The Pennsylvania State University
The Graduate School
Department of Statistics

NONPARAMETRIC TECHNIQUES IN FINITE MIXTURE OF REGRESSION MODELS

A Dissertation in
Statistics

by

Mian Huang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2009

The thesis of Mian Huang was read and approved* by the following:

Runze Li
Professor of Statistics
Dissertation Advisor
Chair of Committee

Thomas P. Hettmansperger
Professor of Statistics

David R. Hunter
Associate Professor of Statistics

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

Hoben Thomas
Professor of Psychology

*Signatures are on File in the Graduate School.

Abstract

Nonparametric Techniques in Finite Mixture of Regression Models

Mixture models have been popular in the literature of both statistics and social science. In this dissertation, we propose a new mixture model, namely, nonparametric finite mixture of regression models, which can be viewed as a natural extension of finite mixture of linear regression. In the newly proposed model, it allows both the regression and variance function as functions of covariates, and their functional forms are nonparametric rather than a specified form. We first consider the mixing proportion in the nonparametric finite mixture of regression models is also a nonparametric function of covariates. We develop an estimation procedure for the nonparametric finite mixture of regression models by employing kernel regression, and proposed an algorithm to carry out the estimation procedure by modifying an EM algorithm. We further systematically studied the sampling properties of the newly proposed estimation procedures and the proposed algorithm. We found that the proposed algorithm preserves the ascent property of the EM algorithm in an asymptotic sense. We derive the asymptotic bias and variance of the resulting estimate. We further established the asymptotic normality of the resulting estimate. Monte Carlo simulation studies are conducted to assess the finite sample performance of the resulting estimate. The proposed methodology is illustrated by analysis of a real data example.

We further study the nonparametric finite mixture of regression models with constant mixing proportion. Since the mixing proportion is parametric, while the regression function and variance function for each components are nonparametric, the model indeed is a semiparametric model. To achieve better convergent rate for mixing proportional parameters, we develop an estimation procedures by using back-fitting algorithm. To reduce computational cost, we further suggest one-step back-fitting algorithm, which behaves similar to the gradient ECM algorithm. Thus, the convergence behavior of the proposed algorithm can be analyzed along the lines for the gradient EM algorithm. We studied the asymptotic properties of the resulting estimate. We showed that the resulting estimate for the mixing proportion parameter is root n consistent, and follows an asymptotic normal distribution. We also derived the asymptotic bias and variance for the resulting estimate of the regression function and variance function, and further established their asymptotic normality. Finite sample performance of the proposed procedure is examined by a Monte Carlo simulation study. The proposed procedure is demonstrated by analysis of a real data example.

As the advent of data collection technology and data storage device, researchers are able to collect functional data without much cost. In this dissertation, we studied mixture models for functional data. More specifically, we proposed mixtures of Gaussian processes for functional data. The proposed model is a natural extension of mixture of high-dimensional normals. We develop an estimation procedure to the mean and covariance function of mixture of Gaussian processes by using kernel regression. The proposed methodology is empirically justified by simulation and illustrated by an analysis of the supermarket data.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First of all, I would like to express my sincere gratitude to my advisor, Dr. Runze Li, for his for his insight, guidance, encouragement, patience, expertise, and for his support in so many aspects. Many of his ideas and discussions were essential in the progress of this thesis. He also provided financial support for my research. Second, I want to thank my thesis committee, Dr. Bruce Lindsay, Dr. Thomas Hettmansberger, Dr. David R. Hunter and Dr. Hoben Thomas, for their precious time and valuable suggestions in improving the contents of this thesis. I also appreciate the help from Dr. John Dziak, for his effort in improving the clarity and correctness of this thesis. Thank you all for your support and suggestions throughout my studies.

This thesis research has been supported by a National Science Foundation grants DMS 0348869 and National Institute on Drug Abuse grants R21 DA024260 and P50 DA10075. The content is solely the responsibility of the author and does not necessarily represent the official views of the NIDA or the NIH.

Chapter 1

Introduction

Mixture models have been widely used in econometrics and social science, and the theory for mixture models has been well studied (Lindsay, 1995). As a class of useful models of the mixture models, finite mixtures of linear regression models have received increasing attention in the literature since its introduction in Goldfeld and Quandt (1976). For example, there are applications in econometrics and marketing (Wedel and DeSarbo, 1993; Frühwirth-Schnatter, 2001; Rossi et al., 2005), in epidemiology (Green and Richardson, 2002), and in biology (Wang et al., 1996). Bayesian approaches for mixture regression models are summarized in Frühwirth-Schnatter (2005). Many efforts have been made to these models and their extensions such as finite mixture of generalized linear models, and comprehensively summarized in McLachlan and Peel (2000).

1.1 Nonparametric Finite Mixture of Regression Models

Motivated by an empirical analysis of US housing index data, we propose a new class of mixture models, namely, nonparametric finite mixture of regression models. The US housing index data contains the monthly SP-Case Shiller House Price Index (HPI) change and United States GDP growth rate from January 1990 to December 2006. It is known that in the literature of economic research, HPI is a measure of a nation's housing cost, and GDP is a measure of the size of

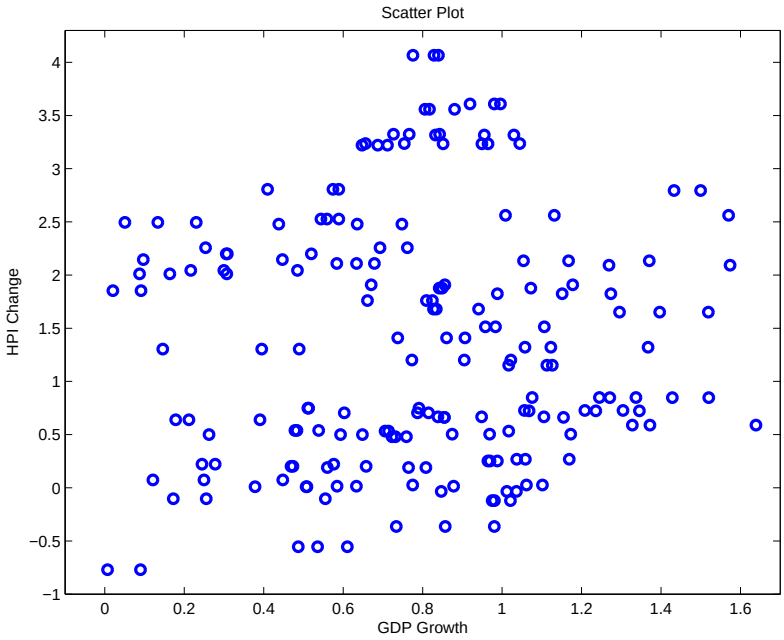


Figure 1.1: Scatterplot of US Housing Index Data

a nation's economy. It is of interest to study the relationship between GDP growth rate and HPI change. As expected, the impact of GDP growth rate on HPI may have different patterns in different macroeconomic cycles. However, it may be difficult to specify the macroeconomic cycles exactly. Figure 1.1 depicts the scatterplot of the GDP growth rate versus the HPI change. From Figure 1.1, it seems that the variability of HPI at any given GDP growth rate seems to be large, and there is no clear relationship between the GDP growth rate and the HPI change.

To interpret the large variability of HPI at a given GDP growth rate, we consider the framework of mixture models, which are powerful tools for data analysis when the population consists of several sub-populations. From Figure 1.1, we may consider a three-component mixture of regression models to fit

this data set. However, a mixture of linear regression may not be appropriate here because the relationship between the GDP growth rate and the HPI changes seems to be nonlinear for the data points with HPI changes lying between 1 and 3 percent. To reduce approximation error and model bias, we may consider a nonparametric regression function rather than a linear regression function for each component.

In Chapters 3, we propose nonparametric mixture of regression models. Specifically, let $\mathcal{C}$ be a latent class variable, and assume that conditioning on $X = x$, $\mathcal{C}$ has a discrete distribution $P(\mathcal{C} = c|X = x) = \pi_c(x)$ for $c = 1, 2, \cdots, C$. Conditioning on $\mathcal{C} = c$ and $X = x$, Y follows a normal distribution with mean $m_c(x)$ and variance $\sigma_c^2(x)$, where $m_c(\cdot)$ and $\sigma_c(\cdot)$ are unknown but smooth functions. In other words, conditioning on $X = x$, the response variable Y follows a finite mixture of normals

$$\sum_{c=1}^{C} \pi_c(x) N\{m_c(x), \sigma_c^2(x)\}, \tag{1.1}$$

where it is assumed that C is fixed. We refer to model (1.1) as a nonparametric finite-mixture-of-regression model because both $m_c(\cdot)$ and $\sigma_c^2(\cdot)$ are nonparametric. For example, we may fit the housing price index data by a nonparametric mixture of regression models with $C = 3$. Detailed analysis of the housing price index will be given in Chapter 3.

When $\sigma_c^2(x)$ is constant, and $m_c(x)$ is linear in x, model (1.1) reduces to a finite mixture of linear regression models (Goldfeld and Quandt, 1976). When $C = 1$, model (1.1) is a nonparametric regression model. Thus, model (1.1) can be regarded as a natural extension of both nonparametric regression models and finite mixture of linear regression models. Compared with the finite mixture of linear regression models, the newly proposed models relax the linearity

assumption on the regression function, and allow that the regression function in each of components is unknown but smooth functions of its covariates. It is easy to adapt the conventional constraints imposed on the finite mixture of linear models for the proposed models so that they are identifiable, and the corresponding likelihood function is bounded. For further references of these two issues, see Hening (2000) and Hathaway (1985).

In Chapter 3, we develop an estimation procedure for the unknown functions in nonparametric mixture of regression model via a local likelihood approach. It is desirable to estimate the curves by evaluating the resulting estimate over set of grid points in x. A naive implementation of the EM algorithm would not ensure that the labels match correctly at different grid points. This is similar to the issue of label switching problem in mixture modeling. We modify the EM algorithm (Dempster et al., 1977) to simultaneously maximize the local likelihood functions for the proposed nonparametric mixture of regression model at set of grid points. The modified EM algorithm works well in our simulation and real data example. We further demonstrate that the proposed EM algorithm preserves the monotone ascent property of the EM algorithm. We further systematically studied the sampling properties of the newly proposed estimation procedures and the proposed algorithm. We found that the proposed algorithm preserves the ascent property of the EM algorithm in an asymptotic sense. We derive the asymptotic bias and variance of the resulting estimates. We further established the asymptotic normality of the resulting estimates. Monte Carlo simulation studies are conducted to assess the finite sample performance of the resulting estimate. The proposed methodology is illustrated by an analysis of a real data example.

In Chapter 4, we further study the nonparametric finite mixture of regres-

sion models with constant mixing proportion. Specifically, let $\mathcal{C}$ be a latent class variable with a discrete distribution $P(\mathcal{C} = c) = \pi_c$ for $c = 1, 2, \cdots, C$. Conditioning on $\mathcal{C} = c$, the relationship between X and Y follows a nonparametric regression model,

$$Y = m_c(X) + \sigma_c(X)\epsilon, \tag{1.2}$$

where $\epsilon \sim N(0, 1)$, $m_c(\cdot)$ and $\sigma_c(\cdot)$ are unknown but smooth functions. In other words, conditioning on x,

$$Y \sim \sum_{c=1}^{C} \pi_c N\{m_c(x), \sigma_c^2(x)\}. \tag{1.3}$$

Compared with nonparametric mixture of regression model (1.1), model (1.3) indeed is a semi-parametric model because π_c is unknown parameter rather than unknown nonparametric function of x. Thus, we may derive a more efficient estimate for π_c.

To achieve a better convergence rate for the mixing proportion parameters, we develop an estimation procedure using a back-fitting algorithm. To reduce computational cost, we further suggest a one-step back-fitting algorithm, which behaves similar to the gradient ECM algorithm. Thus, the convergence behavior of the proposed algorithm can analyzed along the lines for the gradient EM algorithm. We studied the asymptotic properties of the resulting estimate. We showed that the resulting estimate for the mixing proportion parameter is root n consistent, and follows an asymptotic normal distribution. We also derived the asymptotic bias and variance for the resulting estimate of the regression function and variance function, and further established their asymptotic normality. Finite sample performance of the proposed procedure is examined by a Monte Carlo simulation study. The proposed procedure is demonstrated by analysis of a real data example.

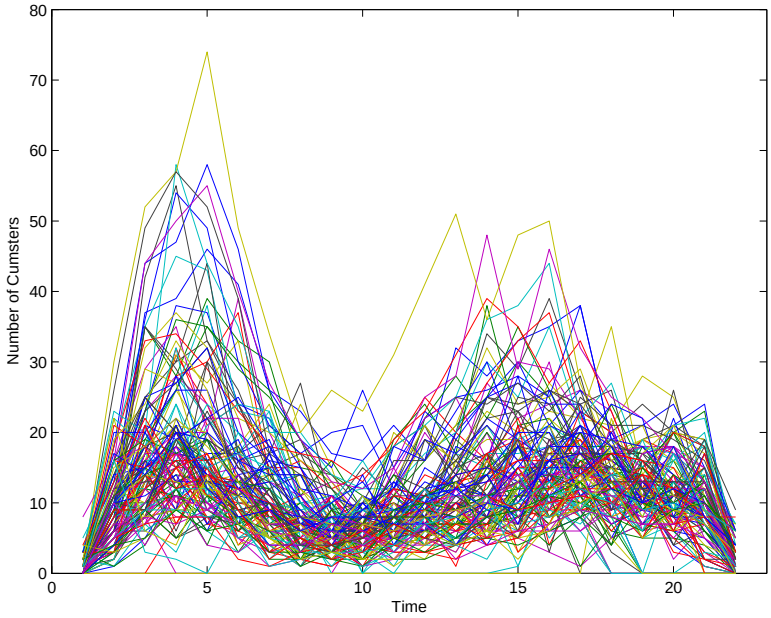


Figure 1.2: Plot of supermarket data.

1.2 Mixture Models for Functional Data

With modern data collection technology, one may easily collect a set of curves over an interval. As an illustration of functional data, Figure 1.2 depicts the plot of a set of collected curves. This data set contains the number of customers who visited a particular supermarket on each of 139 days. For each day, the number of customers shopping in the supermarket is observed every half hour from 7:00am to 5:30pm. Thus, there are 22 observations for each day. The collected time was coded as 1 for 7:00am, 2 for 7:30am, and so on. In the analysis of this data in Chapter 5, we regard each day as one subject. Thus, we have a total of 139 subjects.

Figure 1.2 shows that the variability may be large in certain time periods. Naively, we may consider a 22-dimensional multivariate Gaussian mixture for the data analysis. The challenge with a high-dimensional mixture of normals

is to estimate the covariance matrix. From our own limited experience, the resulting estimate for the covariance matrix of high-dimensional is likely ill-conditioned. This is undesirable. As an alternative, we consider a mixture of Gaussian processes to model this data set. Compared with the high-dimensional mixture normals, mixtures of Gaussian processes naturally take into account the smoothness of the mean function and covariance function over times.

In Chapter 5, we consider mixture of Gaussian processes, which are defined as follows. Let $\mathcal{C}$ be a latent class variable with a discrete distribution $P(\mathcal{C} = c) = \pi_c$ for $c = 1, 2, \cdots, C$. Given $\mathcal{C} = c$, $\{X(t), t \in T\}$ follows a Gaussian process with mean $\mu_c(t)$ and covariance function $\text{Cov}\{X(s), X(t)\} = G_c(s, t)$. Here $\mu_c(t)$ is a smooth function of t, and $G_c(s, t)$ is a positive definite and bivariate smooth function of s and t. Thus, the path of $X(t)$ indeed is a smooth function. In practice, the observed functional curve $Y(t), t \in T$, may not be smooth, as depicted in Figure 1.2. Therefore, it is common in the literature to assume that the observed process $Y(t)$ is

$$Y(t) = X(t) + \epsilon(t),$$

where $\epsilon(t) \sim N(0, \sigma^2)$ is measurement error. In the literature, it is also assumed that $\epsilon(t)$ and $\epsilon(s)$ are independent for $t \neq s$. For reference, see Yao et al. (2003), and Yao et al. (2005).

