The Pennsylvania State University
The Graduate School

SOME CONTRIBUTIONS TO NONPARAMETRIC MODELING WITH CORRELATED DATA

A Thesis in
Statistics
by
Yan Li

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2008

The thesis of Yan Li was reviewed and approved* by the following:

Runze Li
Associate Professor of Statistics
Thesis Advisor, Chair of Committee

Naomi S. Altman
Associate Professor of Statistics

David R. Hunter
Associate Professor of Statistics

Quan Li
Associate Professor of Political Sciences

Bruce G. Lindsay
Willaman Professor of Statistics
Head of the Department of Statistics

*Signatures are on file in the Graduate School.

Abstract

In many statistical applications, data are collected over time and are likely correlated. Theory and empirical studies have shown that ignoring the underlying correlation structure may lead to a less accurate local smoothing estimator. In this dissertation, we investigate how to incorporate the correlation information into the estimation of the nonparametric regression model and the varying-coefficient model. Under the assumption that the error process is an auto-regressive (AR) process, we propose profile least squares techniques to estimate the mean function in the nonparametric regression model and the functional coefficients in the varying-coefficient model respectively. The asymptotic distribution of the proposed estimator under regularity conditions shows that the profile least squares method is asymptotically as efficient as the local linear method with i.i.d. data. Further, we apply the SCAD variable selection procedure (Fan and Li, 2001) to select the order of the AR error process. Extensive Monte Carlo simulation studies are conducted to compare the finite sample performance of the proposed procedures with the existing methods. The simulation results show that the newly proposed procedures can dramatically improve the accuracy of the naive local linear estimates with a working-independent error structure. We also apply the proposed methodology to two real data sets from economic and environmental disciplines. In addition, we extend the profile

least squares estimation to nonparametric regression with multiple responses. The simulation results imply that our method can work with multiple responses as well.

Table of Contents

List of Figures

List of Tables

Acknowledgments

During my long journey through PhD study at Penn State University, I have met a lot of wonderful people and received tons of helps from them. I could not have accomplished my study without their substantial assistance.

My greatest thanks go to my advisor Dr. Runze Li. I could not have had a better advisor than him. His lectures motivated my interest in local modeling areas and led me to explore challenging research topics in depth. His expertise in nonparametric regression and variable selection played an important role during my research. He not only provided me with a lot of insightful suggestions when I was struggling, but also taught me many practical approaches to deal with problems. He spent a considerable amount of time on my dissertation and diligently went through my papers even during winter break. In addition, Dr. Li gave me a lot of helpful suggestions in my job hunting and career development. I regard his guidance as my life-time treasure.

I am also very grateful to my committee members. Dr. Naomi Altman, from the Statistics Department, generously provided me with suggestions on academic writing and was incredibly patient with me during the coordination process of my defense date. Dr. David Hunter, from the Statistics Department, flew hundreds of miles from Paris to attend my defense in person. Dr. Quan Li, from the Political

Science Department, and Dr. Jingzhi Huang, from the Finance Department, took their time from their busy schedules to serve on my committee.

Besides my committee members, I would like to acknowledge the department head Dr. Bruce Lindsay and all other faculty memebers in our department who have taught me or encouraged me in the past few years.

I also owe a particular debt to our department administrative and technology support staff. They provided me great facilities when I was preparing my oral defense. And I need to sincerely thank my fellow graduate students, my friends and community volunteers. They made my life in Happy Valley joyful and unforgettable.

Finally I want to express my deep gratitude to my family overseas. Although they were not with me, they gave me tremendous support beyond description.

This thesis research was supported by grants from the National Science Foundation (DMS 0348869, CCF 0430349 and DMS 0722351) and a grant from the National Institute of Health (1R21 DA024260). I appreciate the financial support from these grants.

Chapter 1

Introduction

Regression techniques are among the most useful tools for data analysis. Parametric regression models, including linear regression models and generalized linear models, provide a parsimonious description of the relationship between the dependent variables and the independent variables. However, they may have large model approximation error and therefore introduce large modeling bias in practice. Thus, nonparametric modeling has become more and more popular during the last two decades (Wahba, 1990; Wand and Jones, 1995; and Fan and Gijbels, 1996). Various nonparametric models such as the local smoothing, spline smoothing and orthogonal series have been proposed in the statistical literature. In this dissertation, all the methods we will propose are based on local linear smoothing.

1.1 Nonparametric regression models

Consider the nonparametric regression model

$$y_t = m(x_t) + \epsilon_t, \tag{1.1}$$

where ϵ_t is a random error. We do not impose any specific functional form on the mean function $m(\cdot)$ so that model (1.1) has great flexibility in interpreting the complicated data set. Such a model with independent errors ϵ_t has been well studied in the literature, and properties of the nonparametric estimator of the regression function $m(\cdot)$ have been also intensively investigated. For example, see Stone (1997, 1980, 1982), Cleveland (1979), Fan and Gijbels (1992), Ruppert and Wand (1994), etc. However, data obtained from various research fields often violate the assumption of independent errors. So it is of great interest to improve the accuracy of the parameter estimation by including the correlation information into estimation procedures.

Altman (1990) and Hart(1991) pioneered the work of model (1.1) with correlated errors under the fixed design, i.e., $x_t = t/n$ or $x_t = (t - 0.5)/n$. In this case, the kernel-based nonparametric estimator has a variance proportional to the long run variance of $\{x_t\}$. Both Altman(1990) and Hart(1991) assumed that the correlation between ϵ_t and ϵ_s was of the form $\rho_n(|t-s|)$, and addressed the issue of how to select a bandwidth adjusting the correlation structure $\rho_n(|t - s|)$, which is assumed to be known or can be estimated from the observed data. In addition, Opsomer, Wang and Yang (2001) gave a good review on nonparametric regression with correlated errors under the fixed design framework and revised Ruppert, Sheather and Wand(1995) direct plug-in bandwidth for local polynomial regression estimator. More detailed descriptions of their work can be found in the later literature review chapter.

In this dissertation, we consider the situation in which x_t is a random design. More specifically, it is assumed that (x_t, y_t), $t = 1, 2, \cdots$, is a sequence of strictly stationary random vectors. Thus, (x_t, y_t), $t = 1, 2, \cdots$, is identically distributed. We are interested in how to incorporate the correlation into the local linear regression estimation procedure when the data are correlated. This issue has been studied in

Xiao, Linton, Carroll and Mammen (2003), in which the authors proposed a new estimation procedure based on a pre-whitening transformation of the dependent variable that must be estimated from the data. They also established the asymptotic distribution of their estimator under weak conditions, but they did not address the critical issues related to practical implementation, such as the selection of the smoothing parameter of their nonparametric regression. They assumed that the error process was an invertible linear process, i.e., a moving average (MA) process with order infinite, but they did not discuss how to determine the order of the error process.

Assuming that the error process is an autoregressive (AR) process of order d, we propose a new estimation procedure for the nonparametric regression using the profile least squares techniques (Speckman, 1988; Fan and Huang, 2005). We show that the resulting estimator has the same asymptotic bias and variance as the local linear estimator with independently and identically distributed observations. We further suggest using penalized profile least squares with the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) to determine the order of the AR process. We address the issue of selecting the bandwidth in profile least squares estimation and selecting the tuning parameter in penalized profile least squares estimation. We compare the finite sample performance of Xiao, Linton, Carroll and Mammem (2003) method, the profile least squares method and the penalized profile least square method in the simulation study, where various conditions of the mean function with distinguished shapes, different sample size, different size of noise, different AR orders, and various AR coefficients are considered. As an outcome, penalized profile least squares method is always the best one to reduce the fitting error and it is very close to the oracle result when the sample size is large. The application in an economic data set also shows that the penalized profile method can identify a

suitable AR order and remove the underlying correlation pattern effectively.

1.2 Varying-coefficient models

As a sophisticated alternative to the nonparametric model (1.1), the varying-coefficient model has been used widely to explore the fine structure in data. Since the systematic studies by Hastie and Tibshirani (1993), the statistical estimation and inference procedures for the varying-coefficient model have been studied intensively. The kernel method (Hastie and Tibshirani 1993; Wu and Chiang, 2000) can be easily implemented to estimate the functional coefficients, but it will undersmooth the underlying coefficient when the coefficients admit different degrees of smoothness. The spline method (Hoover *et. al.*, 1998; Brumback and Rice, 1998) has a better performance since it includes multiple smoothing parameters. However, the spline method is very computationally intensive. To overcome these difficulties, Cai, Fan and Li (2001) developed the one-step local maximum likelihood estimator (MLE). Fan and Zhang (2000) proposed the two-step least squares estimator, and Cai (2003) proposed the two-step likelihood estimator particularly for the coefficient with different degree of smoothness. Some important results of these studies are described in Chapter 2.

The varying-coefficient model allows its coefficients to vary over one or more covariates of particular interest, while retaining its coefficients which have the same interpretation as those in the linear regression model. Due to this nice feature, varying coefficient models have been used for functional data (Faraway, 1997), longitudinal and clustered data (Hoover *et. al.*, 1998; Fan and Zhang, 2000; Wu and Chiang, 2000; He *et. al.*, 2002; Yao *et. al.*, 2005; Qu and Li, 2006; Fan *et. al.*, 2005), survival data (Fahrmeir and Klinger, 1998; Kanermann *et. al.*, 2005, 2006;

Yan and Fine, 2005) and time series data (Cai *et. al.*, 2001; Cai, 2007; Huang and Shen, 2004).

To be consistent with the nonparametric model, we will focus on the varying coefficient model with auto-correlated data in this dissertation. Suppose that a random sample $(u_t, x_{t1}, \cdots, x_{tp}, y_t)$, $t = 1, \cdots, n$, is collected from the varying coefficient model

$$y_t = \alpha_0(u_t)x_{t0} + \alpha_1(u_t)x_{t1} + \cdots + \alpha_p(u_t)x_{tp} + \epsilon_t, \tag{1.2}$$

where the random error ϵ_t is from an AR process with order d. As usual, we set x_{t0} to be 1 to include the intercept $\alpha_0(\cdot)$. In the absence of x-covariates, the model reduces to a nonparametric model (1.1).

Many authors have suggested the estimation method based on the local linear smoothing for the varying-coefficient model. However, to our best knowledge, there is no work discussing how to select the AR order in such a setting. Stimulated by the profile idea employed in model (1.1) with AR errors, we propose the profile least squares estimator for the varying coefficient model with AR errors. Under regularity conditions, the resulting estimator of the AR coefficient β_j can reach the semiparametric efficiency bound, while the asymptotic distribution of the functional coefficient $\alpha_i(\cdot)$ is identical to that of the local linear estimator for i.i.d. data. In order to select the AR order, we introduce the penalized profile least square method with the SCAD penalty function. Our simulation results under different sampling schemes and various AR conditions all indicate that the profile least squares method can improve the estimator's accuracy and that incorporating the order selection procedure into the estimation makes an even larger improvement. In implementing the proposed estimation procedure, we use the bandwidth that minimizes the mean square error (MSE) of the functional coefficient estimators and the BIC criterion

to choose the tuning parameter when the penalty function is involved in penalized profile least squares estimation.

Finally, we apply the proposed method to two data sets that are collected over a period of time. One is the U.S. macroeconomic data, where the unemployment rate is considered as u_t while the interest rate and the gross domestic product growth rate are regarded as x_{t1} and x_{t2} in model (1.2) respectively. After the strong autocorrelation is removed by the profile least square procedure and an appropriate AR order is determined by the penalized profile least squares method, the estimates of functional coefficients are smoother. In the second example, we investigate the effect of three air pollutants on the total number of admissions in a hospital at Hong Kong over two years. By applying the penalized profile least squares estimator, the autocorrelation in the resulting residuals is gone and a much smoother estimator of the functional coefficients is obtained.

1.3 The structure of this dissertation

The remainder of this dissertation is organized as follows. In Chapter 2, we provide a detailed literature review of the nonparametric and varying-coefficient models. Some existing estimation and bandwidth selection methods for independent and correlated data will be covered.

Chapter 3 focuses on the nonparametric regression model with autocorrelated data. We use the penalized profile least squares with the SCAD penalty to estimate $m(\cdot)$ and select the AR order simultaneously. Extensive Monte Carlo simulation studies are conducted to examine the finite sample performance of the proposed procedure and to compare the performance of the proposed method with the existing ones. The proposed method is also applied to a U.S. macroeconomic data set to

reveal the relationship between the housing price index change and the unemployment rate. Regularity conditions and technical proofs of the asymptotic property of the profile least squares estimator are given at the end of this chapter.

In Chapter 4, the varying-coefficient model with autocorrelated errors is systematically studied. We propose the profile least squares method to estimate the functional coefficients. Based on this method, we suggest the penalized profile least squares estimator with the SCAD penalty to select the AR order. We carry out a series of the Monte Carlo simulations to compare the profile least squares method and the penalized profile least square method with the local linear method ignoring the correlation structure in finite samples. We discuss bandwidth selection and tuning parameter selection issues in this chapter. In addition, we apply the proposed method to two real data examples that represent two different sampling schemes. Lastly, we derive the asymptotic distribution of the AR coefficient estimators and the functional coefficient estimators under regularity conditions.

Chapter 5 extends the profile least squares estimation for model (1.1) with AR errors to the nonparametric regression model with multiple responses and vector autoregressive (VAR) errors. Monte Carlo simulations are conducted to demonstrate on how the proposed method works under various situations. Finally future research directions and possible extensions are discussed.

Chapter 2

Literature review

In this chapter, we will give a selective literature review of nonparametric regression and the varying coefficient models. Although many estimation methods are available for each model, we will emphasize local polynomial smoothing because the new estimation procedures proposed in this dissertation are based on the concept of local smoothing. Bandwidth selection to control model complexity will also be discussed. The aim of this chapter is to outline the framework of the well known nonparametric models and to introduce the existing work done by other scholars.

2.1 Nonparametric regression models

Parametric regression, especially the polynomial regression, is one of the most popular statistical techniques in various disciplines. It is used to explore the relationship between dependent and independent variables and make a reliable prediction based on the estimated model. Let $(X_t, Y_t), t = 1, \ldots, n$ be the observed data. Note that both X and Y may be multidimensional data. However, we will consider only the univariate model here for notation simplicity. A general p^{th}-order polynomial model

is:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p + \epsilon \tag{2.1}$$

where ϵ is the random error term and assumed to be independently and identically distributed. The mean of the response variable Y should satisfy:

$$E(Y|X = x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p \tag{2.2}$$

Usually a scatter plot is drawn first to determine the appropriate order of the polynomial. The linear regression can be regarded as a special case of the lower order polynomial regression. The sum of residual squares is minimized to determine the regression coefficients $\beta_0, \ldots, \beta_p$. We also need to conduct model diagnosis to check the assumptions.

Fan and Gijbels (1996) pointed out that polynomial regressions suffered from a few drawbacks, one of which is that the polynomial functions are not very flexible in modeling the features of some complicated data. Another is that an individual observation can have a large influence on remaining parts of the curve. A third one is that the polynomial degree cannot be controlled continuously.

To fix the drawbacks of traditional regression, data analysts have introduced spline smoothing, wavelets and local polynomial regression methods. In this dissertation, we focus on local modeling. Compared to traditional polynomial regression (2.1), local polynomial regression relaxes the assumptions on the form of the regression function and allows the data to suggest a suitable model. Moreover, it can deal with functional data and longitudinal data easily.

2.1.1 Independent data

Assume that we have independently and identically distributed data $(X_i, Y_i), i = 1, \ldots, n$. Let (X, Y) denote the generic form of the observations. The conditional mean and variance are defined as:

$$m(x) = E(Y|X = x) \text{ and } \sigma^2(x) = Var(Y|X = x)$$

Usually, the estimator of $m(x)$ and its ν^{th} derivative $m^{(\nu)}(x)$ are of our main interest. In order to assess how good the fit is, we can consider either:

$$MSE_\nu(x) = E[\{\widehat{m}_\nu(x) - m^{(\nu)}(x)\}^2|\mathbb{X}]$$

where $\mathbb{X} = (X_1, \ldots, X_n)$, or

$$MISE = \int MSE(x)w(x)\,dx$$

where $w(.)$ is a weight function. The MSE-criterion is aimed to estimate $m^{(\nu)})(x)$ at point x and the MISE-criterion is used to assess the performance of the whole curve estimate.

Before local polynomial regression received the full attention, local constant fit method, namely the kernel estimation, was used often because its implementation was easy. Two popular examples of the kernel estimates are: Nadaraya-Watson (1964) estimator,

$$\widehat{m}_h(x) = \frac{\sum_{t=1}^n K_h(X_t - x)Y_t}{\sum_{t=1}^n K_h(X_t - x)}$$

and Gasser-Müller (1979) estimator:

$$\widehat{m}_h(x) = \sum_{t=1}^n \int_{s_{t-1}}^{s_t} K_h(u - x)\,du\,Y_t \text{ with } s_t = \frac{x_t + x_{t+1}}{2}, x_0 = -\infty, x_{n+1} = +\infty.$$

where the function K is the kernel function, h is the bandwidth and $K_h(.) = K(\cdot/h)/h$. The most commonly used kernel functions are the Gaussian kernel

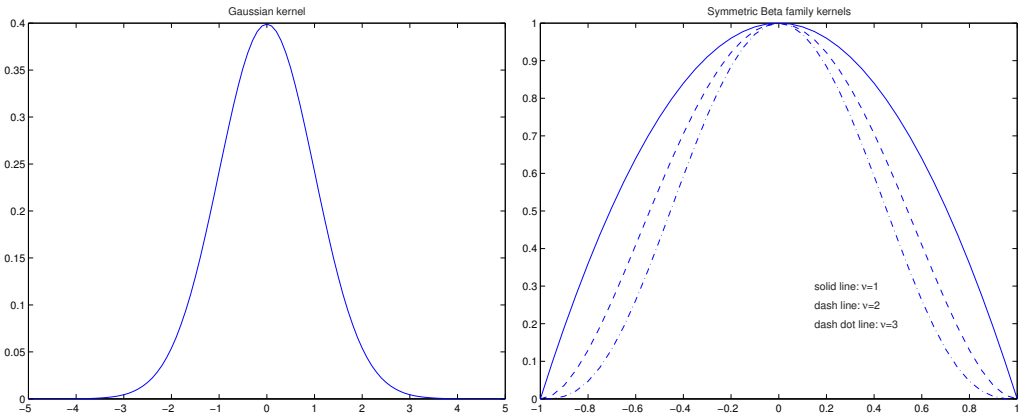

Figure 2.1. Figures of Kernel Functions

$K(t) = \frac{1}{\sqrt{2\pi}} e^{(-t^2/2)}$ and the symmetric Beta family $K(t) = \frac{1}{Beta(\frac{1}{2},\nu+1)}(1-t^2)_+^\nu, \nu = 0, 1, \ldots$. (see Figure 2.1)

The kernel method is essentially a locally weighted average estimator, which suffers from the large order of bias at the boundary and low minimax efficiency. As a better alternative, the local polynomial fit can get rid of the drawbacks of the kernel estimation.

Since the local polynomial fit was systematically studied by Stone (1977,1980,1982) and Cleveland (1979), a lot of work has been done to refine this method. The benchmark work was given by Fan (1992, 1993a), Fan and Gijbels (1992), and Ruppert and Wand (1994). They provided a detailed picture of how to get a robust estimator and proved the asymptotic distribution of the estimator. Thereafter, more and more researchers have realized the beauty of local modeling.

Since the local polynomial regression for i.i.d. data is the building block of our proposed method in the later chapters, let us outline the basic framework of the local polynomial regression. By using Taylor's expansion in the neighborhood of x, we obtain:

$$m(z) \approx \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!}(z-x)^j \equiv \sum_{j=0}^{p} \beta_j (z-x)^j$$

where $\beta_j = \frac{m^{(j)}(x)}{j!}$. The above expression shows that $m(z)$ can be locally fitted by a simple polynomial function. If we minimize the locally weighted square loss function

$$\sum_{t=1}^{n}\{Y_t - \sum_{j=0}^{p}\beta_j(X_t - x)^j\}^2 K_h(X_t - x), \tag{2.3}$$

where $K(.)$ denotes a kernel function and h is a bandwidth, we can get the estimator $\widehat{m}^{(\nu)}(x) = \nu!\widehat{\beta}_\nu$. When $p = 1$, the local polynomial fit is local linear regression. The estimator can be explicitly expressed as

$$\widehat{m}(x) = \frac{\sum_{t=1}^{n} w_t Y_t}{\sum_{t=1}^{n} w_t} \tag{2.4}$$

with $w_t = K_h(X_t - x)\{S_{n,2} - (X_t - x)S_{n,1}\}$, $S_{n,j} = \sum_{t=1}^{n} K_h(X_t - x)(X_t - x)^j$.

Table 2.1 summarizes the asymptotic bias and variance of the Nadaraya-Watson estimator, the Gasser–Müller estimator and the local linear estimator. It is noted that the local linear estimator outperforms the two kernel estimators in terms of bias and variance. Moreover, Cheng, Fan and Marron (1993) showed that the local linear fit is more efficient in correcting the boundary bias than these two kernel estimators.

Table 2.1. Pointwise asymptotic bias and variance of kernel regression smoothers. Taken from Fan(1992). Here, $b_n = \frac{1}{2}\int_{-\infty}^{\infty} u^2 K(u)\,\mathrm{d}u h^2$ and $V_n = \frac{\sigma^2}{f(x)nh}\int_{-\infty}^{\infty} K^2(u)\,\mathrm{d}u$.

Method	Bias	Variance
Nadaraya-Watson	$(m''(x) + \frac{2m'(x)f'(x)}{f(x)})b_n$	V_n
Gasser-Müller	$m''(x)b_n$	$1.5V_n$
Local linear	$m''(x)b_n$	V_n

In practice, we use matrix notation often for conciseness. Denote $m(x_0) = E(Y|X = x_0)$. Let

$$y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x_0) & \cdots & (X_n - x_0)^p \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_n \end{pmatrix}$$

Minimizing the weighted least squares $\sum_{t=1}^{n}\{Y_t - \sum_{j=0}^{p}\beta_j(X_t - x_0)^j\}^2 K_h(X_t - x_0)$ is equivalent to

$$\min_{\beta}(y - X\beta)^T W(y - X\beta)$$

where $W = \text{diag}\{K_h(X_1 - x_0), \ldots, K_h(X_n - x_0)\}$. The estimator $\widehat{\beta} = (X^T W X)^{-1} X^T W y$ indicates that $\widehat{\beta}$ is a weighted linear estimator.

Furthermore we can obtain:

$$E(\widehat{\beta}|X) = (X^T W X)^{-1} X^T W m = \beta + (X^T W X)^{-1} X^T W r$$
$$m = [m(X_1), \cdots, m(X_n)]^T, r = m - X\beta$$
$$\text{Var}(\widehat{\beta}|X) = (X^T W X)^{-1}(X^T \Sigma X)(X^T W X)^{-1}$$
$$\Sigma = diag\{K_h^2(X_1 - x_0)\sigma^2(X_1), \cdots, K_h^2(X_n - x_0)\sigma^2(X_n)\}$$

Theorem 3.1 of Fan and Gijbels (1996) states the asymptotic bias and variance of the local polynomial estimator $\widehat{m}_\nu(x_0)$. We quote Theorem 3.1 in the special situation when the local linear regression is used to estimate the mean function $m(x_0)$ as below:

Theorem 2.1.1. *Assume that $f(x_0) > 0$ and $m''(\cdot)$ is continuous in a neighborhood of x_0. Further, assume that $h \to 0$ and $nh \to \infty$. Then the asymptotic conditional*

bias of $\widehat{m}(x_0)$ *is given by*

$$bias\{\widehat{m}(x_0)|X\} = \frac{1}{2}m''(x_0)h^2 \int x^2 K(x)\,\mathrm{d}x$$

The asymptotic conditional variance of $\widehat{m}(x_0)$ *is given by*

$$Var\{\widehat{m}(x_0)|X\} = \frac{\sigma^2 \int K^2(x)\,\mathrm{d}x}{nhf(x_0)}$$

The above theorem is taken from Fan and Gijbels (1996). Hence a pointwise confidence interval for $\widehat{m}(x_0)$ can be constructed. Now three important issues in the implementation naturally arise:

1. Bandwidth selection

   The bandwidth h is crucial in controlling the overall fit performance and the model complexity. If h is small, the modeling bias will be small but the variance will be large. When $h \to 0$, the regression estimation essentially interpolates the data points. If h is large, the local linear estimation is the same as the traditional linear regression and yields a large modeling bias. Since h can vary from 0 to $+\infty$, we have a rich family of models to meet different requirements. For example, if the overall trend is of interest, a smooth estimator (i.e. large h) is desirable. If the local extremum is of interest, a less smooth estimator (i.e. small h) is preferred. However, an objective criterion such as MSE or MISE is more often used to balance the trade-off between bias and variance and get an optimal bandwidth.

   Bandwidth selection is a well-studied topic in local polynomial regression. Besides a *constant* bandwidth, there are some more flexible bandwidths for choices, such as *local variable* bandwidth (changing with x_0) and *global variable*

bandwidth (changing with data).

A number of bandwidth selection procedures have been created for practical use. For example, the *Rule of thumb* bandwidth, the *least squares cross validation* bandwidth and the *plug-in* bandwidth selector, etc. See Chapter 4 of Fan and Gijbels (1996) for more details.

2. The order of the polynomial

   It is natural to ask the order of the appropriate polynomial before fitting the data. If a large order of polynomial is used, it will reduce the bias but require more computation. For a flat region, a local constant or linear fit will be recommended. For peaks and valleys, a cubic polynomial usually suffice to produce a good result. Fan(1992) argued that polynomial fits with odd orders outperformed those with even orders at the interior design points. By his argument, take $p = \nu + 1$ or $\nu + 3$ for best. Here ν is the derivative order of $m(x)$.

3. The kernel function

   The kernel function $K(.)$ is another factor that will influence the local polynomial fit. The most frequently used kernels are:

   Gaussian kernel $K(z) = \frac{1}{\sqrt{2\pi}} exp(-\frac{z^2}{2})$

   Epanechnikov kernel $K(z) = \frac{3}{4}(1 - z^2)_+$

   Uniform kernel $K(z) = 1_{[-0.5,0.5]}(z)$

   Considering linear minimax risk, Fan (1992) proved that the Epanechnikov kernel with a certain bandwidth will achieve the optimal efficiency in the class of all regression functions whose second derivative is bounded by a constant in

the neighborhood of x_0. We will use Epanechnikov kernel when implementing the local smoothing throughout this dissertation.

2.1.2 Correlated data

In the last section, we reviewed local polynomial regression for independent data. However, we often encounter correlated data in practice, such as financial data, temporal data and spatial data, etc. If we ignore the correlation and apply the same nonparametric model as we deal with independent data, it may not yield a satisfactory result. Therefore, it is necessary to adapt the existing algorithm to serve the correlated data better.

Two sampling schemes for X_t are studied in statistical literature. The first one is the fixed design, where X_t are regularly spaced observations. For example, $X_t = t/n$ or $X_t = (t-0.5)/n$ where n is the sample size. Much work has been done under this design, especially with respect to adjusting the bandwidth on the kernel estimator. The other sampling scheme is the random design, where X_t is a stationary stochastic process from a non-degenerate distribution. We are more interested in the random design in this dissertation. However, some important existing methods under both designs are introduced in this section.

Among various corrective strategies for the correlated errors, we can group them into two major categories: adjusting the bandwidth or decorrelating the errors.

2.1.2.1 Adjust the bandwidth

When the data are correlated, the regular bandwidth selector for the local linear estimation breaks down. Many authors have developed adaptive selection methods by incorporating the knowledge of the correlation structure into the regular bandwidth

selection for i.i.d. data.

Altman(1990) studied the model

$$y = m(x) + \epsilon,$$

where $m(\cdot)$ is a smooth function, x is an equally spaced point in $[0, 1]$ and ϵ is assumed to come from a stationary process with covariance function $E(\epsilon_{n,s}, \epsilon_{n,t}) = \sigma^2 \rho_n(|s - t|)$. $\rho_n(k)$ is a correlation function depending on sample size n. The variance matrix of the errors is denoted by Σ_n. She used the kernel estimators of Pristley and Chao (1972), which would result the following estimate:

$$\widehat{m}_{h,n}(x) = \sum_{t=0}^{n} w_{h,n}(x, t) y_{n,t} \tag{2.5}$$

where the weights are

$$w_{h,n}(x, t) = \frac{K\{(x - x_{n,t})/h\}}{nh}$$

K is the kernel function satisfying:

A: K is symmetric about 0.

B: K has support only on $(-\frac{1}{2}, \frac{1}{2})$.

C: K is *Lipschitz* continuous of any order $\alpha > 0$.

Besides these three conditions on the kernel $K(\cdot)$, two additional conditions are assumed for the summable autocovariance:

D: $\sum_{i=1}^{k/2} |\rho_n(i)|$ converges as n and $k \to \infty$.

Under this condition, the sum of the correlation, S_ρ is well defined. Let $S_{\rho_n}(k) = \sum_{j=1}^{k/2} \rho_n(j)$. Then $S_\rho = \lim_{n,k\to\infty} \rho_n(k)$.

E: $\sum_{t=1}^{k/2} t\rho_n(t) = o(k)$ as n and $k \to \infty$.

Let's denote $\mu_j = \int x^j K(x)\,dx, \nu_j = \int x^j K^2(x)\,dx$. Altman(1990) used the Mean Squared Error (MSE) defined by

$$MSE(x,h,n) = E(\widehat{m}_{h,n}(x) - m(x))^2$$

to assess the good-of-fit. An optimal bandwidth h minimizes the MSE value. Recall that Σ_n is the covariance matrix and $w_{h,n}$ is the weight. Then

$$MSE(x,h,n) = (w'_{h,n}(x,.)m(.) - m(x))^2 + w'_{h,n}(x,.)\Sigma_n w_{h,n}(x,.) \qquad (2.6)$$

Notice that the square of the bias only depends on the sample size n. But the correlation of the errors can affect the variance term, $w'_{h,n}(x,.)\Sigma_n w_{h,n}(x,.)$. Altman (1990) proved for $h/2 \le x \le 1 - h/2$,

$$MSE(x,h,n) = (h^p m^{(p)}(x)\mu_p/p!)^2 + \sigma^2(\nu_0/nh)(1+2S_\rho) + o(1/nh) + o(h^{2p}), \quad (2.7)$$

where S_ρ is defined in Condition D.

Consequently, the asymptotically optimal bandwidth at x is

$$h_n = \frac{(p!)^2\sigma^2\nu_0(1+2S_\rho)}{p\mu_p^2(m^{(p)}(x))^2}^{1/(2p+1)} n^{-1/(2p+1)} \qquad (2.8)$$

When $S_\rho = 0$, equation (2.8) is just the optimal bandwidth formula for independent errors situation. If $S_\rho > 0\,(< 0)$, the optimal bandwidth for the correlated errors is greater (smaller) than the one for independent errors.

The commonly data-driven bandwidth selections are based on *CV* (Allen 1974; Geisser 1975; Stone 1974), *GCV* (Craven and Wahba, 1979) and *Mallow's* C_L (Mal-

low, 1973). All those criteria can be regarded as the estimators of expected squared prediction error (ESPE).

$$\text{ESPE}(x,h,n) = E(y_{new}(x) - \widehat{m}_{h,n}(x))^2 = \sigma^2 + MSE(x,h,n)$$

The squared residual, $r^2(t,h,n) = (y_{n,t} - \widehat{m}_{h,n}(t/n))^2$, is an biased estimator of ESPE. Its expectation is

$$E[r^2(t,h,n)] = \sigma^2 + MSE(t/n,h,n) - 2\sigma^2 w_{t,n}(t/n,t) - V_2(t/n,h,n)$$

where $V_2(t/n,h,n) = 2\sigma^2 \sum_{s\neq 0} w_{h,n}(t/n, s+t)\rho_n(s)$.

In the region $-\frac{h}{2} < \frac{i}{n} < \frac{h}{2}$, Mallow's C_L is defined by

$$r^2_{CL}(t,h,n) = r^2(t,h,n) + 2\widehat{\sigma}^2 w_{h,n}(t/n,t)$$

where $\widehat{\sigma}^2$ is some unbiased estimator of σ^2.

