$$

$\partial\phi_m/\partial\beta = \partial\phi(X_m;\beta_0)/\partial\beta$, and $\varrho_0(n;\theta_0)$ and $\varrho_0'(m;\theta_0)$ are respectively the Fourier coefficients of $k_0(\lambda;\theta)^{-1}$ and $\partial k_0(\lambda;\theta_0)^{-1}/\partial\theta$,

$$\Psi := [\psi_{ij}] = \Big[\sum_{r=1}^{2\times q}\rho_r(i)\rho_r(j)\Big]_{i,j=1,\ldots,q},$$

$\rho_r(i) = \rho(r+i) + \rho(r-i) - 2\rho(r)\rho(i)$ and $\rho(r) = \mathrm{corr}(\xi_t, \xi_{t+r})$.

PROOF: Note that $\partial Q_T(\hat\beta_T, \hat\theta_T)/\partial(\beta,\theta) = \mathbf{0}$. By Taylor's expansion, we have

$$T^{1/2}\frac{\partial Q_T(\beta_0,\theta_0)}{\partial(\beta,\theta)} = -\Big\{\frac{\partial^2 Q_T(\tilde\beta_T,\tilde\theta_T)}{\partial(\beta,\theta)^\top\partial(\beta,\theta)}\Big\}T^{1/2}\big\{(\hat\beta_T,\hat\theta_T) - (\beta_0,\theta_0)\big\},$$

where $\|(\tilde{\beta}_T, \tilde{\theta}_T) - (\beta_0, \theta_0)\| < \|(\hat{\beta}_T, \hat{\theta}_T) - (\beta_0, \theta_0)\|$, i.e., $(\tilde{\beta}_T, \tilde{\theta}_T)$ converges to (β_0, θ_0). It can be calculated that

$$\frac{\partial Q_T(\beta_0, \theta_0)}{\partial(\beta, \theta)} = \frac{2\pi}{T} \sum_{t=1}^{T-1} \Big(\frac{\partial I_z(\lambda_t; \beta_0)/\partial\beta}{k_0(\lambda_t; \theta)},\ I_z(\lambda_t, \beta_0) \frac{\partial [k_0(\lambda_t, \theta_0)^{-1}]}{\partial\theta} \Big)$$

which is taken to be a row vector, and that

$$\frac{\partial^2 Q_T(\beta_0, \theta_0)}{\partial(\beta, \theta)^\top \partial(\beta, \theta)} = \frac{2\pi}{T} \sum_{t=1}^{T-1} \begin{pmatrix} \Sigma_{1t} & D_t^\top \\ D_t & \Sigma_{2t} \end{pmatrix} \tag{2.25}$$

where

$$\Sigma_{1t} = \frac{\partial^2 I_z(\lambda_t; \beta_0)/\partial\beta^2}{k_0(\lambda_t; \theta)},$$
$$\Sigma_{2t} = I_z(\lambda_t, \beta_0) \frac{\partial^2 [k_0(\lambda_t, \theta_0)^{-1}]}{\partial\theta^\top \partial\theta},$$

and

$$D_t = \Big(\frac{\partial I_z(\lambda_t; \beta_0)}{\partial\beta} \Big)^\top \frac{\partial [k_0(\lambda_t, \theta_0)^{-1}]}{\partial\theta}.$$

To finish the proof, we need to prove the convergence of the Hessian matrix in (2.25), and the asymptotic normality of $T^{1/2} \partial Q_T(\beta_0, \theta_0)/\{\partial(\beta, \theta)\}$. The last diagonal block on the right hand size of (2.25), i.e. $2\pi T^{-1} \sum_{t=1}^{T-1} \Sigma_{2T}$, appeared in the Whittle likelihood estimation, converges to

$$\Sigma_2 = \frac{\sigma_0^2}{2\pi} \int_{-\pi}^{\pi} k_z(\lambda; \beta_0) \Big\{ \frac{\partial^2 [k_0(\lambda_t, \theta_0)^{-1}]}{\partial\theta^\top \partial\theta} \Big\} d\lambda;$$

see for example Hannan (1973). The other two matrices are new which only appear in our estimation method.

By the second formula (2.15) for $I_z(\lambda;\beta)$, we have

$$\begin{aligned}\frac{\partial I_z(\lambda;\beta)}{\partial\beta} &= \frac{1}{2\pi}\sum_{n=-T+1}^{T-1}\frac{\partial c_z(n;\beta)}{\partial\beta}e^{-in\lambda}\\ &= \frac{1}{2\pi}\sum_{n=-T+1}^{T-1}\Big[\frac{1}{T}\sum_{t=1}^{T-n}\frac{\partial z_{t+n}(\beta)}{\partial\beta}z_t(\beta)+\frac{1}{T}\sum_{t=1}^{T-n}z_{t+n}(\beta)\frac{\partial z_t(\beta)}{\partial\beta}\Big]e^{-in\lambda}.\end{aligned}$$

Thus

$$\frac{\partial I_z(\lambda;\beta_0)}{\partial\beta}=-\frac{1}{2\pi}\sum_{n=-T+1}^{T-1}\Big[\frac{1}{T}\sum_{t=1}^{T-n}\frac{\partial\phi(X_{t+n};\beta_0)}{\partial\beta}\xi_t+\frac{1}{T}\sum_{t=1}^{T-n}\xi_{t+n}\frac{\partial\phi(X_t;\beta_0)}{\partial\beta}\Big]e^{-in\lambda}.$$

Moreover,

$$\begin{aligned}\frac{\partial^2 I_z(\lambda;\beta)}{\partial\beta^\top\partial\beta} &= \frac{1}{2\pi}\sum_{n=-T+1}^{T-1}\frac{\partial^2 c_z(n;\beta)}{\partial\beta^\top\partial\beta}e^{-in\lambda}\\ &= \frac{1}{2\pi}\sum_{n=-T+1}^{T-1}\Big[-\frac{1}{T}\sum_{t=1}^{T-n}\frac{\partial^2\phi(X_{t+n};\beta)}{\partial\beta^\top\partial\beta}z_t(\beta)-\frac{1}{T}\sum_{t=1}^{T-n}z_{t+n}(\beta)\frac{\partial^2\phi(X_t;\beta)}{\partial\beta^\top\partial\beta}\\ &\qquad+\frac{1}{T}\sum_{t=1}^{T-n}\frac{\partial\phi(X_{t+n};\beta)}{\partial\beta^\top}\frac{\partial\phi(X_t;\beta)}{\partial\beta}\\ &\qquad+\frac{1}{T}\sum_{t=1}^{T-n}\frac{\partial\phi(X_t;\beta)}{\partial\beta^\top}\frac{\partial\phi(X_{t+n};\beta)}{\partial\beta}\Big]e^{-in\lambda},\end{aligned}$$

i.e.