In Chapter 5, we systematically study the proposed mixture Gaussian processes. We propose an estimation procedure for the newly proposed model by using back-fitting algorithm. We empirically test the proposed algorithm by a Monte Carlo simulation, and further apply the proposed procedure for analysis of the supermarket data.

1.3 Organization of This Dissertation

The rest of this thesis is organized as follows. Chapter 2 provides a literature review of mixture models, including mixture regression models and an EM algorithm for mixture models, and local modeling methods. We systematically study finite nonparametric mixture regression models in Chapter 3, and finite nonparametric mixture regression models with constant mixing proportion in Chapter 4. For these two classes of nonparametric mixture regression models, we will develop effective estimation procedures using an EM algorithm and local likelihood method. We further study asymptotic properties of the resulting estimates. We derive the convergence rate of the resulting estimate, and further establish its asymptotic normality. We also conduct Monte Carlo simulation to assess the finite sample performance of the proposed methodologies. In Chapter 5, we propose mixture of Gaussian processes model, and develop an estimation procedure for this mixture of Gaussian processes model. The proposed method is illustrated by simulations and an analysis of a real data set. Chapter 6 presents some discussion and ideas for future research.

Chapter 2

Literature Review

In this chapter, we review the literature of mixture models, EM algorithms and their application in mixture regression models. We will also review local polynomial regression, local likelihood methods, and functional data analysis. These materials will be used to develop statistical inference procedures for the newly proposed models in this thesis.

2.1 Mixture Models

Mixture models are associated with mixture density functions, which are convex combinations of component density functions. A mixture model has the density form

$$f(x) = \sum_{c=1}^{C} \pi_c f_c(x),$$

where $\pi_c \geq 0$ and $\sum_{c=1}^{C} \pi_c = 1$. $f_c(x)$ is the c^{th} component density function and π_c is the mixing proportion of the c^{th} component. Parametric mixture models specify a parametric form with some unknown parameters for each component density, and the mixture density function can be written as

$$f(\mathbf{x}|\mathbf{\Phi}) = \sum_{c=1}^{C} \pi_c f_c(\mathbf{x}|\boldsymbol{\beta}_c), \tag{2.1}$$

where $\mathbf{\Phi} = \{\pi_1, \cdots, \pi_{C-1}, \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_C\}$. In this thesis, we consider C is a fixed number, and model (2.1) is a parametric finite mixture model.

Mixture models may be viewed as model-based clustering approaches for data from several homogeneous subgroups with missing grouping identities. In the literature, mixture models were mainly studied from the likelihood point of view. Lindsay (1983a) and Lindsay (1983b) gave some fundamental properties of the maximum likelihood estimator of the mixing distribution, including sufficient conditions for uniqueness of the maximum likelihood estimate. See Lindsay (1995) and and McLachlan and Peel (2000) for a broad review of mixture models.

In practice, there are two major classes of estimation methods. The first one is EM algorithm, which was proposed in Dempster et al. (1977), and systematically studied in McLachlan and Krishnan (1997). The other one is Bayesian methods, especially Markov Chain Monte Carlo estimation. Smith and Roberts (1993) proposes a Gibbs sampling procedure for mixture models. Recent developments in Bayesian analysis include the reverse-jump algorithm (Green, 1995) and the birth-and-death algorithm (Stephens, 2000), which can be used to estimate the number of components. Frühwirth-Schnatter (2005) gives a comprehensive summary of Bayesian analysis for mixture models and Markov switching models.

In general, Bayesian methods provide more information about the unknown parameters, but they are very expensive in terms of computational cost. Since nonparametric smoothing typically requires intensive computation, we will use the EM algorithm for estimation procedures in the proposed models in this dissertation.

2.1.1 EM Algorithm

The EM algorithm provides iterative steps to maximize a likelihood function when some the data are missing. Suppose that the complete data are $\{(x_i, \mathbf{z}_i), i = 1 \cdots, n\}$, which are independent samples from population $(X, \mathbf{Z})$. The observed data is $\{x_i, i = 1 \cdots, n\}$. Denote $\ell(\mathbf{\Phi})$ the log-likelihood function, where $\mathbf{\Phi}$ denotes the unknown parameters in the likelihood model. Let $\mathcal{L}(\mathbf{\Phi})$ be the complete likelihood function if the missing data $\mathbf{z}$ is given.

The EM algorithm consists of two steps: E-step and M-step. In the E step, we compute the expectation of the complete log-likelihood function over the missing data conditioning on the observed data with the given parameters. The expectation is called a Q function. That is, in the E step of the l^{th} iteration, we compute

$$Q(\mathbf{\Phi}|\mathbf{\Phi}^{(l)}) = \mathrm{E}(\mathcal{L}(\mathbf{\Phi})|\mathbf{\Phi}^{(l)}, \mathbf{x}).$$

Then in the M step, we maximize Q function $Q(\mathbf{\Phi}|\mathbf{\Phi}^{(l)})$ with respect to $\mathbf{\Phi}$, and update the parameters $\mathbf{\Phi}$ as

$$\mathbf{\Phi}^{(l+1)} = \arg \max_{\mathbf{\Phi}} Q(\mathbf{\Phi}|\mathbf{\Phi}^{(l)}).$$

The E step and M step will be iterated until algorithm convergence, i.e. the likelihood difference

$$\ell(\mathbf{\Phi}^{(l+1)}) - \ell(\mathbf{\Phi}^{(l)})$$

is sufficient small. The EM algorithm has several advantages such as reliable convergence and ease of use. Theoretical properties show that in every iteration, EM algorithm increases the objective likelihood function $\ell(\mathbf{\Phi})$ by maximizing the Q function $Q(\mathbf{\Phi}|\mathbf{\Phi}^{(l)})$. In particular, we have

$$\ell(\mathbf{\Phi}^{(l+1)}) - \ell(\mathbf{\Phi}^{(l)}) \geq 0$$

for all $l > 0$. However, it cannot guarantee a global MLE, not even a local MLE. It may converge to a saddle point. Wu (1983) and Mclachlan and Krishnan (1997) analyze the convergence behaviors of the EM algorithm. Under fairly general conditions, the EM algorithm can provide global maximum likelihood estimators. For further reference of EM algorithm, see Dempster *et al.* (1977) and Mclachlan and Krishnan (1997).

2.1.2 Mixture of Regression Models

Mixture of regression models are well known as switching regression models in econometrics literature, which were first introduced by Goldfeld and Quandt (1976). Mixture of regression models are appropriate to use when the observations are from several subgroups with missing grouping identities, and in each subgroup, there are two or more variables with linear relationships between these variables. The model setting can be stated as follows. Let $\mathcal{C}$ be a latent class variable with $P(\mathcal{C} = c) = \pi_c$ for $c = 1, 2, \cdots, C$. Suppose that given $\mathcal{C} = c$, the response y depends on x in a linear way:

$$y = \mathbf{x}^T \boldsymbol{\beta}_c + \epsilon_c = \beta_{0c} + \beta_{1c} x + \epsilon_c, \quad \epsilon_c \sim N(0, \sigma_c^2). \tag{2.2}$$

The conditional distribution of Y given $X = x$ can be written as

$$Y|X = x \sim \sum_{c=1}^{C} \pi_c N(\mathbf{x}^T \boldsymbol{\beta}_c, \sigma_c^2), \tag{2.3}$$

where the $\{(\boldsymbol{\beta}_c, \sigma_c^2), c = 1, \cdots, C\}$ are the parameters of each component density, and $\{\pi_c, c = 1, \cdots, C\}$ are the mixing proportions for each component. Since we have the constraint that $\sum_{c=1}^{C} \pi_c = 1$, denote $\pi = (\pi_1, \cdots, \pi_{C-1})$, and $\pi_C = 1 - \sum_{c=1}^{C-1} \pi_c$. Denote $\phi(y|\mu, \sigma^2)$ to be the density function $N(\mu, \sigma^2)$. The conditional

likelihood function of mixture of regression models can be written as

$$f(y|x) = \sum_{c=1}^{C} \pi_c \phi(y|\mathbf{x}^T \boldsymbol{\beta}_c, \sigma_c^2).$$

EM Algorithm for Mixture Regression Models

In a C-component mixture of regression model (2.3), the observations are $\{(x_i, y_i), i = 1, 2, \ldots, n\}$, the unobserved data are the latent variables, which identify the distribution membership. The complete data are $\{(x_i, y_i, \mathcal{C}_i), i = 1, 2, \ldots, n\}$, where $\mathcal{C}_i$ is the component identity of (x_i, y_i) and has a discrete distribution $P(\mathcal{C}_i = c) = \pi_c, c \in \{1, 2, \cdots, C\}$. The log-likelihood function of a mixture of regressions model is:

$$\ell(\boldsymbol{\Phi}) = \sum_{i=1}^{n} \log \left(\sum_{c=1}^{C} \pi_c \phi(y_i|\mathbf{x}_i^T \boldsymbol{\beta}_c, \sigma_c^2) \right), \tag{2.4}$$

where $\boldsymbol{\Phi} = (\boldsymbol{\beta}^T, \boldsymbol{\sigma}^{2T}, \pi^T)^T$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_C^T)^T$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \cdots, \sigma_C^2)^T$, and $\pi = (\pi_1, \cdots, \pi_{C-1})^T$. In practice it is difficult to directly maximize the log-likelihood function (2.4) because of its complicated structure. The EM algorithm provides a convenient iterative way to maximize the log-likelihood. We next illustrate how the EM algorithm is carried out in a mixture of regressions models. We first define random variables

$$z_{ic} = \begin{cases} 1, & \text{if } (x_i, y_i) \text{ is in the } c^{th} \text{ group} \\ 0, & \text{otherwise.} \end{cases}$$

and let $\mathbf{z}_i = (z_{i1}, \cdots, z_{iC})^T$. The complete data are now $\{(x_i, y_i, \mathbf{z}_i), i = 1, 2, \ldots, n\}$. The complete log-likelihood function for (2.4) is

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Phi}) &= \log \left(\prod_{i=1}^{n} \prod_{c=1}^{C} [\pi_c \phi(y_i|\mathbf{x}_i^T \boldsymbol{\beta}_c, \sigma_c^2)]^{z_{ic}} \right) \\ &= \sum_{i=1}^{n} \sum_{c=1}^{C} z_{ic} \left\{ \log \pi_c + \log \phi(y_i|\mathbf{x}_i^T \boldsymbol{\beta}_c, \sigma_c^2) \right\}. \end{aligned}$$

E-step: Given $\mathbf{\Phi}^{(l)}$, compute

$$r_{ic}^{(l)} = E(z_{ic}|\mathbf{\Phi}^{(l)}) = \frac{\pi_c^{(l)}\phi(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_c^{(l)}, \sigma_c^{2(l)})}{\sum_{q=1}^{C}\pi_q^{(l)}\phi(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_q^{(l)}, \sigma_q^{2(l)})}. \tag{2.5}$$

Then we can calculate the Q function

$$\begin{aligned} Q(\mathbf{\Phi}|\mathbf{\Phi}^{(l)}) &= \mathrm{E}(\mathcal{L}(\mathbf{\Phi})|\mathbf{\Phi}^{(l)}) \\ &= \sum_{i=1}^{n}\left\{\sum_{c=1}^{C} r_{ic}^{(l)}\log\pi_c^{(l)} + \sum_{c=1}^{C} r_{ic}^{(l)}\log\phi(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_c^{(l)}, \sigma_c^{2(l)})\right\}, \end{aligned}$$

where $\mathbf{\Phi}^{(l)} = (\boldsymbol{\beta}_c^{(l)}, \sigma_c^{2(l)}, \pi_c^{(l)}|c = 1, \ldots, c)$.

M step: Maximize $Q(\mathbf{\Phi}|\mathbf{\Phi}^{(l)})$ under the following constraints:

$$\sum_{c=1}^{C} r_{ic}^{(l)} = 1, \quad \text{and} \quad \sum_{c=1}^{C}\pi_c^{(l)} = 1.$$

The resulting estimators are

$$\begin{aligned} \pi_c^{(l+1)} &= n^{-1}\sum_{i=1}^{n} r_{ic}, \\ \boldsymbol{\beta}_c^{(l+1)} &= (\mathbf{X}^T R_c^{(l)}\mathbf{X})^{-1}\mathbf{X}^T R_c^{(l)}\mathbf{y}, \\ \sigma_c^{2(l+1)} &= \frac{\sum_{i=1}^{n} r_{ic}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_c)^2}{\sum_{i=1}^{n} r_{ic}}, \end{aligned} \tag{2.6}$$

where $R_c^{(l)} = \mathrm{diag}(r_{1c}^{(l)}, \cdots, r_{nc}^{(l)})$, and

$$\mathbf{X} = \begin{pmatrix} 1, & 1 & \cdots, & 1 \\ x_1, & x_2 & \cdots, & x_n \end{pmatrix}^T.$$

Iteratively update the E-step and the M-step until the algorithm converges.

2.1.3 Choose The Number of Components

Choosing the number of component is a very important issue in mixture models. In literature there are two major approaches. One approach is the penalized likelihood method. A penalized likelihood function is defined as

$$\mathcal{L}_n(\hat{\mathbf{\Phi}}_C) - \lambda(\mathbf{\Phi}_C),$$

where $\mathbf{\Phi}_C$ represents the parameters for a C component mixture model, $\mathcal{L}_n(\hat{\mathbf{\Phi}}_C)$ is the maximum likelihood function, and $\lambda(\mathbf{\Phi}_C)$ is added as a penalty term. There are two popular criteria which can be used as a penalty term, AIC (Akaike, 1974) is given by $\lambda(\mathbf{\Phi}_C) = \dim(\mathbf{\Phi}_C)$, and BIC (Schwarz, 1978) is given by $\lambda(\mathbf{\Phi}_C) = \frac{1}{2}\log(n)\dim(\mathbf{\Phi}_C)$, where $\dim(\mathbf{\Phi}_C)$ is the number of parameters in a C component mixture model. Leroux (1992) proves the weak consistency of the maximum penalized likelihood estimators for the mixing distribution. For other reference, see Chen and Kalbeisch (1996), Lindsay (1995), and McLachlan and Peel (2000).

Another approach for selecting the number of components is Bayesian methods. By assuming some prior distributions, the Bayesian approach provides estimates as well as their posterior distributions. See Green (1995) for the reversible jump Metropolis-Hasting algorithm and Stephens (2000) for the birth-death processes.

2.2 Local Modeling Methods

Regression is one of the most widely used general statistical methods. Applications of regression models can be found in many research fields, including econometrics, social science, medicine, and psychology. A simple linear regression model has the form

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where $\mathrm{E}(\epsilon|x) = 0$ and $\mathrm{Var}(\epsilon|x) = \sigma^2$. If the relationship from the observed data is close to linearity, then the above model can be used and the parameters can be estimated by the least square methods. However, linear regression may not be appropriate to estimate a functional form of a curved relationship, especially

when the unknown function has a complex shape, and cannot be converted to a linear relationship by some transformation. One successful technique to relax the linear assumption is the nonparametric regression model

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

There is a huge literature on nonparametric regression. One approach to nonparametric regression is local modeling. The basic idea of local modeling is to locally estimate the mean function $m(x)$ by a set of parametric models. Nadaraya (1964) and Watson (1964) proposed the Nadaraya-Watson estimator, which is also referred to as the kernel regression estimator. This is a special case of local polynomial regression, or local constant regression estimator. Fan and Gijbels (1996) give a comprehensive account on local polynomial regression.