CV is defined by

$$r^2_{CV}(t,h,n) = r^2(t,h,n)/(1 - w_{h,n}(t/n,t))^2.$$

GCV is defined by

$$r^2_{GCV}(t,h,n) = r^2(t,h,n)/(1 - 1/n\, trW_{h,n})^2.$$

Altman (1990) proposed the **direct** method adjusting the criterion for a nearly unbiased ESPE. If the correlation function ρ_n is known, Mallow's C_L criterion is

$$r^2_{CL,\rho}(t,h,n) = r^2(t,h,n) + 2\widehat{\sigma}^2 \sum_{s=-[nh/2]}^{[nh/2]} w_{h,n}(i/n, s+t)\rho_n(s)$$

Similarly, CV and GCV criteria are modified as

$$r^2_{CV,\rho}(t,h,n) = \frac{r^2(t,h,n)}{[1-\sum_{s=-[nh/2]}^{[nh/2]} w_{h,n}(t/n, s+t)\rho_n]^2}$$

$$r^2_{GCV,\rho}(t,h,n) = \frac{r^2(t,h,n)}{(1-1/n\, trW_{h,n}R_n)^2}$$

where R_n is the correlation matrix.

Hart(1991) studied the Gasser-Müller (1979) kernel estimator for the nonparametric model

$$y_t = m(\frac{t-0.5}{n}) + \epsilon_t, \quad i = t, \ldots, n$$

where $\{\epsilon_1, \ldots, \epsilon_n\}$ is from a stationary process. He incorporated the estimated covariance into a risk estimation procedure for a more efficient smoothing of positively correlated data.

He showed that the cross-validation will choose a kernel estimate that nearly interpolates the data when sufficiently positive correlation exists. Further, he found that the best predictor by cross-validation would depend on neighboring data as well as on the mean response function $m(\cdot)$. He introduced a risk estimation procedure to account for the correlation.

He defined the mean average-squared error (MASE) curve by

$$\begin{aligned} M(h) &= E[\frac{1}{n}\sum_{t=1}^{n}\{\widehat{m}_h(\frac{t-\frac{1}{2}}{n}) - m(\frac{t-\frac{1}{2}}{n})\}^2] \\ &= E[\frac{1}{n}RSS(h)] - c(0) + \frac{2}{n}\{w_0(h)c(0) + 2\sum_{t=1}^{n(h)} w_t(h)c(t)\} \end{aligned}$$

where

$$RSS(h) = \sum_{t=1}^{n} \{\widehat{m}_h(\frac{t-\frac{1}{2}}{n}) - y_t\}^2$$

$$w_t(h) = (n-t) \int_{(t-\frac{1}{2})/nh}^{(t+\frac{1}{2})/nh} K(y)\, dy$$

and $n(h)$ is the largest integer less than $nh + \frac{1}{2}$.

Assume the estimates of the covariances $c(k), k = 0, 1, \ldots,$ are known. Rice (1984) proposed an estimate of $M(h)$ for the uncorrelated errors,

$$\widehat{M}(h) = \frac{1}{n} RSS(h) - \widehat{\sigma}^2 \left\{ 1 - 2 \int_{-1/2nh}^{1/2nh} K(y)\, dy \right\},$$

where $\widehat{\sigma}^2 = (2n)^{-1} \sum_{t=2}^{n} (y_t - y_{t-1})^2$. Hart modified Rice's criterion for correlated errors,

$$\widehat{M}(h; c) = \frac{1}{n} RSS(h) - c(0) + 2n^{-1} \{w_0(h)c(0) + 2 \sum_{t=1}^{n(h)} w_t(h)c(t)\}$$

Substituting $c(k)$ in the above formula with an estimate $\tilde{c}(k)$, the optimal bandwidth minimizes the value of $\widehat{M}(h; \tilde{c})$. Hart provided a step-by-step procedure to estimate $c(k)$ without having an initial estimate of $m(\cdot)$ which can be summarized as below:

Step 1: Define $\Delta_j = y_{j+1} - 2y_j + y_{i-1}$, $j = 2, \ldots, n-1$.

Compute the periodogram

$$I_\Delta(\omega) = \tfrac{1}{T_n} |\sum_{j=2}^{n-1} \Delta_j^* e^{-i\omega j}|^2, \text{ ,} \omega \in [-\pi, \pi],$$

Where

$$\Delta_j^* = t(\tfrac{j-\frac{1}{2}}{n}) \Delta_j$$

$$T_n = 2\pi \sum_{j=2}^{n-1} t^2(\tfrac{j-\frac{1}{2}}{n})$$

and t is a twice differentiable function that vanishes at zero and unity.

Step2: Let $S(\omega_j;\theta)$ denote the spectrum of the process ϵ_j. Then estimate θ by the maximizer of the approximate log-likelihood

$$\tilde{L_n}(\theta) = -\tfrac{2\pi}{n} \sum_{\omega_j \in \Lambda} \{logS(\omega_j;\theta) + \tilde{I}_\epsilon(\omega_j)/S(\omega_j;\theta)\},$$

where $\tilde{I}_\epsilon(\omega) = |1 - e^{i\omega}|^{-4} I_\Delta(\omega), \omega_j = 2\pi j/n, \Lambda = [\delta, \pi]$ and $\delta > 0$. Due to the one-to-one mapping between $c(k)$ and S_ϵ, estimating θ yields an estimate of $c(k)$.

Different from Altman (1990) and Hart (1991) who considered the kernel-based estimator for correlated data, Francisco-Fernández, Opsomer and Vilar-Fernández(2004) studied the local polynomial estimator in a fixed design regression functional model with dependent observations and revised Ruppert, Sheather and Wand (1995) plug-in bandwidth for correlation.

The idea of the plug-in bandwidth is to minimize the asymptotic MISE given by

$$\text{MISE}(h) = \int \text{MSE}(\widehat{m}(x)) f(x)\, \mathrm{d}x$$

for the theoretical optimal bandwidth, where $\widehat{m}(\cdot)$ is the local linear estimator of the mean function $m(\cdot)$.

By Corollary 1 of Francisco-Fernández and Vilar-Fernández(2001),

$$\text{MSE}(\widehat{m}(x)) = \text{bias}^2(\widehat{m}(x)) + \text{Var}(\widehat{m}(x))$$

where the bias of $\widehat{m}(x)$ is same as that for i.i.d. data and

$$\mathrm{Var}(\widehat{m}(x)) = \frac{c(\epsilon)}{nhf(x)} \int K^2(x)\,\mathrm{d}x(1 + o_p(1))$$

where $c(\epsilon) = \sigma^2[c(0) + 2\sum_{k=1}^{\infty} c(k)]$ with $c(k) = \frac{E(\epsilon_t, \epsilon_{t+k})}{\sigma^2}, k = 0, 1, 2, \ldots$.

Correspondingly, the optimal bandwidth is given by

$$h_{\mathrm{opt}} = C(K)[\frac{c(\epsilon)}{b\int [m''(x)]^2 f(x)\,\mathrm{d}x}]^{\frac{1}{5}} = C(K)[\frac{c(\epsilon)}{n\theta_{2,2}}]^{\frac{1}{5}} \tag{2.9}$$

where $\theta_{2,2} = \int [m''(x)]^2 f(x)\,\mathrm{d}x$ and $C(K)$ is a constant depending only on the kernel K.

Two unknown quantities $c(\epsilon)$ and $\theta_{2,2}$ in (2.9) need to be estimated from the data. Francisco-Fernández, Opsomer and Vilar-Fernández(2004) proposed the local polynomial estimation for $\theta_{2,2}$ and the nonparametric estimation for $c(\epsilon)$. Details can be found in their paper.

2.1.2.2 Decorrelate the error

Besides the **direct method** mentioned in the last section, Altman (1990) also provided the **indirect** method to account for the autocorrelated errors. The idea is to impose a transformation on the covariance structure to remove the correlation and then apply the regular bandwidth selection procedures on the uncorrelated error term.

Follow the notation in the last section. Let $r_{\rho^{-1}}(\centerdot, h, n) = R_n^{-1/2} r(\centerdot, h, n)$ where R_n is the correlation matrix. The goodness-of-fit criterion is the total weighted MSE,

$$TSE_{\rho^{-1}}(h, n) = E(\widehat{m}_{h,n}(\centerdot) - m(\centerdot))' R_n^{-1} (\widehat{m}_{h,n}(\centerdot) - m(\centerdot))$$

The totalled Mallow's C_L, CV, GCV are:

$$\sum r^2_{CL,\rho^{-1}} = \sum_{i=0}^{n} r^2_{\rho^{-1}}(i,h,n) + 2\widehat{\sigma}^2\, tr\, W_{h,n}$$

$$\sum r^2_{CV,\rho^{-1}} = \sum_{i=0}^{n} \frac{r^2_{\rho^{-1}}(i,h,n)}{(1-w_{h,n}(i/n,i))^2}$$

$$\sum r^2_{CGV,\rho^{-1}} = \frac{\sum_{i=0}^{n} r^2_{\rho^{-1}}(i,h,n)}{1-1/n\, tr\, W_{h,n}}$$

Following the same idea, Francisco-Fernández and Vilar-Fernández(2002) suggested transforming the model to get the uncorrelated errors but used the local polynomial regression to estimate the mean function $m(\cdot)$. They presented an explicit solution for AR(1) type correlation for its well known covariance structure.

For the random sampling scheme, Xiao *et. al.* (2003) proposed a new method for the correlated errors, which uses a linear transformation on the original regression model that yields an uncorrelated filtered model. Xiao *et. al.* claimed that their method can be applied to a general autoregressive moving average (ARMA) model theoretically and a wide range of nonparametric estimators, including the kernel estimator and the local polynomial estimator.

Suppose $(X_1, Y_1), \ldots, (X_T, Y_T)$ are the realizations of the model $Y_t = m(X_t) + u_t, t = 1, \ldots, T$. The stationary residual u_t has mean 0 and an invertible linear process representation $u_t = \sum_{j=0}^{\infty} c_j \epsilon_{t-j}$ where ϵ_{n-j} are independent identically distributed with mean 0 and variance σ^2_ϵ.

Let $C(L) = \sum_{j=0}^{\infty} C_j L^j$, where L is the usual lag operator. Inverting $C(L)$, we get

$$C(L)^{-1} = a(L) = a_0 - a_1 L - \cdots - a_j L^j - \cdots = a_0 - \sum_{j=1}^{\infty} a_j L^j \tag{2.10}$$

Without loss of generality, define $a_0 = c_0 = 1$. Then $a(L)u_t = \epsilon_t$. Apply $a(L)$ to

the regression model, we obtain, $a(L)Y_t = a(L)m(X_t) + \epsilon_t$. Note the error term ϵ_i is now uncorrelated.

Rewrite $\underline{Y}_t = m(X_t) + \epsilon_t$ where $\underline{Y}_t = Y_t - \sum_{j=1}^{\infty} a_j(Y_{t-j} - m(X_{t-j}))$. If $\underline{Y}_t$ is known, the regular local nonparametric model can be used here. However, in practice, $\underline{Y}_t$ is unknown. The authors provided the following algorithm to estimate it:

1. Obtain a preliminary consistent estimate of $m(\cdot)$ by a local p^{th} order polynomial smoothing Y_t on X_t with corresponding kernel K_0 and bandwidth h_0. Denote the preliminary estimates as $\widehat{m}(X_t)$ and calculate the estimated residuals $\widehat{u}_t = Y_t - \widehat{m}(X_t)$.

2. Let $\tau = \tau(T)$ be some truncation parameter suitably small relative to the sample size T but large enough to avoid serious bias. Conduct a τth order autoregression of $\widehat{U}_t$,

$$\widehat{u}_t = \widehat{a}_1\widehat{u}_{t-1} + \cdots + \widehat{a}_\tau\widehat{u}_{t-\tau} + residual$$

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

$$\widehat{\mathbf{A}}_\tau = (\widehat{\mathbf{U}'}_\tau\widehat{\mathbf{U}}_\tau)^{-1}\widehat{\mathbf{U}'}_\tau\widehat{\mathbf{u}}$$

where $\widehat{\mathbf{u}} = (\widehat{u}_\tau, \ldots, \widehat{u}_T)'$ and $\widehat{\mathbf{U}}_\tau$ is the $(T-\tau) \times \tau$ matrix of regressors with typical element $\widehat{u}_{t-j}$.

3. Construct an approximation of $\underline{Y}_t$ by

$$\widehat{\underline{Y}}_t = Y_t - \sum_{j=1}^{\tau} \widehat{a}_j(Y_{t-j} - \widehat{m}(X_{t-j}))$$

2.2 Varying-coefficient models

As the modern computers become more powerful, a number of sophisticated applications have been explored. Motivated by the proportional hazards model used in survival analysis, Hastie and Tibshirani (1993) formulated the *Varying-coefficient* models in a systematic fashion.

A varying-coefficient model has the form

$$Y = \beta_0 + X_1\beta_1(U_1) + \cdots + X_p\beta_p(U_p) + \epsilon, \tag{2.11}$$

where $U_1, \ldots, U_p$ are covariates that change the effect of $X_1, \ldots, X_p$ via the unspecified functions $\beta_1(.), \ldots, \beta_p(.)$. The interaction between U_i and X_i gives us the flexibility of interpreting the model and avoids the "Curse of Dimensionality" problem.

Hastie and Tibshirani (1993) listed a few specific varying-coefficient models. Here, we describe the three most commonly used ones:

1. When $\beta_i(U_i) = \beta_i$ (the constant function), then the varying coefficient model (2.11) is just the usual linear model

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \epsilon$$

2. When $X_i = c_i$ (a fixed value), then model (2.11) becomes a generalized additive model

$$Y = \beta_0 + c_1\beta_1(U_1) + \cdots + c_p\beta_p(U_p) + \epsilon$$

3. When $U_1 = \cdots = U_p = U$ such as time or temperature, model (2.11) can be regarded as the conditionally parametric model given $U = u$. Actually,

this model is the one that has received the most attention in the statistics literature. The new estimation procedures that will be presented later are based on this model.

For general varying-coefficient model (2.11), our task is to estimate $\beta_1(.), \ldots, \beta_p(.)$ by minimizing

$$E\{Y - \sum_{i=1}^{p} X_i\beta_i(U_i)\}^2$$

Conditioning on U_i, the solution is obtained by taking the derivative with respect to β_i:

$$E[X_i\{Y - \sum_{i=1}^{p} X_i\beta_i(U_i)\}|U_i] = 0$$

Therefore,

$$\beta(U_i) = \frac{E[X_i\{Y - \sum_{j\neq i} X_j\beta_j(U_j)\}|U_j]}{E(X_i^2|U_i)}, i = 1, \ldots, p$$

We need to solve p equations simultaneously for $\beta(.)$.

Similar to the generalized linear model (GLIM), a generalized varying-coefficient model is defined by

$$\eta(U, \mathbf{X}) = g\{m(U, \mathbf{X})\} = \beta_0(U_0) + X_1\beta_1(U_1) + \cdots + X_p\beta_p(U_p) \tag{2.12}$$

where $g(\cdot)$ is a link function that transforms the mean function $m(u, \mathbf{X})$ to a linear predictor.

Many existing models can be derived from model (2.12). For example, when the coefficient functions $\beta_i(\cdot)$ are all constant, model (2.12) is just a GLIM (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989)

$$\eta(u, \mathbf{X}) = g\{m(u, \mathbf{X})\} = \beta_0 + X_1\beta_1 + \cdots + X_p\beta$$

If only the intercept $\beta_0()$ depends on the covariate U and other coefficients are constants, model (2.12) becomes the generalized partially linear model (Carroll, Fan, Gijbels and Wand, 1997)

$$\eta(u, \mathbf{X}) = g\{m(u, \mathbf{X})\} = \beta_0(U) + X_1\beta_1 + \cdots + X_p\beta$$

2.2.1 Independent Data

Many estimation methods can be employed to estimate the varying-coefficient model. For instance, kernel estimation, local spline smoothing, local linear regression, etc. Since our proposed estimation procedures for the correlated data use the local smoothing technique, we focus on this method in this section.

Assume that $(U_1, \mathbf{X}_1, Y_1), \ldots, (U_n, \mathbf{X}_n, Y_n)$ are a random sample from

$$Y_t = X_{t0}\beta_0(U_t) + X_{t1}\beta_1(U_t) + \cdots + X_{tp}\beta_p(U_t) + \epsilon_t, \quad t = 1, \ldots, n \tag{2.13}$$

where $\beta_0(\cdot), \ldots, \beta_p(\cdot)$ are smooth functions of U and ϵ_t is i.i.d. error. Usually $X_{t0} \equiv 1$.

We use the local linear regression to estimate the functional coefficient $\beta_i(\cdot)$. In a neighborhood of u_0, we can approximate $\beta_i(U)$ by a linear function

$$\beta_i(U) \approx \beta_i(u_0) + \beta_i'(u_0)(U - u_0) \equiv a_i + b_i(U - u_0), \quad i = 1, \ldots, p$$

This leads to the locally weighted least squared function

$$\sum_{t=1}^{n}[Y_t - \sum_{i=0}^{p}\{a_i + b_i(U_t - u_0)\}X_{ti}]^2 K_h(U_t - u_0) \tag{2.14}$$

where K is a kernel function, h is a bandwidth and $K(\cdot) = K(\cdot/h)/h$.

We minimize (2.14) with respect to $\{(a_i, b_i), i = 0, \ldots, p\}$ and get the estimator $\{\widehat{a}_i, i = 0, \ldots, p\}$, which is the estimate of the coefficient function $\{\beta_i(u_0), i = 0, \ldots, p\}$ at the given point u_0.

We refer to this local linear procedure as the one-step method. This procedure is good for the situation when the coefficient functions admit the same degree of smoothness. Otherwise, the local linear approximation may not be accurate and may yield a significant bias. Fan and Zhang (1999) showed that the bias of the one-step estimator is $O(h^2)$ and the variance of the one-step estimator is $O((nh)^{-1})$, which cannot reach the optimal convergence rate $O(n^{-\frac{8}{9}})$.

In order to fit $\beta_i(.)$ accounting for the different degrees of smoothness, Fan and Zhang (1999) proposed a two-step least-squares estimation procedure.

Let's assume that $\beta_p(.)$ is smoother than the rest of $\beta_i(.), (i = 0, \ldots, p-1)$. Specifically, assume that $\beta_p(.)$ has a bounded fourth derivative. Then

$$\beta_p(u) \approx a_p + b_p(U - u_0) + c_p(U - u_0)^2 + d_p(U - u_0)^3$$

The weighted least squares problem becomes

$$\sum_{t=1}^{n}[Y_t - \sum_{i=0}^{p-1}\{a_i + b_i(U_t - u_0)\}X_{ti} - \{a_p + b_p(U_t - u_0) - c_p(U_t - u_0)^2 - d_p(U_t - u_0)^3\}X_{tp}]^2 K_h(U_t - u_0)$$

The coefficients we want to estimate are $\widehat{a}_{i,1}, \widehat{b}_{i,1}, (i = 0, \ldots, p-1)$ and $\widehat{a}_{p,1}$, $\widehat{b}_{p,1}$, $\widehat{c}_{p,1}$, $\widehat{d}_{p,1}$. First, use a small bandwidth to get an initial estimate of $\beta_0(.), \ldots, \beta_{p-1}(.)$. Usually, the initial estimate is undersmoothed because a small bias is desirable. Next we will carry out the two-step estimation procedure.

Step 1: Substitute the initial estimate into the local weighted least-squares

$$\sum_{t=1}^{n}[Y_t - \sum_{i=0}^{p}\{a_i + b_i(U_t - u_0)\}X_{ti}]^2 K_{h0}(U_t - u_0)$$

Minimize it to obtain a preliminary estimate $\widehat{\beta}_{0,0}(u_0), \ldots, \widehat{\beta}_{p,0}(u_0)$.

Step 2: Substitute $\widehat{\beta}_{0,0}(u_0), \ldots, \widehat{\beta}_{p-1,0}(u_0)$ and use a local cubic fit to estimate $\beta_p(.)$, i.e, minimize

$$\sum_{t=1}^{n}[Y_t - \sum_{i=1}^{p-1} \widehat{\beta}_{i,0}(U_i)X_{ti} - \{a_p + b_p(U_t - u_0) - c_p(U_t - u_0)^2 - d_p(U_t - u_0)^3\}X_{tp}]^2 K_h(U_t - u_0)$$

with respect to a_p, b_p, c_p, d_p.

Note that the bandwidth in these two steps are different. Fan and Zhang (1999) showed that the two-step estimation method is not sensitive to the initial bandwidth h_0 from their practical experience. Moreover, they argued that the bias of the two-step estimator is of $O(h_2^4)$ and the variance is of $O\{(nh_2)^{-1}\}$ provided $h_0 = o(h_2^2), nh_0/\log h_0 \to \infty$ and $nh_0^3 \to \infty$. Thus the optimal convergence rate $O(n^{-\frac{8}{9}})$ is achieved.

For the generalized varying-coefficient model, we can estimate the coefficients $\beta_i(\cdot)$ through the local likelihood approach. Consider the generalized varying-coefficient model as below

$$g\{E(Y|U, X_0, \cdots, X_p)\} = \sum_{i=0}^{p} \beta_i(U)X_i$$

where $g(\cdot)$ is a link function. By a Taylor expansion, when U is in a neighborhood of u_0,

$$\beta_i(U) \approx a_i + b_i(U - u_0)$$

Then we can find its local maximum likelihood estimate (MLE) by using an iterative algorithm. The challenge is that we have to solve hundreds of local like-

lihood equations for distinct values of u_0. If cross-validation is used to select the bandwidth, it will burden the computation even more.

Cai, Fan and Li (2000) proposed a one-step local MLE procedure to reduce the computational cost and showed that their one-step estimate is asymptotically as efficient as the fully iterative MLE. In addition, the hypothesis test regarding to whether some coefficients are actually varying or constant can be carried out based on the sampling properties.

Their main idea is to replace the iterative local MLE by the one-step Newton-Raphson estimator. Define a local likelihood

$$\ell(\beta) = \sum_{t=1}^{n} \log(L[g^{-1}\{\sum_{i=0}^{p}(a_i + b_i(U_t - u_0))X_{ti}\}, Y_t]K_h(U_t - u_0)),$$

where $\beta = (a_0, \cdots, a_p, b_0, \cdots, b_p)^T$, $L(\cdot, \cdot)$ is the conditional likelihood function of Y given $\mathbf{X}$, and $K_h(.) = K(./h)/h$ is a given kernel function and h is the bandwidth. $\ell'(\beta)$ and $\ell''(\beta)$ denote the gradient and Hessian matrix of $\ell(\beta)$. Given an initial estimator $\widehat{\beta_0} = \widehat{\beta_0}(u_0) = (\widehat{a}(u_0)^T, \widehat{b}(u_0)^T)$, one-step Newton-Raphson algorithm produces the updated estimator,

$$\widehat{\beta_{os}} = \widehat{\beta_0} - \{\ell''(\widehat{\beta_0})\}^{-1}\ell'(\widehat{\beta_0})$$

Usually, we use the least-squares estimates as the initial estimator.

It is worth to point out that the one-step local MLE and two-step least-squares estimation are not the ultimate methods in varying-coefficient modelling field. Although they have their advantages, they also bear drawbacks. If the Hessian matrix $l''(\widehat{\beta_0})$ is nearly singular, the one-step local MLE will be in trouble. It is not easy to find the asymptotic distribution of the two steps estimators. Therefore, there are many modified methods to improve estimation. For example, Tang and Wang

(2005) proposed a comparable one-step least-squares procedure which can handle the different degrees of smoothness as a two-step estimation.

Bandwidth selection is crucial in the local linear fitting for the varying-coefficient models. The *least squares cross validation* method and the *plug-in* method discussed in the nonparametric models can be adapted in the varying-coefficient context by replacing the corresponding estimators with the ones obtained by the local linear regression for the varying-coefficient model.

When the estimate of the functional coefficient $\beta_i(\cdot)$ is obtained, we naturally want to test whether the estimated coefficient is really varying over the covariate U. Consider the following hypothesis test:

$$H_0 : \beta_i(U) = \beta_{i0} \quad \text{versus} \quad H_1 : \beta_i(U) \neq \beta_{i0} \quad \text{for some } U,\ i = 0, \ldots, p$$

where β_{i0} is an unknown constant.

Denote $\ell(H_0)$ and $\ell(H_1)$ to be the log-likelihood under H_0 and H_1 respectively. The log-likelihood statistics is $T = 2\{\ell(H_1) - \ell(H_0)\}$. We expect a small T if H_0 is true but a large T if H_1 holds. For parametric models, T follows an asymptotic χ^2-distribution with degrees of freedom $f - r$, where f and r are the dimensions of parameter spaces of H_1 and H_0. However, the parameter space under H_1 in our hypothesis test can be infinitely large. The traditional likelihood ratio test can not be applied for this hypothesis testing problem. To solve this problem, Can, Fan and Li (2000) proposed the generalized likelihood ratio (GLR) test and showed that the Wilk's type of phenomenon holds.

2.2.2 Longitudinal data

Longitudinal data often occur in medical and epidemiological studies. Observations are obtained from n independent subjects each repeatedly measured at a set of distinct time points. Longitudinal data are assumed to be independent between subjects but can be correlated within subjects. The traditional parametric methods such as multivariate linear regression, analysis of variance and general linear models can be used to analyze longitudinal data. However, parametric models always raise the problems of restricted assumptions and misspecification. Let us consider the varying coefficient model $Y(t) = X(t)^T\beta(t) + \epsilon(t)$ in the longitudinal scenario.

Assume $\{(t_{ij}, Y_{ij}, X_{ij}^T) : 1 \leq i \leq n, 1 \leq j \leq n_i\}$,where $X_{ij}^T = (1, X_{ij}^{(1)}, \ldots, X_{ij}^{(k)})^T$ denotes the covariates. Then the corresponding varying coefficient model is

$$Y_i(t_{ij}) = X_i(t_{ij})^T\beta(t_{ij}) + \epsilon_i(t_{ij}) \tag{2.15}$$

where $\epsilon_i(t)$ is a zero-mean stochastic process.

Hoover, Rice, Wu and Yang (1998) suggested two computationally straightforward estimators of $\beta(t)$: smoothing spline and locally weighted polynomial. But smoothing spline is not preferred because it is very computationally intensive, even for a longitudinal data set of modest size.

Hoover, Rice, Wu and Yang (1998) proposed a local estimation method. The estimator of $\beta(t)$ can be obtained by minimizing

$$L(t) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}[Y_{ij} - X_i^T(t_{ij})b(t)]^2K(\frac{t - t_{ij}}{h}) \tag{2.16}$$

with respect to $b(t)$. In the matrix form,

$$L(t) = \sum_{i=1}^{n} (Y_i - X_i\beta(t))^T K_i(t;h)(Y_i - X_i\beta(t))$$

where

$$X = \begin{pmatrix} 1 & X_{i1}(t_{i1}) & \cdots & X_{ik}(t_{i1}) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{i1}(t_{in_i}) & \cdots & X_{ik}(t_{in_i}) \end{pmatrix}, y = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix},$$

and $K_i(.,.)$ is a diagonal weight matrix such that

$$K_i(t;h) = diag(K(\frac{t - t_{i1}}{h}), \ldots, K(\frac{t - t_{in_i}}{h}))$$

Assume that $(\sum_{i=1}^{n} X_i^T K_i(t;h) X_i)$ is invertible, then

$$\widehat{\beta}(t;h) = (\sum_{i=1}^{n} X_i^t K_i(t;h) X_i)^{-1} (\sum_{i=1}^{n} X_i^t Y_i)$$

Since the subjects are independent but the measurements within subjects are correlated, Rice and Silverman (1991) suggested using a "leave-one-subject-out" cross-validation rather than single response cross-validation.

Note that (2.16) uses one set bandwidth to estimate all $(k+1)$ curves of $\beta(t)$. It may not be appropriate if $\beta(t)$ admits different degrees of smoothness. Wu and Chiang (2000) proposed two componentwise local least squares estimators based on two weight schemes.

Write (2.15) in matrix form: $Y(t) = X^T(t)\beta(t) + \epsilon(t)$. Multiple both sides by X and take the expectations:

$$\beta(t) = (E_{XX^T})^{-1} E[XY(t)], \text{ given } E_{XX^T} \text{ is invertible.}$$

Let e_{rl} be the $(r,l)^{\text{th}}$ element of $E^{-1}_{XX^T}$. Then, for $r = 0, \ldots, k$,

$$\beta_r(t) = E[(\sum_{l=0}^{k} e_{rl} X^{(l)}) Y(t)] \tag{2.17}$$

Estimate E_{XX^T} by $n^{-1}\sum_{i=1}^{n}(X_i X_i^T)$. $\widehat{e}_{rl}$ is a natural estimator of e_{rl}. Assign weight $(nn_i)^{-1}$ to each measurement of the ith subject and minimize:

$$L_r(t) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\{(\frac{1}{nn_i})[(\sum_{l=0}^{k}\widehat{e}_{rl}X_i^{(l)})Y_{ij} - b_r(t)]^2 K_r(\frac{t-t_{ij}}{h_r})\} \tag{2.18}$$

with respect to $b_r(t)$. By the local least squares theory,

$$\tilde{\beta}_r(t) = \frac{\sum_{i=1}^{n}\{n_i^{-1}\sum_{j=1}^{n_i}[(\sum_{l=0}^{k}\widehat{e}_{rl}X_i^{(l)})Y_{ij}K_r((t-t_{ij})/h_r)]\}}{\sum_{i=1}^{n}\{n_i^{-1}\sum_{j=1}^{n_i}[K_r((t-t_{ij})/h_r)]\}} \tag{2.19}$$

If the uniform weight N^{-1} is assigned to all the measurements, (2.18) will yield an alternative local least square estimator by minimizing:

$$L_r^*(t) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\{(\frac{1}{N})[(\sum_{l=0}^{k}\widehat{e}_{rl}X_i^{(l)})Y_{ij} - b_r(t)]^2 K_r(\frac{t-t_{ij}}{h_r})\} \tag{2.20}$$

The corresponding estimator is:

$$\tilde{\beta}_r^*(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}[(\sum_{l=0}^{k}\widehat{e}_{rl}X_i^{(l)})Y_{ij}K_r((t-t_{ij})/h_r)]}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}[K_r((t-t_{ij})/h_r)]} \tag{2.21}$$

Alternatively, Fan and Zhang (2000) proposed a two-step estimation procedure to overcome the drawback caused by having one set of smoothing parameter when the coefficient functions admit different degrees of smoothness. Their idea is as follows:

"First calculate the raw estimates of the coefficient functions via fitting a stan-

dard linear model and then smooth the raw estimates to obtain the smooth estimates of the coefficient functions by using one of the existing smoothing techniques."

Their method is fast to implement. In the first step, we just need to solve a simple linear model with independent errors at a particular point. So only the data falling in the neighborhood will be used. In the second step, it is just a one-dimensional smoothing problem.

Specifically speaking, in raw estimate step, let $t_j, j = 1, 2, \ldots, T$ denote the distinct time points among $\{t_{ij}, j = 1, 2, \ldots, n_i; i = 1, \ldots, n\}$. For each t_j, let N_j denote the number of subjects y_{ij} observed at t_j. Collect all X_{ij} and Y_{ij} that are in N_j and denote them by $\tilde{X}_j$ and $\tilde{Y}_j$ respectively. Then (2.15) becomes:

$$\tilde{Y}_j = \tilde{X}_j \beta(t_j) + \tilde{e}_j$$

where $E(\tilde{e}_j) = 0, cov(\tilde{e}_j) = \gamma(t_j, t_j) I_{n_j}$.

Then the estimator of $\beta(t_j)$ is $b(t_j) = (\tilde{X}_j^T \tilde{X}_j)^{-1} \tilde{X}_j^T \tilde{Y}_j$. For $r = 0, \ldots, k$, let $b_r(t_j)$ be the rth component of $b(t_j)$. Then $b_r(t_j) = e_{r,k+1}^T (\tilde{X}_j^T \tilde{X}_j)^{-1} \tilde{X}_j^T \tilde{Y}_j$ where $e_{r,k+1}$ denotes a *k+1*dimensional unit vector with 1 at its rth entry.

Usually, the raw estimators will under-smooth the coefficient functions. So we need to refine the raw estimates. The intuitive way is to smooth them over time. Assume we already have raw estimates $\{(t_j, b_r(t_j)), j = 1, \ldots, T\}$ from the first step. In light of local polynomial fit, a smooth estimator is:

$$\widehat{\beta_r^{(q)}}(t) = \sum_{j=1}^{T} w_r(t_j, t) b_r(t_j) \tag{2.22}$$

where the weights $w_r(t_j, t)$ can be constructed as follows: Let $C_j = (1, t_j - t, \ldots, (t_j - t)^p)^T, j = 1, \ldots, T$, $C = (C_1, C_2, \ldots, C_T)^T$ and $W = diag(W_1, W_2, \ldots, W_T)$ with

$W_j = K_h(t_j - t)$.