$$\begin{aligned}\frac{\partial^2 I_z(\lambda;\beta_0)}{\partial\beta^\top\partial\beta} &= \frac{1}{2\pi}\sum_{n=-T+1}^{T-1}\Big[-\frac{1}{T}\sum_{t=1}^{T-n}\frac{\partial^2\phi(X_{t+n};\beta_0)}{\partial\beta^\top\partial\beta}\xi_t-\frac{1}{T}\sum_{t=1}^{T-n}\xi_{t+n}\frac{\partial^2\phi(X_t;\beta_0)}{\partial\beta^\top\partial\beta}\\ &\qquad+\frac{1}{T}\sum_{t=1}^{T-n}\frac{\partial\phi(X_{t+n};\beta_0)}{\partial\beta^\top}\frac{\partial\phi(X_t;\beta_0)}{\partial\beta}\\ &\qquad+\frac{1}{T}\sum_{t=1}^{T-n}\frac{\partial\phi(X_t;\beta_0)}{\partial\beta^\top}\frac{\partial\phi(X_{t+n};\beta_0)}{\partial\beta}\Big]e^{-in\lambda}.\end{aligned}$$

Since $z_t=y_t-\phi(X_t;\beta)=\xi_t+\phi(X_t;\beta_0)-\phi(X_t;\beta)$, letting $\Delta(X_t;\beta)=\phi(X_t;\beta_0)-\phi(X_t;\beta)$, we have

$$\gamma_z(k;\beta) = \operatorname{cov}(\xi_t+\Delta(X_t;\beta),\ \xi_{t+k}+\Delta(X_t;\beta))$$

$$
\begin{aligned}
&= \operatorname{cov}(\xi_k, \xi_{t+k}) + \operatorname{cov}(\xi_t, \Delta(X_{t+k};\beta)) + \operatorname{cov}(\Delta(X_t;\beta), \xi_{t+k}) \\
&\quad + \operatorname{cov}(\Delta(X_t;\beta), \Delta(X_{t+k};\beta)).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\frac{\partial \gamma_z(k;\beta)}{\partial \beta} &= \frac{\partial \operatorname{cov}(\xi_t, \Delta(X_{t+k};\beta))}{\partial \beta} + \frac{\partial \operatorname{cov}(\Delta(X_t;\beta), \xi_{t+k})}{\partial \beta} + \frac{\partial \operatorname{cov}(\Delta(X_t;\beta), \Delta(X_{t+k};\beta))}{\partial \beta} \\
&= \operatorname{cov}\Big(\xi_t, \ \frac{\partial \Delta(X_{t+k};\beta)}{\partial \beta}\Big) + \operatorname{cov}\Big(\frac{\partial \Delta(X_t;\beta)}{\partial \beta}, \xi_{t+k}\Big) \\
&\quad + \operatorname{cov}\Big(\frac{\partial \Delta(X_t;\beta)}{\partial \beta}, \Delta(X_{t+k};\beta)\Big) + \operatorname{cov}\Big(\Delta(X_t;\beta), \frac{\partial \Delta(X_{t+k};\beta)}{\partial \beta}\Big)
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial^2 \gamma_z(k;\beta)}{\partial \beta^\top \partial \beta} &= \operatorname{cov}\Big(\xi_t, \ \frac{\partial^2 \Delta(X_{t+k};\beta)}{\partial \beta^\top \partial \beta}\Big) + \operatorname{cov}\Big(\frac{\partial^2 \Delta(X_t;\beta)}{\partial \beta^\top \partial \beta}, \xi_{t+k}\Big) \\
&\quad + \operatorname{cov}\Big(\frac{\partial^2 \Delta(X_t;\beta)}{\partial \beta^\top \partial \beta}, \Delta(X_{t+k};\beta)\Big) + \operatorname{cov}\Big(\frac{\partial \Delta(X_t;\beta)}{\partial \beta}, \frac{\partial \Delta(X_{t+k};\beta)}{\partial \beta}\Big) \\
&\quad + \operatorname{cov}\Big(\Delta(X_t;\beta), \frac{\partial^2 \Delta(X_{t+k};\beta)}{\partial \beta^\top \partial \beta}\Big) + \operatorname{cov}\Big(\frac{\partial \Delta(X_t;\beta)}{\partial \beta}, \frac{\partial \Delta(X_{t+k};\beta)}{\partial \beta}\Big)^\top.
\end{aligned}
$$

Noting that $\Delta(X_t;\beta_0) = 0$ and $\partial \Delta(X_t;\beta)/\partial \beta = -\partial \phi(X_t;\beta)/\partial \beta$, we have

$$
\gamma_z(k;\beta) = \operatorname{cov}(\xi_k, \xi_{t+k}) = E(\xi_k \xi_{t+k}),
$$

$$
\begin{aligned}
\frac{\partial \gamma_z(k;\beta_0)}{\partial \beta} &= -\operatorname{cov}\Big(\xi_t, \ \frac{\partial \phi(X_{t+k};\beta_0)}{\partial \beta}\Big) - \operatorname{cov}\Big(\frac{\partial \phi(X_t;\beta_0)}{\partial \beta}, \xi_{t+k}\Big) \\
&= -E\Big(\xi_t \ \frac{\partial \phi(X_{t+k};\beta_0)}{\partial \beta}\Big) - E\Big(\frac{\partial \phi(X_t;\beta_0)}{\partial \beta} \xi_{t+k}\Big),
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 \gamma_z(k;\beta_0)}{\partial \beta^\top \partial \beta} &= -\operatorname{cov}\Big(\xi_t, \ \frac{\partial^2 \phi(X_{t+k};\beta_0)}{\partial \beta^\top \partial \beta}\Big) - \operatorname{cov}\Big(\frac{\partial^2 \phi(X_t;\beta_0)}{\partial \beta^\top \partial \beta}, \xi_{t+k}\Big) + \\
&\quad \operatorname{cov}\Big(\frac{\partial \phi(X_t;\beta_0)}{\partial \beta}, \frac{\partial \phi(X_{t+k};\beta_0)}{\partial \beta}\Big) + \operatorname{cov}\Big(\frac{\partial \phi(X_t;\beta_0)}{\partial \beta}, \frac{\partial \phi(X_{t+k};\beta_0)}{\partial \beta}\Big)^\top
\end{aligned}
$$

$$= -E\Big(\xi_t \frac{\partial^2\phi(X_{t+k};\beta_0)}{\partial\beta^\top\partial\beta}\Big) - E\Big(\frac{\partial^2\phi(X_t;\beta_0)}{\partial\beta^\top\partial\beta}\xi_{t+k}\Big) + E\Big(\frac{\partial\phi(X_t;\beta_0)}{\partial\beta^\top}\frac{\partial\phi(X_{t+k};\beta_0)}{\partial\beta}\Big) + E\Big(\frac{\partial\phi(X_{t+k};\beta_0)}{\partial\beta^\top}\frac{\partial\phi(X_t;\beta_0)}{\partial\beta}\Big).$$