2.2.1 Local Polynomial Regression

Suppose we are interested in estimating the mean function $m(x)$ at the point x_0. In local polynomial regression, we first apply a Taylor expansion to $m(x)$ in a neighborhood of x_0:

$$m(x) \approx \mathbf{x}^T\boldsymbol{\beta} = \beta_0 + \beta_1(x - x_0) + \cdots + \beta_p(x - x_0)^p,$$

where $\mathbf{x} = \{1, x - x_0, \cdots, (x - x_0)^p\}^T$, $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)^T$, and $\beta_j = m^{(j)}(x_0)/j!$. Intuitively, points close to x_0 will have more influence on the $m(x_0)$ estimate, while points far from x_0 will have less influence. This suggests a weighted regression model, which puts more weight on the points near x_0, and less weight on the points far from x_0. This goal can be achieved by minimizing a weighted polynomial regression:

$$\sum_{i=1}^{n}\{y_i - \beta_0 - \cdots - \beta_p(x_i - x_0)^p\}^2 K_h(x_i - x_0), \tag{2.7}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ and $K(\cdot)$ is a kernel function which will control the weights of the points at different locations. In general, a symmetric kernel function K satisfies the following conditions

$$K(u) \geq 0, \quad \int K(u)du = 1, \quad \int uK(u)du = 0, \quad \int u^2K(u)du > 0.$$

The resulting estimator is called the local polynomial regression estimator. For convenience, denote

$$\begin{aligned} W &= \text{diag}\{K_h(x_1 - x), \ldots, K_h(x_n - x)\}, \\ \mathbf{X} &= \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} 1, & x - x_1, & \cdots, & (x - x_1)^p \\ \vdots & \vdots & \cdots, & \vdots \\ 1, & x - x_n, & \cdots, & (x - x_n)^p \end{pmatrix}. \end{aligned}$$

Then the solution to the locally weighted least squares problem (2.7) is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T W \mathbf{X})^{-1}\mathbf{X}^T W \mathbf{y}, \tag{2.8}$$

$$\hat{m}(x) = \mathbf{e}_1^T \times (\mathbf{X}^T W \mathbf{X})^{-1}\mathbf{X}^T W \mathbf{y}, \tag{2.9}$$

where $\mathbf{y} = (y_1, \cdots, y_n)^T$, and $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$ is a $1 \times (p+1)$ vector with first entry one and others zero. Furthermore, we can obtain an estimate of the q^{th} $(q < p)$ derivative of $m(x)$:

$$\hat{m}^{(q)}(x) = q!\mathbf{e}_{q+1}^T(\mathbf{X}^T W_{x_0}\mathbf{X})^{-1}\mathbf{X}^T W \mathbf{y},$$

where $\mathbf{e}_{q+1}$ is a $1 \times (p+1)$ vector with $(q+1)^{th}$ entry one and others zero.

Kernel estimators and local linear regression are special cases of local polynomial regression. When $p = 0$, the local polynomial regression estimator reduces to a local constant estimator, and (2.8) becomes:

$$\hat{m}(x) = \frac{\sum_{i=1}^n K_h(x_i - x)y_i}{\sum_{i=1}^n K_h(x_i - x)},$$

Table 2.1: Asymptotic biases and variances

Method	Bias	Variance
NW estimator	$m''(x) + \frac{2m'(x)f'(x)}{f(x)} b_n$	V_n
Local linear	$m''(x) b_n$	V_n

$b_n = \frac{1}{2} h^2 \int u^2 K(u) du, V_n = \frac{\sigma^2(x)}{f(x)nh} h^2 \int K(u) du$

which is the same as the N-W estimator proposed by Nadaraya (1964) and Watson (1964). When $p = 1$, the local polynomial regression estimator is called a local linear estimator. Asymptotic bias and variance of the two estimators are summarized in Table 2.1. From Table 2.1, the local linear estimator has a more concise form of asymptotic bias than N-W estimator, while the asymptotic variances are the same. In addition, the local linear estimator has several good properties, such as automatic correction of boundary effects (Fan and Gijbels, 1992; Cheng, Fan and Marron, 1997), design adaptivity, and best asymptotic efficiency by minimax criteria (Fan, 1993).

The choice of the smoothing parameter h, referred to as bandwidth, is an important issue in local polynomial regression. The smoothing parameter h controls the weights of the observation points used to estimate of regression function. For a good estimate of the unknown function, bandwidth cannot be too large or too small, since a tradeoff occurs between the bias and the variance of the resulting estimate. Theoretically, the optimal bandwidth can be obtained by minimizing mean integrated square error (MISE) or an asymptotic leading term of MISE (Simonoff, 1998). In practice, data driven methods can be used for bandwidth selection, such as cross-validation (CV) criterion. Denote by $\mathcal{D}$ as the full data set. We then partition $\mathcal{D}$ into a training set $\mathcal{R}_j$ and test set $\mathcal{T}_j$,

$j = 1, \cdots, N$. Then we have $\mathcal{D} = \mathcal{T}_j \cup \mathcal{R}_j$. Denote $\hat{\mathbf{\Phi}}_{\mathcal{R}_j}$ the estimate based on the training set and $\hat{y}$ the fitted value on the test set $\mathcal{T}_j$. The CV criterion has the form

$$\text{CV} = \sum_{j=1}^{N} \sum_{x_i \in \mathcal{T}_j} \{y_i - \hat{y}(x_i)\}^2. \tag{2.10}$$

We choose bandwidth when CV is minimized.

The choice of kernel function is not crucial for the selection of bandwidth. A commonly used kernel is the Gaussian kernel:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

The symmetric beta family are also popular choices,

$$K(x) = \frac{1}{\text{Beta}(1/2, \gamma + 1)} (1 - x^2)_+^{\gamma}, \gamma = 0, 1, 2, \cdots,$$

where $\text{Beta}(\cdot, \cdot)$ denotes a beta function, and the support of the beta function family is $[-1, 1]$. When $\gamma = 0$, the kernel function becomes the uniform kernel. When $\gamma = 1$, the kernel function becomes Epanechnikov kernel $K(x) = 0.75(1 - x^2)$. Symmetric kernel functions are preferred because they yield less biased estimates. Marron and Nolan (1988) shows that estimation of a regression function is not sensitive to the choice of kernel function.

2.2.2 Local Likelihood Estimation

Tibshirani and Hastie (1987) first extended the idea of nonparametric regression to likelihood based regression models. Local likelihood techniques were developed for generalized linear models in Fan *et al.* (1995), hazard regression models in Fan *et al.* (1997) and estimating equations in Carroll *et al.* (1998). To introduce the idea of local likelihood estimation, we review an example in Fan

et al. (1998), which shows the connection between local polynomial regression and local likelihood estimation.

Assume the observed data $\{(x_i, y_i), i = 1, \cdots, n\}$ are independent random samples from population (X, Y), and (x, y) follows a normal regression model

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

where $\epsilon \sim N(0, \sigma^2)$, and $m(x)$ is an unknown mean function. Conditioning on $X = x$, the density function of Y can be written as

$$\phi(y|m(x), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\{y - m(x)\}^2\right]. \tag{2.11}$$

Suppose we are interested in estimating $m(x)$ at x_0. Similar to local polynomial regression, $m(x)$ is locally approached by a polynomial around x_0:

$$m(x) \approx \mathbf{x}^T\boldsymbol{\beta} = \beta_0 + \beta_1(x - x_0) + \cdots + \beta_p(x - x_0)^p.$$

We then consider a kernel weighted log-likelihood, which puts more weight on the points in a neighborhood of x_0 and less weight on the points far from x_0. This kernel weighted log-likelihood is called a local likelihood. Let $\mathbf{x}_i = (1, x_i - x_0, \cdots, (x_i - x_0)^p)^T$. The local likelihood function for normal regression model is

$$\begin{aligned} \ell(\boldsymbol{\beta}) = & - \log(\sqrt{2\pi\sigma^2}) \sum_{i=1}^{n} K_h(x_i - x_0) \\ & - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \{y_i - \mathbf{x}_i^T\boldsymbol{\beta}\}^2 K_h(x_i - x_0). \end{aligned}$$

Maximizing the above local likelihood function is equivalent to minimizing

$$\sum_{i=1}^{n} (y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 K_h(x_i - x_0),$$

which yields the local polynomial regression estimator.

Now we can generally describe local likelihood estimation. Suppose we have independent observed data $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ from population (X, Y), and (x_i, y_i) has log-likelihood $l\{m(x_i), y_i\}$, where $m(x)$ is an unknown function of interest. Similar to local polynomial estimation, we apply a Taylor expansion to $m(x)$ in a neighborhood of x_0:

$$m(x) \approx \mathbf{x}^T\boldsymbol{\beta} = \beta_0 + \beta_1(x - x_0) + \cdots + \beta_p(x - x_0)^p,$$

then maximize a local likelihood function

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} l\{\mathbf{x}_i^T\boldsymbol{\beta}, y_i\}K_h(x_i - x_0) \tag{2.12}$$

with respect to $\boldsymbol{\beta}$. Suppose the solution is $\hat{\boldsymbol{\beta}}$, then we have $\hat{m}(x_0) = \hat{\beta}_0$. From the above discussion we can see that local likelihood estimation is a natural extension of local polynomial estimation in likelihood based models.

Example 2.1: Fan et al. (1998) gives an application of the local likelihood method in nonparametric logistic regression. The detail is given below. Suppose we have independent data $(x_1, y_1), \cdots, (x_n, y_n)$ form population (X, Y). Conditioning on $X = x$, Y has Bernoulli distribution with success rate $p(x)$

$$P(Y = 1|X = x) = p(x), \;\; P(Y = 0|X = x) = 1 - p(x).$$

The function of interest is $m(x)$, which relates to $p(x)$ in the following way:

$$p(x) = \frac{\exp\{m(x)\}}{1 + \exp\{m(x)\}}.$$

Then the pointwise log-likelihood is

$$l\{y, m(x)\} = \log[p(x)^y\{1 - p(x)\}^{1-y}] = ym(x) - \log(1 + e^{m(x)}).$$

To estimate $m(x)$ at x_0, we apply a Taylor expansion of $m(x)$ as before, then maximize a local log-likelihood

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n}[y_i\{\beta_0+\cdots+\beta_p(x_i-x_0)^p\} \\ &- \log(1+\exp\{\beta_0+\cdots+\beta_p(x_i-x_0)^p\})]K_h(x_i-x_0).\end{aligned}$$

The maximization of $\ell(\boldsymbol{\beta})$ can be achieved by a Newton-Raphson algorithm:

$$\begin{aligned}\boldsymbol{\beta}^{(l+1)} &= \boldsymbol{\beta}^{(l)} - \left(\frac{\partial^2\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)^{-1}\frac{\partial\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} \\ &= \boldsymbol{\beta}^{(l)} + (\mathbf{X}D^{(l)}W\mathbf{X}^T)^{-1}\mathbf{X}^TW(\mathbf{y}-\mathbf{p}^{(l)}),\end{aligned}$$

where

$$\begin{aligned}W &= \text{diag}\{K_h(x_1-x),\ldots,K_h(x_n-x)\}, \\ \mathbf{p}^{(l)} &= \left\{p^{(l)}(x_1),\cdots,p^{(l)}(x_n)\right\}^T, \\ D^{(l)} &= \text{diag}\left\{p^{(l)}(x_1)\{1-p^{(l)}(x_1)\},\cdots,p^{(l)}(x_n)\{1-p^{(l)}(x_n)\}\right\}.\end{aligned}$$

and

$$p^{(l)}(x_i) = \frac{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}^{(l)}\right)}{1+\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}^{(l)}\right)}.$$

Example 2.2: Fan *et al.* (1998) also point out that local likelihood method can be applied to likelihood models based on Poisson distribution and gamma distribution, without showing details of the estimation procedure. Now we derive the results for nonparametric Poisson regression. Suppose we have independent observed data $\{(x_1,y_1),\cdots,(x_n,y_n)\}$ from population (X,Y). Given $X=x$, Y has a Poisson distribution:

$$P(Y=y|X=x) = \frac{\exp\{-\lambda(x)\}\{\lambda(x)\}^y}{y!}.$$

The parameter of interest is $m(x)$, which relates to $\lambda(x)$ in the following way:

$$\lambda(x) = \exp\{m(x)\}.$$

Then the pointwise log-likelihood function without a normalizing constant $\log(y!)$ is

$$l\{y, m(x)\} = -\lambda(x) + y\log(\lambda(x)) = ym(x) - \exp\{m(x)\}.$$

To estimate $m(x)$ at x_0, we use a Taylor expansion as before, then maximize a local log-likelihood function

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} [y_i\{\beta_0 + \cdots + \beta_p(x_i - x_0)^p\} \\ &- \exp\{\beta_0 + \cdots + \beta_p(x_i - x_0)^p\}] K_h(x_i - x_0). \end{aligned}$$

The maximization of $\ell(\boldsymbol{\beta})$ can be achieved by a Newton-Raphson algorithm:

$$\begin{aligned} \boldsymbol{\beta}^{(l+1)} &= \boldsymbol{\beta}^{(l)} - \left(\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right)^{-1} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= \boldsymbol{\beta}^{(l)} + (\mathbf{X} D^{(l)} W \mathbf{X}^T)^{-1} \mathbf{X}^T W (\mathbf{y} - E^{(l)}), \end{aligned}$$

where

$$\begin{aligned} W &= \operatorname{diag}\left\{K_h(x_1 - x), \cdots, K_h(x_n - x)\right\}, \\ E^{(l)} &= \left\{\exp(\mathbf{x}_1^T \boldsymbol{\beta}^{(l)}), \cdots, \exp(\mathbf{x}_n^T \boldsymbol{\beta}^{(l)})\right\}^T, \\ D^{(l)} &= \operatorname{diag}(E^{(l)}). \end{aligned}$$

The asymptotic normality of local likelihood estimates in the generalized linear model setting is given by Fan *et al.* (1995); in the hazard model setting are given by Fan *et al.* (1997); and in the local estimating equation setting are given by Carroll *et al.* (1998). For further reference, see Fan and Gijbels (1996).

2.3 Functional Data

Functional data analysis has received much attention in the literature; see Ramsay and Silverman (1997) for a comprehensive summary. The basis of functional data analysis consists of two parts: the estimation of the mean function and of the covariance structure. Among many approaches, functional principal component (FPC) analysis serves as a key technique in functional data analysis. Rice and Silverman (1991) studies the spline smoothing in FPC analysis; Staniswalis and Lee (1998) and Yao et al. (2003) focus on kernel-based smoothing for irregular and sparse longitudinal data, Yao et al. (2005) further gives the asymptotic properties for eigenfunctions and eigenvalues.

2.3.1 Functional Principal Component Analysis

Suppose we have a collection of functional curves $\{X_i(t), i = 1 \cdots, n; t \in T\}$, which are random samples from a smooth random function $X(t)$. We write its mean function as $\mathrm{E}X(t) = \mu(t)$, and covariance function as $\mathrm{Cov}\{X(s), X(t)\} = G(s,t)$. By the Karhunen-Loève theorem, we can further assume there is a linear combination of orthogonal eigenfunctions for individual curve representation, $X_i(t) = \mu(t) + \sum_q \xi_{iq} v_q(t), t \in T$, where ξ_{iq} are considered as independent random variables with $\mathrm{E}\xi_{iq} = 0, \mathrm{Var}(\xi_{iqc}) = \lambda_q$. In classical FPC analysis, covariance function $G(s,t)$ can be orthogonal expanded in terms of eigenvalues λ_q, and eigenfunctions v_q, that is $G(s,t) = \sum_q \lambda_q v_q(t) v_q(s)$. We assume λ_q is a nondecreasing sequence, $\lambda_q \geq \lambda_{q+1}, q \geq 1$; and the sum of the sequence is finite, $\sum_q \lambda_q < \infty$.