Then:

$$w_{p+1}(t_j, t) = e_{p+1}^T (C^T W C)^{-1} C_j W_j,\ j = 1, \ldots, T.$$

2.2.3 Time series data

For time dependent data, linear models such as the autoregressive moving average (ARMA) model have been used for a long time. However, it has been found that the linear models sometimes cannot capture complicated nonlinear features.

A number of successful nonlinear models arising from practice have been introduced. For example, the autoregressive conditional heteroscedastic (ARCH) model is concerned with modelling the high volatility of the time series data:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_p x_{pt} + \sigma_t \epsilon_t$$

where $\{\epsilon_t\}$ is Gaussian white noise with mean 0 and variance 1 and $\{\sigma_t\}$ is positive and time varying. Usually $\{\sigma_t\}$ is modelled to be an autoregressive process.

Motivated by economic and ecological data, Tiao and Tsay (1994) and Tong (1990) developed the threshold autoregressive model(TAR):

$$x_t = \phi_1^{(i)} x_{t-1} + \cdots + \phi_p^{(i)} x_{t-p} + \epsilon_t^{(i)} \quad \text{if } x_{t-d} \in \Omega_i,\ i = 1, \ldots, k$$

where $\{\Omega_i\}$ form a non-overlapping partition of the real line. This model was applied to the well-known Canadian lynx data by Tong (1990) and turned out to have a nice interpretation in the ecological sense.

Later, Chen and Tsay (1993) proposed the functional-coefficient autoregressive model (FAR):

$$x_t = f_1(X_{t-1}^*) x_{t-1} + \cdots + f_p(X_{t-1}^*) x_{t-p} + \epsilon_t \tag{2.23}$$

where $X^*_{t-1} = (x_{t-i_1}, x_{t-i_2}, \ldots, x_{t-i_k})'$,$\{\epsilon_t\}$ is a sequence of i.i.d random variables. This model outperforms the threshold model (TAR) because it can depict the gradual change of the coefficient function.

Chen and Tsay (1993) proposed an iterative algorithm to estimate the coefficient function. For simplicity, we consider $X^*_{t-1} = U$ (i.e., a single threshold variable). Our target is to estimate the function $f_i(.)$ at various values U. Consider the simple linear regression: $x_t = a_1 x_{t-1} + \cdots + a_p x_{t-p} + \epsilon_t$ with $t = t_1, t_2, \ldots, t_k$ and $a_i = f_i(U)$. We can use ordinary least squares (OLS) to estimate a_i. Denote such an estimate of $f_i(u)$ by $\widehat{f_i}(u)$. By plotting $\widehat{f_i}(U)$ versus x, we can obtain an estimate of the functional form of $f_i(.)$.

Notice that model (2.23) is a varying-coefficient model except that the response variable is also from $\{x_i\}$. Cai, Fan and Yao (2000) used the general varying-coefficient fitting technique to estimate $f_i(.)$. Approximate $f_i(.)$ locally at u_0 by a linear function $f_i(U) \approx a_i + b_i(U - u_0)$.We need to minimize the sum of weighted squares:

$$\sum_{k=1}^{n} [x_k - \sum_{i=1}^{p} a_i + b_i(U_k - u_0)x_{k-i}]^2 K_h(U_k - u_0)$$

Considering the structure of the stationary time series data, Cai, Fan and Yao (2000) suggested a modified multifold cross-validation criterion to select the bandwidth. Let m and Q be two positive integers such that $n > mQ$. We choose h that minimizes the average mean squared error (AMS):

$$AMS(h) = \textstyle\sum_{q=1}^{Q} AMS_q(h) \quad \text{for } q = 1, \ldots, Q,$$
$$AMS_q(h) = \frac{1}{m} \textstyle\sum_{n-qm+1}^{n-qm+m} \{x_i - \sum_{j=1}^{p} \widehat{f}_{j,p}(U_i)x_{i-j}\}^2$$

In practical implementation, they suggest $m = [0.1n], Q = 4$. This bandwidth selection procedure will save computational cost significantly compared with leave-one-out cross validation.

Recently Cai (2005) studied a time-varying coefficient model with a time trend function and serially correlated errors:

$$Y_i = \beta_{i0} + \sum_{j=1}^{p} \beta_{ij} x_{ij} = \mathbf{X}_i^T \boldsymbol{\beta}_i + \epsilon_i \tag{2.24}$$

where $\mathbf{X}_i = (1, x_{i1}, \ldots, x_{ip})^T$,$\boldsymbol{\beta}_i = (\beta_{i0}, \ldots, \beta_{ip})^T$, $\beta_{ij} = \beta_j(t_i)$ with $t_i = i/n$, $E(\epsilon_i|\mathbf{X}_i) = 0$ and $\mathrm{Var}(\epsilon_i|\mathbf{X}_i) = \sigma_i^2(\mathbf{X}_i)$. By using the first order Taylor approximation in a neighborhood of a fixed time point $t \in [0, 1]$,

$$\beta_j(t_i) \approx a_j + b_j(t_i - t), \quad j = 0, \ldots, p$$

Then (2.24) is approximated by

$$Y_i \approx \mathbf{Z}_i^T \boldsymbol{\theta} + u_i \quad \text{where} \quad \mathbf{Z}_i = \begin{pmatrix} \mathbf{X}_i \\ \mathbf{X}_i(t_i - t) \end{pmatrix} \text{ and } \quad \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\beta}(t) \\ \boldsymbol{\beta}'(t) \end{pmatrix} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

To take account of the autocovariance of $\{\epsilon_i\}$, the adjusted locally weighted sum of squares is

$$(\mathbf{Y} - \mathbf{Z}\boldsymbol{\theta})^{\mathbf{T}} \mathbf{W}^{\frac{1}{2}} \boldsymbol{\Sigma}^{-1} \mathbf{W}^{\frac{1}{2}} (\mathbf{Y} - \mathbf{Z}\boldsymbol{\theta}) \tag{2.25}$$

where $\mathbf{Y}$ and $\mathbf{X}$ are obtained by stacking Y_i and $\mathbf{Z}_i$, $W^{\frac{1}{2}}$ is a diagonal matrix whose i^{th} diagonal element is $K_h^{\frac{1}{2}}(t_i - t)$ and Σ is the covariance of $\{\epsilon_i\}$.

The minimizer of (2.25) with respect to $\boldsymbol{\theta}$ will provide the local linear estimate for $\boldsymbol{\beta}(t)$.

Chapter 3

Nonparametric regression models for data with AR errors

Suppose that $(x_t, y_t), t = 1, \cdots, n$ is a random sample from the nonparametric regression model

$$y_t = m(x_t) + \epsilon_t, \tag{3.1}$$

where error process ϵ_t is a auto-correlated random error with mean zero and finite variance.

3.1 A new estimation procedure

The autoregressive (AR) error ϵ_t can be represented as a linear process

$$\epsilon_t = \beta_1 \epsilon_{t-1} + \cdots + \beta_d \epsilon_{t-d} + \eta_t,$$

where η_t is independently and identically distributed random error with mean zero and variance σ^2. The order d can be large, and the selection of the order d will be

discussed in this section as well. If the values for ϵ_t were available, then we could work on the following partially linear model

$$y_t = m(x_t) + \beta_1 \epsilon_{t-1} + \cdots + \beta_d \epsilon_{t-d} + \eta_t.$$

In practice, ϵ_t is not available, but it may be estimated by $\widehat{\epsilon}_t = y_t - \widehat{m}_I(x_t)$, where $\widehat{m}_I(\cdot)$ is a local linear estimate of $m(\cdot)$ based on (3.1) without considering the AR error structure. We will address the issue of bandwidth selection for $\widehat{m}(\cdot)$ later.

Replacing ϵ_t's with $\widehat{\epsilon}_t$'s, we have

$$y_t = m(x_t) + \mathbf{e}_t^T \boldsymbol{\beta} + \eta_t, \tag{3.2}$$

where $\mathbf{e}_t = (\widehat{\epsilon}_{t-1}, \cdots, \widehat{\epsilon}_{t-d})^T$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)^T$. In Section 3.1.1, we propose a new estimation procedure for $m(\cdot)$ and $\boldsymbol{\beta}$ based on model (3.2). We further propose an order selection procedure for the AR series by using the penalized profile least squares method in Section 3.1.2.

3.1.1 Profile least squares estimate

For the partially linear model (3.2), there exist various estimation procedures, including partial spline estimate (Wahba, 1984; Heckman, 1986; Engle *et al.*, 1986), partial residual method (Speckman, 1998) and profile least squares or likelihood method (Severini and Wong, 1992). Here we will employ the profile least squares techniques to estimate $\boldsymbol{\beta}$ and $m(\cdot)$.

For given $\boldsymbol{\beta}$, denote $y_t^* = y_t - \mathbf{e}_t^T \boldsymbol{\beta}$ for $t = d+1, \cdots, n$. Then

$$y_t^* = m(x_t) + \eta_t \tag{3.3}$$

which is a one-dimensional nonparametric model. We may employ existing linear smoothers, such as local polynomial regression and smoothing splines (Gu and Kim, 2002), to estimate $m(\cdot)$. Here we use the local linear regression. For a given x_0, we locally approximate the regression function

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) \widehat{=} a + b(x - x_0)$$

for x in the local neighborhood of x_0. Thus, the local linear estimate of $m(\cdot)$ is the minimizer of the following weighted least squares function

$$(\widehat{a}, \widehat{b})^T = \text{argmin}_{(a,b)} \sum_{t=d+1}^{n} \{y_t^* - a - b(x_t - x_0)\}^2 K_h(x_t - x_0),$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ is a scaled kernel function of kernel $K(\cdot)$ with bandwidth h. It is clear that the local linear estimate is linear in terms of $\mathbf{y}^* = (y_{d+1}^*, \cdots, y_n^*)^T$. Let $\widehat{\mathbf{M}} = (\widehat{m}(x_{d+1}), \cdots, \widehat{m}(x_n))^T$. Then $\widehat{\mathbf{M}}$ can be represented by

$$\widehat{\mathbf{M}} = S_h \mathbf{y}^*, \tag{3.4}$$

where S_h is a $(n-d)\times(n-d)$ smoothing matrix depending on x_t's and the bandwidth only.

Substituting $m(x_t)$ in (3.3) with $\widehat{m}(x_t)$, we obtain a synthetic linear regression model

$$(I - S_h)\mathbf{y} = (I - S_h)\mathbf{E}\boldsymbol{\beta} + \boldsymbol{\eta},$$

where I is the identity matrix, $\mathbf{E} = (\mathbf{e}_{d+1}, \cdots, \mathbf{e}_n)^T$ and $\boldsymbol{\eta} = (\eta_{d+1}, \cdots, \eta_n)^T$. Thus,

the profile least squares estimator for $\boldsymbol{\beta}$ and $\mathbf{M}$ is

$$\widehat{\boldsymbol{\beta}} = \{\mathbf{E}^T(I - S_h)^T(I - S_h)\mathbf{E})^{-1}\mathbf{E}^T(I - S_h)^T(I - S_h)\mathbf{y}, \tag{3.5}$$

and

$$\widehat{\mathbf{M}} = S_h(\mathbf{y} - \mathbf{E}\widehat{\boldsymbol{\beta}}), \tag{3.6}$$

respectively.

The asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ and the asymptotic bias and variance of $\widehat{m}(x_0)$ are given in the following theorem. Denote $\mu_i = \int x^i K(x)\,\mathrm{d}x$ and $\nu_i = \int x^i K^2(x)\,\mathrm{d}x$.

Theorem 1. *Suppose that Conditions A—G listed in Section 3.3 hold. Then*

(A) The asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ is given by

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{L} N(0, \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}^{-1})$$

where $\mathbf{f}_t = (\epsilon_{t-1}, \ldots, \epsilon_{t-d})^T$ and $\sigma^2 = Var(\eta_t)$.

(B) The asymptotic distribution of $\widehat{m}(x_0, \widehat{\boldsymbol{\beta}})$, conditioning on $x_1, \cdots, x_n$, is given by

$$\sqrt{nh}\{\widehat{m}(x_0, \widehat{\boldsymbol{\beta}}) - m(x_0) - \frac{1}{2}\mu_2 m''(x_0)h^2\} \xrightarrow{L} N(0, \frac{\nu_0\sigma^2}{f(x_0)}),$$

where $f(x)$ is the density function of x.

Note that the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ is the same as that of the Yule-Walker estimator for the AR model:

$$\epsilon_t = \beta_1\epsilon_{t-1} + \cdots + \beta_d\epsilon_{t-d} + \eta_t.$$

(see Theorem 8.1.1 of Brockwell and Davis, 1991). In other words, Theorem 1 implies that $\widehat{\boldsymbol{\beta}}$ is as efficient as if one knew the true regression function $m(\cdot)$ in advance. The asymptotic bias and variance of $\widehat{m}(\cdot, \widehat{\boldsymbol{\beta}})$ are the same as those of the local linear regression for independently and identically distributed observations, respectively. This implies that the proposed profile least squares estimate is very efficient.

3.1.2 SCAD procedure for the AR process

To implement the profile least squares estimation procedure, we have to determine the order of AR process. In practice, we may start with a large order AR process, and then apply a variable selection procedure to select its order. The penalized likelihood procedure with the smoothly clipped absolute deviation (SCAD) penalty was proposed for variable selection in parametric models in Fan and Li (2001). The SCAD procedure is distinguished from traditional variable selection procedures, such as the stepwise regression and best subset selection with the AIC and BIC, in that it selects significant variables and estimates their coefficients simultaneously. Thus, it can be directly applied for high-dimensional data analysis. The SCAD procedure was further developed for the partially linear model with longitudinal data in Fan and Li (2004). In this section, we apply the SCAD procedure to determine the complexity of AR process.

The SCAD penalized least squares function is defined to be

$$\frac{1}{2}\sum_{t=d+1}^{n}\{y_t - m(x_t) - \mathbf{e}_t^T\boldsymbol{\beta}\}^2 + n\sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|)$$

where $p_\lambda(|\beta|)$ is the SCAD penalty with a tuning parameter λ, defined by

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta|, & \text{if } 0 \le |\beta| < \lambda; \\ \frac{(a^2-1)\lambda^2-(|\beta|-a\lambda)^2}{2(a-1)}, & \text{if } \lambda \le |\beta| < a\lambda; \\ \frac{(a+1)\lambda^2}{2}, & \text{if } |\beta| \ge a\lambda. \end{cases}$$

Fan & Li (2001) suggested fixing $a = 3.7$ from a Bayesian argument. Figure 3.1 depicts the SCAD penalty with $\lambda = 1$.

Applying the profile technique for penalized least squares, we can derive the penalized profile least squares estimate, the minimizer of the following penalized

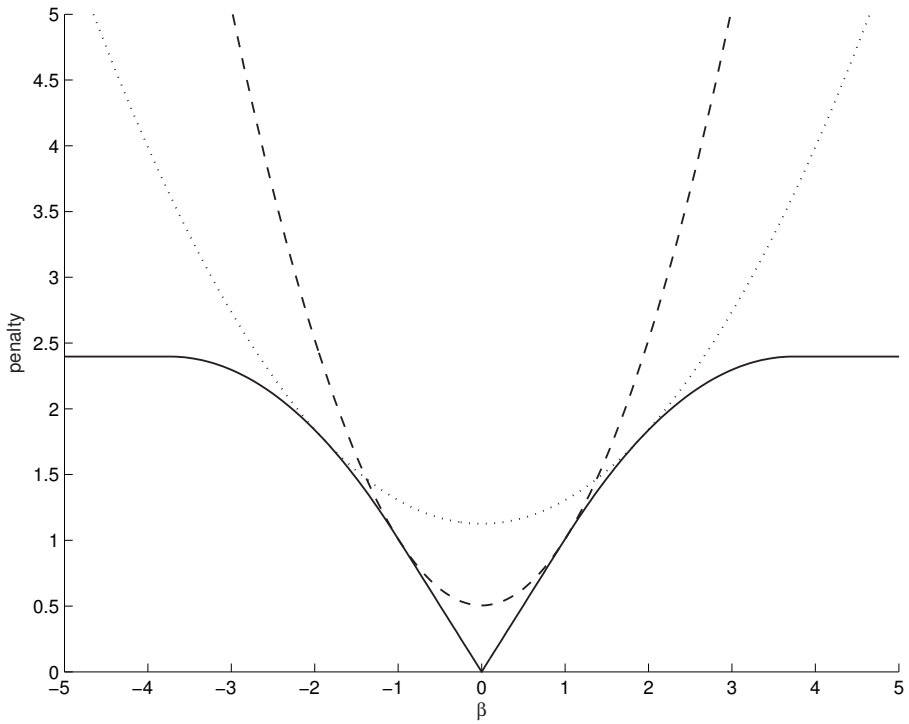

Figure 3.1. The Scad penalty function and its local quadratic approximation

least squares criterion

$$\frac{1}{2}\|(I-S_h)\mathbf{y}-(I-S_h)\mathbf{E}\boldsymbol{\beta}\|^2+n\sum_{j=1}^{d}p_{\lambda_j}(|\beta_j|). \tag{3.7}$$

As demonstrated in Fan and Li (2004), with the proper choice of the tuning parameter, the resulting estimate contains some exact zero coefficients. This is equivalent to excluding the corresponding terms from the selected model and reducing model complexity. Since the SCAD penalty function is a nonconvex function over $[0,\infty)$, it is challenging to minimize the SCAD penalized profile least squares function. Following Fan and Li (2004), we employ the local quadratic approximation (LQA) for the SCAD penalty function. Suppose we can get an estimate $\beta_j^{(k)}$ in the k^{th} step that is close to the true β_j. If $|\beta_j^{(k)}|$ is close to 0, then set $\widehat{\beta}_j=0$. Otherwise, the SCAD penalty can be locally approximated by a quadratic function as

$$[p_{\lambda_j}(|\beta_j|)]'=p'_{\lambda_j}(|\beta_j|)\cdot\operatorname{sgn}(\beta_j)\approx\{p'_{\lambda_j}(|\beta_j^{(k)}|)/|\beta_j^{(k)}|\}\beta_j$$

This is equivalent to

$$p_{\lambda_j}(|\beta_j|)\approx p_{\lambda_j}(|\beta_{j0}|)+\frac{1}{2}\{p'_{\lambda_j}(|\beta_j^{(k)}|)/|\beta_j^{(k)}|\}(\beta_j^2-\beta_j^{(k)2})$$

With the aid of LQA, we may employ the following iterative ridge regression to find the minimizer of (3.7):

$$\boldsymbol{\beta}^{(k+1)}=\{\mathbf{E}^T(I-S_h)^T(I-S_h)\mathbf{E}+n\Sigma_\lambda(\boldsymbol{\beta}^{(k)})\}^{-1}\mathbf{E}^T(I-S_h)^T(I-S_h)\mathbf{y} \tag{3.8}$$

where $\Sigma_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(k)})=\operatorname{diag}\{p'_{\lambda_1}(|\beta_1^{(k)}|)/|\beta_1^{(k)}|,\ldots,p'_{\lambda_d}(|\beta_d^{(k)}|)/|\beta_d^{(k)}|\}$ for nonvanished $\boldsymbol{\beta}^{(k)}$.

3.1.3 Tuning parameter selection and bandwidth selection

In this section, we address how to determine the tuning parameter $\boldsymbol{\lambda}$ in the SCAD procedure and how to select a bandwidth for the profile least squares estimation procedure, two important issues in the practical implementation of the proposed methodology.

Tuning parameter selection. In the implementation of the SCAD procedure, we need to choose the tuning parameter $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_d)$. Following the suggestion of Wang, Li and Tsay (2007), we use the BIC selector to find the optimal $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_d)$. From (3.8), we define the effective number of parameters of the penalized least squares estimator (3.8) to be

$$e(\boldsymbol{\lambda}) = \mathrm{tr}[\{\tilde{D} + \Sigma_{\boldsymbol{\lambda}}(\widehat{\beta})\}^{-1}\tilde{D}]$$

where $\tilde{D} = \mathbf{E}^T(I - S_h)^T(I - S_h)\mathbf{E}$ for nonzero $\widehat{\beta}$.

The BIC score is defined to be

$$BIC(\boldsymbol{\lambda}) = \log\left\{\frac{RSS(\boldsymbol{\lambda})}{n}\right\} + e(\boldsymbol{\lambda})\frac{\log n}{n}$$

where $RSS(\boldsymbol{\lambda}) = \|(I - S_h)\mathbf{y} - (I - S_h)\mathbf{E}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|^2$ is the residual sum of squares with $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$, the penalized profile least squares estimate of $\boldsymbol{\beta}$ with tuning parameter $\boldsymbol{\lambda}$.

It is challenging to minimize $BIC(\boldsymbol{\lambda})$ over a d-dimensional space of $\boldsymbol{\lambda}$. Heuristically the magnitude of λ_j is proportional to the standard error of the profile least squares estimate of β_j. In our implementation, we set $\boldsymbol{\lambda} = \lambda\,\mathrm{se}(\widehat{\beta}_{UP})$, where $\mathrm{se}(\widehat{\beta}_{UP})$ is the standard error of the unpenalized profile least squared estimates and λ is a scalar variable. Thus, the original d-dimensional optimization becomes a 1-dimensional problem. We minimize $BIC(\lambda)$ over a grid of points evenly distributed

in the interval $[\frac{0.1}{\sqrt{n}}, \frac{2\sqrt{\log n}}{\sqrt{n}}]$, and set $\widehat{\lambda} = \text{argmin}_{\lambda} BIC(\lambda)$.

Bandwidth selection. Xiao *et. al.* (2003) pointed out it was challenging to select a bandwidth for their procedure, and the authors simply used the rule of thumb bandwidth, $h = 1.06\, S_X\, n^{-\frac{1}{5}}$, to prewhite the AR process, where S_X is the standard error of x_t. Note that $h = 1.06\, S_X\, n^{-\frac{1}{5}}$ is the rule of thumb bandwidth for kernel density estimate (Silverman, 1986), and we doubt that it may be a careless error. From our limited simulation experience, the bandwidth is not appropriate for the regression problem. Here we propose a bandwidth selector for the profile least squares estimate.

We use local linear regression to get the initial estimate $\widehat{m}_I(\cdot)$ with the plug-in bandwidth selector (Ruppert, Sheather and Wand, 1995), pretending the data are independent. Since model (3.2) is a partially linear model, we can use an existing bandwidth selector for partially linear models from the literature. Here we suggest using the proposal of Fan and Li (2004). Specifically, we calculate the difference-based estimate for $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}_{dbe}$. Plugging-in the difference-based estimate in (3.3), and further applying the plug-in bandwidth selector, we can select an appropriate bandwidth for the profile least squares procedures. The selected bandwidth is used for the SCAD procedure in (3.7).

3.2 Numerical comparison and application

In this section, we investigate the finite sample performance of the proposed procedures by Monte Carlo simulation, and compare the performance of proposed proce-

dures with existing ones by the mean squares error, defined by

$$\text{MSE}\{\widehat{m}(\cdot)\} = \frac{1}{n}\sum_{t=1}^{n}\{\widehat{m}(x_t) - m(x_t)\}^2$$

We summarize our simulation results in terms of relative MSE (RMSE), defined by the ratio of the MSE of an estimation procedure to the MSE of $\widehat{m}_I(\cdot)$, the estimate of $m(\cdot)$ pretending that the error ϵ_t are independent. We report the percentage of accuracy gain, defined by $(1 - RMSE) * 100\%$.

Example 1. In this example, a random sample of size n, either $n = 100$ or $n = 500$, is generated from

$$y_t = m(x_t) + \epsilon_t.$$

In this example, we consider two scenarios for $m(x)$. The first one is

$$m(x) = 4\cos(2\pi x),$$

and the second one is

$$m(x) = \exp(2x).$$

The mean function $m(x)$ is not monotone in the first scenario, while it is monotone in the second scenario. The error process ϵ_t is an AR process of order $d = 10$ or $d = 20$, i.e.,

$$\epsilon_t = \sum_{j=1}^{d}\beta_j\epsilon_{t-j} + \eta_t,$$

where $\eta_t \sim N(0, \sigma^2)$ with $\sigma = 0.5$ or 1. In our simulation we consider two situations: the first one is $\beta_1 = 0.5$, or 0.7, and all other β_j's equal 0, the second one is $\beta_1 = 0.5$, $\beta_2 = 0.4$ or $\beta_1 = 0.7$, $\beta_2 = 0.2$ and all others equal 0. In the first situation, the

error process indeed is an AR(1), while in the second situation, the error process is an AR(2). The number of replication is 500.

To understand how the sampling scheme of x_t affects the proposed procedure, we consider three sampling schemes in our simulation.

I. x_t is independent and identically distributed according to the uniform distribution over $[0, 1]$.

II. u_t is independent and identically distributed according to the standard normal distribution for $t = 1, 2, ...$ Let $x_t = \Phi\{(au_t + bu_{t-1})/\sqrt{a^2 + b^2}\}$ for $t = 2, 3, \cdots$, where $\Phi(u)$ is the cumulative distribution function of the standard normal distribution. Thus, x_t is 1-dependent process. In our simulation, we take $a = 0.9$ and $b = 0.1$.

III. x_t is a fixed design point evenly distributed over $[0, 1]$, i.e., $x_t = (t - 0.5)/n$.

For each sampling scheme, three methods—Xiao, Linton, Carroll and Mammen (2003) method (XLCM), profile least squares method (Profile) and penalized profile least square method with SCAD penalty function (SCAD)— are compared with regard to the efficiency improvement. In addition, the oracle procedure substituting the true autoregressive coefficient and order is listed as a benchmark.

For sample scheme I, the covariate x_t is independently and identically distributed, only the random error is correlated. Tables 3.1 and 3.2 summarize the simulation for sampling scheme I for $d = 10$ and 20, respectively. The overall pattern for $d = 10$ and $d = 20$ is the same, although the gain in term of RMSE with $d = 20$ is slightly more than that with $d = 10$. For both $d = 10$ and 20, the SCAD procedures performs better than the XLCM method and the profile least squares method, and its performance is very close to the oracle procedure. The performance of XLCM

Table 3.1. Simulation results for nonparametric models under sampling scheme I when $d = 10$. $(1 - RMSE) * 100\%$ under various conditions is reported for comparison.

(β_1, β_2)	σ	XLCM	Profile	SCAD	Oracle	XLCM	Profile	SCAD	Oracle
		$m(x) = 4\cos(2\pi x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	3.04	1.08	7.05	8.71	12.27	12.35	13.72	14.07
(0.7,0)	0.5	14.84	13.63	17.72	18.46	26.03	26.08	27.00	27.18
(0.5,0.4)	0.5	17.18	17.05	18.59	19.12	30.32	30.41	30.73	30.70
(0.7,0.2)	0.5	19.96	19.92	20.99	21.67	33.08	33.16	33.47	33.50
(0.5,0)	1	1.16	1.07	5.65	6.47	10.91	10.87	12.48	12.72
(0.7,0)	1	12.87	11.21	15.62	15.78	24.08	24.01	25.10	25.18
(0.5,0.4)	1	15.54	15.23	16.73	17.09	27.68	27.71	28.07	28.09
(0.7,0.2)	1	18.39	18.09	19.42	19.66	30.44	30.47	30.75	30.82
		$m(x) = \exp(2x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	3.79	4.70	5.00	4.21	4.39	3.57	6.53	5.91
(0.7,0)	0.5	12.26	11.29	13.45	14.80	16.00	15.79	16.40	16.47
(0.5,0.4)	0.5	13.48	12.31	13.79	14.33	22.50	22.38	22.70	22.75
(0.7,0.2)	0.5	16.07	14.99	16.06	16.63	25.00	24.87	25.18	25.23
(0.5,0)	1	5.53	3.85	6.90	6.71	6.40	6.02	8.13	8.18
(0.7,0)	1	14.24	13.27	14.39	15.25	15.66	15.69	16.75	16.64
(0.5,0.4)	1	13.60	12.36	14.06	14.51	22.20	22.09	22.42	22.47
(0.7,0.2)	1	16.41	15.29	16.56	16.59	24.64	24.52	24.78	24.90

method and the profile least squares procedure is very close to each other, and no one dominates the other one.

When the sample size is large, such as $n = 500$, the performance of the XLCM method, the profile least squares method and the SCAD procedure are very close to each other, although the SCAD procedure is slightly better than the other two. The gain for these three methods in terms of RMSE with a large sample is more than the one with the smaller sample size ($n = 100$). This is expected because with

Table 3.2. Simulation results for nonparametric models under sampling scheme I when $d = 20$. $(1 - RMSE) * 100\%$ under various conditions is reported for comparison.

(β_1, β_2)	σ	XLCM	Profile	SCAD	Oracle	XLCM	Profile	SCAD	Oracle
		$m(x) = 4\cos(2\pi x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	1.24	1.74	6.22	8.71	10.84	11.00	13.62	14.06
(0.7,0)	0.5	11.35	8.88	17.69	19.10	25.14	25.24	27.01	27.14
(0.5,0.4)	0.5	15.03	14.45	18.24	19.40	27.34	27.42	28.07	28.07
(0.7,0.2)	0.5	17.44	17.86	20.74	21.94	30.17	30.25	30.79	30.84
(0.5,0)	1	4.03	4.98	6.38	7.66	9.53	9.44	12.45	12.70
(0.7,0)	1	9.29	9.03	15.63	16.18	23.24	23.12	25.08	25.13
(0.5,0.4)	1	13.27	12.11	16.19	17.18	27.39	27.46	28.12	28.13
(0.7,0.2)	1	16.25	15.17	18.92	19.80	30.18	30.25	30.80	30.85
		$m(x) = \exp(2x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	0.19	0.98	2.87	3.12	5.05	4.44	5.81	6.10
(0.7,0)	0.5	4.38	4.60	7.02	8.88	15.42	14.48	16.54	16.45
(0.5,0.4)	0.5	11.12	8.83	13.26	14.34	21.94	21.73	22.32	22.37
(0.7,0.2)	0.5	12.28	14.39	15.67	17.09	24.48	24.25	24.78	24.86
(0.5,0)	1	5.06	3.83	6.25	6.71	5.72	4.75	7.76	7.97
(0.7,0)	1	3.54	4.20	7.32	7.90	15.31	14.24	16.52	16.42
(0.5,0.4)	1	11.00	8.53	13.33	14.29	22.05	21.85	22.45	22.50
(0.7,0.2)	1	14.11	13.86	15.67	16.88	24.51	24.27	24.79	24.89

the large sample size, all three methods can estimate $\boldsymbol{\beta}$ more accurately. Because of this, the decorrelation method works better.

Simulation results for sampling scheme II are summarized in Tables 3.3 and 3.4. Under this sampling scheme, the result for $d = 10$ is sometimes better than that for $d = 20$ in terms of RMSE. The overall pattern of Tables 3.3 and 3.4 is similar to that in Tables 3.1 and 3.2, although the sampling scheme II is different from the sampling scheme I in that the covariate x_t is dependent in the sample scheme II,

Table 3.3. Simulation results for nonparametric models under sampling scheme II when $d = 10$. $(1 - RMSE) * 100\%$ under various conditions is reported for comparison.