By (A6), it can be proved that,

$$\sum_{n=0}^{\infty}\Big\|\frac{\partial^l\gamma_z(n;\beta_0)}{\partial\beta^l}\Big\| < \infty$$

for $l = 0, 1, 2$, where $\partial^0\gamma_z(n;\beta_0)/\partial\beta^0 := \gamma_z(n;\beta_0)$. By the convergence theory of trigonometric series and the uniqueness of the Fourier representation, we have

$$\frac{\partial^l k_z(\lambda;\beta_0)}{\partial\beta^l} = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\frac{\partial^l\gamma_z(n;\beta_0)}{\partial\beta^l}e^{-ij\lambda}, \quad \text{for } l = 0, 1, 2,$$

where $\partial\beta^2$ stands for $\partial\beta^\top\partial\beta$. It is not difficult to see that as $T \to \infty$

$$\begin{aligned}
\sqrt{T}\Big\{\frac{T}{T-|k|}c_z(k;\beta_0) - \gamma_z(k;\beta_0)\Big\} &\to N\Big(0,\ \Xi_k^{(0)}(\beta_0,\theta_0)\Big),\\
\sqrt{T}\Big\{\frac{T}{T-|k|}\frac{\partial c_z(k;\beta_0)}{\partial\beta} - \frac{\partial\gamma_z(k;\beta_0)}{\partial\beta}\Big\} &\to N\Big(\mathbf{0}_{1\times p},\ \Xi_k^{(1)}(\beta_0,\theta_0)\Big),\\
\sqrt{T}\Big\{\frac{T}{T-|k|}\frac{\partial^2 c_z(k;\beta_0)}{\partial\beta^\top\partial\beta} - \frac{\partial^2\gamma_z(k;\beta_0)}{\partial\beta^\top\partial\beta}\Big\} &\to N\Big(\mathbf{0}_{p\times p},\ \Xi_k^{(2)}(\beta_0,\theta_0)\Big),
\end{aligned}$$

which means $\partial^l I_z(\lambda;\beta_0)/\partial\beta^l$ is connected with $\partial^l k_z(\lambda;\beta_0)/\partial\beta^l$ in a similar way for $l = 1, 2$ as for $l = 0$. Following almost the same proof of Lemma 2.6.2, $2\pi T^{-1}\sum_{t=1}^{T-1}\Sigma_{1t}$ and $2\pi T^{-1}\sum_{t=1}^{T-1}D_t$ converge to respectively

$$\Sigma_1 = \frac{\sigma_0^2}{2\pi}\int_{-\pi}^{\pi}\frac{\partial^2 k_z(\lambda_t,\beta_0)/\partial\beta^2}{k_0(\lambda;\theta_0)}d\lambda$$

and

$$D = \frac{\sigma_0^2}{2\pi}\int_{-\pi}^{\pi}\left(\frac{\partial k_z(\lambda_t,\beta_0)}{\partial\beta}\right)^\top\frac{\partial[k_0(\lambda_t,\theta_0)^{-1}]}{\partial\theta}d\lambda.$$

As such we have

$$\frac{\partial^2 Q_T(\beta_0,\theta_0)}{\partial(\beta,\theta)^\top\partial(\beta,\theta)} \to \begin{pmatrix} \Sigma_1 & D^\top \\ D & \Sigma_2 \end{pmatrix} = \Omega \ \text{ a.s. as } T\to\infty.$$

Let $f(\lambda;\beta,\theta) = k_z(\lambda;\beta)/k_0(\lambda;\theta)$, which is the spectral density function of $\{1-\theta(B)\}^{-1}\{y_t-\phi(X_t,\beta)\}$. It is easy to verify that

$$\begin{aligned}\Omega &= \frac{\sigma_0^2}{2\pi}\int_{-\pi}^{\pi} f(\lambda;\beta_0,\theta_0)\Big\{\frac{\partial^2[f(\lambda;\beta_0,\theta_0)^{-1}]}{\partial(\beta,\theta)^\top\partial(\beta,\theta)}\Big\}d\lambda \\ &= \frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\frac{\partial\log f(\lambda;\beta_0,\theta_0)}{\partial(\beta,\theta)}\right)\left(\frac{\partial\log f(\lambda;\beta_0,\theta_0)}{\partial(\beta,\theta)}\right)^\top d\lambda,\end{aligned}$$

where the last equation follows from Whittle (1951). It is worthy pointing out that the Ω shares the same form as the covariance matrix for Whittle's estimator in the ARMA model; see Hannan (1973) for details.

Next we prove the normality of

$$\begin{aligned}T^{1/2}\frac{\partial Q_T(\beta_0,\theta_0)}{\partial(\beta,\theta)} &= \frac{2\pi}{T^{1/2}}\sum_{t=1}^{T-1}\Big(\frac{\partial I_z(\lambda_t;\beta_0)/\partial\beta}{k_0(\lambda_t;\theta_0)},\ I_z(\lambda_t,\beta_0)\frac{\partial[k_0(\lambda_t,\theta_0)^{-1}]}{\partial\theta}\Big) \\ &\stackrel{def}{=} (A_{1T},A_{2T}),\end{aligned}$$

where the second term A_{2T} exists in linear Whittle's estimation, which is proved to be asymptotically standard normal with limit variance-covariance matrix

$$\lim_{T\to\infty} E\{A_{2T}^\top A_{2T}\} = \Psi := [\psi_{ij}]_{i,j=1,\ldots,q},$$

where $\psi_{ij} = \sum_{r=1}^{2\times q}\rho_r(i)\rho_r(j)$, $\rho_r(i) = \rho(r+i)+\rho(r-i)-2\rho(r)\rho(i)$ and $\rho(r) = \text{corr}(\xi_t,\xi_{t+r})$; see the Theorem 3 of Hannan and Heyde (1972) and the Theorem 2 of Hannan (1973) for details.

We then prove the asymptotic normality of A_{1T}. By (A5) and (A6), we have

$$A_{1T} = \sqrt{T}\Big\{\frac{2\pi}{T}\sum_{t=1}^{T-1}\frac{\partial I_z(\lambda_t;\beta_0)/\partial\beta}{k_0(\lambda_t;\theta_0)}\Big\} = \sqrt{T}\int_{-\pi}^{\pi}\frac{\partial I_z(\lambda;\beta_0)}{\partial\beta}k_0(\lambda;\theta_0)^{-1}d\lambda + O(T^{1/2-\alpha}),$$

where α is defined in (A5). Following from Lemma 2.6.3, since (β_0, θ_0) minimizes the integral, the corresponding derivative is a zero vector, i.e. $\int_{-\pi}^{\pi}\frac{\partial k_z(\lambda,\beta_0)/\partial\beta}{k_0(\lambda;\theta_0)}d\lambda = \mathbf{0}$. So we can write A_{1T} as

$$A_{1T} = \sqrt{T}\int_{-\pi}^{\pi}\Big\{\frac{\partial I_z(\lambda;\beta_0)}{\partial\beta} - \frac{\partial k_z(\lambda;\beta_0)}{\partial\beta}\Big\}k_0(\lambda;\theta_0)^{-1}d\lambda + O(T^{1/2-\alpha}).$$