In the model setting, we model functional data in a closed and bounded interval T. This conventional time index variable does not necessary mean that our model is limited to time intervals; it can also be equally adapted to other

variables besides time, to widen its applications. For the observed data, we assume there are uncorrelated measurement errors with mean 0 and variance σ^2 for all the curves. Suppose the observations are $\{(y_{ij}, t_{ij}), j = 1, \cdots, N_i; i = 1, \cdots, n\}$, where $y_{ij} \equiv y(t_{ij})$. In the i^{th} curve, we have

$$y_{ij} = \mu(t_{ij}) + \sum_{q=1}^{\infty} \xi_{iq} v_q(t_{ij}) + \epsilon_{ij}, \tag{2.13}$$

where $\mathrm{E}(\epsilon_{ij}) = 0, \mathrm{Var}(\epsilon_{ij}) = \sigma^2$. Denote $\Sigma(s,t) = \mathrm{Cov}(y(s), y(t))$, then it is obvious that $\Sigma(t,t) = G(t,t) + \sigma^2$, and $\Sigma(s,t) = G(s,t)$, if $s \neq t$. We consider $N_i \equiv N$ as fixed for balanced data. Note that this method can be applied to sparse and irregular designs.

The first step is to estimate the mean function $\mu(t)$. Since $\mu(t)$ does not have a specific form, we introduce local linear regression (Fan and Gijbels, 1996), a well established nonparametric smoothing technique. Using a Taylor's expansion, $\mu(t)$ can be approximated by a linear function in a neighborhood of t_0:

$$\mu(t) \approx \beta_0 + \beta_1(t - t_0) \equiv \mathbf{t}^T \boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$. Let $K_h(\cdot) = h^{-1}K(\cdot/h)$ be a rescaled kernel for a kernel function $K(\cdot)$ and a bandwidth h. We apply a local linear smoother on the pooled data from all curves, minimizing

$$\sum_{i=1}^{n} \sum_{j=1}^{N_i} \{y_{ij} - \beta_0 - \beta_1(t_{ij} - t_0)\}^2 K_{h_1}(t_{ij} - t_0), \tag{2.14}$$

with respect to β_0 and β_1. Let $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_0, \tilde{\beta}_1)$ be a solution of maximizing (2.14), then we have $\tilde{\mu}(t_0) = \tilde{\beta}_0$. Once the estimates of mean functions are obtained, covariance functions can be estimated by nonparametric surface smoothing mehtod. Follow Staniswalis and Lee (1998) and Yao et al. (2005), we estimate

$G_c(s,t)$ by minimizing

$$\sum_{i=1}^{n} \sum_{1\le j\neq l\le N} [G_i(t_{ij},t_{il}) - g(\boldsymbol{\beta},s,t,t_{ij},t_{il})]^2 K_{h_2}(t_{ij}-s,t_{il}-t), \tag{2.15}$$

where $G_i(t_{ij},t_{il}) = (Y_{ij} - \tilde{\mu}_c(t_{ij}))(Y_{il} - \tilde{\mu}_c(t_{il}))$ are the 'raw elements' of the covariance function, and $K_{h_2}(t_{ij}-s,t_{il}-t) = K_{h_2}(t_{ij}-s)K_{h_2}(t_{il}-t)$. Note that the diagonal elements of covariance function are $\Sigma(t,t) = G(t,t) + \sigma^2$. Thus, the diagonal of the 'raw elements' should be excluded for covariance function estimation. For a local linear surface smoother, we choose $g(\boldsymbol{\beta},s,t,t_{ij},t_{il}) = \beta_0+\beta_1(t_{ir}-s)+\beta_2(t_{il}-t)$, and minimize (2.15) with respect to $\boldsymbol{\beta} = (\beta_0,\beta_1,\beta_2)^T$, then we have $\hat{G}(s,t) = \hat{\beta}_0$. The eigenvalues and eigenfunctions can be derived by

$$\int_T \hat{G}(s,t)\hat{v}_q(s)ds = \hat{\lambda}_q\hat{v}_q(t), \tag{2.16}$$

where the eigenfunctions $\hat{v}_q(t)$ satisfy $\int_T \hat{v}_q^2(t)dt = 1$, and $\int_T \hat{v}_p(t)\hat{v}_{qc}(t)dt = 0$ if $p \neq q$. The estimation is based on an orthogonal expansion of $\hat{G}(s,t)$, as illustrated in Rice and Silverman (1991). The projection scores ξ_{iqc} is then estimated by $\hat{\xi}_{iq} = \int_T (y(t) - \tilde{\mu}(t))\hat{v}_q(t)dt$.

The functional PCA method was well studied in Ramsay and Silverman (1997). Principal component analysis for sparse functional and longitudinal data is developed in James *et al.* (2000). The asymptotic results can be found at Yao *et al.* (2005) and Hall *et al.* (2006) for principal component functions, Yao (2007) for the local polynomial estimators of mean function and covariance function.

2.3.2 Clustering Analysis for Functional Data

Classical methods for functional data clustering include *regularization* methods and *filtering* methods. Regularization methods transform the data at

fine grid points of the time interval, and then clustering methods such as K-mean algorithm and EM algorithm are applied to the regularized data. Filtering methods project the observations onto a finite basis, and then the clustering are based on basis coefficients. James and Sugar (2003) provides a mixture-model-based clustering approach to sparse functional data, where curves are represented by cubic splines. Sugar and James (2003) discuss how to select the number of cluster in functional clustering analysis. In genetic research, model based functional clustering methods have been used in Luan and Li (2003), Luan and Li (2004). Spline bases are used to model the mean function in their methods. Bayesian approaches for functional data clustering are studied in Heard et al. (2006), and Ma and Zhong (2008).

Chapter 3

Nonparametric Mixture of Regression Models

3.1 Introduction

In this chapter, we propose *nonparametric mixtures of regression models*. Compared with finite mixture of linear regression models, the newly proposed models relax the linearity assumption on the regression function, and allow that the regression function in each of components is an unknown but smooth function of its covariates. We consider the situation in which the mixing proportion, the mean functions and the variance functions all are nonparametric. Using a kernel regression technique, we develop an estimation procedure for the unknown functions in the nonparametric mixture of regression models via local likelihood approach. The sampling properties of the proposed estimation procedure are investigated. We derive the asymptotic bias and variance of the local likelihood estimates, and establish its asymptotic normality.

Although one may naively implement an EM algorithm for maximizing the local likelihood function. However, it is desirable to estimate the whole curves by evaluating the resulting estimate over set of grid points. The naive implementation of the EM algorithm does not ensure that the component labels will match correctly at different grid points. This is similar to label switching problem in older applications of mixture modeling. We modify the EM algorithm (Dempster, Laird and Rubin, 1977) to simultaneously maximize the local likeli-

hood functions for the proposed nonparametric mixture of regression model at set of grid points. The modified EM algorithm works well in our simulation and in a real data example. We further study the ascent property of the proposed EM algorithm.

We derive a standard error formula for the resulting estimate by the conventional sandwich formula. A bandwidth selector is proposed for the local likelihood estimate using a multi-fold cross-validation method. A simulation study is conducted to examine the performance of the proposed procedures and test the accuracy of the proposed standard error formula. We further demonstrate the proposed model and estimation procedure by an empirical analysis of US housing price index data.

3.2 Estimation Procedure and its Sampling Properties

Suppose that $\{(x_i, y_i), i = 1, \cdots, n\}$ are random samples from the population (X, Y). Throughout this chapter, we assume X is univariate. The proposed methodology and theoretical results can be extended to multivariate X, but the extension is less useful due to the "curse of dimensionality". Let $\mathcal{C}$ be a latent class variable, and assume that conditioning on $X = x$, $\mathcal{C}$ has a discrete distribution $P(\mathcal{C} = c|X = x) = \pi_c(x)$ for $c = 1, 2, \cdots, C$. Conditioning on $\mathcal{C} = c$ and $X = x$, Y follows a normal distribution with mean $m_c(x)$ and variance $\sigma_c^2(x)$, where $m_c(\cdot)$ and $\sigma_c(\cdot)$ are unknown but smooth functions. In other words, conditioning on $X = x$, the response variable Y follows a finite mixture of normals

$$\sum_{c=1}^{C} \pi_c(x) N\{m_c(x), \sigma_c^2(x)\}. \tag{3.1}$$

In this chapter we assume that C is fixed, and refer to model (3.1) as a nonparametric finite mixture of regression models because $\pi_c(\cdot)$, $m_c(\cdot)$ and $\sigma_c^2(\cdot)$ are all

nonparametric. When $\pi_c(x)$ and $\sigma_c^2(x)$ are constants, and $m_c(x)$ is linear in x, model (3.1) reduces to a finite mixture of linear regression models (Goldfeld and Quandt, 1976). When $C = 1$, model (3.1) is a nonparametric regression models. See Chapter 2 and Fan and Gijbels (1996) for a review of nonparametric regression model. Thus, model (3.1) can be regarded as a natural extension of nonparametric models and finite mixture of linear regression models. It is easy to adapt the conventional constraints imposed on the finite mixture of linear models for the proposed models so that they are identifiable, and the corresponding likelihood function is bounded. For further references on these two issues, see Hening (2000) and Hathaway (1985).

Since $\pi_c(x)$, $m_c(\cdot)$ and $\sigma_c^2(\cdot)$ are nonparametric, we need nonparametric smoothing techniques for (3.1). In this thesis, we will employ kernel regression techniques for model (3.1). Suppose we want to estimate the unknown functions at x_0. In kernel regression, we first use local constants π_{c0}, σ_{c0}^2, and m_{c0} to approximate $\pi_c(x_0)$, $\sigma_c^2(x_0)$, and $m_c(x_0)$. Let $K_h(\cdot) = h^{-1}K(\cdot/h)$ be a rescaled kernel for a kernel function $K(\cdot)$ and a bandwidth h. Further, denote $\phi(y|\mu, \sigma^2)$ to be the density function $N(\mu, \sigma^2)$. Then, the corresponding local log-likelihood function for data $\{(x_i, y_i), i = 1, 2, \cdots, n\}$ is

$$\ell_n(\pi, \boldsymbol{\sigma}_0^2, \mathbf{m}_0) = \sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \pi_{c0} \phi(y_i | m_{c0}, \sigma_{c0}^2) \right\} K_h(x_i - x_0), \tag{3.2}$$

where $\mathbf{m}_0 = (m_{10}, \cdots, m_{C0})^T$, $\boldsymbol{\sigma}_0^2 = (\sigma_{10}^2, \cdots, \sigma_{C0}^2)^T$, $\pi = (\pi_{10}, \cdots, \pi_{C-1,0})^T$, and $\pi_{C0} = 1 - \sum_{c=1}^{C-1} \pi_{c0}$. One may also apply local linear regression techniques for estimation of $\pi_c(\cdot)$, $m_c(\cdot)$, and $\sigma_c^2(\cdot)$. Local linear regression has several nice properties. It is design-adaptive and possesses high statistical efficiency in an asymptotic minimax sense (Fan, 1993). It automatically corrects edge effects (Fan and Gijbels, 1992; Cheng, Fan and Marron, 1997). However, we do not have

a closed form solution for local linear regression of variance function in normal likelihood function setting. See Fan and Yao (1998) for estimation of conditional variance function under the setting of a simple nonparametric regression model. In this chapter, we consider local constant estimation because it has a closed form solution, and because it is convenient for deriving asymptotic properties.

Let $\{\tilde{\pi}, \tilde{\boldsymbol{\sigma}}^2, \tilde{\mathbf{m}}\}$ be the solution of maximizing the local likelihood function (3.2). Then the estimates for $\pi_c(x_0)$, $\sigma_c^2(x_0)$, and $m_c(x_0)$ are

$$\tilde{\pi}_c(x_0) = \tilde{\pi}_{c0}, \quad \tilde{\sigma}_c^2(x_0) = \tilde{\sigma}_{c0}^2, \quad \text{and} \quad \tilde{m}_c(x_0) = \tilde{m}_{c0}.$$

3.2.1 Asymptotic Properties

We next study the asymptotic properties of $\tilde{\pi}_c(x_0)$, $\tilde{\sigma}_c^2(x_0)$ and $\tilde{m}_c(x_0)$. Let $\boldsymbol{\theta} = (\pi^T, \boldsymbol{\sigma}^{2T}, \mathbf{m}^T)^T$, and denote

$$\eta(y|\boldsymbol{\theta}) = \sum_{c=1}^{C} \pi_c \phi\left\{y|m_c, \sigma_c^2\right\}, \quad \text{and} \quad \ell(\boldsymbol{\theta}, y) = \log \eta(y|\boldsymbol{\theta}).$$

Let $\boldsymbol{\theta}(x_0) = \{\pi^T(x_0), \boldsymbol{\sigma}^2(x_0)^T, \mathbf{m}(x_0)^T\}^T$, and define

$$\eta_1\{\boldsymbol{\theta}(x), y\} = \frac{\partial \eta\{y|\boldsymbol{\theta}(x)\}}{\partial \boldsymbol{\theta}}, \quad \eta_2\{\boldsymbol{\theta}(x), y\} = \frac{\partial^2 \eta\{y|\boldsymbol{\theta}(x)\}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T},$$
$$q_1\{\boldsymbol{\theta}(x), y\} = \frac{\partial \ell\{\boldsymbol{\theta}(x), y\}}{\partial \boldsymbol{\theta}}, \quad q_2\{\boldsymbol{\theta}(x), y\} = \frac{\partial^2 \ell\{\boldsymbol{\theta}(x), y\}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T},$$
$$\mathcal{I}(x_0) = -\mathrm{E}[q_2\{\boldsymbol{\theta}(X), Y\}|X = x_0], \quad \text{and} \quad \Lambda(x) = \int_Y q_1\{\boldsymbol{\theta}(x_0), y\}\eta\{y|\boldsymbol{\theta}(x)\}dy.$$

Denote $\gamma_n = (nh)^{-1/2}$,

$$\hat{m}_c^* = \sqrt{nh}\{\tilde{m}_{c0} - m_c(x_0)\},$$
$$\tilde{\sigma}_c^{2*} = \sqrt{nh}\{\tilde{\sigma}_{c0}^2 - \sigma_c^2(x_0)\},$$
$$\tilde{\pi}_c^* = \sqrt{nh}\{\tilde{\pi}_{c0} - \pi_c(x_0)\},$$
$$\tilde{\pi}_C^* = \sqrt{nh}\{\tilde{\pi}_{C0} - \pi_C(x_0)\} = \sqrt{nh}[1 - \sum_{c=1}^{C-1}\{\tilde{\pi}_{c0} - \pi_c(x_0)\}].$$

Let $\tilde{\mathbf{m}}^* = (\tilde{m}_1^*, \cdots, \tilde{m}_C^*)^T$, $\tilde{\boldsymbol{\sigma}}^{2*} = (\tilde{\sigma}_1^{2*} \cdots, \tilde{\sigma}_C^{2*})^T$, and $\tilde{\pi}^* = (\tilde{\pi}_1^*, \cdots, \tilde{\pi}_{C-1}^*)^T$, and $\tilde{\boldsymbol{\theta}}^* = \{(\tilde{\pi}^*)^T, (\tilde{\boldsymbol{\sigma}}^{2*})^T, (\tilde{\mathbf{m}}^*)^T\}^T$. The asymptotic bias, variance, and normality of the resulting estimate are given in the following theorem. The proof is given in section 3.5.