(β_1, β_2)	σ	XLCM	Profile	SCAD	Oracle	XLCM	Profile	SCAD	Oracle
		$m(x) = 4\cos(2\pi x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	4.82	3.06	8.20	10.30	12.94	13.09	13.85	14.06
(0.7,0)	0.5	16.33	14.92	18.42	20.28	25.95	26.05	26.58	26.71
(0.5,0.4)	0.5	19.39	19.47	20.93	21.46	30.43	30.54	30.65	30.71
(0.7,0.2)	0.5	20.15	20.22	21.24	22.04	33.12	33.25	33.35	33.41
(0.5,0)	1	2.21	3.17	5.57	7.72	11.86	11.90	12.81	12.86
(0.7,0)	1	17.32	16.33	19.93	20.45	24.01	24.02	24.67	24.65
(0.5,0.4)	1	18.10	17.81	18.72	19.44	27.66	27.70	27.86	27.88
(0.7,0.2)	1	23.29	24.23	25.38	25.41	30.29	30.35	30.44	30.52
		$m(x) = \exp(2x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	2.11	4.15	6.05	7.85	4.14	3.79	4.67	4.34
(0.7,0)	0.5	13.63	13.21	14.46	15.10	14.91	14.65	15.69	15.55
(0.5,0.4)	0.5	16.65	15.99	16.64	16.90	20.84	20.68	20.86	20.87
(0.7,0.2)	0.5	19.28	18.77	19.20	19.67	23.46	23.32	23.42	23.50
(0.5,0)	1	6.29	5.80	7.80	10.34	5.76	5.68	6.74	6.51
(0.7,0)	1	14.17	14.69	15.63	15.99	16.99	17.12	17.77	17.64
(0.5,0.4)	1	17.07	16.41	17.11	17.46	20.51	20.35	20.62	20.60
(0.7,0.2)	1	19.93	19.50	20.17	20.44	22.87	22.72	22.91	22.98

while it is independent in the sample scheme I. On the whole, the SCAD procedure performs best among the three methods in the comparison, and its performance is very close to the oracle procedure. The performances of Xiao *et. al.* method and the profile least squares procedure are similar, and no one dominates the other one.

In a summary, the performance of the proposed profile least squares procedure and the SCAD procedures seems not to rely on the sampling scheme of the covariate x_t.

Table 3.4. Simulation results for nonparametric models under sampling scheme II when $d = 20$. $(1 - RMSE) * 100\%$ under various conditions is reported for comparison.

(β_1, β_2)	σ	XLCM	Profile	SCAD	Oracle	XLCM	Profile	SCAD	Oracle
		$m(x) = 4\cos(2\pi x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	0.70	2.10	6.33	10.05	11.70	12.07	13.85	14.00
(0.7,0)	0.5	13.23	10.62	19.41	20.97	25.07	25.38	26.52	26.57
(0.5,0.4)	0.5	17.22	17.54	20.52	21.87	27.19	27.36	27.79	27.76
(0.7,0.2)	0.5	19.55	20.48	23.52	24.42	29.89	30.09	30.46	30.47
(0.5,0)	1	1.45	1.80	6.41	8.35	10.74	10.94	12.85	12.89
(0.7,0)	1	13.19	10.44	19.45	20.06	23.38	23.56	24.81	24.76
(0.5,0.4)	1	16.53	16.54	19.51	20.41	27.26	27.43	27.87	27.83
(0.7,0.2)	1	19.11	19.35	22.27	22.93	29.97	30.17	30.54	30.55
		$m(x) = \exp(2x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	0.23	0.96	1.17	3.82	1.62	0.95	2.54	2.59
(0.7,0)	0.5	9.30	8.83	12.66	13.69	13.37	12.75	13.83	13.74
(0.5,0.4)	0.5	14.74	13.65	15.91	16.57	20.16	19.96	20.54	20.42
(0.7,0.2)	0.5	17.48	16.62	18.93	19.48	22.60	22.41	22.91	22.88
(0.5,0)	1	3.39	3.38	7.23	8.19	1.27	2.40	5.97	5.97
(0.7,0)	1	9.72	7.03	14.44	14.74	12.54	11.76	13.13	13.03
(0.5,0.4)	1	15.31	14.24	16.73	17.23	20.21	20.02	20.60	20.48
(0.7,0.2)	1	17.79	16.88	19.49	19.89	22.70	22.53	22.99	23.03

Although the sampling scheme III does not satisfy the regularity conditions and is not the focus of this paper, we include this scheme to demonstrate that the proposed method may work well for this sampling scheme.

For the sampling scheme III, $\{x_t\}$ is a fixed design. As demonstrated in Altman (1990) and Hart (1991) for kernel regression estimator, the ordinary bandwidth selector will tend to undersmooth the true regression function when the error is positively correlated. From our simulation experience, the local linear regression es-

Table 3.5. Simulation results for nonparametric models under sampling scheme III when $d = 10$. $(1 - RMSE) * 100\%$ under various conditions is reported for comparison.

(β_1, β_2)	σ	XLCM	Profile	SCAD	Oracle	XLCM	Profile	SCAD	Oracle
					$m(x) = 4\cos(2\pi x)$				
			$n = 100$				$n = 500$		
(0.5,0)	0.5	9.32	9.74	18.58	23.12	16.87	18.56	27.06	27.56
(0.7,0)	0.5	21.36	21.47	29.87	32.48	37.68	39.90	43.27	43.66
(0.5,0.4)	0.5	18.80	17.51	19.62	20.67	39.56	43.33	44.95	45.03
(0.7,0.2)	0.5	21.03	23.33	24.72	26.09	42.17	45.80	47.33	47.33
(0.5,0)	1	8.64	9.24	22.00	27.44	20.63	22.74	28.96	29.53
(0.7,0)	1	21.30	24.60	32.78	36.28	40.03	42.54	46.96	47.74
(0.5,0.4)	1	21.32	24.68	25.29	28.71	39.80	43.72	45.29	45.59
(0.7,0.2)	1	21.55	25.14	27.08	30.24	42.39	36.15	47.64	47.78
					$m(x) = \exp(2x)$				
			$n = 100$				$n = 500$		
(0.5,0)	1	8.10	11.40	23.54	30.50	24.75	27.09	36.27	37.19
(0.7,0)	1	21.16	25.27	33.64	37.63	40.97	43.53	48.58	49.38
(0.5,0.4)	1	21.74	25.58	28.08	32.05	39.80	43.74	45.52	45.66
(0.7,0.2)	1	21.71	25.77	28.67	32.11	42.41	46.18	47.83	47.85
(0.5,0)	1	7.63	11.18	24.39	31.24	25.50	27.85	37.29	38.51
(0.7,0)	1	21.09	25.26	33.83	37.84	41.03	43.60	48.57	49.50
(0.5,0.4)	1	21.95	25.98	27.51	32.59	39.83	43.79	45.40	45.71
(0.7,0.2)	1	21.66	25.80	28.41	32.28	42.40	46.16	47.68	47.84

timator also suffers from this difficulty: the plug-in bandwidth proposed by Ruppert, Sheather and Wand (1995) always picks a small bandwidth which undersmooths the fitting. Many authors have proposed various adjustment methods to bandwidth selection to overcome this problem in the fixed design. For example, Altman (1990) suggested revising CV and GCV criteria by incorporating the estimate of the covariance structure. Hart (1991) proposed a risk estimation procedure. For simplicity, a ratio based on the pilot study is multiplied by the plugging-in bandwidth to adjust

Table 3.6. Simulation results for nonparametric models under sampling scheme III when $d = 20$. $(1 - RMSE) * 100\%$ under various conditions is reported for comparison.

(β_1, β_2)	σ	XLCM	Profile	SCAD	Oracle	XLCM	Profile	SCAD	Oracle
		$m(x) = 4\cos(2\pi x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	8.55	9.26	9.98	22.40	14.80	17.44	22.01	22.13
(0.7,0)	0.5	22.23	12.36	25.90	32.25	36.07	39.24	43.11	43.56
(0.5,0.4)	0.5	18.92	14.06	19.61	20.73	38.78	43.27	44.93	45.04
(0.7,0.2)	0.5	21.62	19.53	24.69	26.13	41.41	45.84	47.32	47.35
(0.5,0)	1	11.98	12.18	19.63	27.62	15.40	19.18	27.92	28.69
(0.7,0)	1	23.75	22.30	30.95	35.91	37.31	41.14	46.50	47.43
(0.5,0.4)	1	21.86	21.41	25.30	28.64	38.94	43.77	45.39	45.67
(0.7,0.2)	1	22.51	20.72	27.79	30.47	41.56	45.64	47.09	47.51
		$m(x) = \exp(2x)$							
		$n = 100$				$n = 500$			
(0.5,0)	0.5	13.59	10.84	22.80	31.60	16.81	21.09	33.93	35.17
(0.7,0)	0.5	24.11	18.28	31.17	37.78	37.81	41.74	47.79	48.81
(0.5,0.4)	0.5	21.93	21.32	27.18	32.04	38.85	43.73	45.51	45.71
(0.7,0.2)	0.5	22.07	20.98	28.39	32.21	41.47	46.11	47.77	47.85
(0.5,0)	1	13.02	10.31	25.29	32.52	16.94	21.33	34.45	38.94
(0.7,0)	1	23.90	18.61	33.02	38.10	37.85	41.84	47.82	48.94
(0.5,0.4)	1	21.67	21.47	27.31	32.43	38.87	43.76	45.43	45.73
(0.7,0.2)	1	21.84	21.18	28.77	32.35	41.54	46.20	47.71	47.92

the undersmoothness in our simulation study,

Because $\{x_t\}$ is the fixed design in scheme III, the results of Table 3.5 and 3.6 are quite different from Table 3.1 — 3.4 , although all XLCM, Profile, SCAD methods improve the estimation efficiency as expected. The magnitude of the gain at the same correlation level is much more significant than that in sampling schemes I and II, especially when $\beta_1 = 0.7$ in the $AR(1)$ model. The profile least squares procedure is similar to XLCM method. More interestingly, all three methods have more gain for

the monotone regression function $m(x) = \exp(2x)$ than the non-monotone function $m(x) = 4\cos(2\pi x)$ only in this scheme.

Example 2. In this example, we illustrate the proposed methodology by an analysis of a data set about U.S. macroeconomics, collected from January 1980 to December 2006 on a monthly basis. Our interest here is to investigate the relationship between the unemployment rate and house price index change.

In the past few years, housing prices in U.S. have shown a strong upward trend, although the bubble warning always exists. Many home buyers who do not have sound credit history nor sufficient financial capability have become home owners with the help of sub-prime mortgages. They have a high level of interest payments but believe the property will keep appreciating. In the meantime, the mortgage agent packages the debt and sells it to other institutional investors. This long chain prospers and works well when the housing market is booming. However, when the house prices began to plummet in spring 2007, borrowers had to default and many houses went to foreclosure. Consequently, a number of big financial institutions that have heavy investment in sub-prime mortgage market claimed billions of dollars in write-offs due to the crisis.

In this example, we are interested in the effect of the unemployment rate on the housing price. By classical economic theory, the unemployment rate is an important indicator of the overall economy. If many people claim unemployment, purchasing power is definitely hurt. However, to our best knowledge, there is not much literatures that studies the relationship between the unemployment rate and the housing market in a quantitative manner. Motivated by the sub-prime mortgage turmoil and the recent suspicion of recession, it is believed that the historical data might shed some interesting insights on how these two indices are related. Thus, we take

the unemployment rate as the covariate x and the House Price Index Change as the response variable y, and consider the following model

$$y_t = m(x_t) + \epsilon_t. \tag{3.9}$$

Initial estimate and residual analysis. We ignore the correlation of the random errors temporarily and estimate (3.9) by the conventional local linear model as Fan and Gijbels (1996). The Ruppert, Sheather and Wand (1995) direct plug-in bandwidth is 0.2969.

When the initial estimate $\tilde{m}(x_t)$ is obtained, we can estimate the residual ϵ_t as $y_t - \tilde{m}(x_t)$. There is an obvious correlation pattern present in the autocorrelation plot of $\widehat{\epsilon}_t$ (Figure 3.2 (a)). The partial-autocorrelation plot (Figure 3.2 (b)) indicates an autoregressive model and the first lag effect is most outstanding. Furthermore, the Ljung-Box-Pierce test used to check the autocorrelation pattern for white noise has the P-value less than 0.0001. It also verifies that autocorrelation exists in $\widehat{\epsilon}_t$.

From a conservative point of view, we suspect that the house price might have a year lag. So we assume AR(12) model on errors and employ the penalized profile least square method to select the AR order and estimate $m(\cdot)$ simultaneously. The plug-in bandwidth in the profile least squares estimation is 0.2140. By the BIC criterion, the optimal tuning parameter used in the order selection procedure is 0.000019.

As an result, the AR(1) model with a strong autocorrelation coefficient 0.9438 is most appropriate. It means that the house price has only one month lag, which agrees with the partial-autocorrelation plot in Figure 3.2 (b). After accounting for the autocorrelation, the correlogram of $\widehat{\eta}_t$ does not have any significant pattern. (See Figure 3.2 (c) and (d)) In addition, the P-value in the Ljung-Box-Pierce test at

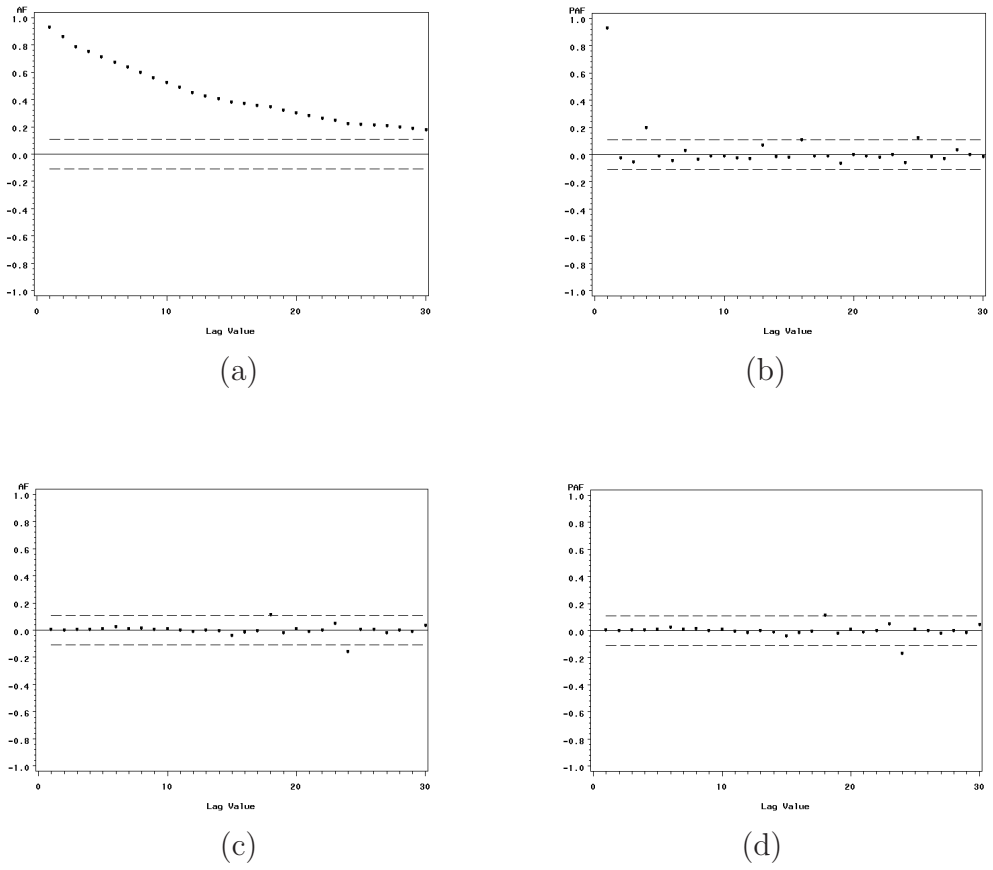

Figure 3.2. Correlogram of residual $\widehat{\epsilon}_t$ and $\widehat{\eta}_t$ for macroeconomic data. Plot (a) and (b) are the autocorrelation and partial autocorrelation for $\widehat{\epsilon}_t$. Plot (c) and (d) are the autocorrelation and partial autocorrelation for $\widehat{\eta}_t$. In each plot, the upper and lower dashed lines represent 95% confidence interval.

the first 24 lags, 0.9134, also shows that the autocorrelation has been successfully removed.

Final model. By applying the penalized profile least squares estimation method, the relationship between the House Price Index Change and the unemployment rate turns out to be

$$\widehat{y}_t = \widehat{m}(x_t) + 0.9438\widehat{\epsilon}_{t-1} \tag{3.10}$$

where $\widehat{m}(\cdot)$ is displayed in Figure 3.3. The penalized profile least square approach

yields a smoother estimate than the conventional local linear regression because it takes the correlation into account. As expected, the unemployment rate has a negative correlation with house price index change. But this effect is most significant when the unemployment varies between 4% and 5% or between 8% and 10%.

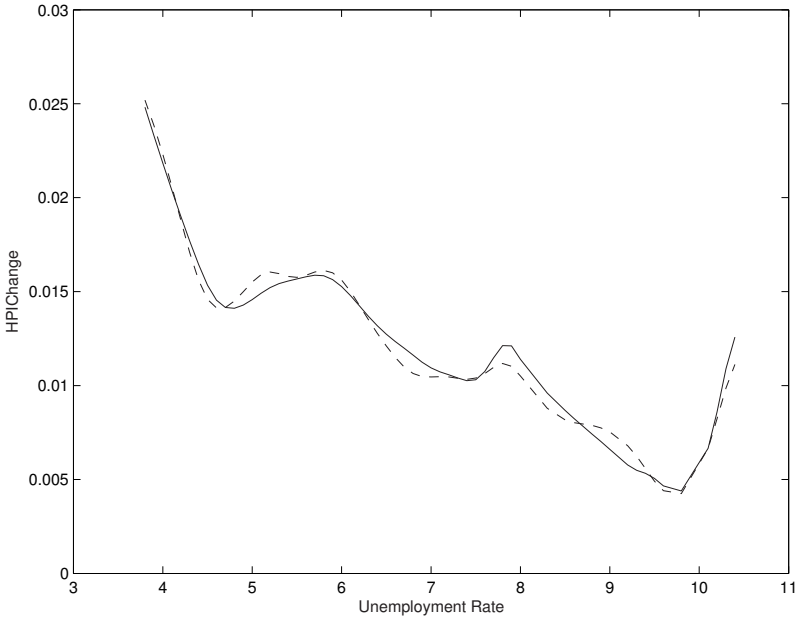

Figure 3.3. Estimation of $m(\cdot)$ for macroeconomic data. Dashed curves are the initial estimates; Solid curves are the penalized profile least squares estimate.

3.3 Proofs

3.3.1 Preliminaries

To present the regularity conditions, we need the following definitions for a sequence of random vectors $\{\mathbf{z}_t, t = 0, \pm 1, \pm 2, \cdots\}$. The following notation and definitions are adopted from Chapter 2 of Fan and Yao (2003).

Definition 1. *A sequence of random vectors* $\{\mathbf{z}_t, t = 0, \pm 1, \pm 2, \cdots\}$ *is said to be*

strictly stationary if $\{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$ *and* $\{\mathbf{z}_{1+k}, \cdots, \mathbf{z}_{1+n}\}$ *have the same joint distributions for any integer* $n \geq 1$ *and any integer* k.

Denote $\mathcal{F}_i^j$ to be the σ-algebra generated by events $\{\mathbf{z}_t, i \leq t \leq j\}$, and $\mathcal{L}^2(\mathcal{F}_i^j)$ consists of $\mathcal{F}_i^j$-measurable random variables with finite second moment. Intuitively, $\mathcal{F}_i^j$ assembles all information on the sequence collected between time i and j. Define

$$\alpha(n) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_n^\infty} |P(A)P(B) - P(AB)| \tag{3.11}$$

Definition 2. *A sequence of random vectors* $\{\mathbf{z}_t, t = 0, \pm 1, \pm 2, \cdots\}$ *is said to be* α*-mixing if it is strictly stationary and* $\alpha(n) \to 0$ *as* $n \to \infty$.

3.3.2 Regularity conditions and proofs

To make the argument concise, denote $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)^T$ with $\mathbf{f}_t = (\epsilon_{t-1}, \ldots, \epsilon_{t-d})^T$, and $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_n)^T$ with $\mathbf{e}_t = (\widehat{\epsilon}_{t-1}, \ldots, \widehat{\epsilon}_{t-d})^T$. Define $\mathbf{\Delta} = \mathbf{E} - \mathbf{F}$. Our proof follows the same strategy as that in Fan and Huang (2005). The following conditions are imposed to facilitate the proof and are adopted from Fan and Huang (2005). They are not the weakest possible conditions.

A. The random variable x_t has a bounded support Ω. Its density function $f(.)$ is Lipschitz continuous and bounded away from 0 on its support.

B. There is an $s > 2$ such that $E||\mathbf{f}_t||^s < \infty$ and for some $\xi > 0$ such that $n^{1-2s^{-1}-2\xi} h \to \infty$.

C. $m(\cdot)$ has a continuous second derivative in $x \in \Omega$.

D. The function $K(\cdot)$ is a bounded symmetric density function with bounded support $[-M, M]$, satisfying the Lipschitz condition.

E. $nh^8 \to 0$ and $nh^2/(\log n)^2 \to \infty$.

F. $\sup_{x\in\Omega} |\widehat{m}_I(x) - m(x)| = o_p(n^{-\frac{1}{4}})$ where $\widehat{m}_I(x_t)$ is obtained by local linear regression pretending that data are i.i.d.

G. The sequence of random vectors (x_t, ϵ_t), $t = 1, 2, \cdots$, is a strictly stationary and satisfies the following condition for α-mixing processes: assume that for some $\delta > 2$ and $a > 1 - 2/\delta$,

$$\sum_l l^a [\alpha(l)]^{1-2/\delta} < \infty, \quad E|\epsilon_1|^\delta < \infty, \quad g_{x_1|\epsilon_1}(x|\epsilon) \le A_1 < \infty$$

Lemma 3.3.1 is taken from Lemma 6.1 of Fan and Yao (2003) and will be used in our proof repeatedly.

Lemma 3.3.1. *Let $(x_1, \epsilon_1), \ldots, (x_n, \epsilon_n)$ be a strictly stationary sequence satisfying the mixing condition $\alpha(l) \le cl^{-\tau}$ for some $c > 0$ and $\tau > 5/2$. Assume further that for some $s > 2$ and interval $[a, b]$,*

$$E|\epsilon_t|^s < \infty \quad \textit{and} \quad \sup_{\forall x\in[a,b]} \int |\epsilon_t|^s g(x, \epsilon) \mathrm{d}\epsilon < \infty,$$

where g denote the joint density of (x_t, ϵ_t). In addition, Condition G holds, and the conditional density $g_{x_1,x_l|\epsilon_1,\epsilon_l}(x_1, x_l|\epsilon_1, \epsilon_l) \le A_2 < \infty, \forall l \ge 1$. Let K satisfy Condition D. Then

$$\sup_{x\in[a,b]} |\frac{1}{n}\sum_{i=1}^{n} \{K_h(x_i - x)\epsilon_i - E[K_h(x_i - x)\epsilon_i]\}| = O_p(\left\{\frac{\log n}{nh}\right\}^{1/2})$$

provided that $h \to 0$, for some $\xi > 0$, $n^{1-2s^{-1}-2\xi}h \to \infty$ and $n^{(\tau+\frac{3}{2})(s^{-1}+\xi)-\frac{\tau}{2}+\frac{5}{4}}h^{-\frac{\tau}{2}-\frac{5}{4}} \to 0$.

Lemma 3.3.2. *Under Conditions A—G, it follows that*

$$\frac{1}{n}\mathbf{F}^T(I-S)^T(I-S)\mathbf{F} \xrightarrow{P} E(\mathbf{f}\mathbf{f}^T).$$

Proof. Denote W_x be a $n \times n$ diagonal matrix with j-th diagonal element $K_h(x_j - x)$ and

$$D_x = \begin{pmatrix} 1 & \frac{x_1-x}{h} \\ \vdots & \vdots \\ 1 & \frac{x_n-x}{h} \end{pmatrix}$$

Then the smoothing matrix $\mathbf{S}$ for the local linear regression can be expressed as

$$\mathbf{S} = \begin{pmatrix} [1,0]\{D_{x_1}^T W_{x_1} D_{x_1}\}^{-1} D_{x_1}^T W_{x_1} \\ \vdots \\ [1,0]\{D_{x_n}^T W_{x_n} D_{x_n}\}^{-1} D_{x_n}^T W_{x_n} \end{pmatrix}$$

where

$$D_x^T W_x D_x = \begin{pmatrix} \sum_{i=1}^n K_h(x_i - x) & \sum_{i=1}^n (x_i - x) K_h(x_i - x)/h \\ \sum_{i=1}^n (x_i - x) K_h(x_i - x)/h & \sum_{i=1}^n (x_i - x)^2 K_h(x_i - x)/h^2 \end{pmatrix}$$

The generic element of matrix $D_x^T W_x D_x$ is in the form of $\sum_{i=1}^n (\frac{x_i - x}{h})^j K_h(x_i - x)$, $j = 0, 1, 2$. Denote $S_{n,j} = \sum_{i=1}^n (\frac{x_i - x}{h})^j K_h(x_i - x)$. By using the formula $S_{n,j} = E(S_{n,j}) + O_p(\sqrt{\mathrm{Var}(S_{n.j})})$, it is easy to show that if j is even,

$$\begin{aligned} S_{n,j} &= n \int v^j K(v) f(x + hv) \mathrm{d}v + O_p(\sqrt{nE\{(x_1 - x)^{2j} K_h^2(x_1 - x)\}}) \\ &= nf(x)\mu_j + O_p(h^2 + 1/\sqrt{nh}) \end{aligned}$$

Because of the symmetry of the kernel function, for any odd numbered j, $\mu_j = 0$

and then $S_{n,j} = O_p(h+1/\sqrt{nh})$. Indeed, with Lemma 3.3.1, it can be further shown that for even j,

$$S_{n,j} = nf(x)\mu_j + O_p(h^2 + \sqrt{\log n/nh}),$$

and for odd j,

$$S_{n,j} = O_p(h + \sqrt{\log n/nh})$$

holds uniformly in x.

Therefore,

$$\frac{1}{n}D_x^T W_x D_x = \begin{pmatrix} f(x)(1+O_p(h^2+\sqrt{\log n/nh})) & O_p(h+\sqrt{\log n/nh}) \\ O_p(h+\sqrt{\log n/nh}) & f(x)\mu_2(1+O_p(h^2+\sqrt{\log n/nh})) \end{pmatrix}$$

holds uniformly in x.

Since $h + \sqrt{\log n/nh} = o_p(1)$, we can regard the above matrix as being approximately diagonal. Then its inverse is

$$\{\frac{1}{n}D_x^T W_x D_x\}^{-1} = \begin{pmatrix} \{f(x)\}^{-1}(1+O_p(h^2+\sqrt{\frac{\log n}{nh}})) & O_p(h+\sqrt{\frac{\log n}{nh}}) \\ O_p(h+\sqrt{\frac{\log n}{nh}}) & \{f(x)\mu_2\}^{-1}(1+O_p(h^2+\sqrt{\frac{\log n}{nh}})) \end{pmatrix}$$

holds uniformly in x.

Similarly, by Lemma 3.3.1 and the assumption of independence between the process ϵ_t and the process x_t, it follows that

$$\frac{1}{n}D_x^T W_x \mathbf{F} = \begin{pmatrix} O_p(h^2+\sqrt{\frac{\log n}{nh}}) \\ O_p(h+\sqrt{\frac{\log n}{nh}}) \end{pmatrix}$$

holds uniformly in x.

Consequently,

$$[1,0]\{\frac{1}{n}D_x^T W_x D_x\}^{-1}\{\frac{1}{n}D_x^T W_x \mathbf{F}\}$$

$$= [1,0] \begin{pmatrix} \{f(x)\}^{-1}(1+O_p(h^2+\sqrt{\frac{\log n}{nh}})) & O_p(h+\sqrt{\frac{\log n}{nh}}) \\ O_p(h+\sqrt{\frac{\log n}{nh}}) & \{f(x)\mu_2\}^{-1}(1+O_p(h^2+\sqrt{\frac{\log n}{nh}})) \end{pmatrix} \begin{pmatrix} O_p(h^2+\sqrt{\frac{\log n}{nh}}) \\ O_p(h+\sqrt{\frac{\log n}{nh}}) \end{pmatrix}$$

$$= \{f(x)\}^{-1} O_p(h^2+\sqrt{\frac{\log n}{nh}})(1+o_p(1)) = o_p(1)$$

Substituting this result into the smoothing matrix S, we have

$$S\mathbf{F} = \begin{pmatrix} [1,0]\{D_{x_1}^T W_{x_1} D_{x_1}\}^{-1} D_{x_1}^T W_{x_1}\mathbf{F} \\ \vdots \\ [1,0]\{D_{x_n}^T W_{x_n} D_{x_n}\}^{-1} D_{x_n}^T W_{x_n}\mathbf{F} \end{pmatrix} = \begin{pmatrix} o_p(1) \\ \vdots \\ o_p(1) \end{pmatrix}.$$

Thus,

$$\mathbf{F} - S\mathbf{F} = \mathbf{F}\{1+o_P(1)\}.$$

Finally, by the WLLN,

$$\frac{1}{n}\mathbf{F}^T(I-S)^T(I-S)\mathbf{F} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{f}_i\mathbf{f}_i^T\right)\{1+o_p(1)\}^2 \xrightarrow{P} E(\mathbf{f}\mathbf{f}^T)$$

□

Lemma 3.3.3. *Under Conditions A—G, we have*

$$\frac{1}{n}\mathbf{E}^T(I-S)^T(I-S)\mathbf{E} \xrightarrow{P} E(\mathbf{f}\mathbf{f}^T).$$

Proof. Since $\boldsymbol{\Delta} = \mathbf{E} - \mathbf{F}$, the generic element of $\boldsymbol{\Delta}$ is of the form $m(x_t) - \widehat{m}(x_t)$, which is of order $o_P(n^{-1/4})$ uniformly in x by Condition F. Thus, $\boldsymbol{\Delta} = o_P(n^{-1/4})$. Therefore

$$\frac{1}{n}\mathbf{E}^T(I-S)^T(I-S)\mathbf{E} = \frac{1}{n}(\mathbf{F}+\boldsymbol{\Delta})^T(I-S)^T(I-S)(\mathbf{F}+\boldsymbol{\Delta})$$

By using a similar argument to the proof of Lemma 3.3.2, it can be shown that

$$\frac{1}{n}\mathbf{E}^T(I-S)^T(I-S)\mathbf{E} = \frac{1}{n}\mathbf{F}^T(I-S)^T(I-S)\mathbf{F} + o_P(1)$$

Thus, Lemma 3.3.3 follows by Lemma 3.3.2.