Let $q_T(\lambda;\theta) > 0$ be the Cesàro sum of the Fourier series of $k_0(\lambda;\theta)^{-1}$ taken to T terms. Then by (A5), we have (see Zygmund (1959), pp. 91)

$$\sup_{\lambda}|k_0(\lambda;\theta_0)^{-1} - q_T(\lambda;\theta_0)| < O(T^{-\alpha}),$$

and

$$A_{1T} = \sqrt{T}\int_{-\pi}^{\pi}\Big\{\frac{\partial I_z(\lambda;\beta_0)}{\partial\beta} - \frac{\partial k_z(\lambda;\beta_0)}{\partial\beta}\Big\}q_T(\lambda;\theta_0)d\lambda + O(T^{1/2-\alpha}).$$

The Cesàro sum of the Fourier series of $\partial k_z(\lambda;\beta_0)/\partial\beta$, denoted by $k'_\beta(\lambda;\beta_0)$ for short, taken to T terms is

$$S_c[k'_\beta] = \frac{1}{2\pi}\sum_{n=-T+1}^{T-1}\Big(1 - \frac{|n|}{T}\Big)\frac{\partial\gamma_z(n;\beta_0)}{\partial\beta}e^{-in\lambda}.$$

Similarly, we have

$$\sup_{\lambda}\|k'_\beta(\lambda;\beta_0) - S_c[k'_\beta]\| < O(T^{-\alpha}),$$

thus

$$\begin{aligned} A_{1T} &= \frac{T^{1/2}}{2\pi}\int_{-\pi}^{\pi}\sum_{n=-T+1}^{T-1}\Big\{\frac{\partial c_z(n;\beta_0)}{\partial\beta}-\Big(1-\frac{|n|}{T}\Big)\frac{\partial\gamma_z(n;\beta_0)}{\partial\beta}\Big\}e^{-in\lambda}q_T(\lambda;\theta_0)d\lambda+O(T^{1/2-\alpha}) \\ &= T^{1/2}\sum_{n=-T+1}^{T-1}\Big\{\frac{\partial c_z(n;\beta_0)}{\partial\beta}-\Big(1-\frac{|n|}{T}\Big)\frac{\partial\gamma_z(n;\beta_0)}{\partial\beta}\Big\}\Big(1-\frac{|n|}{T}\Big)\varrho_0(n;\theta_0)+O(T^{1/2-\alpha}). \end{aligned}$$

Define

$$c_z'(n)=\frac{\partial c_z(n;\beta_0)}{\partial\beta}=-\frac{1}{T}\sum_{t\in\mathscr{T}_n}\frac{\partial\phi(X_{t+n};\beta_0)}{\partial\beta}\xi_t-\frac{1}{T}\sum_{t\in\mathscr{T}_n}\xi_{t+n}\frac{\partial\phi(X_t;\beta_0)}{\partial\beta} \qquad (2.26)$$

and

$$\tilde\gamma_z'(n)=\Big(1-\frac{|n|}{T}\Big)\frac{\partial\gamma_z(n;\beta_0)}{\partial\beta}=-\frac{1}{T}\sum_{t\in\mathscr{T}_n}E\Big[\frac{\partial\phi(X_{t+n};\beta_0)}{\partial\beta}\xi_t\Big]-\frac{1}{T}\sum_{t\in\mathscr{T}_n}E\Big[\xi_{t+n}\frac{\partial\phi(X_t;\beta_0)}{\partial\beta}\Big], \qquad (2.27)$$

where $\mathscr{T}_n=\{1,2,...,T-n\}$ if $n\geq 0$ and $\mathscr{T}_n=\{-n,1-n,...,T\}$ if $n<0$. Since

$$\begin{aligned} T\times\mathrm{cov}(c_z'(n),c_z'(m)) &= T\Big\{E[c_z'(n)^\top c_z'(m)]-\tilde\gamma_z'(n)^\top\tilde\gamma_z'(m)\Big\} \\ &= \frac{1}{T}\sum_{t\in\mathscr{T}_n}\sum_{s\in\mathscr{T}_m}\Big\{E\Big[\frac{\partial\phi_{t+n}}{\partial\beta^\top}\frac{\partial\phi_{s+m}}{\partial\beta}\xi_t\xi_s\Big]-E\Big[\frac{\partial\phi_{t+n}}{\partial\beta}\xi_t\Big]^\top E\Big[\frac{\partial\phi_{s+m}}{\partial\beta}\xi_s\Big] \\ &\quad+E\Big[\frac{\partial\phi_{t+n}}{\partial\beta^\top}\frac{\partial\phi_s}{\partial\beta}\xi_t\xi_{s+m}\Big]-E\Big[\frac{\partial\phi_{t+n}}{\partial\beta}\xi_t\Big]^\top E\Big[\frac{\partial\phi_s}{\partial\beta}\xi_{s+m}\Big] \\ &\quad+E\Big[\frac{\partial\phi_t}{\partial\beta^\top}\frac{\partial\phi_{s+m}}{\partial\beta}\xi_{t+n}\xi_s\Big]-E\Big[\frac{\partial\phi_t}{\partial\beta}\xi_{t+n}\Big]^\top E\Big[\frac{\partial\phi_{s+m}}{\partial\beta}\xi_s\Big] \\ &\quad+E\Big[\frac{\partial\phi_t}{\partial\beta^\top}\frac{\partial\phi_s}{\partial\beta}\xi_{t+n}\xi_{s+m}\Big]-E\Big[\frac{\partial\phi_t}{\partial\beta}\xi_{t+n}\Big]^\top E\Big[\frac{\partial\phi_s}{\partial\beta}\xi_{s+m}\Big]\Big\}, \end{aligned}$$

by changing the suffixes ($r=t-s$), we have

$$\begin{aligned} &T\times\mathrm{cov}(c_z'(n),c_z'(m)) \\ &=\frac{1}{T}\sum_{r\in\mathscr{R}_{m,n}}\sum_{s\in\mathscr{S}_{m,n,r}}\Big\{E\Big[\frac{\partial\phi_{s+r+n}}{\partial\beta^\top}\frac{\partial\phi_{s+m}}{\partial\beta}\xi_{s+r}\xi_s\Big]-E\Big[\frac{\partial\phi_{s+r+n}}{\partial\beta^\top}\xi_{s+r}\Big]E\Big[\frac{\partial\phi_{s+m}}{\partial\beta}\xi_s\Big] \end{aligned}$$