Theorem 1. *Suppose that conditions (A)—(F) in section 5 hold. Then it follows that*

$$\sqrt{nh}\{\gamma_n \tilde{\boldsymbol{\theta}}^* - \mathcal{B}(x_0) + o_p(h^2)\} \xrightarrow{D} N\left\{0, f^{-1}(x_0)\nu_0 \mathcal{I}^{-1}(x_0)\right\},$$

where $f(\cdot)$ is the marginal density function of X,

$$\mathcal{B}(x_0) = \mathcal{I}^{-1}(x_0)\left\{\frac{f'(x_0)\Lambda'(x_0)}{f(x_0)} + \frac{1}{2}\Lambda''(x_0)\right\}\kappa_2 h^2,$$

$\kappa_l = \int u^l K(u)\, du$, and $\nu_l = \int u^l K^2(u)\, du$.

3.2.2 An Effective EM Algorithm

For a given x_0, one may maximize the local likelihood function (3.2) using an EM algorithm easily. However, in practice we typically want to evaluate the estimated functions at a set of grid points over an interval of x. This requires us to maximize the local likelihood function (3.2) at different grid points. This imposes some challenges because the labels in the EM algorithm may change at different grid points, and therefore, the resulting estimated curve may be mixed up. This is similar to the label switching problem occurs when one conducts a bootstrap to get a confidence interval for parameters in a mixture model. Thus, a naive implementation of the EM algorithm may fail to yield smooth estimated curves. An illustration is given in Figure 3.1.

In this section, we propose an effective EM algorithm to deal with this issue. The key idea is that in the M-step of the EM algorithm, we update the

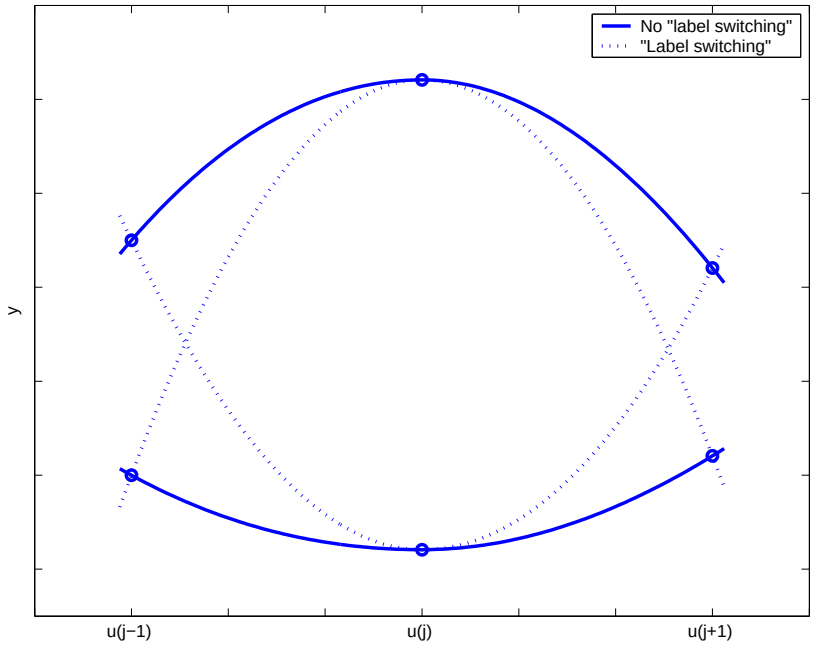


Figure 3.1: Naive implementation of EM algorithm at each location may lead to"label switching" type problem, which yields possible solutions in both solid and dash lines.

estimated curves at all grid points for the given label in the E-step. Define Bernoulli random variables

$$z_{ic} = \begin{cases} 1, & \text{if } (x_i, y_i) \text{ is in the } c^{th} \text{ group,} \\ 0, & \text{otherwise.} \end{cases}$$

and let $\mathbf{z}_i = (z_{i1}, \cdots, z_{iC})^T$. The complete data are $\{(x_i, y_i, \mathbf{z}_i), i = 1, 2, \cdots, n\}$, and the complete log-likelihood function is

$$\sum_{i=1}^{n} \sum_{c=1}^{C} z_{ic} \left[\log \pi_c(x_i) + \log \phi\{y_i | m_c(x_i), \sigma_c^2(x_i)\} \right].$$

In the l-th step of the EM algorithm iteration, we have $m_c^{(l)}(\cdot)$, $\sigma_c^{2(l)}(\cdot)$, and $\pi^{(l)}(\cdot)$. In the E-step, the expectation of the latent variable z_{ic} is given by

$$r_{ic}^{(l)} = \frac{\pi_c^{(l)}(x_i)\phi\{y_i | m_c^{(l)}(x_i), \sigma_c^{2(l)}(x_i)\}}{\sum_{c=1}^{C} \pi_c^{(l)}(x_i)\phi\{y_i | m_c^{(l)}(x_i), \sigma_c^{2(l)}(x_i)\}}. \tag{3.3}$$

Let $\{u_1, \cdots, u_N\}$ be a set of grid points at which the estimated functions are evaluated, where N is the number of grid points. In the M-step, we maximize

$$\sum_{i=1}^{n}\sum_{c=1}^{C} r_{ic}^{(l)} \left[\log \pi_{c0}(x_0) + \log \phi\{y_i | m_{c0}(x_0), \sigma_{c0}^2(x_0)\}\right] K_h(x_i - x_0), \tag{3.4}$$

for $x_0 = u_i$, $i = 1, \cdots, N$. In practice, if n is not very large, one may directly set the observed $\{x_1, \cdots, x_n\}$ to be the grid points. In such case, $N = n$.

The maximization (3.4) is equivalent to maximizing for $c = 1, \cdots, C$,

$$\sum_{i=1}^{n} r_{ic}^{(l)} \log \pi_{c0}(x_0) K_h(x_i - x_0), \tag{3.5}$$

and

$$\sum_{i=1}^{n} r_{ic}^{(l)} \log \phi\{y_i | m_{c0}(x_0), \sigma_{c0}^2(x_0)\} K_h(x_i - x_0), \tag{3.6}$$

separately. The solution for (3.5) is, for $x_0 \in \{u_j, j = 1, \cdots, N\}$,

$$\pi_{c0}^{(l+1)}(x_0) = \frac{\sum_{i=1}^{n} r_{ic}^{(l)} K_h(x_i - x_0)}{\sum_{i=1}^{n} K_h(x_i - x_0)}. \tag{3.7}$$

The closed form solution for (3.6) is, for $x_0 \in \{u_j, j = 1, \cdots, N\}$,

$$m_{c0}^{(l+1)}(x_0) = \sum_{i=1}^{n} w_{ci}^{(l)}(x_0) y_i / \sum_{i=1}^{n} w_{ci}^{(l)}(x_0), \tag{3.8}$$

$$\sigma_{c0}^{2(l+1)}(x_0) = \frac{\sum_{i=1}^{n} w_{ci}^{(l)}(x_0)(y_i - m_{c0}^{(l+1)}(x_0))^2}{\sum_{i=1}^{n} w_{ci}^{(l)}(x_0)}, \tag{3.9}$$

where $w_{ci}^{(l)}(x_0) = r_{ic}^{(l)} K_h(x_i - x_0)$. Furthermore, we update $\pi_c(x_i)$, $m_c(x_i)$ and $\sigma_c^2(x_i)$, $i = 1, \cdots, n$ by linearly interpolating $\pi_c^{(l+1}(u_j)$, $\beta_{c0}^{(l+1)}(u_j)$ and $\sigma_{c0}^2(u_j)$, $j = 1, \cdots, N$, respectively. With initial values of π_c, $m_c(\cdot)$ and $\sigma^2(\cdot)$, the proposed estimation procedure is summarized in the following algorithm.

An EM algorithm:

Initial Value: Obtain an initial value for $\pi_c(\cdot)$, $m_c(\cdot)$ and $\sigma_c^2(\cdot)$ by conducting a mixture of polynomial regressions. Denote the initial value by $\pi_c^{(1)}(\cdot)$, $m_c^{(1)}(\cdot)$ and $\sigma_c^{2(1)}(\cdot)$. Set $l = 1$.

E-step: Use (3.3) to calculate $r_{ic}^{(l)}$ for $i = 1, \cdots, n$, and $c = 1, \cdots, C$.

M-step: For $c = 1, \cdots, C$ and $j = 1, \cdots, N$, evaluate $\pi_c^{(l+1)}(u_j)$ in (3.7), $m_c^{(l+1)}(u_j)$ in (3.8) and $\sigma_{c0}^{2(l+1)}(u_j)$ in (3.9). Further obtain $\pi_c^{(l+1)}(x_i)$, $m_c^{(l+1)}(x_i)$ and $\sigma_c^{2(l+1)}(x_i)$ using linear interpolation.

Iteratively update the E-step and the M-step with $l = 1, 2, \cdots$, until the algorithm converges.

It is well known that an ordinary EM algorithm for parametric models possesses an ascent property, which is a desired property. The ascent property guarantees that the likelihood function is nondecreasing as we update the parameters in the EM iterations. The effective EM algorithm can be regarded as an extension of the EM algorithm from parametric models to nonparametric models. Thus, it is of interest to study whether the modified EM algorithm still preserves the ascent property.

Let $\boldsymbol{\theta}^{(l)}(\cdot) = \{\pi^{(l)}(\cdot), \boldsymbol{\sigma}^{2(l)}(\cdot), \mathbf{m}^{(l)}(\cdot)\}$ be the l-th step estimated functions in the proposed EM algorithm. We rewrite the local likelihood function (3.2) as

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, y_i) K_h(x_i - x_0). \tag{3.10}$$

Theorem 2. *For any given point x_0 in the interval of x, suppose that $\boldsymbol{\theta}^{(l)}(\cdot)$ has continuous first derivative. as $h \to 0$ $n \to \infty$, it follows*

$$\liminf_{n\to\infty} \; n^{-1} \left[\ell_n\{\boldsymbol{\theta}^{(l+1)}(x_0)\} - \ell_n\{\boldsymbol{\theta}^{(l)}(x_0)\} \right] \geq 0 \tag{3.11}$$

in probability.

The proof of Theorem 2 is given in Section 3.4. Theorem 2 implies that when the sample size n is large enough, the proposed algorithm possess the ascent property at any given x_0. We evaluate estimated curve at a set of grid points $\{u_1, \cdots, u_N\}$. For large n, we have

$$\ell_n\{\boldsymbol{\theta}^{(l+1)}(u_j)\} - \ell_n\{\boldsymbol{\theta}^{(l)}(u_j)\} \geq 0.$$

3.3 Simulation and Application

In this section, we address some practical implementation issues such as standard error formula and bandwidth selection for nonparametric mixture of regression model. To assess the performance of the estimates of the unknown regression functions $m_c(x)$, we consider the square root of the average square errors (RASE) for mean functions,

$$\text{RASE}_m^2 = N^{-1} \sum_{c=1}^{C} \sum_{j=1}^{N} \{\hat{m}_c(u_j) - m_c(u_j)\}^2,$$

where $\{u_j, j = 1, \cdots, N\}$ is the grid points at which the unknown functions $m_c(\cdot)$ are evaluated. For simplification, the grid points are taken evenly on the range of the x-variable. Similarly, we can define RASE for variance functions $\sigma_c^2(x)$s and proportion functions $\pi_c(x)$s, denoted by RASE_σ and RASE_π, respectively. In simulation, we set $N = 100$.

3.3.1 Standard Error Formula

Define the fitted value for the i-th observation as a weight sum of the estimated means,

$$\hat{y}_i = \sum_{c=1}^{C} r_{ic} \hat{m}_c(x_i),$$

where r_{ic} are the posterior of the identities when the effective EM algorithm converges. Then, the residual is $e_i = y_i - \hat{y}_i$. Rewrite the estimate of $m_c(x)$ in the proposed algorithm as

$$\hat{m}_c(x) = (\mathbf{E}^T W_c \mathbf{E})^{-1} \mathbf{E}^T W_c \mathbf{y},$$

where $\mathbf{E}$ is a $n \times 1$ vector with all entries equal to 1; $W_c = \text{diag}\{w_{c1}, \cdots, w_{cn}\}$ with $w_{ci} = r_{ic} K_h(x_i - x_0)$. We consider the following approximate standard error formula for $\hat{m}_c(x)$:

$$\widehat{\text{Var}}\{\hat{m}_c(x)\} = (\mathbf{E}^T W_c \mathbf{E})^{-1} \mathbf{E}^T W_c \widehat{\text{Cov}}(\mathbf{y}) W_c \mathbf{E} (\mathbf{E}^T W_c \mathbf{E})^{-1}, \tag{3.12}$$

where $\widehat{\text{Cov}}(\mathbf{y}) = \text{diag}\{e_1^2, e_2^2, \cdots, e_n^2\}$, a diagonal matrix consisting of the residuals squares e_i^2. Furthermore, (3.12) can be written as

$$\widehat{\text{Var}}\{\hat{m}_c(x)\} = \frac{\sum_{i=1}^n w_{ic}^2 e_i^2}{(\sum_{i=1}^n w_{ic})^2}. \tag{3.13}$$

The accuracy of this formula will be tested in section 3.3.3.

3.3.2 Bandwidth Selection

Bandwidth selection is fundamental to nonparametric smoothing. In practice, data driven methods such as cross-validation (CV) can be used to choose the bandwidth. Denote $\mathcal{D}$ to be the full data set. We then partition $\mathcal{D}$ into a training set $\mathcal{R}_j$ and test set $\mathcal{T}_j$, $\mathcal{D} = \mathcal{T}_j \cup \mathcal{R}_j$ $j = 1, \cdots, J$. We use the training set $\mathcal{R}_j$ to obtain the estimates $\{\hat{m}_c(\cdot), \hat{\sigma}_c^2(\cdot), \hat{\pi}_c(\cdot)\}$. Then we can estimate $m_c(x)$, $\sigma_c^2(x)$ and $\pi_c(x)$ for the data points belong to the corresponding

test set. For $(x_l, y_l) \in \mathcal{T}_j$,

$$\begin{aligned}
\hat{m}_c(x_l) &= \frac{\sum_{\{i:x_i \in \mathcal{R}_j\}} r_{ic} K_h(x_i - x_l) y_i}{\sum_{\{i:x_i \in \mathcal{R}_j\}} r_{ic} K_h(x_i - x_l)}, \\
\hat{\sigma}_c^2(x_l) &= \frac{\sum_{\{i:x_i \in \mathcal{R}_j\}} r_{ic} K_h(x_i - x_l)(y_i - \hat{m}_c(x_i))^2}{\sum_{\{i:x_i \in \mathcal{R}_j\}} r_{ic} K_h(x_i - x_l)}, \\
\hat{\pi}_c(x_l) &= \frac{\sum_{\{i:x_i \in \mathcal{R}_j\}} r_{ic} K_h(x_i - x_l)}{\sum_{\{i:x_i \in \mathcal{R}_j\}} K_h(x_i - x_l)}.
\end{aligned}$$

Based on the estimated $\hat{m}_c(x_l)$ of test set $\mathcal{T}_j$, we again calculate the posterior memberships in test set $\mathcal{T}_j$. For $(x_l, y_l) \in \mathcal{T}_j, c = 1, \cdots, C$,

$$r_{lc} = \frac{\hat{\pi}_c(x_l)\phi\{y_l|\hat{m}_c(x_l), \hat{\sigma}_c^2(x_l)\}}{\sum_{q=1}^{C} \hat{\pi}_q(x_l)\phi\{y_l|\hat{m}_q(x_l), \hat{\sigma}_q^2(x_l)\}}.$$

Now we can implement regular CV criterion in this mixture model

$$CV = \sum_{j=1}^{J} \sum_{l \in \mathcal{T}_j} (y_l - \hat{y}_l)^2, \tag{3.14}$$

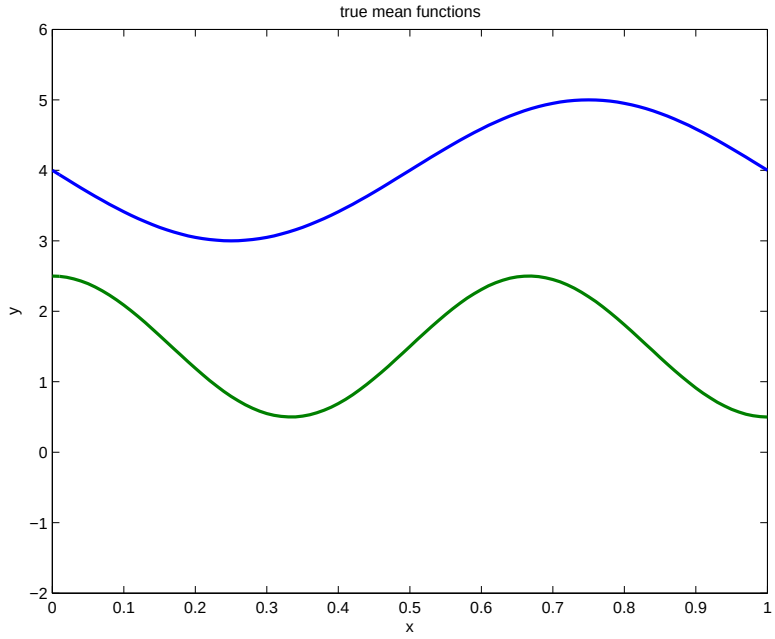
where $\hat{y}_l = \sum_{c=1}^{C} r_{lc}\, \hat{m}_c(x_l)$ is the predicted value of y_l in the test set $\mathcal{T}_j$.