□

Lemma 3.3.4. *Suppose Conditions A—G hold. It follows*

$$\frac{1}{\sqrt{n}}\mathbf{F}^T(I-\mathbf{S})^T(I-\mathbf{S})\mathbf{M} = o_p(1)$$

Proof. It is noted that

$$\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\mathbf{M} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathbf{f}_i-(S\mathbf{f})_i][m(x_i)-[1,0]\{D_{x_i}^TW_{x_i}D_{x_i}\}^{-1}D_{x_i}^TW_{x_i}\mathbf{M}] \tag{3.12}$$

Similar to the argument in the proof of Lemma 3.3.2, we can show that

$$[1,0]\{\frac{1}{n}D_x^TW_xD_x\}^{-1}\{\frac{1}{n}D_x^TW_x\mathbf{M}\} = m(x)(1+O_p(h^2+\sqrt{\log n/nh}))$$

holds uniformly in $x \in \Omega$. Plugging this into (3.12), it follows that

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\mathbf{M}\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathbf{f}_i-(S\mathbf{f})_i][m(x_i)-m(x_i)(1+O_p(h^2+\sqrt{\log n/nh}))]\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{f}_im(x_i)[1+o_p(1)]O_p(h^2+\sqrt{\log n/nh})
\end{aligned}$$

Note that $E\{\mathbf{f}_im(x_i)\} = 0$, and the covariance matrix for $\{\mathbf{f}_im(x_i)\}$ is finite. Thus, using $R = E(R) + O_p(\sqrt{\mathrm{Var}(R)})$, it follows that $\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\mathbf{M} =$

$o_p(1)$. □

Lemma 3.3.5. *Under Conditions A—G, we have*

$$\frac{1}{\sqrt{n}}\mathbf{E}^T(I-S)^T(I-S)\mathbf{M} = o_p(1)$$

Proof. Since $\mathbf{E} = \mathbf{F} + \mathbf{\Delta}$, we can break $\frac{1}{\sqrt{n}}\mathbf{E}^T(I-S)^T(I-S)\mathbf{M}$ into two terms: $\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\mathbf{M}$, which is $o_P(1)$ by Lemma 3.3.4, and $\frac{1}{\sqrt{n}}\mathbf{\Delta}^T(I-S)^T(I-S)\mathbf{M}$, which is also $o_P(1)$ as $\mathbf{\Delta} = o_P(n^{-1/4})$. □

Lemma 3.3.6. *Suppose that Conditions A—G hold. We have*

$$\frac{1}{\sqrt{n}}\mathbf{E}^T(I-S)^T(I-S)\mathbf{\Delta}\boldsymbol{\beta} = o_p(1)$$

Proof. This is a direct result from the proof of Lemma 3.3.3. □

Lemma 3.3.7. *Under Conditions A—G, let $\eta = (\eta_1, \cdots, \eta_n)^T$. Then*

$$\sqrt{n}[\mathbf{F}^T(I-S)^T(I-S)\mathbf{F}]^{-1}\mathbf{F}^T(I-S)^T(I-S)\eta \to N(0, \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}^{-1})$$

Proof. We observe that

$$\mathbf{F}^T(I-S)^T(I-S)\eta = \sum_{i=1}^{n}\mathbf{f}_i[\eta_i - [1,0]\{D_{x_i}^T W_{x_i} D_{x_i}\}^{-1}D_{x_i}^T W_{x_i}\eta][1+o_p(1)] \quad (3.13)$$

By using Lemma 3.3.1 on $\{x_i, \eta_i\}$, we can show that

$$\begin{aligned}
&[1,0]\{\frac{1}{n}D_x^T W_x D_x\}^{-1}\{\frac{1}{n}D_x^T W_x\eta\} \\
&= [1,0]\begin{pmatrix} \{f(x)\}^{-1}(1+O_p(h^2+\sqrt{\frac{\log n}{nh}}) & O_p(h+\sqrt{\frac{\log n}{nh}})) \\ O_p(h+\sqrt{\frac{\log n}{nh}}) & \{f(x)\mu_2\}^{-1}(1+O_p(h^2+\sqrt{\frac{\log n}{nh}})) \end{pmatrix}\begin{pmatrix} O_p(h^2+\sqrt{\frac{\log n}{nh}}) \\ O_p(h+\sqrt{\frac{\log n}{nh}}) \end{pmatrix} \\
&= o_p(1)
\end{aligned}$$

Then $\eta_i - [1,0]\{D_{x_i}^T W_{x_i} D_{x_i}\}^{-1} D_{x_i}^T W_{x_i}\eta = \eta_i\{1+o_p(1)\}$. Plugging this in (3.13), we obtain that

$$\mathbf{F}^T(I-S)^T(I-S)\eta = \sum_{i=1}^{n} \mathbf{f}_i\eta_i\{1+o_p(1)\}$$

Since $E(\mathbf{f}_i\eta_i) = 0$, $\mathrm{Var}(\mathbf{f}_i\eta_i) = \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\} < \infty$, and $E(\mathbf{f}_i\eta_i\mathbf{f}_j\eta_j) = 0$ for $i \neq j$ since η_i is independent of $\mathbf{f}_i$. By Central Limit Theorem for strictly stationary sequence (see Theorem 2.21 of Fan and Yao, 2003),

$$\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\eta \xrightarrow{L} N(0, \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}).$$

By Lemma 3.3.2, $\frac{1}{n}\mathbf{F}^T(I-S)^T(I-S)\mathbf{F} \xrightarrow{P} E(\mathbf{f}\mathbf{f}^T)$. Applying the Slutsky theorem, it follows that

$$\sqrt{n}[\mathbf{F}^T(I-S)^T(I-S)\mathbf{F}])^{-1}\mathbf{F}^T(I-S)^T(I-S)\eta \xrightarrow{L} N(0, \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}^{-1}).$$

□

Lemma 3.3.8. *Under Conditions A—G, we have*

$$\sqrt{n}[\mathbf{E}^T(I-S)^T(I-S)\mathbf{E}])^{-1}\mathbf{E}^T(I-S)^T(I-S)\eta \xrightarrow{L} N(0, \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}^{-1})$$

Proof. Since $\mathbf{E} = \mathbf{F} + \mathbf{\Delta}$, we may write $\mathbf{E}^T(I-S)^T(I-S)\eta = \mathbf{F}^T(I-S)^T(I-S)\eta + \mathbf{\Delta}^T(I-S)^T(I-S)\eta$. Note that $\mathbf{\Delta} = o_P(n^{-1/4})$ by Condition F, it can be shown that

$$\frac{1}{\sqrt{n}}\Delta^T(I-S)^T(I-S)\eta = o_p(1).$$

Furthermore, we have shown in the last lemma that $\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\eta \to N(0, \sigma^2 E(\mathbf{f}\mathbf{f}^T))$. So $\frac{1}{\sqrt{n}}\mathbf{E}^T(I-S)^T(I-S)\eta \to N(0, \sigma^2 E(\mathbf{f}\mathbf{f}^T))$ as well. The proof is

completed by the Slutsky theorem and Lemma 3.3.3. □

Proof of Theorem 1

Let us first show the asymptotic normality of $\widehat{\boldsymbol{\beta}}$. According to the expression of $\widehat{\boldsymbol{\beta}}$ in (3.5), we can break $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ into the sum of the following three terms (a), (b) and (c)

$$(a)\widehat{=}\sqrt{n}[\{\mathbf{E}^T(I-S)^T(I-S)\mathbf{E}\}^{-1}\mathbf{E}^T(I-S)^T(I-S)\mathbf{M}]$$
$$(b)\widehat{=}\sqrt{n}[\{\mathbf{E}^T(I-S)^T(I-S)\mathbf{E}\}^{-1}\mathbf{E}^T(I-S)^T(I-S)\boldsymbol{\Delta}\beta]$$
$$(c)\widehat{=}\sqrt{n}[\{\mathbf{E}^T(I-S)^T(I-S)\mathbf{E}\}^{-1}\mathbf{E}^T(I-S)^T(I-S)\eta]$$

Term (a) is a product of $[\frac{\{\mathbf{E}^T(I-S)^T(I-S)\mathbf{E}\}}{n}]^{-1}$ and $[\frac{\mathbf{E}^T(I-S)^T(I-S)\mathbf{M}}{\sqrt{n}}]$. From Lemmas 3.3.3 and 3.3.5, the asymptotic properties of these two terms lead to the conclusion that $(a) = o_p(1)$. Similarly, applying Lemmas 3.3.3 and 3.3.6 on two product components of term (b) results in $(b) = o_p(1)$ as well. In addition, Lemma 3.3.8 states that term (c) converges to $N(0, \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}^{-1})$. Put three terms together and we get the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$.

Next we derive the asymptotic bias and variance of $\widehat{m}(\cdot)$. By Lemmas 3.3.1— 3.3.8, we have

$$\widehat{m}(x_0, \widehat{\boldsymbol{\beta}}) = [1, 0]\{D_{x_0}^T W_{x_0} D_{x_0}\}^{-1} D_{x_0}^T W_{x_0}(\mathbf{M} + \eta)\{1 + o_P(1)\}$$

Note that $E(\eta|\mathcal{X}) = 0$, where $\mathcal{X} = (x_1, \cdots, x_n)$. Thus, So

$$E\{\widehat{m}(x_0, \widehat{\boldsymbol{\beta}})|\mathcal{X}\} = [1, 0]\{D_{x_0}^T W_{x_0} D_{x_0}\}^{-1} D_{x_0}^T W_{x_0}\mathbf{M}\{1 + o_p(1)\}$$

which is same as the conditional expected mean for the local linear regression derived in Fan and Gijbels(1992). We quoted their theorem in Section 2.1.1. Hence the asymptotic bias is $\frac{1}{2}m''(x_0)h^2 \int x^2 K(x)$.

Regarding to the asymptotic variance of $\widehat{m}(\cdot)$, conditioning on $x_1, \cdots, x_n$,

$$\mathrm{Var}[\widehat{m}(x_0, \widehat{\boldsymbol{\beta}})|\mathcal{X}] = [1,0]\{D_{x_0}^T W_{x_0} D_{x_0}\}^{-1} D_{x_0}^T W_{x_0} \mathrm{Var}\{\eta\} W_{x_0} D_{x_0} \{D_{x_0}^T W_{x_0} D_{x_0}\}^{-1} [1,0]^T$$

Using the same argument as that in the proof of Lemma 3.3.2, we have

$$\mathrm{Var}[\widehat{m}(x_0, \widehat{\boldsymbol{\beta}})|\mathcal{X}] = \frac{\sigma^2}{nhf(x_0)} \int K^2(x)\, \mathrm{d}x$$

As to the asymptotic normality,

$$\widehat{m}(x_0, \widehat{\boldsymbol{\beta}}) - E\{\widehat{m}(x_0, \widehat{\boldsymbol{\beta}})|\mathcal{X}\} = [1,0]\{D_{x_0}^T W_{x_0} D_{x_0}\}^{-1} D_{x_0}^T W_{x_0} \eta\{1 + o_P(1)\}$$

Thus, conditioning on $\mathcal{X}$, the asymptotic normality can be established using the Center Limit Theorem since η_i are independent and identically distributed with mean zero and variance σ^2.

Chapter 4

Varying-coefficient models with AR errors

Suppose that a response variable y_t along with its covariates $\{u_t, x_{t1}, \cdots, x_{tp}\}$ is collected from the varying coefficient model

$$y_t = \alpha_0(u_t) + \alpha_1(u_t)x_{t1} + \cdots + \alpha_p(u_t)x_{tp} + \epsilon_t, \tag{4.1}$$

where $\{\alpha_0(\cdot), \alpha_1(\cdot), \ldots, \alpha_p(\cdot)\}$ are unknown regression coefficient functions and ϵ_t is assumed to be an autoregressive (AR) series. The order of ϵ_t can be large. We will discuss how to estimate the coefficients $\alpha_i(\cdot)$ and select the AR order in this chapter.

4.1 A new estimation procedure

The AR error ϵ_t can be represented as

$$\epsilon_t = \beta_1\epsilon_{t-1} + \cdots + \beta_d\epsilon_{t-d} + \eta_t,$$

where $\{\eta_t\}$ is independently and identically distributed random error with mean zero and variance σ^2. Thus, the model (4.1) can be written as

$$y_t = \alpha_0(u_t) + \alpha_1(u_t)x_{t1} + \cdots + \alpha_p(u_t)x_{tp} + \beta_1\epsilon_{t-1} + \cdots + \beta_d\epsilon_{t-d} + \eta_t. \tag{4.2}$$

If the values for ϵ_t were available, then the coefficient functions $\alpha_j(\cdot)$ and AR coefficients β_j can be estimated directly by using existing estimation procedures for semiparametric varying-coefficient partially linear models. See, for example, Fan and Huang (2005) and Fan, Huang and Li (2007). In practice, ϵ_t is not available, but it may be estimated by $\widehat{\epsilon}_t = y_t - \tilde{\alpha}_0(u_t) - \tilde{\alpha}_1(u_t)x_{t1} - \cdots - \tilde{\alpha}_p(u_t)x_{tp}$, where $\{\tilde{\alpha}_i(\cdot), i = 1, \cdots, p\}$ is obtained by the nonparametric method pretending that the errors are independent. In this dissertation, we employ the local linear estimator to estimate $\alpha_i(\cdot)$. Detailed implementation can be found in Section 2.2.1. or Fan and Zhang (1999).

For simplicity of presentation, denote $\boldsymbol{\alpha}(u_t) = (\alpha_0(u_t), \cdots, \alpha_p(u_t))^T$, $\mathbf{X}_t = (1, x_{t1}, \cdots, x_{tp})^T$ and $\mathbf{e}_t = (\widehat{\epsilon}_{t-1}, \cdots, \widehat{\epsilon}_{t-d})^T$. Replacing ϵ_t's with $\widehat{\epsilon}_t$'s, model (4.2) becomes

$$y_t = \boldsymbol{\alpha}(u_t)^T\mathbf{X}_t + \mathbf{e}_t^T\boldsymbol{\beta} + \eta_t, \tag{4.3}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)^T$.

4.1.1 Profile least squares estimate

We will use the profile least squares estimation procedure proposed by Fan and Huang (2005) to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}(\cdot)$.

For given $\boldsymbol{\beta}$, denote $y_t^* = y_t - \mathbf{e}_t^T\boldsymbol{\beta}$ for $t = d+1, \cdots, n$. Then

$$y_t^* = \sum_{i=0}^{p} \alpha_i(u_t)x_{ti} + \eta_t \tag{4.4}$$

which is a varying coefficient model. We can employ local linear smoothers to estimate $\{\alpha_i(\cdot), i = 0, \cdots, p\}$. Specifically for a given u_0, we locally approximate the coefficient function as

$$\alpha_i(u) \approx \alpha_i(u_0) + \alpha_i'(u_0)(u - u_0) \widehat{=} a_i + b_i(u - u_0).$$

Local linear regression is used to estimate the local parameter $\{(a_i, b_i), i = 0, \cdots, p\}$ via minimizing the following weighted least squares function

$$\sum_{t=d+1}^{n} [y_t^* - \sum_{i=1}^{p}\{a_i - b_i(u_t - u_0)\}x_{ti}]^2 K_h(u_t - u_0),$$

with respect to $\{(a_i, b_i), i = 0, \cdots, p\}$. Denote the resulting estimate by $\{(\widehat{a}_i, \widehat{b}_i), i = 0, \cdots, p\}$. Then,

$$\widehat{\alpha}_j(t_0) = \widehat{a}_j, \quad \text{and} \quad \widehat{\alpha}_j'(t_0) = \widehat{b}_j$$

for $j = 0, \cdots, p$.

It is clear that the local linear estimate of $\mathbf{M} = (\boldsymbol{\alpha}(u_{d+1})^T\mathbf{X}_{d+1}, \cdots, \boldsymbol{\alpha}(u_n)^T\mathbf{X}_n)$ is linear in terms of $\mathbf{y}^* = (y_{d+1}^*, \cdots, y_n^*)^T$. Let $\widehat{\mathbf{M}}$ be the estimator of $\mathbf{M}$. Then it can be represented as

$$\widehat{\mathbf{M}} = S_h\mathbf{y}^*, \tag{4.5}$$

where S_h is a $(n-d) \times (n-d)$ smoothing matrix depending on $\{u_t, \mathbf{X}_t\}$'s and the bandwidth only.

Substituting for $\mathbf{M}(u_t, \mathbf{X}_t)$ by $\widehat{\mathbf{M}}(u_t, \mathbf{X}_t)$ in the matrix form of (4.3), we obtain

a synthetic linear regression model

$$(I - S_h)\mathbf{y} = (I - S_h)\mathbf{E}\boldsymbol{\beta} + \boldsymbol{\eta},$$

where I is the identity matrix, $\mathbf{E} = (\mathbf{e}_{d+1}, \cdots, \mathbf{e}_n)^T$ and $\boldsymbol{\eta} = (\eta_{d+1}, \cdots, \eta_n)^T$. Thus, the profile least squares estimators for $\boldsymbol{\beta}$ and $\mathbf{M}$ are

$$\widehat{\boldsymbol{\beta}} = \{\mathbf{E}^T(I - S_h)^T(I - S_h)\mathbf{E}\}^{-1}\mathbf{E}^T(I - S_h)^T(I - S_h)\mathbf{y}, \tag{4.6}$$

and

$$\widehat{\mathbf{M}} = S_h(\mathbf{y} - \mathbf{E}\widehat{\boldsymbol{\beta}}), \tag{4.7}$$

respectively.

The asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\alpha}_i(u_0), i = 0, \ldots, p$ are given in the following theorem. Denote $\mu_l = \int u^l K(u)\,\mathrm{d}u$ and $\nu_l = \int u^l K^2(u)\,\mathrm{d}u$.

Theorem 2. *Suppose that Conditions A—G listed in Section 4.3 hold. Then*

(A) The asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ *is*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{L} N(0, \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}^{-1})$$

where $\mathbf{f}_t = (\epsilon_{t-1}, \ldots, \epsilon_{t-d})^T$ *and* $\sigma^2 = Var(\eta_t)$.

(B) The asymptotic distribution of $\widehat{\alpha}_i(u_0, \widehat{\boldsymbol{\beta}}), i = 0, \ldots, p$, *conditioning on* $u_1, \cdots, u_n$ *and* $\mathbf{X}_1, \cdots, \mathbf{X}_n$, *is*

$$\sqrt{nh}\{\widehat{\alpha}_i(u_0, \widehat{\boldsymbol{\beta}}) - \alpha_j(u_0) - \frac{1}{2}\mu_2\alpha_i''(u_0)h^2\} \xrightarrow{L} N(0, \frac{\nu_0\sigma^2}{g(u_0)}),\ i = 0, \ldots, p.$$

where $g(u)$ *is the density function of* u.

According to Fan and Huang (2005), $\sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}^{-1}$ is the semiparametric efficiency bound for general varying-coefficient partially linear model. In addition, the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ is the same as that of Yule-Walker estimator for the AR model:

$$\epsilon_t = \beta_1\epsilon_{t-1} + \cdots + \beta_d\epsilon_{t-d} + \eta_t.$$

(see Theorem 8.1.1 of Brockwell and Davis, 1991). So Theorem 2 (a) implies that $\widehat{\boldsymbol{\beta}}$ is as efficient as if the one knew the true functional coefficients $\alpha_i(\cdot)$'s in advance. Theorem 2 (b) indicates that $\widehat{\alpha}_i(\cdot, \widehat{\boldsymbol{\beta}})$ shares the same asymptotic bias and variance as those of the local linear regression for independently and identically distributed observations. This result indicates that our proposed profile least squares estimate is very effective.

4.1.2 The selection of the AR order

Regarding the varying-coefficient model (4.2), we may start from a large AR order and need to establish an algorithm to select the appropriate order. Motivated by the variable selection mechanism in parametric regression, we add a penalty term onto the squared loss function as below:

$$\frac{1}{2}\sum_{t=d+1}^{n}\{y_t - \alpha_0(u_t) - \alpha_1(u_t)x_{t1} - \cdots - \alpha_p(u_t)x_{tp} - \mathbf{e}_t^T\boldsymbol{\beta}\}^2 + n\sum_{j=1}^{d}\lambda_j p_j(|\beta_j|) \quad (4.8)$$

where $p_j(\cdot)$ is the penalty function and λ_j is the tuning parameter to control the model complexity. For ease of presentation, we denote $\lambda_j p_j(|\beta_j|)$ by $p_{\lambda_j}(|\beta_j|)$.

With a proper choice of penalty function and λ_j, we expect to get some exact zero estimates by minimizing (4.8) with respect to $\boldsymbol{\beta}$. This is equivalent to removing the corresponding term from the original model. However, it is challenging to minimize

(4.8) directly because the functional coefficient $\alpha_l(\cdot)$ has not been parameterized. Using the profiling technique as introduced in the previous section, we can substitute the functional coefficient by a linear form of $\boldsymbol{\beta}$ and get the following profile least squared loss function:

$$\frac{1}{2}(\mathbf{y} - \mathbf{E}^T\boldsymbol{\beta})^T(I - \mathbf{S}_h)^T(I - \mathbf{S}_h)(\mathbf{y} - \mathbf{E}^T\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|) \tag{4.9}$$

There are various choices for the penalty function $p_{\lambda_j}(\cdot)$. We want the penalty function (1) to force the nonsignificant estimates of β_j to zero automatically (2) to keep the large estimates of β_j unbiased (3) to be continuous. Some commonly used penalty functions such as the family of L_q penalties ($q \geq 0$) do not satisfy these desired properties. Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty that meets all these criteria so that we use it as our preferred penalty in this paper. The derivative of the SCAD penalty is defined by

$$p'_\lambda(\beta) = \lambda\{\mathbf{I}(\beta \leq \lambda) + \frac{(\mathbf{a}\lambda - \beta)_+}{(\mathbf{a}-\mathbf{1})\lambda}\mathbf{I}(\beta > \lambda)\}$$

for some $a > 0$ and $\beta > 0$. From a Bayesian point of view, Fan & Li (2001) suggested fixing $a = 3.7$. Figure 4.1 depicts a sample of the derivative of the SCAD penalty with $\lambda = 2$ and $a = 3.7$. From the plot, we find that the SCAD penalty has two knots at λ and $a\lambda$ and poses a different amount of penalty according to the magnitude of β.

In practice, the minimization of the SCAD penalized profile least squares is not easy because it is irregular at the origin and does not have second derivative at some points. To solve this difficulty, we take the local quadratic approximation (LQA) for the SCAD penalty function suggested by Fan and Li (2004). Suppose we can

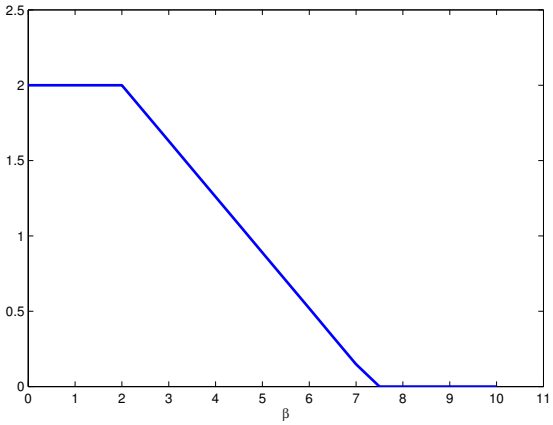

Figure 4.1. The derivative of the Scad penalty function with $\lambda = 2$ and $a = 3.7$

get an estimate $\beta_j^{(k)}$ in the k^{th} step that is close to the true β_j. If $|\beta_j^{(k)}|$ is close to 0, then we set $\widehat{\beta_j} = 0$. Otherwise, the SCAD penalty is locally approximated by a quadratic function as

$$[p_{\lambda_j}(|\beta_j|)]' = p'_{\lambda_j}(|\beta_j|) \cdot \text{sgn}(\beta_j) \approx \{p'_{\lambda_j}(|\beta_j^{(k)}|)/|\beta_j^{(k)}|\}\beta_j$$

Then we can employ Newton-Raphson algorithm to minimize (4.9). In practice, we use the following iterative ridge regression to find the minimizer of (4.9):

$$\boldsymbol{\beta}^{(k+1)} = \{\mathbf{E}^T(I - S_h)^T(I - S_h)\mathbf{E} + n\Sigma_\lambda(\boldsymbol{\beta}^{(k)})\}^{-1}\mathbf{E}^T(I - S_h)^T(I - S_h)\mathbf{y} \quad (4.10)$$

where $\Sigma_\lambda(\boldsymbol{\beta}^{(k)}) = \text{diag}\{p'_{\lambda_1}(|\beta_1^{(k)}|)/|\beta_1^{(k)}|, \ldots, p'_{\lambda_d}(|\beta_d^{(k)}|)/|\beta_d^{(k)}|\}$ for nonvanished $\boldsymbol{\beta}^{(k)}$. The unpenalized profile least squares estimator is taken as the initial value to update $\boldsymbol{\beta}^{(1)}$.

The other important issue in the implementation is to select the tuning parameter λ_j. Since minimization of (4.9) with respect to $(\lambda_1, \ldots, \lambda_d)$ is a high dimensional optimization problem, it can be challenging. But the magnitude of of λ_j is believed

to be proportional to the standard error of the estimate of β_j. So we set $\lambda_j = \lambda \, \mathrm{se}(\widehat{\beta}_j)$ where $\mathrm{se}(\widehat{\beta}_j)$ is the standard error of the unpenalized least square estimates. Now the original d-dimensional optimization reduces to a 1-dimensional problem. We can minimize the BIC or GCV score to find the optimal λ. Here we use the BIC selector.

Define the effective number of parameters of the penalized least square estimator (4.10) to be

$$e(\lambda) = \mathrm{tr}[\{\tilde{D} + \Sigma_\lambda(\widehat{\boldsymbol{\beta}})\}^{-1}\tilde{D}]$$

where $\tilde{D} = \mathbf{E}^T(I - \mathbf{S}_h)^T(I - \mathbf{S}_h)\mathbf{E}$ for nozero $\widehat{\boldsymbol{\beta}}$.

Then the BIC score can be calculated by

$$BIC(\lambda) = \log(\frac{RSS}{n}) + e(\lambda)\frac{\log n}{n}$$

where $RSS = \|(I - S_h)\mathbf{y} - (I - S_h)\mathbf{E}\widehat{\boldsymbol{\beta}}\|^2$ is the residual sum of squares given λ. In practice, λ is taken from a range of $[\frac{0.1}{\sqrt{n}}, \frac{2\sqrt{\log n}}{\sqrt{n}}]$ and the minimizer of the BIC score is selected as the tuning parameter.

4.2 Simulation studies and applications

Example 1. In this section, we aim to examine the finite sample performance of the proposed procedures by Monte Carlo simulation. We will compare the performance of the proposed procedures with the local linear estimator without considering the error structure with respect to the mean squares errors, which is defined by

$$\mathrm{MSE}\{\boldsymbol{\alpha}(\cdot)\} = \frac{1}{n}\sum_{t=1}^{n}\sum_{i=0}^{p}\{\widehat{\alpha}_i(u_t) - \alpha_i(u_t)\}^2$$

We summarize our simulation results in terms of relative MSE (RMSE), defined by the ratio of the MSE of an estimation procedure to the MSE of $\widehat{\boldsymbol{\alpha}}_I(\cdot)$, where $\widehat{\alpha}(\cdot)$ is the estimate of $\boldsymbol{\alpha}_I(\cdot)$ pretending that the error ϵ_t is independent. We report the percentage of accuracy gain, defined by $(1 - RMSE) * 100\%$ for comparison purposes.

In this example, a random sample of size n, either $n = 100$ or $n = 500$, is generated from

$$y_t = \alpha_0(u_t) + \alpha_1(u_t)x_{t1} + \alpha_2(u_t)x_{t2} + \epsilon_t.$$

where $\alpha_0(u) = 3u^2 - 2u + 1, \alpha_1(u) = \cos(2\pi u), \alpha_2(u) = 2\sin(2\pi u)$ are the functional coefficients with same degree of smoothness and $\{u_t, x_{t1}, x_{t2}\}$ is the covariate with its own distribution that will be considered in three distinguished schemes as follows:

I. $\{u_t\}$ is independently and identically distributed according to the uniform distribution over $[0, 1]$. $\{x_{t1}, x_{t2}\}$ follows a multivariate normal distribution with mean vector $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and covariance $\begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$.

II. $\{v_t\}$ is independently and identically distributed according to the standard normal distribution for $t = 1, 2, ...$ Let $u_t = \Phi\{(av_t + bv_{t-1})/\sqrt{a^2 + b^2}\}$ for $t = 2, 3, \cdots, n$ where $\Phi(v)$ is the cumulative distribution function of the standard normal distribution. Thus, $\{u_t\}$ is 1-dependent process. $\{x_{t1}, x_{t2}\}$ is generated as a multidimensional 1-dependent process as well. Assume $\{z_{t1}, z_{t2}\}$ to be multivariate normally distributed with mean vector $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and covariance $\begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$. $\{x_{t1}, x_{t2}\} = c\{z_{t1}, z_{t2}\} + d\{z_{t-1,1}, z_{t-1,2}\}$ with $c^2 + d^2 = 1$ for $t = 2, 3, \cdots, n$. In our simulation, we take $a = 0.9, b = 0.1, c = 0.8$ and

$d = 0.6$.

III. $\{u_t\}$ is the fixed design point evenly distributed over $[0, 1]$, i.e., $u_t = (t-0.5)/n$, all other settings are same as those in scheme II.

The error process ϵ_t is an AR process with order $d = 10$ or 20, i.e.,

$$\epsilon_t = \sum_{j=1}^{d} \beta_j \epsilon_{t-j} + \eta_t,$$

where $\eta_t \sim N(0, \sigma^2)$ with $\sigma = 0.5$ or 1. We consider two situations: the first one is AR(1) model with $\beta_1 = 0.5$, or 0.7, and all other β_j's equal 0, the second one is AR(2) model with $\beta_1 = 0.5$, $\beta_2 = 0.4$ or $\beta_1 = 0.7$, $\beta_2 = 0.2$ and all others equal 0. The number of replications is 500.

For each sampling scheme, two methods, the least squares method (Profile) and the penalized profile least square method with the SCAD penalty function (SCAD) are compared with the oracle estimator. The oracle result is obtained by substituting the true autoregressive coefficients and the autoregressive order.

Bandwidth Selection: For a given bandwidth, we can form the mean square of errors as below:

$$\text{MSE}(h) = \frac{1}{n} \sum_{t=1}^{n} \sum_{i=0}^{p} \{\widehat{\alpha}_i(u_t) - \alpha_i(u_t)\}^2$$

where $\widehat{\alpha}_i(\cdot)$ is the regular local linear estimator pretending that data are i.i.d. or the profile least square estimator respectively. We find the bandwidth that minimizes MSE(h) in a pilot study. In the simulation table 4.1—4.6, h_1 and h_2 denote the bandwidths in the initial estimation and profile least squares estimation correspondingly.

Table 4.1. Simulation results for the varying-coefficient model under sampling scheme I when $d = 10$. $(1 - RMSE) * 100\%$ is summarized for comparison.

(β_1, β_2)	σ	h_1	h_2	Profile	SCAD	Oracle	h_1	h_2	Profile	SCAD	Oracle
				$n = 100$					$n = 500$		
(0.5,0)	0.5	0.200	0.150	13.87	17.09	19.04	0.125	0.100	26.67	27.14	27.97
(0.7,0)	0.5	0.175	0.175	17.29	21.35	21.90	0.125	0.100	30.15	31.62	31.75
(0.5,0.4)	0.5	0.225	0.200	22.74	24.04	25.99	0.125	0.100	39.07	39.51	39.50
(0.7,0.2)	0.5	0.175	0.175	29.09	30.21	32.48	0.125	0.100	42.64	43.12	43.05
(0.5,0)	1	0.200	0.200	6.99	7.40	11.95	0.125	0.125	16.67	17.49	17.39
(0.7,0)	1	0.250	0.225	18.17	21.62	21.66	0.125	0.125	34.02	34.59	34.46
(0.5,0.4)	1	0.275	0.225	23.36	25.07	25.66	0.125	0.125	41.69	41.95	42.02
(0.7,0.2)	1	0.275	0.225	26.96	28.01	29.22	0.125	0.125	45.62	45.85	45.85

Table 4.2. Simulation results for the varying-coefficient model under sampling scheme I when $d = 20$. $(1 - RMSE) * 100\%$ is summarized for comparison.

(β_1, β_2)	σ	h_1	h_2	Profile	SCAD	Oracle	h_1	h_2	Profile	SCAD	Oracle
				$n = 100$					$n = 500$		
(0.5,0)	0.5	0.125	0.175	10.38	18.76	22.26	0.100	0.100	25.42	27.35	27.61
(0.7,0)	0.5	0.175	0.175	11.75	20.48	21.90	0.125	0.100	29.96	31.60	31.75
(0.5,0.4)	0.5	0.200	0.175	20.63	25.11	27.76	0.150	0.100	39.75	40.52	40.52
(0.7,0.2)	0.5	0.200	0.175	24.00	28.37	31.08	0.150	0.100	43.00	43.42	43.39
(0.5,0)	1	0.250	0.225	4.51	9.79	11.42	0.150	0.125	13.31	15.38	15.33
(0.7,0)	1	0.250	0.200	11.71	19.54	20.05	0.150	0.125	29.35	30.96	30.85
(0.5,0.4)	1	0.200	0.225	24.71	28.33	30.29	0.150	0.125	38.02	38.72	38.77
(0.7,0.2)	1	0.225	0.225	26.02	30.25	31.53	0.150	0.125	41.82	42.48	42.45

Simulation results for sampling scheme I are summarized in Table 4.1 and 4.2. Under this sampling scheme, the covariates $\{u_t\}$ and $\{x_{t1}, x_{t2}\}$ are independent. Our proposed methods can effectively improve the estimation accuracy, especially when the correlation is strong. The SCAD procedure always outperforms the profile least squares method. This superiority is more significant when the sample size is small. On the other hand, when $n = 100$, the profile least squares estimations at $d = 10$ are better than those at $d = 20$. This outcome is reasonable because $d = 20$

is farther away from the true AR order and then results in more biases. However, the SCAD estimations at $d = 10$ and $d = 20$ are close to each other. This implies that the estimation result will not be sensitive to the assumption of the AR order as long as an order selection procedure is conducted.

When the sample size is large, such as $n = 500$, both the profile least squares method and the SCAD method have larger gain than those in a moderate sample such as $n = 100$. In addition, they are close to each other and the oracle estimator. The difference between $d = 10$ and $d = 20$ is not remarkable. This is expected because the estimate should be more accurate as the sample size increases.