$$
\begin{aligned}
&+ E\Big[\frac{\partial \phi_{s+r+n}}{\partial \beta^\top}\frac{\partial \phi_s}{\partial \beta}\xi_{s+r}\xi_{s+m}\Big] - E\Big[\frac{\partial \phi_{s+r+n}}{\partial \beta^\top}\xi_{s+r}\Big]E\Big[\frac{\partial \phi_s}{\partial \beta}\xi_{s+m}\Big] \\
&+ E\Big[\frac{\partial \phi_{s+r}}{\partial \beta^\top}\frac{\partial \phi_{s+m}}{\partial \beta}\xi_{s+r+n}\xi_s\Big] - E\Big[\frac{\partial \phi_{s+r}}{\partial \beta^\top}\xi_{s+r+n}\Big]E\Big[\frac{\partial \phi_{s+m}}{\partial \beta}\xi_s\Big] \\
&+ \ E\Big[\frac{\partial \phi_{s+r}}{\partial \beta^\top}\frac{\partial \phi_s}{\partial \beta}\xi_{s+r+n}\xi_{s+m}\Big] - E\Big[\frac{\partial \phi_{s+r}}{\partial \beta^\top}\xi_{s+r+n}\Big]E\Big[\frac{\partial \phi_s}{\partial \beta}\xi_{s+m}\Big]\Big\},
\end{aligned}
$$

where $\#\mathscr{S}_{m,n,r} = T - |n| - |r|$ and

$$
\mathscr{R}_{m,n} = \begin{cases}
\{-(T-m-1), ..., T-n-1\}, & \text{for } m \geq 0,\ n \geq 0, \\
\{-(T-m+n), ..., T-1\}, & \text{for } m \geq 0,\ n < 0, \\
\{-(T-1), ..., T-n+m\}, & \text{for } m < 0,\ n \geq 0, \\
\{-(T+n), ..., T+m\}, & \text{for } m < 0,\ n < 0.
\end{cases}
$$

By stationarity of $\partial \phi_t/\partial \beta$ and ξ_t, we have

$$
\begin{aligned}
&T \times \operatorname{cov}(c'_z(n), c'_z(m)) \\
&= \sum_{r \in \mathscr{R}_{m,n}} \frac{T-|n|-|r|}{T}\Big\{E\Big[\frac{\partial \phi_{r+n}}{\partial \beta^\top}\frac{\partial \phi_m}{\partial \beta}\xi_r\xi_0\Big] - E\Big[\frac{\partial \phi_{r+n}}{\partial \beta}\xi_r\Big]^\top E\Big[\frac{\partial \phi_m}{\partial \beta}\xi_0\Big] \\
&\quad + E\Big[\frac{\partial \phi_{r+n}}{\partial \beta^\top}\frac{\partial \phi_0}{\partial \beta}\xi_r\xi_m\Big] - E\Big[\frac{\partial \phi_{r+n}}{\partial \beta}\xi_r\Big]^\top E\Big[\frac{\partial \phi_0}{\partial \beta}\xi_m\Big] \\
&\quad + E\Big[\frac{\partial \phi_r}{\partial \beta^\top}\frac{\partial \phi_m}{\partial \beta}\xi_{r+n}\xi_0\Big] - E\Big[\frac{\partial \phi_r}{\partial \beta}\xi_{r+n}\Big]^\top E\Big[\frac{\partial \phi_m}{\partial \beta}\xi_0\Big] \\
&\quad + \ E\Big[\frac{\partial \phi_r}{\partial \beta^\top}\frac{\partial \phi_0}{\partial \beta}\xi_{r+n}\xi_m\Big] - E\Big[\frac{\partial \phi_r}{\partial \beta}\xi_{r+n}\Big]^\top E\Big[\frac{\partial \phi_0}{\partial \beta}\xi_m\Big]\Big\} \\
&:= \sum_{r \in \mathscr{R}_{m,n}} \frac{T-|n|-|r|}{T}\Phi_T(m,n,r).
\end{aligned}
$$

Following by (A6) and Proposition 2.6.1, we have

$$
\sum_{r \in \mathscr{R}_{m,n}} |\Phi_T(m,n,r)| \ \leq \ \sum_{r \in \mathscr{R}_{m,n}} \Big\{\Big|E\Big[\frac{\partial \phi_{r+n}}{\partial \beta^\top}\frac{\partial \phi_m}{\partial \beta}\xi_r\xi_0\Big] - E\Big[\frac{\partial \phi_{r+n}}{\partial \beta}\xi_r\Big]^\top E\Big[\frac{\partial \phi_m}{\partial \beta}\xi_0\Big]\Big|
$$

$$
\begin{aligned}
&+ \left| E\Big[\frac{\partial \phi_{r+n}}{\partial \beta^\top}\frac{\partial \phi_0}{\partial \beta}\xi_r\xi_m\Big] - E\Big[\frac{\partial \phi_{r+n}}{\partial \beta}\xi_r\Big]^\top E\Big[\frac{\partial \phi_0}{\partial \beta}\xi_m\Big]\right| \\
&+ \left| E\Big[\frac{\partial \phi_r}{\partial \beta^\top}\frac{\partial \phi_m}{\partial \beta}\xi_{r+n}\xi_0\Big] - E\Big[\frac{\partial \phi_r}{\partial \beta}\xi_{r+n}\Big]^\top E\Big[\frac{\partial \phi_m}{\partial \beta}\xi_0\Big]\right| \\
&+ \left| E\Big[\frac{\partial \phi_r}{\partial \beta^\top}\frac{\partial \phi_0}{\partial \beta}\xi_{r+n}\xi_m\Big] - E\Big[\frac{\partial \phi_r}{\partial \beta}\xi_{r+n}\Big]^\top E\Big[\frac{\partial \phi_0}{\partial \beta}\xi_m\Big]\right|\Bigg\} \\
\le & \sum_{r\in\mathscr{R}_{m,n}} 8\alpha_{r-|n-m|}^{\delta/(2+\delta)}\Bigg\{\Big\{E\Big|\frac{\partial \phi_{r+n}}{\partial \beta^\top}\xi_r\Big|^{(2+\delta)}E\Big|\frac{\partial \phi_m}{\partial \beta}\xi_0\Big|^{(2+\delta)}\Big\}^{1/(2+\delta)} \\
&+ \Big\{E\Big|\frac{\partial \phi_{r+n}}{\partial \beta^\top}\xi_r\Big|^{(2+\delta)}E\Big|\frac{\partial \phi_0}{\partial \beta}\xi_m\Big|^{(2+\delta)}\Big\}^{1/(2+\delta)} \\
&+ \Big\{E\Big|\frac{\partial \phi_r}{\partial \beta^\top}\xi_{r+n}\Big|^{(2+\delta)}E\Big|\frac{\partial \phi_m}{\partial \beta}\xi_0\Big|^{(2+\delta)}\Big\}^{1/(2+\delta)} \\
&+ \Big\{E\Big|\frac{\partial \phi_r}{\partial \beta^\top}\xi_{r+n}\Big|^{(2+\delta)}E\Big|\frac{\partial \phi_0}{\partial \beta}\xi_m\Big|^{(2+\delta)}\Big\}^{1/(2+\delta)}\Bigg\} \\
= & \sum_{r\in\mathscr{R}_{m,n}} 16\alpha_{r-|n-m|}^{\delta/(2+\delta)}\Bigg\{\Big\{E\Big|\frac{\partial \phi_n}{\partial \beta^\top}\xi_0\Big|^{(2+\delta)}E\Big|\frac{\partial \phi_m}{\partial \beta}\xi_0\Big|^{(2+\delta)}\Big\}^{1/(2+\delta)} \\
&+ \Big\{E\Big|\frac{\partial \phi_0}{\partial \beta^\top}\xi_n\Big|^{(2+\delta)}E\Big|\frac{\partial \phi_0}{\partial \beta}\xi_m\Big|^{(2+\delta)}\Big\}^{1/(2+\delta)}\Bigg\},
\end{aligned}
$$

where $|\cdot|$, "$X^{2+\delta}$" and "$\le$" operate on each component of the matrixes.