3.3.3 Simulation Study

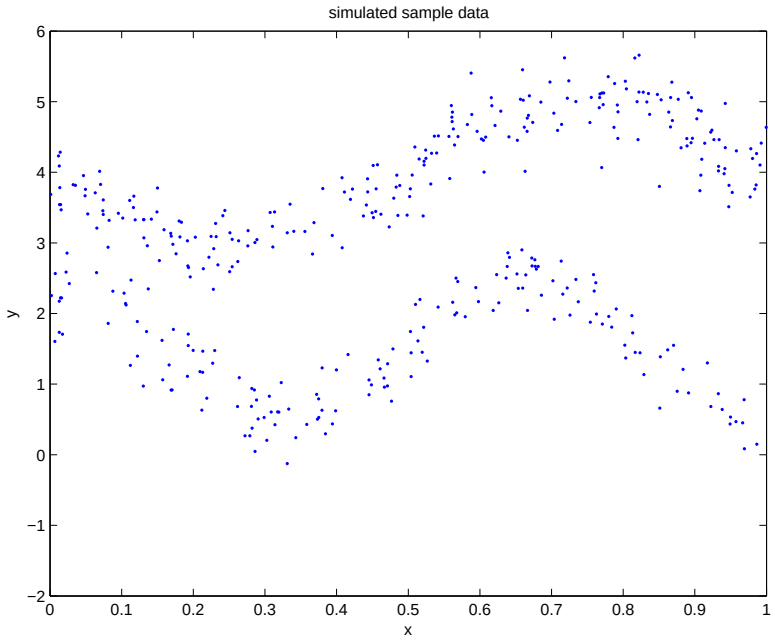
In the following example, we conduct a simulation for a 2-component nonparametric mixture of regressions model with

$$\begin{aligned}
&\pi_1(x) = \exp(0.5x)/\{1 + \exp(0.5x)\}, \text{ and } \pi_2(x) = 1 - \pi_1(x), \\
&m_1(x) = 4 - \sin(2\pi x), \text{ and } m_2(x) = 1.5 + \cos(3\pi x), \\
&\sigma_1(x) = 0.25\exp(0.5x), \text{ and } \sigma_2(x) = 0.3\exp(-0.2x).
\end{aligned}$$

For each of the samples, 500 simulations were conducted with sample sizes $n = 200, 400, 800$. The predictor x is generated from one dimension uniform distribution in $[0, 1]$. The Epanechnikov kernel is used in our simulation. Figure 3.2 shows the plots of true mean functions with a typical sample data.



(a)



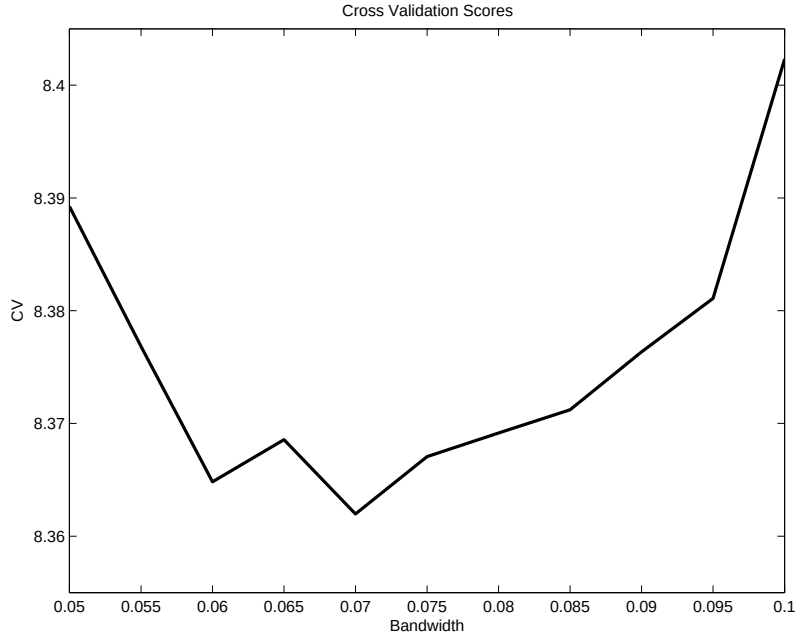
(b)

Figure 3.2: (a) Plot of true mean functions for the two components in the simulation; (b) A typical sample of simulated data (n=400)

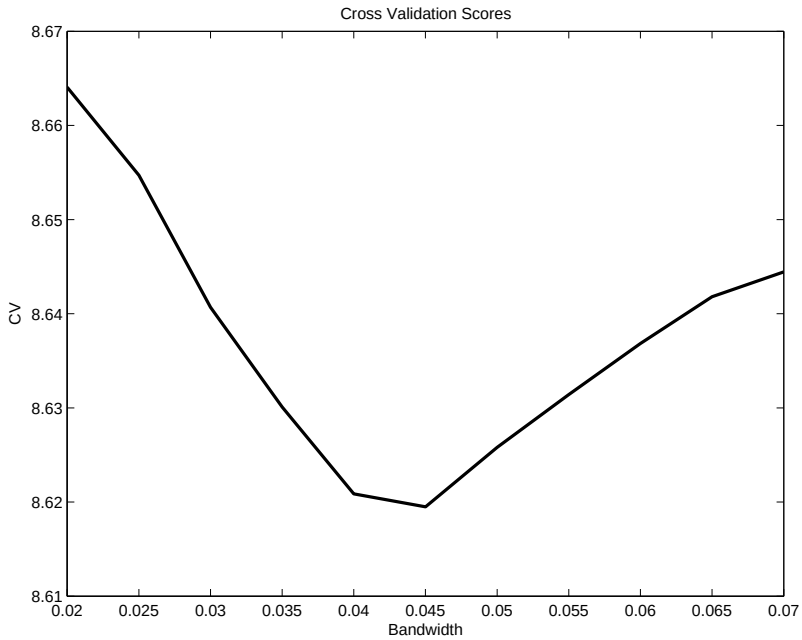
It is well known that the EM algorithm may be trapped by a local maximizer. Thus, the EM algorithm may be sensitive to the initial value. To obtain a good initial value, we first fit a mixture of polynomial regression models, which gives the estimates of mean functions $\bar{m}_c(x)$, and parameters $\bar{\sigma}_c^2$, $\bar{\pi}_c$. Then we set the initial values $m_c^{(1)}(x) = \bar{m}_c(x)$, $\sigma^{2(1)}(x) = \bar{\sigma}_c^2$, and $\pi_c^{(1)}(x) = \bar{\pi}_c$. In our simulation, we first generate several simulation data sets for a given sample size, and then use the CV bandwidth selectors to choose a bandwidth for each data set. This provides us an idea about the optimal bandwidth for a given sample size. To demonstrate that the proposed procedure works quite well for a wide range of bandwidths, we consider three different bandwidths: half of the selected bandwidth, the selected bandwidth, and twice of the selected bandwidth, which corresponds to the under-smoothing, appropriate-smoothing and over-smoothing, respectively. Table 3.1 displays the mean and standard deviation of RASEs over 500 simulations. From Table 3.1, the proposed procedure performs quite well for all three different bandwidths.

We next test the accuracy of the standard error formulas. Table 3.2 summarizes the simulation results for the unknown functions $m_c(x)$ at points $0.25, 0.5, 0.75$. The standard deviation of 500 estimates, denoted by SD, can be viewed as the true standard errors. We then calculate the mean and standard deviation of the estimated standard error using the proposed standard error formulas (3.12), denoted by MSD(Std). The result shows that the proposed sandwich formula works reasonably well because the difference between the true value and the estimate is less than twice of the standard error of the estimate.

We now illustrate the performance of the proposed procedure by using a typical simulated sample, which is selected to be the one with the median of RASE_m in the 500 simulation. For this data set, we use the cross-validation (CV)



(a)



(b)

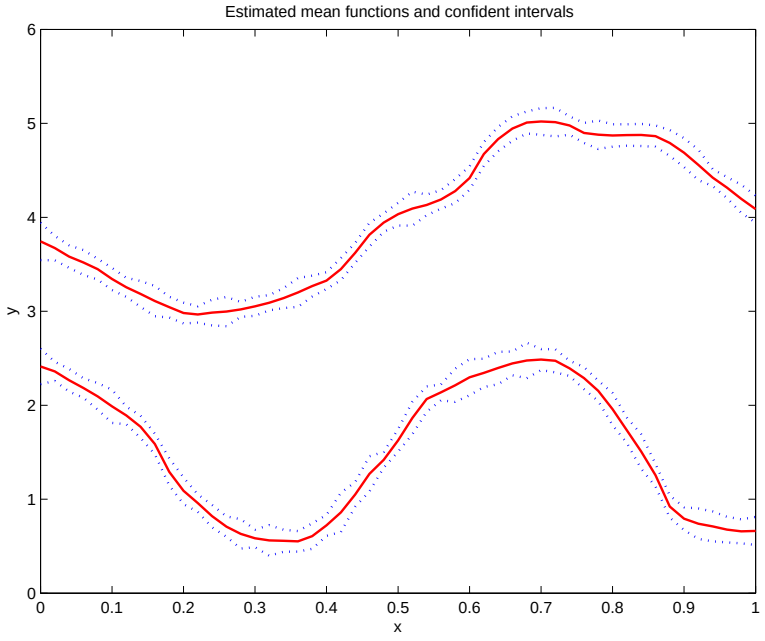
Figure 3.3: Cross-validation error versus the bandwidth: (a) n=400; (b) n=800.

Table 3.1: RASEs: Mean and Standard Deviation

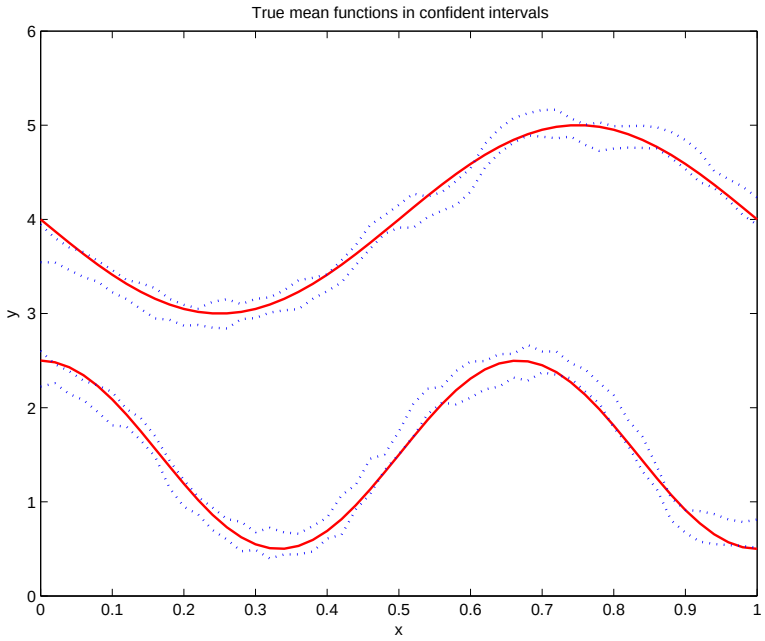
		RASE_m	RASE_σ	RASE_π
n	h	Mean(Std)	Mean(Std)	Mean(Std)
200	0.16	0.1007(0.0476)	0.0081(0.0060)	0.0092(0.0066)
	0.08	0.0330(0.0177)	0.0045(0.0027)	0.0113(0.0042)
	0.04	0.0588(0.0573)	0.0163(0.0572)	0.0203(0.0059)
400	0.12	0.0420(0.0318)	0.0034(0.0004)	0.0053(0.0045)
	0.06	0.0168(0.0065)	0.0027(0.0013)	0.0069(0.0023)
	0.03	0.0320(0.0201)	0.0048(0.0012)	0.0140(0.0033)
800	0.08	0.0112(0.0042)	0.0011(0.0007)	0.0028(0.0016)
	0.04	0.0107(0.0029)	0.0019(0.0005)	0.0051(0.0016)
	0.02	0.0183(0.0030)	0.0037(0.0007)	0.0103(0.0020)

Table 3.2: Standard error of the unknown mean functions

		$m_1(x)$		$m_2(x)$	
n	x	SD	MSD(Std)	SD	MSD(Std)
200 (h=0.08)	0.25	0.0753	0.0691(0.0155)	0.1068	0.0874(0.0219)
	0.50	0.0938	0.0804(0.0185)	0.1315	0.0866(0.0234)
	0.75	0.0977	0.0856(0.0198)	0.1222	0.0860(0.0227)
400 (h=0.06)	0.25	0.0586	0.0580(0.0108)	0.0803	0.0731(0.0152)
	0.50	0.0810	0.0648(0.0121)	0.0921	0.0696(0.0153)
	0.75	0.0735	0.0728(0.0146)	0.0731	0.0710(0.0180)
800 (h=0.04)	0.25	0.0517	0.0511(0.0098)	0.0731	0.0649(0.0113)
	0.50	0.0607	0.0575(0.0109)	0.0713	0.0639(0.0126)
	0.75	0.0653	0.0634(0.0104)	0.0699	0.0628(0.0140)



(a)



(b)

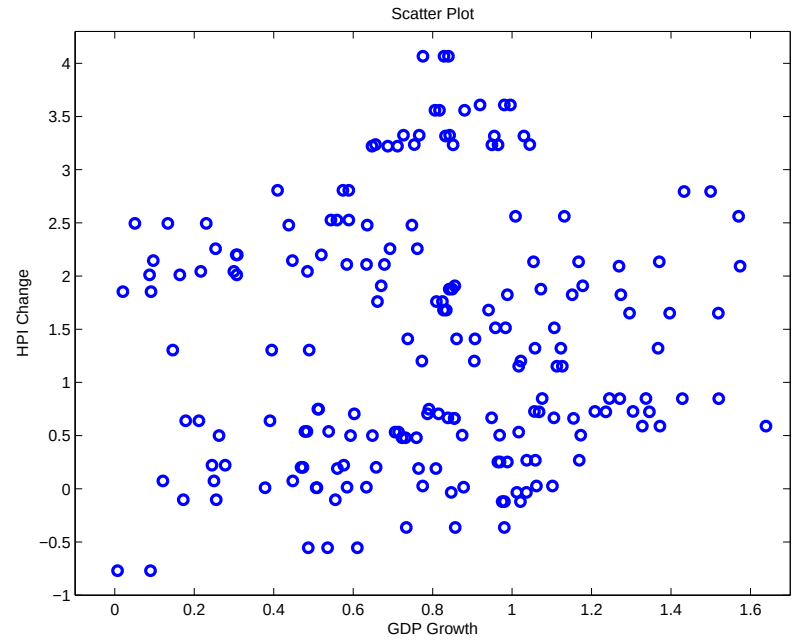
Figure 3.4: (a) The estimated mean functions and 95% pointwise confidence intervals ($n = 400, h = 0.07$); (b) True mean functions and 95% confidence intervals.

criterion proposed in Section 3.2 to select a bandwidth. The cross-validation scores are depicted in Figure 3.3. The CV bandwidth selector yields the bandwidth 0.07 for $n = 400$, and 0.045 for $n = 800$. The resulting estimate with $n = 400$ along with its pointwise confidence interval is depicted in Figure 3.4, from which we can see that the true mean functions lies within the confidence interval. This implies that the proposed estimation procedure performs quite with the moderate sample size.