For sampling II, both $\{u_t\}$ and $\{x_{t1}, x_{t2}\}$ are 1-dependent processes. However, the overall pattern of Table 4.3 and 4.4 is very similar to that in Table 4.1 and 4.2. The SCAD method has a better performance than the profile least squares method all the time. This advantage is more notable when the sample size is small or the assumption of the AR order is much larger than the truth. This finding indicates that the order selection procedure can effectively reduce the fitting bias and improve the estimation accuracy. Furthermore, the profile least squares estimator and the SCAD estimator are close to the oracle result, especially when the sample size is large.

We can conclude that our proposed estimation methods can work well with either independent or 1-dependent situation.

In the last sampling scheme, $\{u_t\}$ are fixed design points that contain a more complicated correlation. Many authors have found that such a sampling design in nonparametric setting can impose great challenges for bandwidth selection and model estimation. From our limited simulation experience, our proposed profile least squares method and the penalized profile least squares method still work well in this situation. The SCAD method is still better than the profile least squares

Table 4.3. Simulation results for the varying-coefficient model under sampling scheme II when $d = 10$. $(1 - RMSE) * 100\%$ is summarized for comparison.

(β_1, β_2)	σ	h_1	h_2	Profile	SCAD	Oracle	h_1	h_2	Profile	SCAD	Oracle
				$n = 100$					$n = 500$		
(0.5,0)	0.5	0.200	0.175	7.62	12.76	13.26	0.100	0.100	13.64	14.53	14.57
(0.7,0)	0.5	0.150	0.150	18.38	21.54	22.11	0.100	0.100	30.56	31.19	31.08
(0.5,0.4)	0.5	0.225	0.175	24.38	24.74	27.07	0.125	0.100	40.67	40.98	41.08
(0.7,0.2)	0.5	0.150	0.175	30.83	31.67	33.37	0.125	0.100	43.54	43.80	43.79
(0.5,0)	1	0.150	0.225	6.12	10.64	11.83	0.175	0.150	16.94	17.97	17.83
(0.7,0)	1	0.275	0.225	16.95	21.29	21.20	0.175	0.150	27.61	28.09	28.32
(0.5,0.4)	1	0.350	0.275	23.70	24.38	25.84	0.200	0.150	39.21	39.40	39.48
(0.7,0.2)	1	0.250	0.275	25.14	26.05	27.68	0.200	0.150	41.17	41.25	41.31

Table 4.4. Simulation results for the varying-coefficient model under sampling scheme II when $d = 20$. $(1 - RMSE) * 100\%$ is summarized for comparison.

(β_1, β_2)	σ	h_1	h_2	Profile	SCAD	Oracle	h_1	h_2	Profile	SCAD	Oracle
				$n = 100$					$n = 500$		
(0.5,0)	0.5	0.200	0.175	1.38	11.68	13.26	0.125	0.100	14.15	16.70	16.62
(0.7,0)	0.5	0.200	0.150	10.68	17.93	18.44	0.125	0.100	26.96	28.67	28.46
(0.5,0.4)	0.5	0.150	0.175	19.19	22.79	25.81	0.125	0.100	39.91	41.00	41.08
(0.7,0.2)	0.5	0.125	0.175	33.84	37.59	39.36	0.125	0.100	42.73	43.81	43.79
(0.5,0)	1	0.200	0.225	1.84	9.85	11.46	0.125	0.150	16.67	13.68	13.74
(0.7,0)	1	0.250	0.225	11.66	20.09	20.40	0.200	0.150	31.80	33.50	33.27
(0.5,0.4)	1	0.325	0.225	23.43	26.39	28.68	0.225	0.150	41.07	41.83	41.86
(0.7,0.2)	1	0.325	0.225	22.15	25.77	27.53	0.225	0.150	42.43	43.09	43.09

method. But the overall gain under this sampling scheme is less than that in scheme I and II.

Example 2. In this example, we illustrate the proposed methodology by an analysis of US macroeconomic data from January 1980 to December 2006. All variables are recorded on a monthly basis. Thus, there are in total 324 observations. It is of interest to investigate how the U.S. macroeconomics factors affect the house price

Table 4.5. Simulation results for the varying-coefficient model under sampling scheme III when $d = 10$. $(1 - RMSE) * 100\%$ is summarized for comparison.

(β_1, β_2)	σ	h_1	h_2	Profile	SCAD	Oracle	h_1	h_2	Profile	SCAD	Oracle
				$n = 100$					$n = 500$		
(0.5,0)	0.5	0.175	0.200	1.01	2.58	5.77	0.125	0.125	1.74	2.11	2.61
(0.7,0)	0.5	0.200	0.250	1.25	4.30	7.24	0.125	0.125	11.47	12.51	12.78
(0.5,0.4)	0.5	0.225	0.200	5.57	6.05	11.52	0.175	0.150	17.98	18.29	20.32
(0.7,0.2)	0.5	0.175	0.200	7.27	8.20	11.53	0.125	0.150	24.74	24.74	25.59
(0.5,0)	1	0.150	0.150	2.37	4.49	5.09	0.175	0.150	5.09	6.58	6.43
(0.7,0)	1	0.250	0.225	5.22	8.16	9.30	0.175	0.150	12.51	13.38	13.91
(0.5,0.4)	1	0.200	0.325	9.47	9.95	12.97	0.200	0.200	20.65	20.70	21.97
(0.7,0.2)	1	0.300	0.275	9.45	10.43	13.91	0.250	0.175	17.32	17.80	20.49

Table 4.6. Simulation results for the varying-coefficient model under sampling scheme III when $d = 20$. $(1 - RMSE) * 100\%$ is summarized for comparison.

(β_1, β_2)	σ	h_1	h_2	Profile	SCAD	Oracle	h_1	h_2	Profile	SCAD	Oracle
				$n = 100$					$n = 500$		
(0.5,0)	0.5	0.150	0.175	1.45	2.42	6.42	0.100	0.125	5.03	7.75	8.05
(0.7,0)	0.5	0.175	0.175	3.38	6.92	8.18	0.150	0.100	9.29	12.16	13.87
(0.5,0.4)	0.5	0.225	0.200	4.57	5.87	11.52	0.200	0.150	18.63	19.61	22.63
(0.7,0.2)	0.5	0.275	0.200	10.65	12.60	16.91	0.200	0.150	17.08	17.58	20.39
(0.5,0)	1	0.275	0.200	0.03	2.09	4.53	0.150	0.150	3.49	6.20	6.37
(0.7,0)	1	0.250	0.250	3.14	8.01	8.50	0.200	0.150	11.94	14.02	14.60
(0.5,0.4)	1	0.300	0.275	7.50	8.94	13.52	0.300	0.175	21.10	21.73	26.36
(0.7,0.2)	1	0.275	0.300	7.52	9.82	12.74	0.300	0.175	19.50	19.44	24.05

index change. To this end, we consider

$$y_t = \alpha_0(u_t) + \alpha_1(u_t)\mathrm{x}_{t1} + \alpha_2(u_t)x_{t2} + \epsilon_t, \tag{4.11}$$

where we take *the house price index change* as response variable y_t, *the unemployment rate* as u_t, *the prime interest rate* as x_{t1} and *the gross domestic product growth rate* as x_{t2}. To reduce rounding error, variable u_t is rescaled into $[0, 1]$.

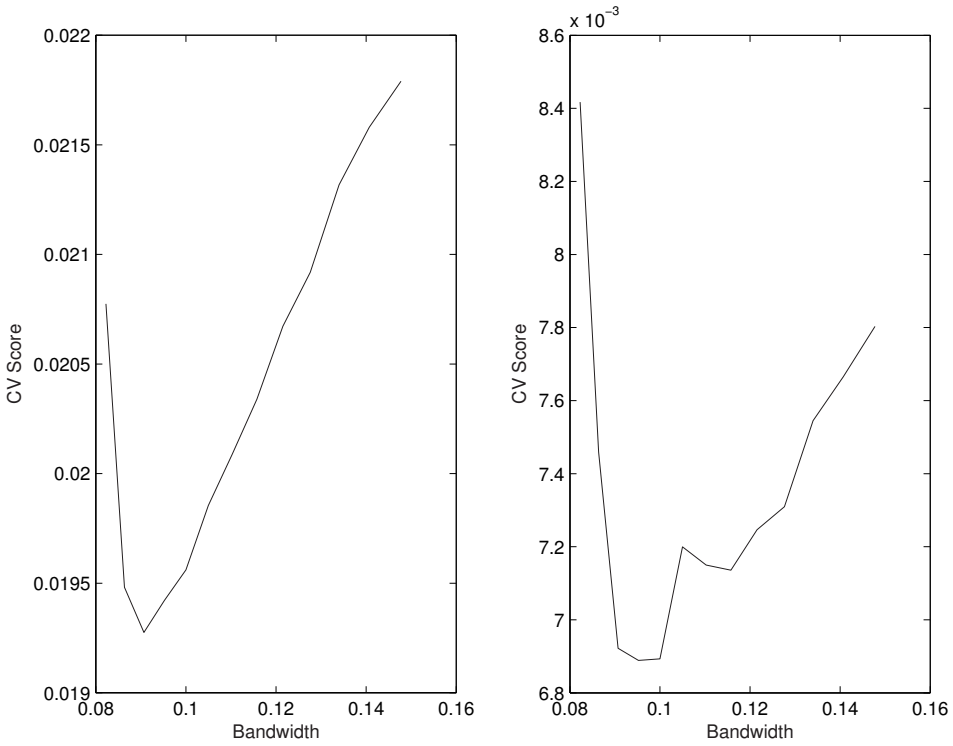

Figure 4.2. Plots of cross validation score for macroeconomic data. Left panel is for CV in the initial estimation; Right panel is for CV in the profile least square estimation

Initial estimate $\tilde{\alpha}_j(\cdot)$. We employ local linear regression to get the initial estimate of $\tilde{\alpha}_j(\cdot)$ ignoring the correlation of the random error. We use a multifold cross validation to select the bandwidth for the initial estimate. We divide the total 324 observations into 9 equally sized groups, whose $j^{\text{th}}, j = 1, \ldots, 9$ group consists of the observations with indices

$$d_j = \{9k + j, k = 0, \ldots, 35\},\ j = 1, \ldots, 9$$

For each j, all observations except those in the j^{th} group are used to fit the model while the j^{th} group is used to test the goodness of the fit. The sum of the

mean squares of residuals over 9 groups is computed.

$$CV(h) = \frac{1}{n}\sum_{j=1}^{9}\sum_{i\in d_j}\{y_i - \widehat{y}_{-d_j}(U_i, \boldsymbol{X}_i)\}^2$$

The cross-validation score versus different bandwidths is plotted in Figure 4.2. The selected bandwidth is 0.0907.

Residual analysis. With the initial estimate, we are able to conduct residual analysis. Figure 4.3 (a) shows the auto-correlation of $\widehat{\epsilon}_t$ and indicates that the random error is strongly auto-correlated. Figure 4.3 (b) shows the partial auto-correlation of $\widehat{\epsilon}_t$ and indicates that an AR model for $\widehat{\epsilon}_t$ will be appropriate.

We consider AR(12) model for $\{\epsilon_t\}$, which implies that the error may have a year lag. Similar to the initial estimation, we divide the data set into 9 equally sized groups again and sum up the squares of the estimation errors over those observations that are excluded from the estimation each time

$$CV(h) = \frac{1}{n}\sum_{j=1}^{9}\sum_{i\in d_j}\{y_i - \widehat{\alpha}_0(u_i) - \widehat{\alpha}_1(u_i)x_{i1} - \widehat{\alpha}_2(u_i)x_{i2} - \mathbf{e_i^T}\widehat{\boldsymbol{\beta}}\}^{\mathbf{2}}$$

where $\widehat{\alpha}_l(\cdot)$ and $\widehat{\boldsymbol{\beta}}$ are obtained by the profile least squares method discussed in Section 4.1. By the CV plot in Figure 4.2, the optimal bandwidth is 0.0952.

Next, we use penalized profile least squares with the SCAD penalty to select the AR order and estimate the coefficients. By minimizing BIC score, the optimal tuning parameter λ is 0.00021.

The AR(1) model with a strong autocorrelation coefficient 0.8974 is selected. This means that the house price has one month lag. After accounting for the auto-correlation, there is no significant pattern left. (See Figure 4.3 (c) and (d))

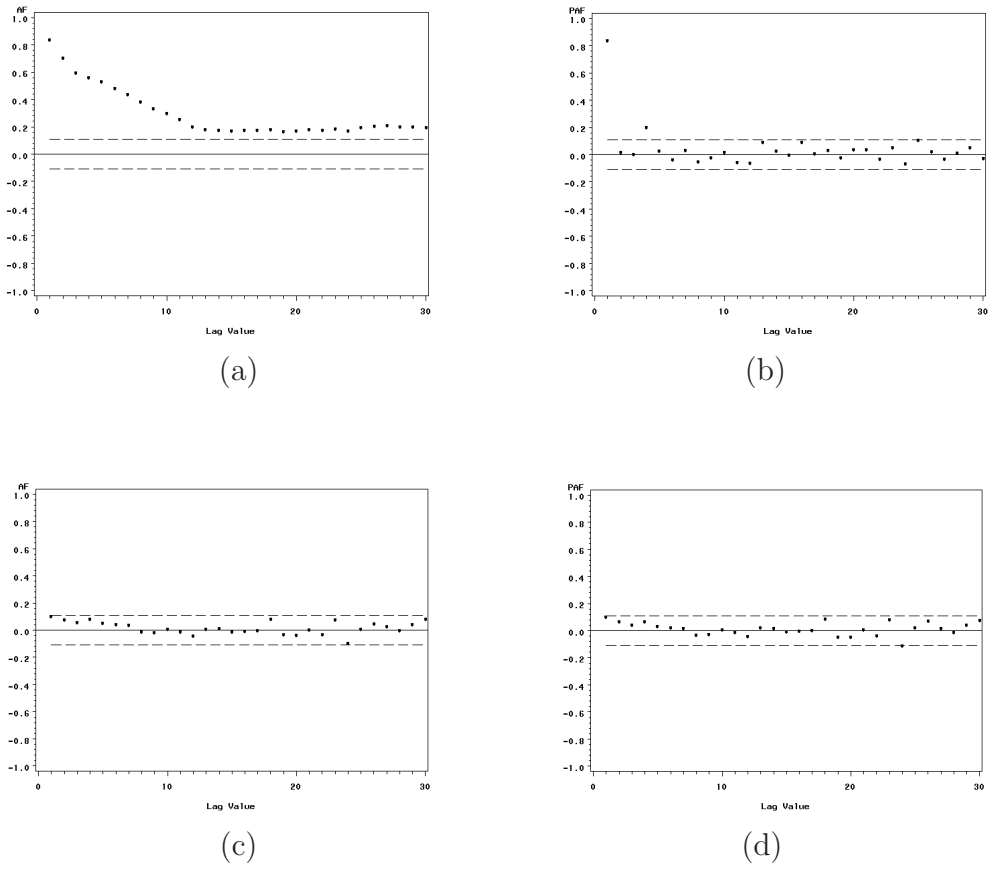

Figure 4.3. Correlogram of residual $\widehat{\epsilon}_t$ and $\widehat{\eta}_t$ for macroeconomic data. Plot (a) and (b) are the autocorrelation and partial autocorrelation for $\widehat{\epsilon}_t$. Plot (c) and (d) are the autocorrelation and partial autocorrelation for $\widehat{\eta}_t$. In each plot, the upper and lower dashed lines represent 95% confidence interval.

Prediction comparison The residual plot suggests that the house price index change has a significant autocorrelation in the short lag and the penalized profile least square estimation method selects $AR(1)$ model. These findings stimulate us to consider the house price index change from a pure $AR(1)$ process without any covariate as a simple alternative of the varying-coefficient model. That is,

$$y_t = \mu + \phi_1(y_{t-1} - \mu) + \omega_t \tag{4.12}$$

where μ is the mean of the autocorrelation process and $\{\omega_t\}$ is from a white noise process.

To make a sound choice between the simple autocorrelation model (4.12) and the varying-coefficient model with the autocorrelated errors (4.11), we hold out the last 24 observations for prediction comparison purpose. The coefficients are estimated based on the rest of 300 observations and the sum of the prediction error squares over the holdout 24 observations is calculated.

The maximum likelihood estimates of model (4.12) is

$$\widehat{\mu} = 0.0144,\ \widehat{\phi_1} = 0.93677$$

Therefore, we get the sum of the prediction error for model (4.12) is

$$\mathrm{PE_{AR}} = \sum_{t=301}^{324} [y_t - \widehat{\mu} - \widehat{\phi_1}(y_{t-1} - \widehat{\mu})]^2 = 0.00548$$

When only the first 300 observations are used to fit the varying-coefficient model (4.11), the bandwidths used in the initial local linear estimation and the profile least squares estimation are 0.0907 and 0.1216 respectively. The profile least squares method selects $AR(1)$ model with the coefficient 0.7912. Interpolating the estimated coefficients $\widehat{\alpha}_0(\cdot), \widehat{\alpha}_1(\cdot), \widehat{\alpha}_2(\cdot)$ at $u_{301} \sim u_{324}$, we can calculate the sum of the prediction error for model (4.11) as

$$\mathrm{PE_{varying}} = \sum_{t=301}^{324} [y_t - \widehat{\alpha}_0(u_t) + \widehat{\alpha}_1(u_t)\mathrm{x}_{t1} + \widehat{\alpha}_2(u_t)x_{t2} - \widehat{\beta_1}\widehat{\epsilon}_{t-1}]^2 = 0.00003246$$

The varying-coefficient model yields a much smaller prediction error than the simple autocorrelation model. It means that incorporating some covariates can predict the house price index change more accurately. Hence we will stick to the

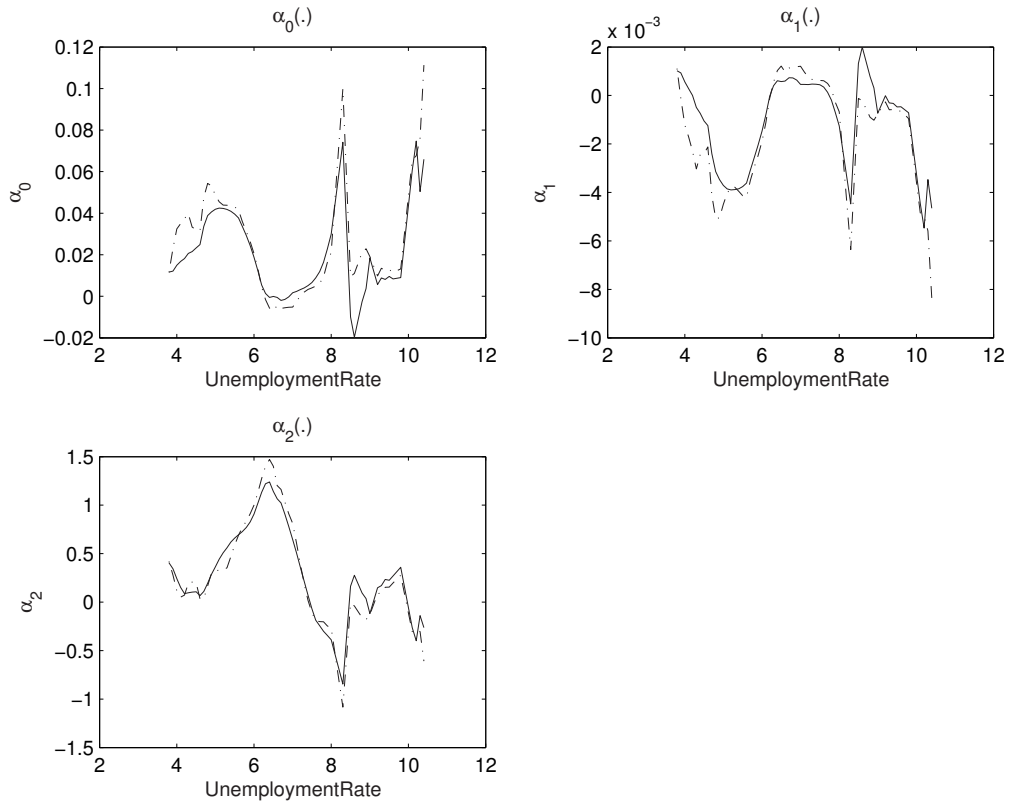

Figure 4.4. Estimation of functional variables for macroeconomic data. Dashed curves are the initial estimates; Solid curves are the penalized profile least squares estimate.

varying-coefficient model (4.11) and report the final estimated model based on all 324 observations.

Final model. The penalized profile least squares estimation procedure yields the final model

$$\widehat{y}_t = \widehat{\alpha}_0(u_t) + \widehat{\alpha}_1(u_t)\mathrm{x}_{t1} + \widehat{\alpha}_2(u_t)x_{t2} + 0.8974\widehat{\epsilon}_{t-1}, \tag{4.13}$$

where $\widehat{\alpha}_j$, $i = 0, 1, 2$ is depicted in Figure 4.4, from which it can be seen that by taking account of the error correlation, the curve is smoother than the initial estimate. In addition, the effects of prime interest rate and GDP growth rate seem to vary over different levels of unemployment rate.

Example 3. The goal of this example is to demonstrate the proposed methods by studying the relationship between levels of pollutants and the number of total hospital admissions for circulatory and respiratory problems. This data set was collected daily in Hong Kong from January 1, 1994 to December 31, 1995, and therefore its sample size n equals 730. We set the logarithm of the daily admission number to be the response variable y_t. Three air pollutants, sulfur dioxide(in $\mu g/m^3$), nitrogen dioxide (in $\mu g/m^3$) and dust (in $\mu g/m^3$) are recorded daily to measure the extent of the air pollution. The scatter plots (Figure 4.5) and the Pearson correlation coefficients suggest a moderate correlation between sulfur dioxide and nitrogen dioxide, and a strong correlation between dust and nitrogen dioxide. Therefore, we take the sum of these three air pollutant values and standardize it as the covariate x_t rather than treat them as three individual covariates.

To study the effects of air pollutants at a given time point, we take the date at which data were collected as u_t, which is rescaled so that its values lies between 0 and 1. Thus, we consider

$$y_t = \alpha_0(u_t) + \alpha_1(u_t)x_t + \epsilon_t$$

Initial estimate $\tilde{\alpha}_j(\cdot)$. Local linear regression is used to obtain the initial estimate of $\alpha_j(\cdot)$. The correlation structure of errors is ignored at this moment. In order to find a suitable bandwidth, we divide the data into 10 equally sized groups and minimize the cross validation score over 10 groups. The $j^{\text{th}}, j = 1, \ldots, 10$ group contains the observations with indices $d_j = \{10k + j, k = 0, \ldots, 72\}$, $j = 1, \ldots, 10$. At a given bandwidth, the observations in the j^{th} group are used to calculate the square of prediction errors based on the estimation of the rest 657 points. Then the

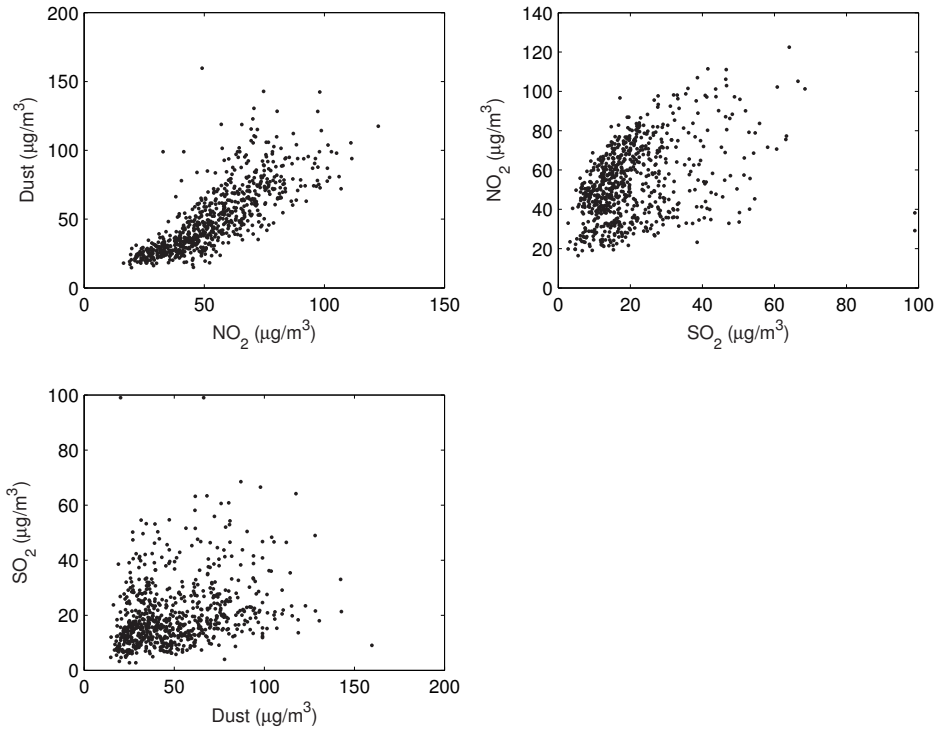

Figure 4.5. Scatter plots between sulfur dioxide, nitrogen dioxide and dust. The Pearson correlation coefficients of dust versus nitrogen dioxide, nitrogen dioxide versus sulfur dioxide, and sulfur dioxide versus dust are 0.7820, 0.4025 and 0.2810 respectively.

cross validation score is computed by

$$CV(h) = \frac{1}{n}\sum_{j=1}^{10}\sum_{i\in d_j}\{y_i - \widehat{y}_{-d_j}(U_i, \boldsymbol{X}_i)\}^2$$

The optimal bandwidth is 0.0338. (See Figure 4.6)

Residual analysis. With the selected bandwidth, we can construct the corresponding residuals $\widehat{\epsilon}_t$. The autocorrelation plot (Figure 4.7 (a)) shows that a quite bit of structure exists in $\{\widehat{\epsilon}_t\}$ while the partial-autocorrelation plot (Figure 4.7 (b)) suggests an autoregressive process with an order larger than 20.

We think that the total hospital admission number might have a month lag.

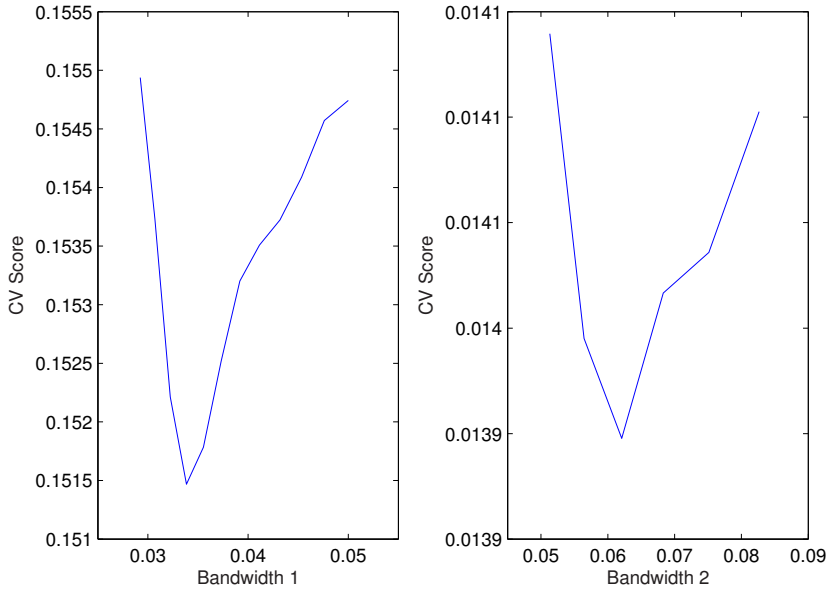

Figure 4.6. Plots of cross validation score for environmental data. Left panel is for CV in the initial estimation; Right panel is for CV in the profile least square estimation

So the AR(30) model is employed in the profile least square method to estimate $\alpha_j(\cdot)$. Similar 10-folded cross validation is used again to select the bandwidth. The predicted value for the observations in the j^{th} group is obtained by the profile least squares method.

$$CV(h) = \frac{1}{n}\sum_{j=1}^{10}\sum_{i\in d_j}\{y_i - \widehat{\alpha}_0(u_i) - \widehat{\alpha}_1(u_i)x_i - \mathbf{e_i^T}\widehat{\boldsymbol{\beta}}\}^{\mathbf{2}}$$

By the CV plot in Figure 4.6, the CV score reaches its minimum at the bandwidth 0.0621. When the optimal bandwidth is determined, we can apply penalized profile least squares with the SCAD penalty to select the AR order and estimate $\alpha_j(\cdot)$ and $\boldsymbol{\beta}$. By the BIC criterion, the tuning parameter λ used in the penalized least squares estimation is 0.0191. AR coefficients at lag 1, 5, 7, 8, 11, 14, 16, 17, 19, 21, 23, 24, 25, 27 and 29 are significant. So AR(29) model is selected, which implies that the

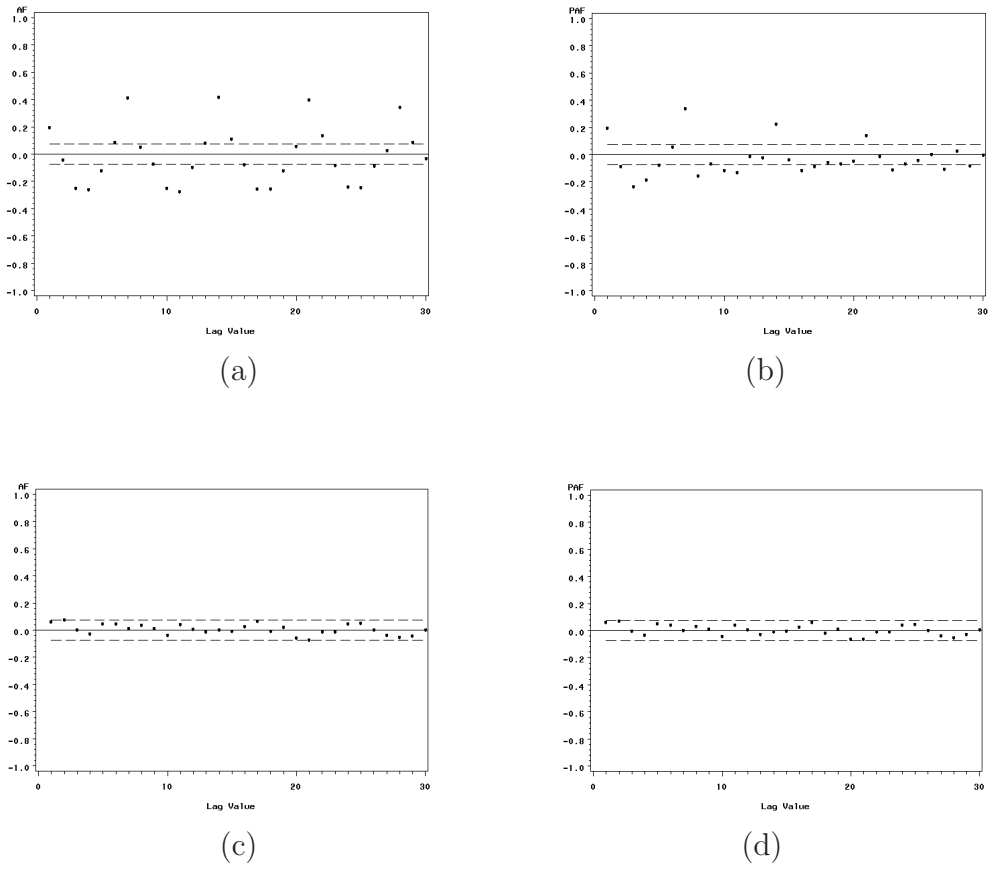

Figure 4.7. Correlogram of residual $\widehat{\epsilon}_t$ and $\widehat{\eta}_t$ for environmental data. Plot (a) and (b) are the autocorrelation and partial autocorrelation for $\widehat{\epsilon}_t$. Plot (c) and (d) are the autocorrelation and partial autocorrelation for $\widehat{\eta}_t$. In each plot, the upper and lower dashed lines represent 95% confidence interval.

number of total hospital admission has 29-day lag. The autocorrelation and partial-autocorrelation plots in Figure 4.7 (c) and (d) confirm that the autocorrelation has been removed and the residual $\widehat{\eta}_t$ looks like a white noise process.