By (A6), $\sum_{r\in\mathscr{R}_{m,n}}|\Phi_T(m,n,r)|$ is a convergent summation as $T\to\infty$ uniformly for m and n, i.e., there exists a constant K_0, such that

$$
\sum_{r\in\mathscr{R}_{m,n}}|\Phi_T(m,n,r)| \le K_0 < \infty \text{ for } m,n \ge 1. \tag{2.28}
$$

Consequently, $\sum_{r\in\mathscr{R}_{m,n}}\Phi_T(m,n,r)$, denoted by $\Phi_T^{(11)}(m,n)$, is a convergent summation. Then we have (see Zygmund(1959), p. 77)

$$
T\mathrm{cov}(c_z'(n),c_z'(m)) = \Phi_T^{(11)}(m,n) + o(1). \tag{2.29}
$$

To employ the small-block and large-block arguments, we first note that by construction of the Cesàro sum, we have for any ε, there exists an M, such that for any $T > M$,

$$\sum_{t=M+1}^{T} \Big(1 - \frac{|t|}{T}\Big)|\varrho_0(t;\theta_0)| < \varepsilon/2.$$

Then, we partition the set $\{-T+1, ..., 0, ..., T-1\}$ into two subsets $S_M = \{-M, ..., 0, ..., M\}$ and $S_{T\setminus M} = \{-T+1, ..., -M-1, M+1, ..., T-1\}$, such that,

$$\sum_{t\in S_{T\setminus M}} \Big(1 - \frac{|t|}{T}\Big)|\varrho_0(t;\theta_0)| < \varepsilon. \tag{2.30}$$

Based on (2.28), (2.29) and (2.30), letting $\kappa(n) = \Big(1 - \frac{|n|}{T}\Big)\varrho_0(n;\theta_0)$ and

$$\tilde{c}_z'(n) = \frac{\partial c_z(n;\beta_0)}{\partial\beta} - \Big(1 - \frac{|n|}{T}\Big)\frac{\partial\gamma_z(n;\beta_0)}{\partial\beta},$$

it follows that

$$\begin{aligned}
& E\left|T^{1/2}\sum_{n\in S_{T\setminus M}}\Big\{\frac{\partial c_z(n;\beta_0)}{\partial\beta} - \Big(1-\frac{|n|}{T}\Big)\frac{\partial\gamma_z(n;\beta_0)}{\partial\beta}\Big\}\Big(1-\frac{|n|}{T}\Big)\varrho_0(n;\theta_0)\right| \\
\leq & \left[E\left(T^{1/2}\sum_{n\in S_{T\setminus M}}\tilde{c}_z'(n)\kappa(n)\right)^2\right]^{1/2} \\
= & \left[E\left(\sum_{n\in S_{T\setminus M}}\sum_{m\in S_{T\setminus M}} T\tilde{c}_z'(n)\tilde{c}_z'(m)\kappa(n)\kappa(m)\right)\right]^{1/2} \\
\leq & \left[\left(\sum_{n\in S_{T\setminus M}}\sum_{m\in S_{T\setminus M}} K_0|\kappa(n)\kappa(m)|\right)\right]^{1/2} \leq \left[K_0\varepsilon^2\right]^{1/2} = \sqrt{K_0}\varepsilon.
\end{aligned}$$

Thus, to study the asymptotic property of A_{1T}, we only need to consider

$$\tilde{A}_{1T} = T^{1/2}\sum_{n=-M}^{M}\Big\{\frac{\partial c_z(n;\beta_0)}{\partial\beta} - \Big(1-\frac{|n|}{T}\Big)\frac{\partial\gamma_z(n;\beta_0)}{\partial\beta}\Big\}\Big(1-\frac{|n|}{T}\Big)\varrho_0(n;\theta_0),$$

where we fix M for each given ε which can be made arbitrarily small. Based on (2.26) and (2.27), we have,

$$\begin{aligned}
\tilde{A}_{1T} &= T^{1/2}\sum_{n=-M}^{M}\Bigg\{\frac{1}{T}\sum_{t\in\mathscr{T}_n}\Big(\frac{\partial\phi(X_{t+n};\beta_0)}{\partial\beta}\xi_t - E\Big[\frac{\partial\phi(X_{t+n};\beta_0)}{\partial\beta}\xi_t\Big]\Big) \\
&\quad + \frac{1}{T}\sum_{t\in\mathscr{T}_n}\Big(\xi_{t+n}\frac{\partial\phi(X_t;\beta_0)}{\partial\beta} - E\Big[\xi_{t+n}\frac{\partial\phi(X_t;\beta_0)}{\partial\beta}\Big]\Big)\Bigg\}\kappa(n) \\
&:= \frac{1}{\sqrt{T}}\sum_{n=-M}^{M}\sum_{t\in\mathscr{T}_n}H(t,n)\kappa(n) \\
&= \frac{1}{\sqrt{T}}\sum_{t=1}^{T-1}\sum_{n=-\min(M,T-t)}^{\min(M,T-t)}H(t,n)\kappa(n) \\
&:= \frac{1}{\sqrt{T}}\sum_{t=1}^{T-1}W_t(M). \qquad (2.31)
\end{aligned}$$

By (A6) and fixing M, W_t is an α-mixing process with mixing coefficients α_j satisfying $\sum_{j=1}^{\infty}\alpha_j^{\frac{\delta}{2+\delta}} < \infty$ for some $\delta > 0$. Moreover, it is easy to see that W_t is stationary and $E(W_t) = 0$. Then following Theorem 2.21 of Fan and Yao (2003), we have

$$\tilde{A}_{1T} \xrightarrow{D} N(0, \tilde{\Phi}_{1T})$$

where $\tilde{\Phi}_{1T} = E\{\tilde{A}_{1T}^{\top}\tilde{A}_{1T}\} = E\{A_{1T}^{\top}A_{1T}\} := \Phi_{1T}$ and

$$\begin{aligned}
\Phi_{1T} &= \sum_{m,n=-T+1}^{T-1}\varrho_0(n;\theta_0)\varrho_0(m;\theta_0)TE\Big[(c_z'(n)-\tilde{\gamma}_z'(n))^{\top}(c_z'(m)-\tilde{\gamma}_z'(m))\Big] + o(1) \\
&= \sum_{m,n=-T+1}^{T-1}\varrho_0(n;\theta_0)\varrho_0(m;\theta_0)\Phi_T^{(11)}(m,n) + o(1).
\end{aligned}$$

Putting what we have discussed together, we have proved that

$$A_{1T} \xrightarrow{D} N(0, \Phi_{1T}).$$

The joint normality of A_{1T} and A_{2T} is seen by noting that both A_{1T} and A_{2T} can be written into summations of stationary α-mixing series like (2.31). So for any unit column vector η, the random variable $A_\eta = (A_{1T}, A_{2T}) \times \eta$ is also a summation of a stationary α-mixing process satisfying the conditions of Theorem 2.21 of Fan and Yao (2003). As such, A_η converges in distribution to a normal distribution for any unit vector η, which implies that

$$T^{1/2} \frac{\partial Q_T(\beta_0, \theta_0)}{\partial(\beta, \theta)} = (A_{1T}, A_{2T}) \xrightarrow{D} N(\mathbf{0}, \Phi_T),$$

where $\Phi_T = E[(A_{1T}, A_{2T})^\top (A_{1T}, A_{2T})]$.