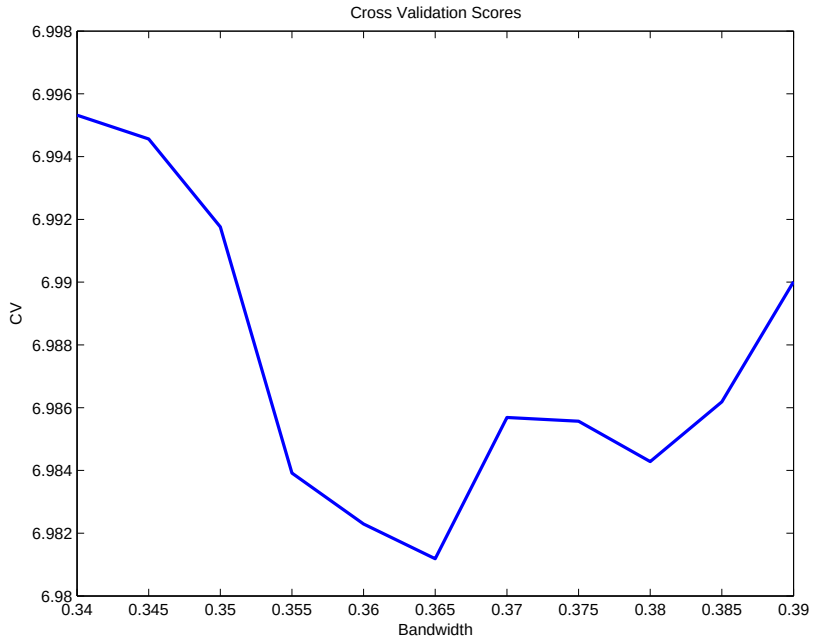
3.3.4 Analysis of US Housing Index Data

We apply the proposed nonparametric mixture of regression model and estimation procedure to analyze a real data set, which contains the SP-Case Shiller House Price Index (HPI) change and United States GDP growth rate from Jan, 1990 to Dec, 2006. It is known that in the literature of economic research, HPI is a measure of a nation's housing cost and GDP is a measure of the size of a nation's economy. It is of interest to investigate the impact of GDP growth rate on HPI change. As expected, the impact of GDP growth rate on HPI may have different pattern in different macroeconomic cycles. In this illustration, we set HPI change to be the response variable, and the GDP growth rate to be predictor. In this analysis, we limit ourselves to the data with positive GDP growth rate. The scatter plot of this data set is depicted in Figure 3.5(a). In our analysis, we consider three component nonparametric mixture of regression models.

We first select the bandwidth by a 5-folder CV selector described in (3.14). The selected optimal bandwidth is 0.365, as shown in Figure 3.5(b). With the optimal bandwidth we fit the data in a 3-component nonparametric mixture of regression models. The estimated mean functions and a hard-clustering result



(a)

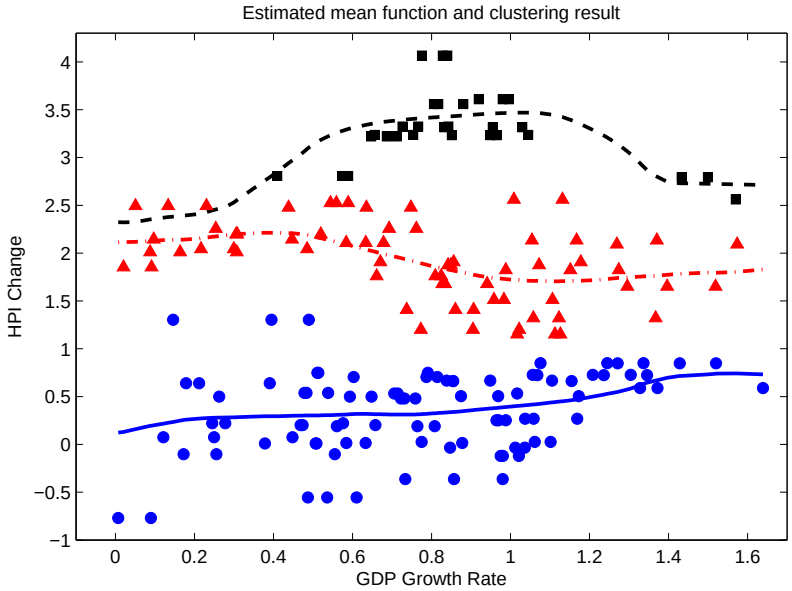


(b)

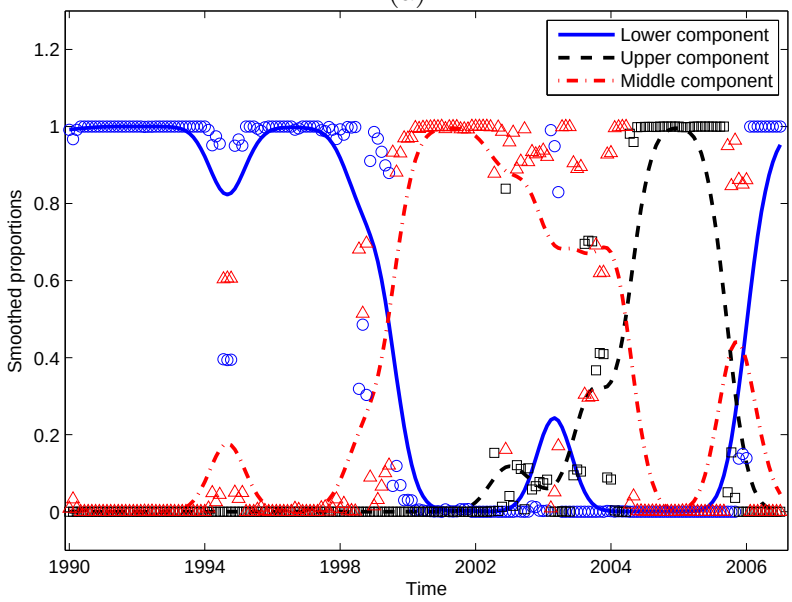
Figure 3.5: (a) Scatterplot of GDP Growth and HPI Change; (b)Cross-validation error versus the bandwidth.

based on posterior estimate of identities are shown in Figure 3.6(a). The dotted points from the lower cluster are mainly from January 1990 to September 1997, and from April 2006 to December 2006. The triangular points in the upper cluster are mainly from October 2003 to September 2005, which is recognized as a serious housing bubble period. In the middle cluster, the observations are mainly from October 1997 to September 2003, during which the economy experienced an internet boom and the following bust. We plot the estimated component identities r_{ic} against time in Figure 3.6(b), along with corresponding smoothed curves. This figure verify our claims that the three components have identified the three distinct time periods. The estimated mixing proportion functions are plotted in Figure 3.7. The 95% pointwise confidence intervals of each component are plotted in Figure 3.8(a). We observe that the mixing proportion function for the "housing bubble component" is close to 0 when GDP growth rate is low (less than 0.4%), and has a dramatic increase as GDP growth rate is modest (between 0.4% and 1.2%). There are overlaps of the confidence intervals (Figure 3.6(b)) in the two upper components, which suggest that they can be considered as one component when the overlaps occur. This result agrees with posterior hard-identity in Figure 3.5(a).

We compare the proposed nonparametric finite mixture of regression model to two other models: a local linear regression model and 3-component mixture of linear regression models. The optimal bandwidth for local linear regression was selected by a plug-in method (Ruppert et al., 1995), and parametric mixture of regression model will be estimated by an EM algorithm. We randomly select 20% of the data to be a test set and the rest as a training set. Estimations for 3 different models are obtained from training set and prediction errors are calculated based on the deviations in test set. Our result shows that



(a)



(b)

Figure 3.6: (a)Clustering result; (b)The estimated component identities against time.

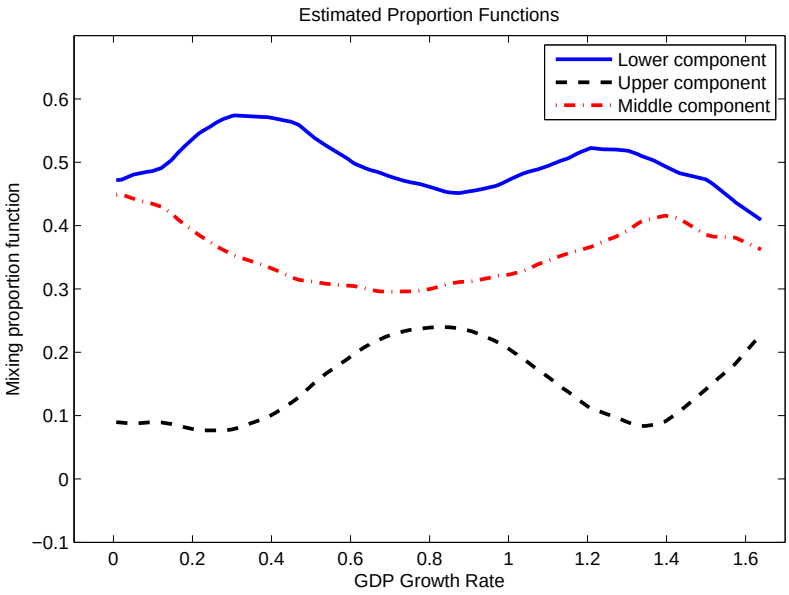
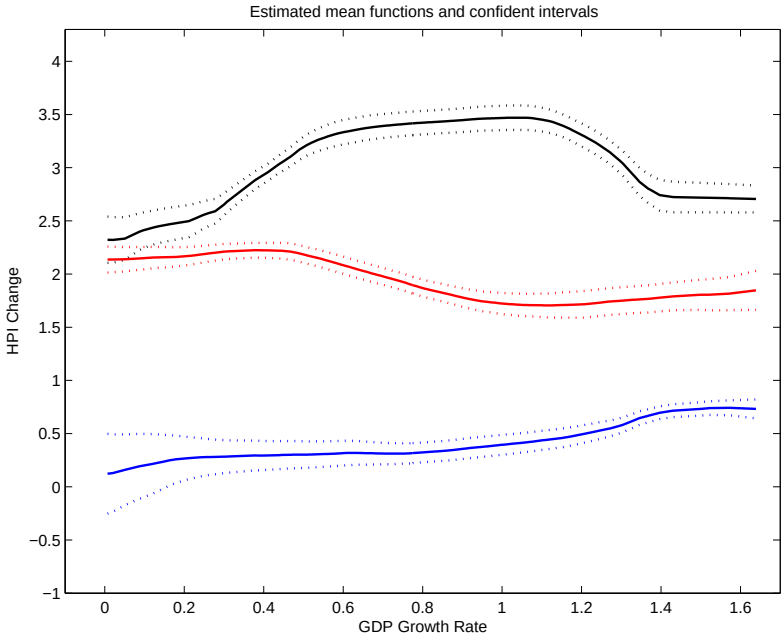
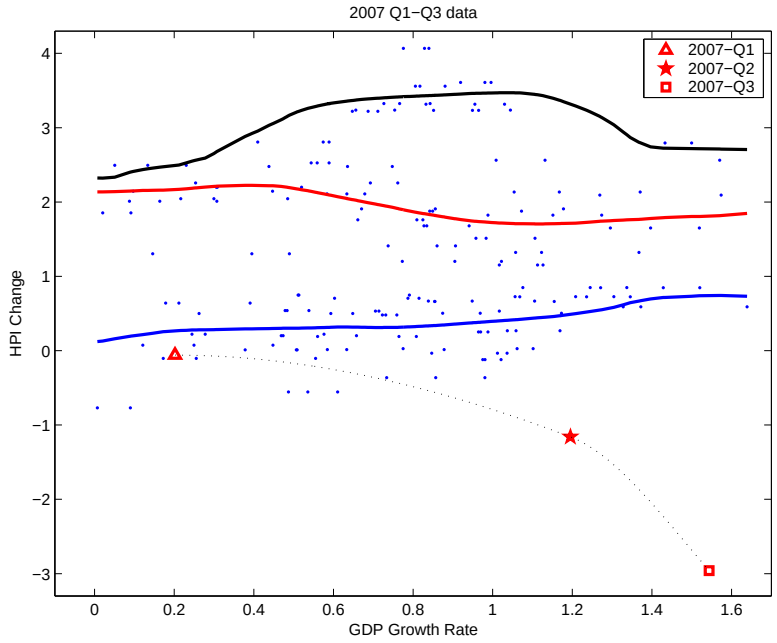


Figure 3.7: Estimated mixing proportion functions.

in 100 simulations, the nonparametric mixture of regression model has average prediction error 1.38×10^{-5}, while average prediction error for local linear regression and a 3-component parametric mixture of regression model are 1.47×10^{-4} and 1.63×10^{-5}. The result means that nonparametric mixture of regression model has 15.4% reduction in prediction error compared to parametric mixture of regression model, and 90.6% reduction in prediction error compared to local linear regression. The improved prediction error demonstrates the advantages of nonparametric finite mixture of regression model in the analysis of US housing data. We finally plot the prediction and true value of first 3 quarters (Q1,Q2, and Q3) of year 2007 data in Figure 3.8(b). From the figure we see that 2007 Q1 data still falls in the cycle of lower HPI change. However, from 2007 Q2 the economic enters a new scenario of decreasing housing price and increasing GDP growth rate which goes out of the historical ranges since 1990.



(a)



(b)

Figure 3.8: (a)95% confidence intervals; (b) True value of HPI change vs GDP growth rate in 2007 Q1-Q3.

3.4 Discussion

In this chapter we proposed nonparametric finite mixture of regression models, and develop estimation procedure for unknown functions. We study the monotone ascent property of the proposed effective EM algorithm and further give asymptotic distribution of the resulting estimates. Young (2007) proposed an "EM-like" algorithm for a mixture of linear regression models with covariate-dependent mixing proportions. This model is similar to a special case of model (3.1) with constant variances and a linear form in the regression functions. However, his estimation procedure does not maximize a local likelihood function.