Final model. We conclude the final model is

$$\widehat{y}_t = \widehat{\alpha}_0(u_t) + \widehat{\alpha}_1(u_t)\mathrm{x}_t + 0.1565\widehat{\epsilon}_{t-1} - 0.0931\widehat{\epsilon}_{t-5} + 0.2005\widehat{\epsilon}_{t-7} - 0.1685\widehat{\epsilon}_{t-8}$$
$$- 0.1494\widehat{\epsilon}_{t-11} + 0.1662\widehat{\epsilon}_{t-14} - 0.1216\widehat{\epsilon}_{t-16} - 0.1141\widehat{\epsilon}_{t-17} - 0.1261\widehat{\epsilon}_{t-19} + 0.1409\widehat{\epsilon}_{t-21}$$

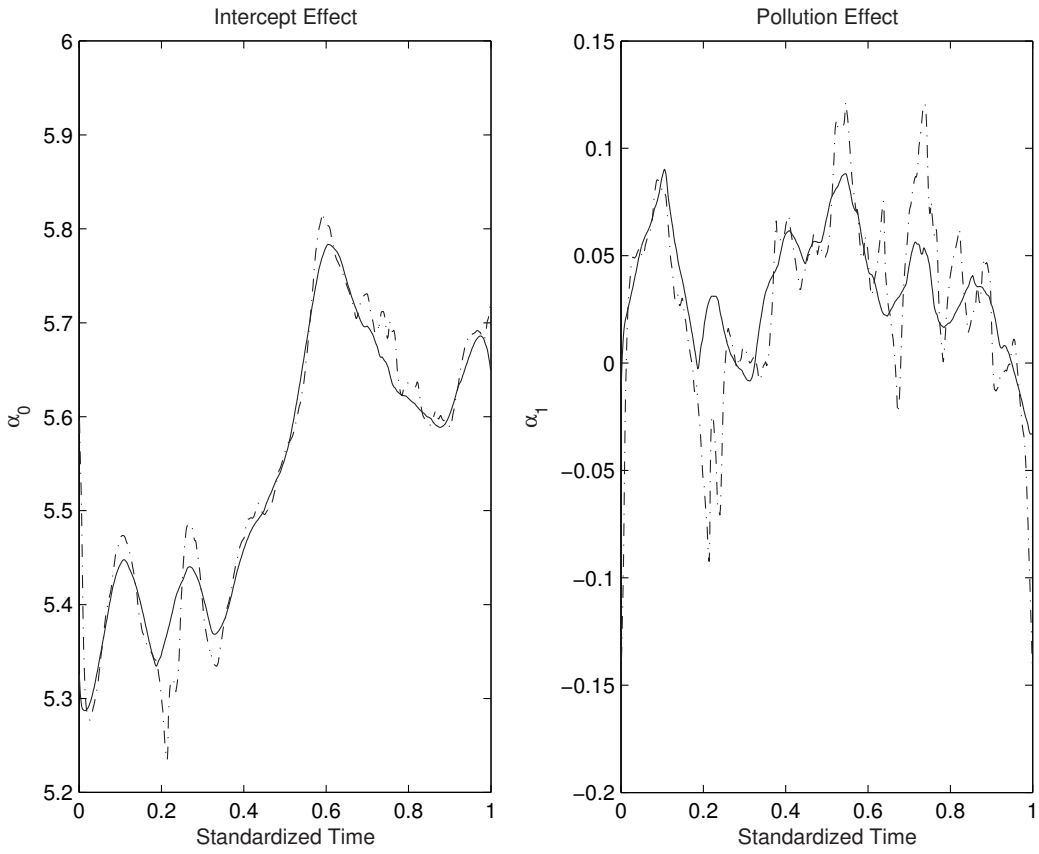

Figure 4.8. Estimation of functional variables for environmental data. Dashed curves are the initial estimates; Solid curves are the penalized profile least squares estimate.

$$- 0.0824\widehat{\epsilon}_{t-23} - 0.1185\widehat{\epsilon}_{t-24} - 0.0930\widehat{\epsilon}_{t-25} - 0.1344\widehat{\epsilon}_{t-27} - 0.0751\widehat{\epsilon}_{t-29}$$

where $\widehat{\alpha}_j$, $i = 0, 1$ is depicted in Figure 4.8. The plot not only shows the association between the number of total hospital admission and the levels of air pollutants but also confirms that the pollution effect varies over time. In addition, the estimate $\alpha_j(\cdot)$ is much smoother when the correlation of errors is accounted.

4.3 Proofs

To make the argument concise, denote $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)^T$ with $\mathbf{f}_t = (\epsilon_{t-1}, \ldots, \epsilon_{t-d})^T$, and $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_n)^T$ with $\mathbf{e}_t = (\widehat{\epsilon}_{t-1}, \ldots, \widehat{\epsilon}_{t-d})^T$, where $\widehat{\epsilon}_t$ is the estimated residual in the initial step when the profile least square method is implemented. Define $\boldsymbol{\Delta} = \mathbf{E} - \mathbf{F}$. Our proof follows the same strategy as that in Fan and Huang (2005). The following conditions are imposed to facilitate the proof and are adopted from Fan and Huang (2005). They are not the weakest possible conditions.

A. The random variable u_t has a bounded support Ω. Its density function $g(.)$ is Lipschitz continuous and bounded away from 0 on its support.

B. There is an $s > 2$ such that $E||\mathbf{f}_t||^s < \infty$ and $E||\mathbf{X}_t||^s < \infty$ and for some $\xi > 0$ such that $n^{1-2s^{-1}-2\xi}h \to \infty$.

C. $\{\alpha_i(\cdot), i = 0, \ldots, p\}$ have continuous second derivatives in $u \in \Omega$.

D. The function $K(\cdot)$ is a bounded symmetric density function with bounded support $[-M, M]$, satisfying a Lipschitz condition.

E. $nh^8 \to 0$ and $nh^2/(\log n)^2 \to \infty$.

F. $\sup_{u_t \in \Omega} |\widehat{\alpha}_i^I(u_t) - \alpha_i(u_t)| = o_p(n^{-\frac{1}{4}})$ for all $i = 0, \ldots, p$, where $\widehat{\alpha}_i^I(u_t)$ is the local linear estimator pretending that data are i.i.d..

G. The sequence of random vector $(u_t, \mathbf{X}_t^T, \epsilon_t)$, $t = 1, 2, \cdots$, is strictly stationary and satisfies the following conditions for α-mixing processes:

$$\sum_l l^a [\alpha(l)]^{1-2/\delta} < \infty, \quad E|\epsilon_1|^\delta < \infty, \quad E|\mathbf{X}_1\mathbf{X}_1^T|^\delta < \infty$$

$$g_{u_1|\epsilon_1}(u|\epsilon) \leq A_1 < \infty \quad g_{u_1|\mathbf{X}_1}(u|\mathbf{X}) \leq A_2 < \infty$$

with some $\delta > 2$ and $a > 1 - 2/\delta$,

The definition of a process being strictly stationary and α-mixing can be found in Section 3.3.1.

Lemma 6.1 of Fan and Yao (2003) has been mentioned and repeatedly used in the proof section of Chapter 3. In this chapter, we adopt this lemma to apply to (u_i, ϵ_i).

Lemma 4.3.1. *Let $(u_1, \epsilon_1), \ldots, (u_n, \epsilon_n)$ be a strictly stationary sequence satisfying the mixing condition $\alpha(l) \leq cl^{-\tau}$ for some $c > 0$ and $\tau > 5/2$. Assume further that for some $s > 2$ and interval $[a, b]$,*

$$E|\epsilon_t|^s < \infty \quad \textit{and} \quad \sup_{\forall x \in [a,b]} \int |\epsilon_t|^s g(u, \epsilon) \mathrm{d}\epsilon < \infty,$$

where g denotes the joint density of (u_t, ϵ_t).

In addition, Condition G holds, and the conditional density $g_{u_1,u_l|\epsilon_1,\epsilon_l}(u_1, u_l|\epsilon_1, \epsilon_l) \leq A_2 < \infty, \forall l \geq 1$. Let K satisfy Condition D. Then

$$\sup_{u \in [a,b]} |\frac{1}{n} \sum_{i=1}^{n} \{K_h(u_i - u)\epsilon_i - E[K_h(u_i - u)\epsilon_i]\}| = O_p(\left\{ \frac{\log n}{nh} \right\}^{1/2})$$

provided that $h \to 0$, for some $\xi > 0, n^{1-2s^{-1}-2\xi} h \to \infty$ and $n^{(\tau+1.5)(s^{-1}+\xi)-\frac{\tau}{2}+\frac{5}{4}} h^{-\frac{\tau}{2}-\frac{5}{4}} \to 0$.

Lemma 4.3.2. *Under Conditions A—G, it follows that*

$$\frac{1}{n} \mathbf{F}^T (I - S)^T (I - S) \mathbf{F} \xrightarrow{P} E(\mathbf{f}\mathbf{f}^T).$$

Proof. Denote W_u be a $n \times n$ diagonal matrix with j-th diagonal element $K_h(u_j - u)$

and

$$D_u = \begin{pmatrix} \mathbf{X}_1^T & \frac{u_1-u}{h}\mathbf{X}_1^T \\ \vdots & \vdots \\ \mathbf{X}_n^T & \frac{u_n-u}{h}\mathbf{X}_n^T \end{pmatrix}$$

Then the smoothing matrix $\mathbf{S}$ for the local linear regression can be expressed as

$$\mathbf{S} = \begin{pmatrix} [\mathbf{X}_1^T, 0]\{D_{u_1}^T W_{u_1} D_{u_1}\}^{-1} D_{u_1}^T W_{u_1} \\ \vdots \\ [\mathbf{X}_n^T, 0]\{D_{u_n}^T W_{u_n} D_{u_n}\}^{-1} D_{u_n}^T W_{u_n} \end{pmatrix}$$

where

$$D_u^T W_u D_u = \begin{pmatrix} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T K_h(u_i - u) & \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T \frac{u_i-u}{h} K_h(u_i - u) \\ \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T \frac{u_i-u}{h} K_h(u_i - u) & \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T (\frac{u_i-u}{h})^2 K_h(u_i - u) \end{pmatrix}$$

Each element of matrix $D_u^T W_u D_u$ is a kernel regression. With Lemma 4.3.1 and the symmetry of kernel function $K(\cdot)$, it follows that

$$\frac{1}{n} D_u^T W_u D_u = \begin{pmatrix} f(u)E(\mathbf{X}\mathbf{X}^T|u)(1 + O_p(h^2 + \sqrt{\frac{\log n}{nh}})) & O_p(h + \sqrt{\frac{\log n}{nh}}) \\ O_p(h + \sqrt{\frac{\log n}{nh}}) & f(u)\mu_2 E(\mathbf{X}\mathbf{X}^T|u)(1 + O_p(h^2 + \sqrt{\frac{\log n}{nh}})) \end{pmatrix}$$

holds uniformly in u.

Since $h + \sqrt{\log n/nh} = o_p(1)$, we can regard the above matrix as being approximately diagonal. Then its inverse is

$$\begin{aligned} &\{\frac{1}{n} D_u^T W_u D_u\}^{-1} \\ &= \begin{pmatrix} [f(u)E(\mathbf{X}\mathbf{X}^T|u)]^{-1}(1 + O_p(h^2 + \sqrt{\frac{\log n}{nh}})) & O_p(h + \sqrt{\frac{\log n}{nh}}) \\ O_p(h + \sqrt{\frac{\log n}{nh}}) & [f(u)\mu_2 E(\mathbf{X}\mathbf{X}^T|u)]^{-1}(1 + O_p(h^2 + \sqrt{\frac{\log n}{nh}})) \end{pmatrix} \end{aligned}$$

holds uniformly in u.

By the independence assumption of $(u_t, \mathbf{X}_t^T)$ and ϵ_t, we get $E(\mathbf{X}_t \mathbf{f}_t^T | u) = 0$. Following a similar argument, we have

$$\frac{1}{n} D_u^T W_u \mathbf{F} = \begin{pmatrix} O_p(h^2 + \sqrt{\frac{\log n}{nh}}) \\ O_p(h + \sqrt{\frac{\log n}{nh}}) \end{pmatrix}$$

holds uniformly in u.

Consequently,

$$\begin{aligned}
&[\mathbf{X}^T, 0]\{\frac{1}{n} D_u^T W_u D_u\}^{-1}\{\frac{1}{n} D_u^T W_u \mathbf{F}\} \\
&= [\mathbf{x}^T, 0] \begin{pmatrix} [f(u)E(\mathbf{X}\mathbf{X}^T|u)]^{-1}(1 + O_p(h^2 + \sqrt{\frac{\log n}{nh}})) & O_p(h + \sqrt{\frac{\log n}{nh}}) \\ O_p(h + \sqrt{\frac{\log n}{nh}}) & [f(u)\mu_2 E(\mathbf{X}\mathbf{X}^T|u)]^{-1}(1 + O_p(h^2 + \sqrt{\frac{\log n}{nh}})) \end{pmatrix} \\
&\begin{pmatrix} O_p(h^2 + \sqrt{\frac{\log n}{nh}}) \\ O_p(h + \sqrt{\frac{\log n}{nh}}) \end{pmatrix} \\
&= \mathbf{X}^T [f(x)E(\mathbf{X}\mathbf{X}^T|u)]^{-1} O_p(h^2 + \sqrt{\frac{\log n}{nh}})(1 + o_p(1)) = o_p(1)
\end{aligned}$$

Substituting this result into the smoothing matrix S, we have

$$S\mathbf{F} = \begin{pmatrix} [\mathbf{X}_1^T, 0]\{D_{u_1}^T W_{u_1} D_{u_1}\}^{-1} D_{u_1}^T W_{u_1} \mathbf{F} \\ \vdots \\ [\mathbf{X}_n^T, 0]\{D_{u_n}^T W_{u_n} D_{u_n}\}^{-1} D_{u_n}^T W_{u_n} \mathbf{F} \end{pmatrix} = \begin{pmatrix} o_p(1) \\ \vdots \\ o_p(1) \end{pmatrix}.$$

Thus,

$$\mathbf{F} - S\mathbf{F} = \mathbf{F}\{1 + o_P(1)\}.$$

Finally, by the WLLN,

$$\frac{1}{n}\mathbf{F}^T(I-S)^T(I-S)\mathbf{F} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{f}_i\mathbf{f}_i^T\right)\{1+o_p(1)\}^2 \xrightarrow{P} E(\mathbf{f}\mathbf{f}^T)$$

□

Lemma 4.3.3. *Under Conditions A—G, we have*

$$\frac{1}{n}\mathbf{E}^T(I-S)^T(I-S)\mathbf{E} \xrightarrow{P} E(\mathbf{f}\mathbf{f}^T).$$

Proof. Since $\mathbf{\Delta} = \mathbf{E} - \mathbf{F}$, the generic element of $\mathbf{\Delta}$ is of the form $\sum_{i=0}^{p}[\widehat{\alpha}_i^I(u_t)x_{ti} - \alpha_i(u_t)x_{ti}]$. By condition F: $\sup_{u\in\Omega}|\widehat{\alpha}_i^T(u_t) - \alpha_i(u_t)| = o_p(n^{-\frac{1}{4}})$ and the assumption that x_{ti} is bounded, $\mathbf{\Delta}$ is of order $o_P(n^{-1/4})$ uniformly in u. We observe that

$$\frac{1}{n}\mathbf{E}^T(I-S)^T(I-S)\mathbf{E} = \frac{1}{n}(\mathbf{F}+\mathbf{\Delta})^T(I-S)^T(I-S)(\mathbf{F}+\mathbf{\Delta})$$

By using an argument similar to the proof of Lemma 4.3.2, it can be shown that

$$\frac{1}{n}\mathbf{E}^T(I-S)^T(I-S)\mathbf{E} = \frac{1}{n}\mathbf{F}^T(I-S)^T(I-S)\mathbf{F} + o_P(1)$$

Thus, Lemma 4.3.3 follows by Lemma 4.3.2.

□

Lemma 4.3.4. *Suppose Conditions A—G hold. It follows*

$$\frac{1}{\sqrt{n}}\mathbf{F}^T(I-\mathbf{S})^T(I-\mathbf{S})\mathbf{M} = o_p(1)$$

Proof. It is noted that

$$\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\mathbf{M}$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathbf{f}_i-(S\mathbf{f})_i][\mathbf{X}_i^T\boldsymbol{\alpha}(u_i)-[\mathbf{X}_i^T,0]\{D_{u_i}^TW_{u_i}D_{u_i}\}^{-1}D_{u_i}^TW_{u_i}\mathbf{M}] \quad (4.14)$$

Similar to the argument in the proof of Lemma 4.3.2, we can show that

$$[\mathbf{X}^T,0]\{\frac{1}{n}D_u^TW_uD_u\}^{-1}\{\frac{1}{n}D_u^TW_u\mathbf{M}\}=\mathbf{X}^T\boldsymbol{\alpha}(u)(1+O_p(h^2+\sqrt{\log n/nh}))$$

holds uniformly in $u\in\Omega$. Plugging this in (4.14), it follows that

$$\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\mathbf{M}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathbf{f}_i-(S\mathbf{f})_i][\mathbf{X}_i^T\boldsymbol{\alpha}(u_i)-\mathbf{X}_i^T\boldsymbol{\alpha}(u_i)(1+O_p(h^2+\sqrt{\frac{\log n}{nh}}))]$$
$$=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{f}_i\mathbf{X}_i^T\boldsymbol{\alpha}(u_i)[1+o_p(1)]O_p(h^2+\sqrt{\frac{\log n}{nh}})$$

Note that $E\{\mathbf{f}_i\mathbf{X}_i^T\boldsymbol{\alpha}(u_i)\}=0$ because of the independence between $\mathbf{f}_i$ and $(u_i,\mathbf{X}_i^T)$, and covariance matrix for $\{\mathbf{f}_i\mathbf{X}_i^T\boldsymbol{\alpha}(u_i)\}$ is finite. Thus, using $R=E(R)+O_p(\sqrt{\mathrm{Var}(R)})$, it follows that $\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\mathbf{M}=o_p(1)$. □

Lemma 4.3.5. *Under Conditions A—G, we have*

$$\frac{1}{\sqrt{n}}\mathbf{E}^T(I-S)^T(I-S)\mathbf{M}=o_p(1)$$

Proof. Since $\mathbf{E}=\mathbf{F}+\boldsymbol{\Delta}$, we can break $\frac{1}{\sqrt{n}}\mathbf{E}^T(I-S)^T(I-S)\mathbf{M}$ into two terms: $\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\mathbf{M}$, which is $o_P(1)$ by Lemma 4.3.4, and $\frac{1}{\sqrt{n}}\boldsymbol{\Delta}^T(I-S)^T(I-S)\mathbf{M}$, which is also $o_P(1)$ as $\boldsymbol{\Delta}=o_P(n^{-1/4})$. □

Lemma 4.3.6. *Suppose that Conditions A—G hold. We have*

$$\frac{1}{\sqrt{n}}\mathbf{E}^T(I-S)^T(I-S)\boldsymbol{\Delta}\boldsymbol{\beta}=o_p(1)$$

Proof. This is a direct result from the proof of Lemma 4.3.5. □

Lemma 4.3.7. *Under Conditions A—G, let* $\eta = (\eta_1, \cdots, \eta_n)^T$. *Then*

$$\sqrt{n}[\mathbf{F}^T(I-S)^T(I-S)\mathbf{F}]^{-1}\mathbf{F}^T(I-S)^T(I-S)\eta \to N(0, \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}^{-1})$$

Proof. We observe that

$$\mathbf{F}^T(I-S)^T(I-S)\eta = \sum_{i=1}^{n} \mathbf{f}_i[\eta_i - [\mathbf{X}_i^T, 0]\{D_{u_i}^T W_{u_i} D_{u_i}\}^{-1} D_{u_i}^T W_{u_i}\eta][1+o_p(1)] \quad (4.15)$$

By using Lemma 4.3.1 on $\{u_i, \eta_i\}$, we can show that

$$\begin{aligned}
&[\mathbf{X}^T, 0]\{\frac{1}{n}D_u^T W_u D_u\}^{-1}\{\frac{1}{n}D_u^T W_u \eta\} \\
&= [\mathbf{X}^T, 0] \begin{pmatrix} [f(u)E(\mathbf{X}\mathbf{X}^T|u)]^{-1}(1+O_p(h^2+\sqrt{\frac{\log n}{nh}})) & O_p(h+\sqrt{\frac{\log n}{nh}}) \\ O_p(h+\sqrt{\frac{\log n}{nh}}) & [f(u)\mu_2 E(\mathbf{X}\mathbf{X}^T|u)]^{-1}(1+O_p(h^2+\sqrt{\frac{\log n}{nh}})) \end{pmatrix} \\
&\begin{pmatrix} O_p(h^2+\sqrt{\frac{\log n}{nh}}) \\ O_p(h+\sqrt{\frac{\log n}{nh}}) \end{pmatrix} = o_p(1)
\end{aligned}$$

Then $\eta_i - [\mathbf{X}_i^T, 0]\{D_{u_i}^T W_{u_i} D_{u_i}\}^{-1} D_{u_i}^T W_{u_i}\eta = \eta_i\{1+o_p(1)\}$. Plugging this in (4.15), we obtain that

$$\mathbf{F}^T(I-S)^T(I-S)\eta = \sum_{i=1}^{n} \mathbf{f}_i\eta_i\{1+o_p(1)\}$$

Since $E(\mathbf{f}_i\eta_i) = 0$, $\mathrm{Var}(\mathbf{f}_i\eta_i) = \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\} < \infty$, and $E(\mathbf{f}_i\eta_i\mathbf{f}_j\eta_j) = 0$ for $i \neq j$ since η_i is independent of $\mathbf{f}_i$. By Central Limit Theorem for strictly stationary sequence (see Theorem 2.21 of Fan and Yao, 2003),

$$\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\eta \xrightarrow{L} N(0, \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}).$$

By Lemma 4.3.2, $\frac{1}{n}\mathbf{F}^T(I-S)^T(I-S)\mathbf{F} \xrightarrow{P} E(\mathbf{f}\mathbf{f}^T)$. Applying Slutsky's theorem, it

follows that

$$\sqrt{n}[\mathbf{F}^T(I-S)^T(I-S)\mathbf{F}])^{-1}\mathbf{F}^T(I-S)^T(I-S)\eta \xrightarrow{L} N(0,\sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}^{-1}).$$

□

Lemma 4.3.8. *Under Conditions A—G, we have*

$$\sqrt{n}[\mathbf{E}^T(I-S)^T(I-S)\mathbf{E}])^{-1}\mathbf{E}^T(I-S)^T(I-S)\eta \xrightarrow{L} N(0,\sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}^{-1})$$

Proof. Since $\mathbf{E}=\mathbf{F}+\mathbf{\Delta}$, we may write $\mathbf{E}^T(I-S)^T(I-S)\eta = \mathbf{F}^T(I-S)^T(I-S)\eta + \mathbf{\Delta}^T(I-S)^T(I-S)\eta$. Note that $\mathbf{\Delta} = o_P(n^{-1/4})$ by Condition F, it can be shown that

$$\frac{1}{\sqrt{n}}\Delta^T(I-S)^T(I-S)\eta = o_p(1).$$

Furthermore, we have shown in the last lemma that $\frac{1}{\sqrt{n}}\mathbf{F}^T(I-S)^T(I-S)\eta \to N(0,\sigma^2E(\mathbf{f}\mathbf{f}^T))$. So $\frac{1}{\sqrt{n}}\mathbf{E}^T(I-S)^T(I-S)\eta \to N(0,\sigma^2E(\mathbf{f}\mathbf{f}^T))$ as well. The proof is completed by the Slutsky theorem and Lemma 4.3.3. □

Proof of Theorem 2

Let us first show the asymptotic normality of $\widehat{\boldsymbol{\beta}}$. According to the expression in $\widehat{\boldsymbol{\beta}}$ in (4.6), we can break $\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ into the sum of the following three terms (a), (b) and (c)

$$(a)\widehat{=}\sqrt{n}[\{\mathbf{E}^T(I-S)^T(I-S)\mathbf{E}\}^{-1}\mathbf{E}^T(I-S)^T(I-S)\mathbf{M}]$$
$$(b)\widehat{=}\sqrt{n}[\{\mathbf{E}^T(I-S)^T(I-S)\mathbf{E}\}^{-1}\mathbf{E}^T(I-S)^T(I-S)\mathbf{\Delta}\beta]$$
$$(c)\widehat{=}\sqrt{n}[\{\mathbf{E}^T(I-S)^T(I-S)\mathbf{E}\}^{-1}\mathbf{E}^T(I-S)^T(I-S)\eta]$$

Term (a) is a product of $[\frac{\{\mathbf{E}^T(I-S)^T(I-S)\mathbf{E}\}}{n}]^{-1}$ and $[\frac{\mathbf{E}^T(I-S)^T(I-S)\mathbf{M}}{\sqrt{n}}]$. From Lemmas 4.3.3 and 4.3.5, the asymptotic properties of these two terms lead to the conclusion that $(a) = o_p(1)$. Similarly, applying Lemmas 4.3.3 and 4.3.6 on two product components of term (b) results in $(b) = o_p(1)$ as well. In addition, Lemma 4.3.8 states that term (c) converges to $N(0, \sigma^2\{E(\mathbf{f}\mathbf{f}^T)\}^{-1})$. Put three terms together and we get the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$.

Next we derive the asymptotic bias and variance of $\widehat{\alpha}_i(\cdot)$. By (4.7) and the arguments in Lemma 4.3.1— 4.3.8, we have

$$\widehat{\alpha}_i(u_0, \widehat{\boldsymbol{\beta}}) = e_{i+1}^T\{D_{u_0}^T W_{u_0} D_{u_0}\}^{-1} D_{u_0}^T W_{u_0}(\mathbf{y} - \mathbf{E}\widehat{\boldsymbol{\beta}}),\ i = 0, \ldots, p$$

where e_{i+1}^T is a $2(p+1) \times 1$ vector consisting of 0's except 1 at the $(i+1)^{\text{th}}$ element. By the matrix format of semiparametric varying-coefficient partially linear model (4.3) and $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}|| = O_p(n^{-\frac{1}{2}})$, it follows

$$\widehat{\alpha}_i(u_0, \widehat{\boldsymbol{\beta}}) = e_{i+1}^T\{D_{u_0}^T W_{u_0} D_{u_0}\}^{-1} D_{u_0}^T W_{u_0}(\mathbf{M} + \eta)\{1 + o_P(1)\},\ i = 0, \ldots, p$$

Note that $E(\eta|\mathcal{U}, \mathcal{X}) = 0$, where $\mathcal{U} = (u_1, \cdots, u_n)$ and $\mathcal{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)$. Thus,

$$E\{\widehat{\alpha}_i(u_0, \widehat{\boldsymbol{\beta}})|\mathcal{U}, \mathcal{X}\} = e_{i+1}^T\{D_{u_0}^T W_{u_0} D_{u_0}\}^{-1} D_{u_0}^T W_{u_0}\mathbf{M}\{1 + o_p(1)\}$$

Since we use a linear function in the neighborhood of u_0 to approximate $\alpha_i(u_t)$, the leading residual term should be $\frac{1}{2}\alpha_i''(u_0)(u_t - u_0)^2$. Therefore

$$E\{\widehat{\alpha}_i(u_0, \widehat{\boldsymbol{\beta}})|\mathcal{U}, \mathcal{X}\} - \alpha_i(u_0, \widehat{\boldsymbol{\beta}}) = e_{i+1}^T\{D_{u_0}^T W_{u_0} D_{u_0}\}^{-1} D_{u_0}^T W_{u_0} H_{u_0} \begin{pmatrix} \frac{1}{2}\alpha_0''(u_0) \\ \vdots \\ \frac{1}{2}\alpha_p''(u_0) \end{pmatrix} (1 + o_p(1))$$

where $H_{u_0} = \begin{pmatrix} (u_1 - u_0)^2 \mathbf{X}_1^T \\ \vdots \\ (u_n - u_0)^2 \mathbf{X}_n^T \end{pmatrix}$.

Following an argument similar to Lemma 4.3.2, we can derive

$$\{\frac{1}{n} D_{u_0}^T W_{u_0} D_{u_0}\}^{-1} \frac{1}{n} D_{u_0}^T W_{u_0} H_{u_0} = h^2 \mu_2 (1 + o_p(1))$$

Hence, the asymptotic bias of $\widehat{\alpha}_i(u_0, \widehat{\boldsymbol{\beta}})$ is $\frac{1}{2}\alpha_i''(u_0)h^2\mu_2$.

Regarding to the asymptotic variance of $\widehat{\alpha}_i(\cdot)$,

$$\mathrm{Var}[\widehat{\alpha}_i(u_0, \widehat{\boldsymbol{\beta}})|\mathcal{U}, \mathcal{X}] = e_{i+1}^T \{D_{u_0}^T W_{u_0} D_{u_0}\}^{-1} D_{u_0}^T W_{u_0} \mathrm{Var}\{\eta\} W_{u_0} D_{u_0} \{D_{u_0}^T W_{u_0} D_{u_0}\}^{-1} e_{i+1}$$

Using the same argument as that in the proof of Lemma 4.3.2, we have

$$\mathrm{Var}[\widehat{\alpha}_i(u_0, \widehat{\boldsymbol{\beta}})|\mathcal{U}, \mathcal{X}] = \frac{\sigma^2}{nhg(u_0)} \int K^2(u)\, \mathrm{d}u$$

As to the asymptotic normality,

$$\widehat{\alpha}_i(u_0, \widehat{\boldsymbol{\beta}}) - E\{\widehat{\alpha}_i(u_0, \widehat{\boldsymbol{\beta}})|\mathcal{U}, \mathcal{X}\} = e_{i+1}^T \{D_{u_0}^T W_{u_0} D_{u_0}\}^{-1} D_{u_0}^T W_{u_0} \eta \{1 + o_P(1)\}$$

Since η_i are independently and identically distributed with mean zero and variance σ^2, the asymptotic normality of $\widehat{\alpha}_i(u_0, \widehat{\boldsymbol{\beta}})$ conditioning on $\mathcal{U}$ and $\mathcal{X}$ can be established as shown in Theorem 2 (b).

Chapter 5

Some extensions and future research directions

In this chapter, we first propose a natural extension of the proposed methodology in Chapters 3 by considering nonparametric regression models with multiple response in Section 5.1. Monte Carlo simulation studies under various autocorrelation situations are conducted to show the finite sample performance of the proposed method. In Section 5.2, we discuss other possible extensions. In Section 5.3, we point out some future research directions.

5.1 Nonparametric regression with multiple responses

In Chapter 3, we proposed the profile least squares method to incorporate the correlation information of the errors into estimation for nonparametric regression. The proposed method outperforms local linear regression without taking the correlated errors into account with respect to the accuracy of mean function estimation. It

is natural to extend the methodology to the multivariate context. Intuitively, simultaneously modeling all response variables can improve the efficiency by putting relevant variables together in an integrated system rather than dealing with them one by one.

As a natural extension of the model studied in Chapter 3, we consider the nonparametric model with K components,

$$\underline{y}_t = \underline{m}(x_t) + \underline{\epsilon}_t \quad t = 1, 2, \ldots, n \tag{5.1}$$

where $\underline{y} = (y_{t,1}, \ldots, y_{t,K})$, $\underline{m}(\cdot) = (m_1(\cdot), \ldots, m_K(\cdot))$ and $\{m_i(\cdot), i = 1, \ldots, K\}$ is a smooth function. In the multivariate setting, $\{\underline{\epsilon}_t\}$ is a vector form of the autoregressive error with order p. We can represent as $\underline{\epsilon}_t = (\epsilon_{t,1}, \ldots, \epsilon_{t,K})$ by

$$\underline{\epsilon}_t = \underline{\epsilon}_{t-1}\Phi_1 + \cdots + \underline{\epsilon}_{t-p}\Phi_p + \underline{\eta}_t \tag{5.2}$$

where Φ_j's are the $K \times K$ autoregressive coefficient matrices and $\underline{\eta}_t$ follows a multivariate normal distribution with mean vector $\underline{0}$ and covariance $\Sigma(> 0)$. The distinction between the vector autoregression (VAR) and the univariate autoregression (AR) is that the coefficient matrices of the VAR allow correlation across different components. The VAR model is a more complicated time series than the AR process.

Motivated by the profile least squares methodology in univariate nonparametric regression, we will adapt an implementable estimation procedure for model (5.1). Simulation results are presented to compare our proposed method with the marginal profile least squares method.