Similar to the calculation of $E\{A_{1T}^\top A_{1T}\}$, we have

$$\begin{aligned} E\{A_{1T}^\top A_{2T}\} &= \sum_{m,n=-T+1}^{T-1} \varrho_0(n;\theta_0)\varrho_0(m;\theta_0) T E\Big[(c_z'(n) - \tilde{\gamma}_z'(n))^\top (c_z(m) - \tilde{\gamma}_z(m))\Big] + o(1) \\ &= 2 \sum_{m,n=-T+1}^{T-1} \varrho_0(n;\theta_0) \Phi_T^{(12)}(m,n) \varrho_0'(m;\theta_0) + o(1), \end{aligned}$$

where $\varrho_0'(m;\theta_0)$ is the coefficient (row) vector of the Fourier series of $\partial k_0(\lambda;\theta_0)^{-1}/\partial\theta$ and

$$\begin{aligned} \Phi_T^{(12)}(m,n) &= 2 \sum_{r \in \mathscr{R}_{m,n}} \Big\{ E\Big[\frac{\partial \phi_{r+n}}{\partial \beta^\top} \xi_m \xi_r \xi_0\Big] - E\Big[\frac{\partial \phi_{r+n}}{\partial \beta} \xi_r\Big]^\top E\Big[\xi_m \xi_0\Big] \\ &\quad + E\Big[\frac{\partial \phi_r}{\partial \beta^\top} \xi_0 \xi_{r+n} \xi_m\Big] - E\Big[\frac{\partial \phi_r}{\partial \beta} \xi_{r+n}\Big]^\top E\Big[\xi_0 \xi_m\Big] \Big\}. \end{aligned}$$

Let $\Phi_{2T} = E\{A_{2T}^\top A_{2T}\} \to \mathbf{I}_{q\times q}$ as $T \to \infty$. We have

$$\Phi_T = E[(A_{1T}, A_{2T})^\top (A_{1T}, A_{2T})]$$

$$= \sum_{m,n=-T+1}^{T-1} \begin{pmatrix} \varrho_0(n;\theta_0)\varrho_0(m;\theta_0)\Phi_T^{(11)}(m,n) & 2\varrho_0(n;\theta_0)\Phi_T^{(12)}(m,n)\varrho_0'(m;\theta_0) \\ 2\varrho_0(n;\theta_0)\varrho_0'(m;\theta_0)^\top\Phi_T^{(12)}(m,n)^\top & \frac{1}{(2T-2)^2}\Phi_{2T} \end{pmatrix} + o(1).$$

The proof of Theorem 2.6.2 is completed by letting $T \to \infty$. □

CHAPTER 3

Conclusion and Future Works

In Chapter 1, by partitioning the sample space into several regions adaptively and fitting a single-index model to each region, we proposed the piecewise single-index model (1.7) as a new dimension reduction approach to improve the estimation efficiency of nonparametric regression. Numerical studies suggest that the approach is able to discover complicated structures in the data and make accurate predictions. Statistical theories of the model has been investigated.

In terms of modeling, the piecewise single-index model has its advantages in three essential aspects. Firstly, the single-index model itself has strong approximation ability, and so does the piecewise single-index model; see Jones (1987).

Secondly, adopting the single-index structure offers a convenient way to identify heterogenous structure by allowing the gradients in each single-index model to take on a unique direction. Thirdly, the model retains the decent estimation efficiency for the univariate nonparametric functions and root-n convergency rate for parameter estimation. On the other hand, the piecewise single-index model extends the popular CART (Breiman et al, 1984) and the piecewise linear model, and suggests a direction for further research in dimension reduction techniques (Li, 1991).

In Chapter 2, we have proposed a modified Whittle likelihood estimation (XWLE) to estimate general nonlinear time series models with serial correlated residuals that follow an MA process. Even in the linear model, some good performance of XWLE is also observed as compared the original WLE in our calculations. Adding MA residuals to an autoregressive model can simplify the model structure as compared to the pure autoregressive counterparts. The necessity of adding the MA residuals is also demonstrated in the real data analysis. Asymptotic properties of the estimator have been investigated. This Chapter only discusses the asymptotic properties under parametric setting. The idea can be easily extended to nonparametric or semiparametric time series models and time series models with exogenous variables, where residuals are serial correlated.

The following are two open problems for future works:

1. It is interesting to connect the piecewise single-index model with the smooth adaptive Gaussian mixtures (SAGM) of Villani et al (2009) for regression density estimation. In SAGM, the partition rule is assumed to be governed by a multinomial logit mixing function which is continuously differentiable with respect to the parameters involved. The Bayesian approach proposed by Villani et al (2009) can not be easily extended to a high dimension case due to the computation complexity of MCMC. Under such partition rule, however, it is possible to estimate the SAGM under a profiled MAVE framework, which can be a promising research direction to pursue.

2. We have only studied the estimation of parametric nonlinear AR models with MA errors. It is interesting to investigate the estimation methods of the semiparametric and nonparametric (nonlinear) AR models (Fan and Yao, 2003) to which we add an MA part.

Bibliography

[1] Anderson, T. W. (1971) *The Statistical Analysis of Time Series*. Wiley, New York.

[2] Box, G. E. P. and Pierce, D. A. (1970) Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, **65**, 1509-1526.

[3] Breiman, L. and Friedman, J. H. (1985) Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* **80**, 580-597.

[4] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984) *Classification and Regression Trees*. Belmont, CA: Wadsworth.

[5] Brockwell, P. J. and Davis, R. A. (1991) *Time Series: Theory and Methods*, second edition. Springer-Verlag, New York.

[6] Chaudhuri, P., Huang, M. C., Loh, W. Y., and Yao, R. (1994) Piecewise polynomial regression trees. *Statistica Sinica* **4**, 143-167.

[7] Chan, K. S. and Tong, H. (1986) On estimating thresholds in autoregressive models, *Journal of Time Series Analysis*, **7**, 179-190.

[8] Chen, R. and Tsay, R. S. (1993) Functional-Coefficient Autoregressive Models. *Journal of the American Statistical Association*, **88**, 298-308.

[9] Chipman, H., Geoge, E., and McCulloch, R. (2002) Bayesian treed models. *Machine Learning* **48**, 303-324.

[10] Cui, X., Härdle, W. and Zhu, L. (1993) The EFM approach for single-index models. *The Annals of Statistics*, **39**, 1658-1688.

[11] Doukhan, P. (1994) *Mixing*. Springer-Verlag, New York.

[12] Fan, J. (1996) Test of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association*, **91**, 674-688.

[13] Fan, J., and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.

[14] Fan, J. and Yao, Q. (2003) *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag, New York.

[15] Friedman, J. H. and Stuetzle, W. (1981) Projection pursuit regression. *Journal of the American Statistical Association*, **76**, 817-823.

[16] Giraitis, L. and Robinson, P. M. (2001) Whittle estimation of ARCH models. *Econometric Theory*, **17**, 608-631.

[17] Gramacy, R. (2009) An R Package for Bayesian Nonstationary, Semiparametric Nonlinear Regression and Design by Treed Gaussian Process Models. *Journal of Statistical Software*, **19**(9).

[18] Gramacy, R. and Lee, H. (2008) Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association* **103**, 1119-1130.

[19] Gramacy, R. and Lian, H. (2012) Gaussian process single-index models as emulators for computer experiments. *Technometrics* **54**, 30-41.

[20] Hannan E. J. (1970) *Multiple Time Series*. John Wiley, New York.

[21] Hannan E. J. (1973) The asymptotic theory of linear time-series models. *Journal of Applied Probability*, **10**, 130-145.

[22] Hannan E. J. and Heyde, C. C. (1972) On limit theorems for quadratic functions of discrete time series. *The Annals of Mathematical Statistics*, **43**, 2058-2066.

[23] Härdle, W., Hall, P. and Ichimura, H. (1993) Optimal smoothing in single-index models. *The Annals of Statistics*, **21**, 157-178.

[24] Härdle, W. and Stoker, T. M. (1989) Investigating smoothing multiple regression by the method of average derivatives. *Journal of the American Statistical Association* **84**, 986-995.

[25] Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer.

[26] Horowitz, J. L. and Hädle, W. (1996) Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association* **91**, 1632-1640.

[27] Hristache, M. , Juditski, A. and Spokoiny, V. (2001) Direct estimation of the index coefficients in a single-index model. *The Annals of Statistics* **29**, 595-623.

[28] Ichimura, H. (1993) Semiparametric least squares (SLS) and weighted SLS estimation of singleindex models. *Journal of Econometrics* **58**, 71-120.

[29] Jones, L. K. (1987) On a conjecture of Huber concerning the convergence of projection pursuit regression. *The Annals of Statistics* **15**, 880-882

[30] Li, K. C. (1991) Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316-327.

[31] Li, K. C. (1992) On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Journal of the American Statistical Association* **87**, 1025-1039.

[32] Li, K. C., Lue, H. H., and Chen, C. H. (2000) Interactive tree-structured regression via principal hessian directions. *Journal of the American Statistical Association* **95**, 547-560.

[33] Ljung, G. M. and Box, G. E. P. (1978) On a measure of lack of fit in time series models. *Biometrika*, **65**, 297-303.

[34] Lu, Z. (1996) Multivariate locally weighted polynomial fitting and partial derivative estimation. *Journal of Multivariate Analysis* **59**, 187-205.

[35] Montanari, A. and Viroli, C. (2011) Dimensionally reduced mixtures of regression models. *Journal of Statistical Planning and Inference* **141**, 1744-1752.

[36] De la Pena, V. H., Lai, T. L. and Shao, Q.-M. (2009) *Self-normalized Processes: limit theory and statistical applications*. Berlin: Springer.

[37] Powell, J. L., Stock, J. H. and Stoker, T. M. (1989) Semiparametric estimation of index coefficients. *Econometrica* **57**, 1403-1430.

[38] Ruppert, D. and Wand, M. P. (1994) Multivariate locally weighted least squares regression, *The Annals of Statistics* **22**, 1346-1370.

[39] Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics* **6**, 461-464.

[40] Sherman, R. P. (1994) Maximal inequalities for degenerate U-processes with applications to optimization estimators. *The Annals of Statistics* **22**, 439-459.

[41] Shimotsu, K. and Phillips, P. C. B. (2005) Exact local Whittle estimation of fractional integration. *The Annals of Statistics*, **33**, 1890-1933.

[42] Silverman, B. M. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

[43] Teräsvirta, T. (1994) Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, **89**, 208-218.

[44] Trenberth, K. E. and Stepaniak, D. (2001) Indices of El Niño evolution. *Journal of Climate*, **14**, 1697-1701.

[45] Tong, H. (1990) *Nonlinear Time Series Analysis: a Dynamical System Approach*. Oxford University Press.

[46] Ubilava, D. and Helmers, C. G. (2013) Forecasting ENSO with a smooth transition autoregressive model. *Environmental Modelling and Software*, **40**, 181-190.

[47] Villani, M., Kohn, R. and Giordani, P. (2009) Regression density estimation using smooth adaptive Gaussian mixtures. *Journal of Econometrics*, **153**, 155-173.

[48] Whittle, P. (1953) The analysis of multiple stationary time series. *Journal of the Royal Statistical Society, Series B*, **15**, 125-139.

[49] Wu, Z., Yu, K. and Yu, Y. (2010) Single-index quantile regression, *Journal of Multivariate Analysis* **101**, 607-1621.

[50] Xia, Y. (2006) Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, **22**, 1112-1137

[51] Xia, Y. (2007) A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics* **35**, 2654-2690.

[52] Xia, Y. and Tong, H. (2011) Feature Matching in Time Series Modeling (with discussion), *Statistical Science* **26**, 21-62.

[53] Xia, Y., Tong, H., Li, W. K., and Zhu, L.-X. (2002) An adaptive estimation of dimension reduction space (with discussions). *Journal of the Royal Statistical Society*, Ser. B **64**, 363-410.

[54] Yao, Q. and Brockwell, P. J. (2006) Gaussian maximum likelihood estimation for ARMA models I: time series. *Journal of Time Series Analysis*, **27**, 857-875.

[55] Yin, X., and Cook, R. D. (2002) Reduction for the conditional kth moment in regression. *Journal of the Royal Statistical Society*, Ser. B **64**, 159-175.

[56] Yin, X. and Cook, R. D. (2005) Direction estimation in single-index regressions. *Biometrika* **92**, 371-384.

[57] Yu, Y. and Ruppert, D. (2002) Penalized spline estimation for partially linear single index models. *Journal of the American Statistical Association* **97**, 1042-1054.

[58] Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **226**, 267C298.

[59] Zhang, C. (2003) Calibrating the degrees of freedom for automatic data smoothing and effective curve checking. *Journal of the American Statistical Association* **98**, 609-628.

[60] Zygmund, A. (1959) *Trigonometric Series*, Vol. I. Cambridge University Press.