5.1.1 Estimation Procedure

For the nonparametric model (5.1), two designs for x_t are studied in the statistical literature: fixed or random. For the fixed design, $x_t = t/n$ and $x_t = (t-0.5)/n$, the regular bandwidth selection for the nonparametric regression fails and the adjusted bandwidth selection mechanisms are proposed to account for the correlation (see Altman, 1990; Hart 1991; Opsomer *et. al.*, 2001). In this chapter, we consider the situation x_t generated from a random non-degenerate distribution. Under this design, the Ruppert, Sheather and Wand (1995) direct plug-in bandwidth still works well, as the simulation results show in Chapter 3.

Substituting $\underline{\epsilon}_t$ with its autoregressive expression (5.2), model (5.1) becomes a multivariate partially linear model as follows:

$$\underline{y}_t = \underline{m}(x_t) + \underline{\epsilon}_{t-1}\Phi_1 + \cdots + \underline{\epsilon}_{t-p}\Phi_p + \underline{\eta}_t, \quad t = p+1, \ldots, n$$

Since $\underline{\epsilon}_t$ is usually not available in practice, we replace it with $\widehat{\underline{\epsilon}}_t = \underline{y}_t - \underline{\tilde{m}}(x_t)$, where $\underline{\tilde{m}}(\cdot)$ is the local linear estimate of $\underline{m}(\cdot)$ pretending that data are i.i.d. Specifically, we employ the univariate local linear regression on each component of data, i.e., $\{x_t, y_{t,k}, (1 \le k \le K)\}$ for $\tilde{m}_k(\cdot)$ and then the vector $\underline{\tilde{m}}(x_t)$ is constructed by stacking $(\tilde{m}_1(x_t), \ldots, \tilde{m}_K(x_t))$ together. Now we can obtain

$$\underline{y}_t = \underline{m}(x_t) + \widehat{\underline{\epsilon}}_{t-1}\Phi_1 + \cdots + \widehat{\underline{\epsilon}}_{t-p}\Phi_p + \underline{\eta}_t, \quad t = p+1, \ldots, n \tag{5.3}$$

Formulating in the matrix format, denote

$$\mathbf{Y} = \begin{pmatrix} \underline{y}_{p+1} \\ \vdots \\ \underline{y}_n \end{pmatrix} = \begin{pmatrix} y_{p+1,1} & \cdots & y_{p+1,K} \\ \vdots & & \vdots \\ y_{n,1} & \cdots & y_{n,K} \end{pmatrix}$$

$$\mathbf{M} = \begin{pmatrix} \underline{m}(x_{p+1}) \\ \vdots \\ \underline{m}(x_n) \end{pmatrix} = \begin{pmatrix} m_1(x_{p+1}) & \ldots & m_K(x_{p+1}) \\ \vdots & & \vdots \\ m_1(x_n) & \ldots & m_K(x_n) \end{pmatrix}$$

$$\mathbf{E} = \begin{pmatrix} \widehat{\underline{\epsilon}}_p & \cdots & \widehat{\underline{\epsilon}}_1 \\ \vdots & & \vdots \\ \widehat{\underline{\epsilon}}_n & \cdots & \widehat{\underline{\epsilon}}_{n-p+1} \end{pmatrix} = \begin{pmatrix} \widehat{\epsilon}_{p,1} & \cdots & \widehat{\epsilon}_{p,K} | & \cdots & |\widehat{\epsilon}_{1,1} & \cdots & \widehat{\epsilon}_{1,K} \\ \vdots & & \vdots & & \vdots & & \vdots \\ \widehat{\epsilon}_{n,1} & \cdots & \widehat{\epsilon}_{n,K} | & \cdots & |\widehat{\epsilon}_{n-p+1,1} & \cdots & \widehat{\epsilon}_{n-p+1,K} \end{pmatrix}$$

$$\mathbf{V} = \begin{pmatrix} \underline{\eta}_{p+1} \\ \vdots \\ \underline{\eta}_n \end{pmatrix} = \begin{pmatrix} \eta_{p+1,1} & \cdots & \eta_{p+1,K} \\ \vdots & & \vdots \\ \eta_{n,1} & \cdots & \eta_{n,K} \end{pmatrix}$$

$$\mathbf{\Phi} = \begin{pmatrix} \Phi_1 \\ \vdots \\ \Phi_p \end{pmatrix} = \begin{pmatrix} \Phi_{11}^{(1)} & \ldots & \Phi_{1K}^{(1)} \\ \vdots & \vdots & \vdots \\ \Phi_{K1}^{(1)} & \ldots & \Phi_{KK}^{(1)} \\ \vdots & \vdots & \vdots \\ \Phi_{11}^{(p)} & \ldots & \Phi_{1K}^{(p)} \\ \vdots & \vdots & \vdots \\ \Phi_{K1}^{(p)} & \ldots & \Phi_{KK}^{(p)} \end{pmatrix}$$

Then, we have

$$\mathbf{Y} = \mathbf{M} + \mathbf{E}\mathbf{\Phi} + \mathbf{V}. \tag{5.4}$$

For the k^{th} element ($1 \leq k \leq K$), equation (5.4) is just a univariate partially linear

model as below

$$\mathbf{Y}_{\cdot,k} = \mathbf{M}_{\cdot,k} + \mathbf{E}\mathbf{\Phi}_{\cdot,k} + \mathbf{V}_{\cdot,k} \tag{5.5}$$

We can apply the profile least squares method described in Chapter 3 to (5.5) to get an estimate of $\mathbf{M}_{\cdot,k}$ and $\mathbf{\Phi}_{\cdot,k}$. By repeating this procedure through 1 to K and stacking the univariate estimates together, we can obtain the profile estimate of $\mathbf{M}$ and $\Phi_1, \ldots, \Phi_p$. Although the univariate profile least squares method is used, considering the AR structure in a vector structure allows us to take advantage of the correlation information from the off-diagonal elements in Φ_i. Therefore the multivariate profile method will be more effective for retrieving the VAR coefficient matrix than the marginal profile method. This feature will be demonstrated in the simulation studies. In addition, since we apply the profile least squares method on each component, we can use different bandwidths according to the degree of smoothness of each component. The implementation does not involve any extra difficulty.

5.1.2 Simulation Results

In this section, we compare the multivariate profile least squares method and the marginal profile least squares method with the local linear method ignoring the correlation structure of errors. The comparisons are conducted via Monte Carlo simulations with respect to the mean square errors defined by

$$\text{MSE}\{\underline{\widehat{m}}(\cdot)\} = \frac{1}{n}\sum_{t=1}^{n}\sum_{k=1}^{K}\{m_k(x_t) - \widehat{m}_k(x_t)\}^2$$

The ratios of the MSE (RMSE) of the multivariate profile least squares method to the local linear method and the marginal profile least squares method to the local

linear method are recorded respectively. We summarize the percentage of accuracy gain $(1 - RMSE) * 100\%$ in the simulation tables.

A random sample of size n, either $n = 100$ or $n = 500$, is generated from a 2-component model:

$$\underline{y}_t = \underline{m}(x_t) + \underline{\epsilon}_t$$

where $m_1(x_t) = 8\cos(2\pi x_t)$ and $m_2(x_t) = \exp(-(2x_t - 1)^2)$.

The error $\{\underline{\epsilon}_t\}$ is generated by a vector autoregressive process with order p ($p = 1$ or 2). When $p = 1$, it is a VAR(1) model defined by

$$\underline{\epsilon}_t = \underline{\epsilon}_{t-1}\Phi_1 + \underline{\eta}_t$$

In our simulation, we consider seven scenarios for Φ_1.

(A) $\Phi_1 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$;

(B) $\Phi_1 = \begin{pmatrix} 0.7 & 0 \\ 0 & 0.7 \end{pmatrix}$;

(C) $\Phi_1 = \begin{pmatrix} 0 & 0.5 \\ 0.5 & 0 \end{pmatrix}$;

(D) $\Phi_1 = \begin{pmatrix} 0 & 0.7 \\ 0.7 & 0 \end{pmatrix}$;

(E) $\Phi_1 = \begin{pmatrix} 0.5 & 0.4 \\ 0.3 & 0.5 \end{pmatrix}$;

(F) $\Phi_1 = \begin{pmatrix} 0.2 & 0.3 \\ 0.7 & 0.5 \end{pmatrix}$;

(G) $\Phi_1 = \begin{pmatrix} 0.2 & -0.3 \\ -0.7 & 0.5 \end{pmatrix}$

For $p = 2$, it is a VAR(2) model defined by

$$\underline{\epsilon}_t = \underline{\epsilon}_{t-1}\Phi_1 + \underline{\epsilon}_{t-2}\Phi_2 + \underline{\eta}_t$$

In our simulation, we consider three cases:

(H) $\Phi_1 = \begin{pmatrix} 0.4 & 0.1 \\ 0.1 & 0.5 \end{pmatrix}, \Phi_2 = \begin{pmatrix} 0.3 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}$;

(I) $\Phi_1 = \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.4 \end{pmatrix}, \Phi_2 = \begin{pmatrix} 0.1 & 0.2 \\ 0.2 & 0.1 \end{pmatrix}$;

(J) $\Phi_1 = \begin{pmatrix} 0.3 & -0.2 \\ -0.5 & 0.4 \end{pmatrix}, \Phi_2 = \begin{pmatrix} 0.1 & 0.2 \\ 0.2 & 0.3 \end{pmatrix}$.

In both situations, the white noise vector $\underline{\eta}_t$ follows a multivariate normal distribution with mean vector $\underline{0}$ and covariance Σ either being $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ or being $\begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix}$. In the simulation tables, we use "Independent" and "Dependent" to represent the two cases in Table 5.1 and 5.2, respectively.

In our simulation, we focus on the random design rather than fixed design for x_t. Two random sampling schemes of $\{x_t\}$ are used to enhance the understanding of the effects of different samplings schemes on model estimations.

Table 5.1. Simulation results for nonparametric models with multiple responses under sampling scheme I. The percentage of accuracy gain $(1 - \text{RMSE}) * 100\%$ is reported.

	Independent				Dependent			
	$n = 100$		$n = 500$		$n = 100$		$n = 500$	
Φ	Marginal	Multi	Marginal	Multi	Marginal	Multi	Marginal	Multi
(A)	8.22	8.62	13.94	13.93	7.93	7.92	12.45	12.46
(B)	12.11	12.11	23.56	23.65	14.33	14.07	17.88	17.97
(C)	-1.45	17.62	0.77	12.12	-0.95	15.5	1.72	10.44
(D)	-1.67	34.06	0.22	23.53	-1.30	37.45	1.61	16.40
(E)	12.19	20.36	9.87	12.00	19.99	27.90	9.22	10.76
(F)	8.12	15.80	11.50	16.68	13.16	21.99	9.12	10.26
(G)	4.15	8.22	29.07	49.15	1.20	5.97	13.24	32.77
(H)	12.05	13.92	8.90	8.99	19.67	21.39	7.23	7.42
(I)	12.40	15.59	8.66	9.17	22.02	25.12	7.29	7.63
(J)	7.42	10.03	15.88	26.76	6.30	8.92	8.57	15.83

I. x_t is independent and identically distributed according to the uniform distribution over $[0, 1]$.

II. u_t is independent and identically distributed according to the standard normal distribution for $t = 1, 2, ...$ Let $x_t = \Phi\{(au_t + bu_{t-1})/\sqrt{a^2 + b^2}\}$ for $t = 2, 3, \cdots$, where $\Phi(u)$ is the cumulative distribution function of the standard normal distribution. Thus, x_t is 1-dependent process. In our simulation, we take $a = 0.9$ and $b = 0.1$.

Each experiment is repeated 500 times. Ruppert, Sheather and Wand (1995) direct plug-in bandwidth selector is used in both local linear regression and profile least squares estimation, as we described in Section 3.1.3.

For sampling scheme I, case (A) and (B) can be regarded as two independent AR processes with order 1, since their off-diagonal elements are 0's. So the performance of the multivariate profile least square estimate should be very close to that of the

Table 5.2. Simulation results for nonparametric models with multiple responses under sampling scheme I. The percentage of accuracy gain $(1 - \text{RMSE}) * 100\%$ is reported.

	Independent				Dependent			
	$n = 100$		$n = 500$		$n = 100$		$n = 500$	
Φ	Marginal	Multi	Marginal	Multi	Marginal	Multi	Marginal	Multi
(A)	7.44	7.35	13.94	13.93	8.23	7.94	12.45	12.46
(B)	13.08	12.99	23.18	23.22	16.04	15.72	17.95	18.00
(C)	-2.70	16.67	0.51	12.55	-2.55	14.10	1.46	10.53
(D)	-2.19	34.22	0.31	23.69	-0.72	39.20	1.81	16.86
(E)	14.31	22.21	9.82	12.17	23.12	31.28	9.46	11.08
(F)	9.11	17.33	11.61	17.43	15.07	24.74	9.36	10.92
(G)	4.40	8.91	28.03	46.87	1.20	6.79	12.87	30.24
(H)	12.98	14.70	8.89	9.12	21.52	22.9	7.31	7.58
(I)	13.60	16.61	8.65	9.35	23.99	26.81	7.35	7.79
(J)	7.83	9.95	15.51	25.18	6.38	8.82	9.50	15.05

marginal profile least square estimate in these two cases. The simulation results confirm this observation. In addition, when the correlation pattern is strong, both methods show larger gain than the local linear regression without considering the correlation structure.

Case (C) and (D) represent the opposing examples of case (A) and (B). In these two cases, the diagonal elements of Φ_1 are 0's, while all correlation information is contained in the off-diagnoal entries. More specifically, $\epsilon_{t,1}$ is correlated with the previous status from the other component $\epsilon_{t-1,2}$ and similar rule applies to $\epsilon_{t,2}$. In the simulation table 5.1, the marginal profile least square turns out a negative gain when $n = 100$ and a trivial gain when $n = 500$. By contrast, the multivariate profile least square method can retrieve the correct VAR coefficient matrix so that it displays a significant improvement in finite samples.

Cases (A)—(D) demonstrate the extreme situation of the possible VAR coefficient matrix. Case (E), (F) and (G) represent the more general scenarios where all

entries contain a certain amount of autocorrelation.

For case (E), the major autocorrelation is from the diagonal elements. In other words, the correlation within each component is dominant over the correlation between the different components. In this case, both the multivariate profile method and the marginal profile method improve the estimation's accuracy, but the former one always outperforms the latter one, especially when the sample size is small.

As to case (F), the dominant correlation for the first component is from the off-diagonal element while the primary correlation for the second component is from the diagonal entry. Therefore the marginal profile least squares estimation should work poorly in one dimension and fairly well in the other one. Although the multivariate profile method is still better than the marginal method, the overall performance of these two methods is worse than that in case (E). In addition, the multivariate method shows a remarkably greater advantage over the marginal one in a moderate-sized sample in case (E) and (F).

The last case for VAR(1), (G), has the negative off-diagonal elements that offset some autocorrelation effect caused by the positive diagonal elements. In other words, the autocorrelation signal in case (G) is weakest among cases (A) to (G). The marginal profile method and the multivariate method show small improvements when $n = 100$. However, the multivariate method exhibits a significant improvement when the sample size increases. This observations suggests that the multivariate method is asymptotically efficient.

Cases (H), (I) and (J) refer to VAR(2) model. For case (H), the primary correlation concentrates in the diagonal elements in both Φ_1 and Φ_2. The multivariate profile method shows a larger improvement than the marginal profile method as expected. This superiority is especially outstanding when the sample size is small.

Regarding to case (I), Φ_1 has the dominant within-component autocorrelation in

both components, while one component of Φ_2 has the dominant within-component correlation and the other component of Φ_2 has the dominant between-component correlation. For the moderate-sized sample, the marginal method yields to a compatible result to its counterpart in case (H) but the multivariate method displays a larger gain. The difference between the marginal profile method and the multivariate profile method becomes less significant in a large-sized sample.

For case (J), Φ_2 is same as that in case (I). But Φ_1 in this case has negative off-diagonal elements that are similar to the setting in case (G). Because the negative coefficients will weaken the autocorrelation, the gain of the marginal and the multivariate methods is less in this case. But both methods are still better than the conventional local linear regression.

The pattern, which is shown in the simulation results under sampling scheme II when the covariate $\{x_t\}$ is from a 1-dependent process, highly agrees with that in sampling scheme I when $\{x_t\}$ is independently distributed. This finding implies that the different random sampling schemes do not affect the results much.

In a summary, the multivariate profile least squares method outperforms the marginal profile least squares method as long as VAR coefficient matrix is not diagonal. This advantage is most significant when the between-component autocorrelation is dominant. If VAR coefficient matrix is diagonal, the multivariate method is just as good as the marginal method.

5.2 Future research directions

In this dissertation, we propose applying profile least squares estimation to the nonparametric regression and the varying-coefficient model with auto-correlated errors. Under regular conditions, our estimator enjoys the same asymptotic properties as

the one obtained by the local linear regression with i.i.d. data. We also extend the profile least squares approach to nonparametric regression with multiple responses and VAR errors. The numerical results show that the multivariate profile method is more efficient than the marginal profile method. The gain is even more remarkable when compared with the local linear regression without taking the error correlation structure into account.

Furthermore, we use the penalized profile least squares method with the SCAD penalty to estimate the coefficients and select the AR order simultaneously in the univariate nonparametric models. The gain of the penalized profile least square method is notably greater than the procedure without order selection.

However, the potential nonparametric regression techniques for correlated data are far beyond what has been presented here. There is still ample room for future research to contribute this area. In this section, we outline some future research directions.

The first possible avenue for future research is to investigate the effect of misspecification of error structure in the nonparametric regression problem. Throughout this dissertation, the errors are assumed from an autoregressive process. In practice, we might not be certain of this assumption.

Secondly, for the varying-coefficient model, we might want to test if the functional coefficients are really varying over different values of the covariates. Fan and Huang (2005) suggested a generalized likelihood ratio test in the semiparametric varying-coefficient model with identical and independent distributed data. Further investigation is needed to test whether a similar test can be adopted and whether Wilks phenomenon holds for the varying-coefficient model with correlated data.

In the previous section of this chapter, we extended the profile least squares method to nonparametric regression with multiple responses and VAR errors. Since

the estimation methodology has been systematically established for the univariate varying-coefficient model with AR errors in Chapter 4, it should not involve extra difficulty for a similar extension to the varying-coefficient model with multiple responses and VAR errors.

The other issue for the multivariate profile least squares estimator is its asymptotic behavior. Whether it follows the property neatly as that in the univariate situation requires more in-depth studies.

Moreover, an VAR order selection algorithm needs to be developed to reduce the estimation bias. In light of the solution we proposed for the univariate nonparametric and varying-coefficient model, we expect that the multivariate penalized profile approach can be adapted. However, we could foresee the challenges caused by the high-dimensional time series.

5.3 Other possible extensions

In this dissertation, we studied the autoregressive errors in the local modelling context that represent a linear relationship between the current status and the previous ones. Beyond the linear domain, a more general model is a functional-coefficient autoregression (FAR) defined as below:

$$\epsilon_t = \beta_1(\boldsymbol{\epsilon}^*_{t-1})\epsilon_{t-1} + \cdots + \beta_p(\boldsymbol{\epsilon}^*_{t-1})\epsilon_{t-p} + \eta_t,$$

where $\boldsymbol{\epsilon}^*_{t-1} = (\epsilon_{t-i_1}, \ldots, \epsilon_{t-i_k})^T$ and $\{\eta_t\}$ is a sequence of i.i.d. random variables.

Chen and Tsay (1993) proposed an iterative algorithm while Cai, Fan and Yao (2000) proposed a local linear estimator for such a FAR model. But they did not combine the FAR errors with the trend function. It is expected that the varying-

coefficient model with the FAR errors will exhibit large flexibility for complicated data.

The other interesting direction is to explore a more general time series structure on errors such as the autoregressive moving average (ARMA) model defined as below:

$$\epsilon_t = \beta_1 \epsilon_{t-1} + \cdots + \beta_p \epsilon_{t-p} + \eta_t + \cdots + \gamma_{t-q} \eta_{t-q},$$

where $\{\eta_t\}$ is i.i.d. $N(0, 1)$. Although Xiao *et. al.* (2003) showed that the pre-whiting method can work for MA errors theoretically, their simulation results did not show a convincing improvement.

The autoregressive conditional heteroscedastic (ARCH) model can also be applied in the local modeling problem. ARCH model is especially useful for the highly volatile data so that it has been used often in finance, such as pricing stocks and options.

Bibliography

[1] Ahmad, I., Leelahanon, S. and Li, Q. (2005). Efficient Estimation of A semiparametric Partially Linear Varying Coefficient Model. *The Annals of Statistics*. **33**, 258–283.

[2] Altman, N. (1990). Kernel Smoothing of Data with Correlated Errors. *Journal of the American Statistical Association*, **85**, 749–759.

[3] Altman, N. (1991). An Iterated Cochrane-Orcutt Procedure for Nonparametric Regression. *Journal of Statistical Computation Simulation*, **40**, 93–108.

[4] Altman, N. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, **46**, 175–185.

[5] Altman, N. and Villarreal, J. (2004). Self-modeling Regression for Longitudinal Data with Time-invariant Covariate. *The Canadian Journal of Statistics*, **32**, 251–268.

[6] Brockwell, P. and Davis, R. (1991). Time Series: Theory and Methods. *Springer-Verlag*, New York.

[7] Brumback, B. and Rice, J. (1998). Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves. *Journal of The American Statistical Association*, **93**, 961–976.

[8] Cai, Z. (2002). A Two-stage Approach to Additive Time Series Models. *Statistica Neerlandica*, **56**, 415–433.

[9] Cai, Z. (2007). Trending Time-Varying Coefficient Time Series Models with Serially Correlated Errors. *Journal of Econometrics*. **136**, 163–188.

[10] Cai, Z., Fan, J. and Li, R. (2000). Effective Estimation and Inferences for Varying-Coefficient Models. *Journal of American Statistical Association*, **95**, 888–902.

[11] Cai, Z., Fan, J. and Yao, Q. (2000). Functional-Coefficient Regression Models for Nonlinear Time Series. *Journal of the American Statistical Association*, **95**, 941–956.

[12] Cai, Z. and Ould-Said, E. (2003). Local M-estimator for Nonparametric Time Series. *Statistics and Probability Letters*. **65**, 433–449

[13] Cai, Z., Yao, Q. and Zhang, W. (2001). Smoothing for Discrete-values Time Series. *Journal of The Royal Statistical Society Series B-Statistical Methodology*. **63**, 357–375.

[14] Carroll, R. J., Lin, X., Linton, O.B. and Mammen, E. (2003). Accounting for Correlation in Marginal Longitudinal Nonparametric Regression. In *Second Seattle Symposium on Biostatistics*. Springer-Verlag, Berlin.

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

[16] Cleveland, W. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

[17] Fahrmeir, L. and Klinger A. (1998). A Nonparametric Multiplicative Hazard Model for Event History Analysis. *Biometrika*. **85**, 581–592.

[18] Fan, J. (2005). A Selective Overview of Nonparametric Methods in Financial Econometrics. *Statistical Science*, **20**, 317–337.

[19] Fan, J. and Gijbels, I. (1996). Local Polynomial Modeling and Its Applications, *Chapman and Hall*, London.

[20] Fan, J. and Huang, T. (2005) Profile Likelihood Inferences on Semiparametric Varying-coefficient Partially Linear Models. *Bernoulli*. **11**, 1031–1059.

[21] Fan, J., Huang, T. and Li, R. (2007) Analysis of Longitudinal Data with Semiparametric Estimation of Covariance Function *Journal of American Statistical Association*, **102**, 632–641.

[22] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of American Statistical Association*, **96**. 1348–1360.

[23] Fan, J. and Li, R. (2004). New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis. *Journal of American Statistical Association*, **99**, 710–723.

[24] Fan, J. and Li, R. (2006). An overview on nonparametric and semiparametric techniques for longitudinal data. In *Frontiers of Statistics* (Fan, J. and Koul, H., eds.). Imperial College Press, 277–304,

[25] Fan, J. and Yao, Q. (2003). Nonlinear Time Series: Nonparametric and Parametric Methods. *Springer-Verlag*, New York.

[26] Fan, J., Yao, Q. and Cai, Z. (2003) Adaptive Varying-ceofficient Linear Models. *Journal of the Royal Statistical Society*, Series B, **65**, 57–80.

[27] Fan, J. and Zhang, W. (1999). Statistical Estimation in Varying Coefficient Models. *the Annals of Statistics*, **27**, 1491–1518.

[28] Fan, J. and Zhang J. (2000). Two-Step Estimation of Functional Linear Models with Applications to Longitudinal Data. *Journal of the Royal Statistical Society*, Series B, **62**, 303–322.

[29] Faraway J. (1997). Regression Analysis for a Functional Response. *Technometrics*. **39**, 254-261.

[30] Fernández, J. and Fernández, M. (2000). Local Polynomial Regression Smoothers with AR-error Structure, *Test*, **11**, 439–464.

[31] Fernández, M. and Fernández, J. (2000). The Polynomial Regression with Estimation with correlated data., *Test*, **11**, 439–464.

[32] Fernández, J. and Fernández, M. (2004). Plug-in Bandwidth Selector for Local Polynomial Regression Estimator with Correlated Errors. *Journal of Nonparametric Statistics*, **16**, 127–151.

[33] Fernández, M. and Opsomer, J. (2004) A Plug-in Bandwidth Selector for Local Polynomial Regression with Correlated Errors. *Journal of Nonparametric Statistics*. **16**, 127–152.

[34] Gu, C. and Kim, Y. (2002). Penalized Likelihood Regression: General Formula and Efficient Approximation. *The Canadian Journal of Statistics* **30**, 619–628.

[35] Hart, J. (1991). Kernel Regression Estimation with Time Series Errors. *Journal of the Royal Statistical Society, Series B,* **53**, 173–187.

[36] Hastie, T. and Tibshirani, R. (1993). Varying-coefficient Models. *Journal of the Royal Statistical Society, Series B.* **55**, 757–796.

[37] Heckman, N. (1986). Spline Smoothing in a Partially Linear Model. *Journal of the Royal Statistical Society, Series B,* **48**, 244–248.

[38] Hoover, D., Rice, J., Wu, C. and Yang, L. (1998). Nonparametric Smoothing Estimates of Time-Varying Coefficient Models with Longitudinal Data. *Biometrika*, **85**, 809–822.

[39] Huang, J. and Shen, H. (2004). Functional Coefficient Regression Models for Non-linear Time Series: A Polynomial Spline Approach. *the Scandinavian Journal of Statistics*, **31**, 515–534.

[40] Kauermann, G. (2005). Penalized Spline Smoothing in Multivariable Survival Models with Varying Coefficients. *Computational Statistics & Data Analysis*. **49**, 169–186.

[41] Kauermann, G. and Khomski, P. (2006). Additive Two-way Hazards Model with Varying Coefficients. *Computational Statistics & Data Analysis*. **51**, 1944–1956.

[42] Liang, K. and Zeger, S. (1986) Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika*, **73**, 13–22.

[43] Lin, X. and Carroll, R.J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association*, **95**, 520–534.

[44] Masry, E. (1996a). Multivariate regression estimation: local polynomial fitting for time series, *Stochastic Processes and Their Applications*, **65**, 81–101.

[45] Masry, E. (1996b). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series*, **17**, 571–599.

[46] Tsai, C. and Naik, P. (2000) Partial Least Squares Estimator for Single-Index Models. *Journal of the Royal Statistical Society*, Series B, **62**, 763–771.

[47] Opsomer, J. (1995) Estimating a Function by Local Linear Regression when the Errors are Correlated. Department of Statistics, Iowa State University. Preprint 95-42.

[48] Opsomer, J., Wang, Y. and Yang, Y. (2001) Nonparametric Regression with Correlated Errors. *Statistical Science*, **16**, 134–153.

[49] Qu, A., Lindsay, B. and Li, B. (2000) Improving Generalized Estimating Equations Using Quadratic Inference Functions. *Biometrika*, **87**, 823–836.

[50] Ruppert, D. and Wand, M. (1994). Multivariate Weighted Least Squares Regression. *The Annals of Statistics*, **22**, 1346–1370.

[51] Ruppert, D., Sheather, S.J. and Wand, M. (1995). An Effective Bandwidth Selector for Local Least Squares Regression. *Journal of American Statistical Association*, **90**, 1257–1270.

[52] Silverman, B. (1986). Density Estimation for Statistics and Data Analysis. *Chapman and Hall*, London.

[53] Severini, T. and Wong, W. (1992). Profile Likelihood and Conditionally Parametric Models. *Annals of Statistics*, **20**, 1768–1802.

[54] Severini, T. and Staniswalis, J. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of American Statistical Association*, **89**, 501–511.

[55] Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statististical Soceity. Series B*, **50**, 413–436.

[56] Stone, C. (1977). Consistent Nonparametric Regression. *The Annals of Statistics*, **5**, 595–645.

[57] Stone, C. (1980). Optimal Global Rates of Convergence for Nonparametric Estimators. *The Annals of Statistics*, **10**, 1040–1053.

[58] Stone, C. (1982). Optimal Rates of Convergence for Nonparametric Estimators. *The Annals of Statistics*, **8**, 1348–1360.

[59] Tang, Q. and Wang, J. (2005) One-step Estimation for Varying Coefficient Models. *Science in China* Series A, **48**, 198–213.

[60] Tiao, G. and Tsay, R. (2002). Some Advances in Nonlinear and Adaptive Modeling in Time Series . *Journal of Forecasting*, **13**, 109–131.

[61] Wahba, G. (1975). Smoothing Noisy Data with Spline Functions. *Numerische Mathematik*, **24**, 383–393.

[62] Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika*, **90**, 43–52.

[63] Wang, H., Li, R. and Tsai, C. (2007). Tuning parameter selectors for the smaoothly clipped absolute deviation method. *Biometrika*, **94**, 553–568.

[64] Wu, C. and Chiang, C. (2000). Kernel Smoothing On Varying Coefficient Models With Longitudinal Dependent Variable. *Statistica Sinica*, **10**, 433–456.

[65] Wu, C., Chiang, C. and Hoover, D. (1998). Asymptotic Cofidence Regions for Kernel Smoothing of A Varying-Coefficient Model with Longitudinal Data. *Journal of the American Statistical Association*, **93**, 1388–1402.

[66] Wu, H. and Liang, H. (2004). Backfitting Random Varying-Coefficient Models with Time-Dependent Smoothing Covariates. *the Scandinavian Journal of Statistics*, **31**, 3-19.

[67] Xiao, Z., Linton, O., Carroll, R.J. and Mammen, E. (2003) More Efficient Local Polynomial Estimation in Nonparametric Regression with Autocorrelated Errors. *Journal of the AMerican Statistical Association*, **98**, 980–992.

[68] Yao,Q. and Tong, H. (2000) Nonparametric Estimation of Ratios of Noise to Signal in Statistical Regrression. *Statistica Sinica*, **10**, 751–770.

[69] Yatchew, A. (1997). An elementary estimator of the partial linear model. *Economics Letters*, **57**, 135–143.

[70] Zeger, S. and Liang, K. (1986) Longitudinal Data-Analysis for Discrete and Continuous Outcomes. *Biometrika*, **42**, 121–130.

[71] Zhang, W (2001). Local Polynomial Fitting in Semivarying Coefficient Model. *Journal of Multivariate Analysis*, **82**, 166–188.

Vita

Yan Li

The Pennsylvania State University
Department of Statistics
326 Thomas Building
University Park, PA 16801

Office: (814) 865-8045
Cell: (814) 777-0805
Fax: (814) 863-7114
Email: yul135@psu.edu

- Education
  - PhD candidate in Statistics at Pennsylvania University
    Expected graduation: May 2008
  - Master of Science in Applied Mathematics at Ohio University
    Graduated: August 2003
  - Bachelor of Science in Computational Mathematics at Nanjing University, China
    Graduated: July 2002
- Work Experience
  - Market Risk Reporting Analyst Intern
    Washington Mutual Inc, Seattle, WA , June 2006 - August 2006
    * Developed models and methodologies for complex operational loss data
  - Graduate Student Consultant
    Penn State University, Department of Statistics, Fall 2004 - Spring 2005
    * Recommended to six clients the appropriate statistical models and presented the analysis results
  - Research Assistant
    Penn State University, Department of Statistics, Summer 2007, Spring 2008
    * Supposed by the National Science Foundation grants and supervised by the advisor, Runze Li
  - Instructor
    Penn State University, Department of Statistics, Summer 2004, Summer 2005, Fall 2006
    * Taught Elementary Probability and Introduction to Statistics courses individually
  - Teaching Assistant
    Penn State University, Department of Statistics, Fall 2003 - Spring 2006
    * Handled three lab sessions for STAT200 classes and administered the quizzes