The performance of the proposed procedures was empirically examined by a Monte Carlo simulation and illustrated by an analysis of a real data example. Throughout the thesis, the computations and calculations are performed in MATLAB, a numerical computing environment and scientific programming language. In general the proposed procedure performs well in simulation and application. To achieve computational stability the bandwidth must not be too small. In our simulation we do observe several computational failures in a 500 simulation with $n = 200$ and $h = 0.04$ (Table 3.1), which did not occur in other settings. Note that in the case where computational failures happen, there are $nh = 8$ expected number of data points within a local window controlled by the bandwidth. This number is larger than 5, which is the number of unknown parameters in local likelihood (3.2) with 2 components. However, for a random sample there may be some locations at which number of data points is less than 5 within the bandwidth, and this may cause a failure in computation. Although we do not have a accurate idea on the general data requirements in an analysis using the proposed model, the mixture of linear regression models may provide a guide on this issue if we look at the dataset locally. Since our estimation proce-

dure depends on the grid points, we may choose grid points that avoid the areas where data points are too sparse.

3.5 Proofs

In this section we outline the key steps of proofs for Theorems 1 and 2. Note that $\boldsymbol{\theta} = (\pi^T, \boldsymbol{\sigma}^{2T}, \mathbf{m}^T)^T$ is a $(3C-1) \times 1$ vector. Whenever necessary, we rewrite $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_{3C-1})^T$ without changing the order of $\pi, \boldsymbol{\sigma}^2$, and $\mathbf{m}$. Correspondingly, we rewrite $\boldsymbol{\eta} = (\boldsymbol{\sigma}^{2T}, \mathbf{m}^T)^T$ as $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_{2C})^T$ if necessary. Otherwise, we will use the same notations as we defined in section 2.

Regularity Conditions

A. The sample $\{(x_i, y_i), i = 1, \cdots, n\}$ is independent and identically distributed from its population (x, y). The support for x, denoted by $\mathcal{X}$, is closed and bounded of R^1. The density $\eta\{y|\boldsymbol{\theta}(x_0)\}$ is identifiable up to a permutation of the mixture components.

B. The marginal density function $f(x)$ is continuous first derivative and positive for $x \in \mathcal{X}$.

C. There exists a function $M(y)$, with $\mathrm{E}\{M(Y)\} < \infty$, such that for all y, and all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}(x_0)$, $|\partial^3 \ell(\boldsymbol{\theta}, y)/\partial\theta_j \partial\theta_k \partial\theta_l| < M(y)$.

D. The unknown functions $\boldsymbol{\theta}(x)$ has continuous second derivative.

E. The kernel function $K(\cdot)$ has a bounded support, and satisfies that

$$\int K(t)dt = 1, \quad \int tK(t)dt = 0, \quad \int t^2 K(t)dt = \kappa_2 > 0.$$

F. The following conditions hold for all i and j.

$$E\left(\left|\frac{\partial \ell(\boldsymbol{\theta}(x_0), Y)}{\partial \theta_j}\right|^3\right) < \infty, \quad \mathrm{E}\left[\left\{\frac{\partial^2 \ell(\boldsymbol{\theta}(x_0), Y)}{\partial \theta_i \partial \theta_j}\right\}^2\right] < \infty.$$

All these conditions are mild conditions and have been used in the literature of local likelihood estimation and mixture models. The following lemma will be used in the proof of Theorem 1.

Lemma 1. *Under Conditions A, C, and F, for x_0 in the interior of $\mathcal{X}$, it holds that*

$$\mathrm{E}[q_1\{\boldsymbol{\theta}(X), Y\}|X = x_0] = 0, \tag{3.15}$$

$$\mathrm{E}[q_2\{\boldsymbol{\theta}(X), Y\}|X = x_0] = -\mathrm{E}[q_1\{\boldsymbol{\theta}(X), Y\}q_1^T\{\boldsymbol{\theta}(X), Y\}|X = x_0]. \tag{3.16}$$

Proof. Conditioning $X = x_0$, Y follows a finite mixtures of normals. Thus, (3.15) holds by some calculations. Furthermore, (3.16) follows by using under regularity conditions C, F and the arguments in Page 39 of McLachlan and Peel (2000) together. This completes the proof of this lemma.

We refer (3.15 and (3.16) to as the local Barlette's first and second identities, respectively. (3.15) implies that $\Lambda(x_0) = 0$.

Proof of Theorem 1. Denote

$$\begin{aligned}
m_c^* &= \sqrt{nh}\{m_{c0} - m_c(x_0)\}, \\
\sigma_c^{2*} &= \sqrt{nh}\{\sigma_{c0}^2 - \sigma_c^2(x_0)\}, \\
\pi_c^* &= \sqrt{nh}\{\pi_{c0} - \pi_c(x_0)\}, \\
\pi_C^* &= \sqrt{nh}\{\pi_{C0} - \pi_C(x_0)\} = \sqrt{nh}[1 - \sum_{c=1}^{C-1}\{\pi_{c0} - \pi_c(x_0)\}].
\end{aligned}$$

Let $\mathbf{m}^* = (m_1^*, \cdots, m_C^*)^T$, $\boldsymbol{\sigma}^{2*} = (\sigma_1^{2*} \cdots, \sigma_C^{2*})^T$, and $\pi^* = (\pi_1^*, \cdots, \pi_{C-1}^*)^T$. Denote $\boldsymbol{\theta}^* = (\pi^{*T}, \boldsymbol{\sigma}^{2*T}, \mathbf{m}^{*T})^T$. Recall that

$$\ell(\boldsymbol{\theta}(x_0), y) = \log \eta\{y|\boldsymbol{\theta}(x_0)\} = \log \left\{ \sum_{c=1}^{C} \pi_c(x_0) \phi\{y|m_c(x_0), \sigma_c^2(x_0)\} \right\}.$$

Let

$$\begin{aligned} \ell(\boldsymbol{\theta}(x_0) + \gamma_n \boldsymbol{\theta}^*, y) &= \log \Bigg\{ \sum_{c=1}^{C} (\pi_c(x_0) + \gamma_n \pi_c^*) \\ &\quad \times \phi\left(y|m_c(x_0) + \gamma_n m_c^*, \sigma_c^2(x_0) + \gamma_n \sigma_c^{2*}\right) \Bigg\}. \end{aligned}$$

Thus, if $\{\tilde{\pi}, \tilde{\boldsymbol{\sigma}}^2, \tilde{\mathbf{m}}\}$ maximizes (3.2), then $\tilde{\boldsymbol{\theta}}^*$ maximizes

$$\ell_n^*(\boldsymbol{\theta}^*) = h \sum_{i=1}^{n} \{\ell(\boldsymbol{\theta}(x_0) + \gamma_n \boldsymbol{\theta}^*, y_i) - \ell(\boldsymbol{\theta}(x_0), y_i)\} K_h(x_i - x_0). \tag{3.17}$$

By the Taylor expansion,

$$\ell_n^*(\boldsymbol{\theta}^*) = \Delta_n \boldsymbol{\theta}^* + \frac{1}{2} \boldsymbol{\theta}^{*T} \Gamma_n \boldsymbol{\theta}^* + \frac{h\gamma_n^3}{6} \sum_{i=1}^{n} R(\boldsymbol{\theta}(x_0), \tilde{\xi}), \tag{3.18}$$

where $\tilde{\xi}$ is a vector between 0 and $\gamma_n \boldsymbol{\theta}^*$, and

$$\begin{aligned} \Delta_n &= \sqrt{\frac{h}{n}} \sum_{i=1}^{n} q_1\{\boldsymbol{\theta}(x_0), y_i\} K_h(x_i - x_0), \\ \Gamma_n &= \frac{1}{n} \sum_{i=1}^{n} q_2\{\boldsymbol{\theta}(x_0), y_i\} K_h(x_i - x_0), \\ R(\boldsymbol{\theta}(x_0), \tilde{\xi}) &= \sum_{j,k,l} \frac{\partial^3 \ell(\boldsymbol{\theta}(x_0) + \tilde{\xi}, y_i)}{\partial\theta_j \partial\theta_k \partial\theta_l} K_h(x_i - x_0). \end{aligned}$$

Denote $\Gamma_n(i,j)$ the (i,j) element of Γ_n. By condition E, it can be shown that

$$\begin{aligned} \mathrm{E}\Gamma_n(i,j) &= \int_Y \int_X \frac{\partial^2 \ell(\boldsymbol{\theta}(x_0), y)}{\partial\theta_i \partial\theta_j} \eta\{y|\boldsymbol{\theta}(x)\} f(x) K_h(x - x_0) dx dy \\ &= f_X(x_0) \int_Y \frac{\partial^2 \ell(\boldsymbol{\theta}(x_0), y)}{\partial\theta_i \partial\theta_j} \eta\{y|\boldsymbol{\theta}(x_0)\} dy + o_p(1). \end{aligned}$$

Therefore, $\mathrm{E}\Gamma_n = -f_X(x_0)\mathcal{I}(x_0) + o_p(1)$. $\mathrm{Var}\{\Gamma_n(i,j)\}$ is dominated by the following term

$$\frac{1}{n}\int_Y\int_X \left\{\frac{\partial^2 \ell(\boldsymbol{\theta}(x_0), y)}{\partial\theta_i \partial\theta_j}\right\}^2 \eta\{y|\boldsymbol{\theta}(x)\} f(x) K_h^2(x - x_0) dxdy,$$

which can be shown to have the order $O_p\{(nh)^{-1}\}$ under condition F. Therefore, we have

$$\Gamma_n = -f_X(x_0)\mathcal{I}(x_0) + o_p(1).$$

By Condition C, the expectation of the absolute value of the last term of (3.18) is bounded by

$$O\left(\gamma_n \mathrm{E}\max_{j,k,l}\left|\frac{\partial^3 \ell(\boldsymbol{\theta}(x_0) + \tilde{\xi}, Y)}{\partial\theta_j\partial\theta_k\partial\theta_l} K_h(x_i - x_0)\right|\right) = O(\gamma_n). \tag{3.19}$$

Thus, the last term of (3.18) is of order $O_p(\gamma_n)$. Therefore, we have

$$\ell_n^*(\boldsymbol{\theta}^*) = \Delta_n \boldsymbol{\theta}^* - \frac{1}{2} f(x_0) \boldsymbol{\theta}^{*T} \mathcal{I}(x_0) \boldsymbol{\theta}^* + o_p(1). \tag{3.20}$$

Using the quadratic approximation lemma (for example, see p.210 of Fan and Gijbels, 1996), we have

$$\hat{\boldsymbol{\theta}}^* = f(x_0)^{-1}\mathcal{I}(x_0)^{-1}\Delta_n + o_p(1). \tag{3.21}$$

To establish asymptotic normality, it remains to calculate the mean and variance of Δ_n, and verify the Lyapounov condition. Note that

$$\begin{aligned}
\mathrm{E}(\Delta_n) &= \sqrt{nh}\int_Y\int_X q_1\{\boldsymbol{\theta}(x_0), y\}\eta\{y|\boldsymbol{\theta}(x)\} f(x) K_h(x - x_0) dxdy \\
&= \sqrt{nh}\int_X \Lambda(x) f(x) K_h(x - x_0) dx.
\end{aligned}$$

Under Conditions C, D and F, $\Lambda(x)$ has continuous second derivative. Thus, using the fact $\Lambda(x_0) = 0$ by Lemma 1 and standard arguments in the kernel regression, it follows that

$$\mathrm{E}(\Delta_n) = \left\{\frac{f'(x_0)\Lambda'(x_0)}{f(x_0)} + \frac{1}{2}\Lambda''(x_0)\right\}\kappa_2 h^2 + o(h^2).$$

For the covariance term of Δ_n, we have

$$\mathrm{Cov}(\Delta_n) \;=\; h\mathrm{E}\left\{q_1\{\boldsymbol{\theta}(x_0),Y\}q_1^T\{\boldsymbol{\theta}(x_0),Y\}K_h^2(X-x_0)\right\}+o_p(1),$$

where its (i,j) element is

$$\begin{aligned}
& h\int_Y\int_X \frac{\partial\ell(\boldsymbol{\theta}(x_0),y)}{\partial\theta_i}\frac{\partial\ell(\boldsymbol{\theta}(x_0),y)}{\partial\theta_j}K_h^2(x-x_0)f(x)\eta\{y|\boldsymbol{\theta}(x)\}dxdy \\
\xrightarrow{P}\;& f(x_0)\nu_0\int_Y \frac{\partial\ell(\boldsymbol{\theta}(x_0),y)}{\partial\theta_i}\frac{\partial\ell(\boldsymbol{\theta}(x_0),y)}{\partial\theta_j}\eta\{y|\boldsymbol{\theta}(x_0)\}dy \\
=\;& -f(x_0)\nu_0\int_Y \frac{\partial^2\ell(\boldsymbol{\theta}(x_0),y)}{\partial\theta_i\partial\theta_j}\eta\{y|\boldsymbol{\theta}(x_0)\}dy.
\end{aligned}$$

The last step holds due to (3.16). Thus, $\mathrm{Cov}(\Delta_n)=f(x_0)\mathcal{I}(x_0)\nu_0+o_p(1)$. In order to establish asymptotic normality for Δ_n, it is necessary to show for any unit vector $\mathbf{d}$,

$$\{\mathbf{d}^T\mathrm{Cov}(\Delta_n)\mathbf{d}\}^{-1/2}\mathbf{d}^T\{\Delta_n-\mathrm{E}(\Delta_n)\}\xrightarrow{D}N(0,1). \tag{3.22}$$

Since $\mathrm{Cov}(\Delta_n)=O_p(1)$, it follows that $\{\mathbf{d}^T\mathrm{Cov}(\Delta_n)\mathbf{d}\}=O_p(1)$. Let $\lambda_i=\mathbf{d}^Tq_1\{\boldsymbol{\theta}(x_0),y_i\}K_h(x_i-x_0)$, then $\mathbf{d}^T\Delta_n=h\gamma_n\sum_{i=1}^n\lambda_i$. Then, it is sufficient to show that $nh^3\gamma_n^3\mathrm{E}(|\lambda_i|^3)=o_p(1)$. By condition F and arguments similar to (3.19), it can be shown that $nh^3\gamma_n^3\mathrm{E}(|\lambda_i|^3)=O_p(\gamma_n)=o_p(1)$, and thus Lyapounov's condition holds for (3.22). By (3.21) and the Slutsky theorem, we have

$$\sqrt{nh}\{\gamma_n\tilde{\boldsymbol{\theta}}^*-\;\mathcal{B}(x_0)+o_p(h^2)\}\xrightarrow{D}N\left\{0,f^{-1}(x_0)\nu_0\mathcal{I}^{-1}(x_0)\right\}. \tag{3.23}$$

Proof of Theorem 2.

We assume the unobserved data $(\mathcal{C}_i,i=1,\cdots,n)$ are random samples from population $\mathcal{C}$, and the complete data $\{(x_i,y_i,\mathcal{C}_i),i=1,2,\cdots,n\}$ are random samples from $(X,Y,\mathcal{C})$. Let $h\{y,\mathcal{C}|\boldsymbol{\theta}(x)\}$ be the joint pdf of $(Y,\mathcal{C})$ given

$X = x$, and $f_X(x)$ be the marginal density of X. Conditioning on $X = x$, Y follows a distribution $f_Y\{y|\boldsymbol{\theta}(x)\}$. Then the conditional density of $\mathcal{C}$ is given by

$$g\{c|y, \boldsymbol{\theta}(x)\} = \frac{h\{y, c|\boldsymbol{\theta}(x)\}}{f_Y\{y|\boldsymbol{\theta}(x)\}}. \tag{3.24}$$

The local log-likelihood function (3.2) can be re-written as

$$\ell_n(\boldsymbol{\theta}_0) = \sum_{i=1}^{n} \log\{f_Y(y_i|\boldsymbol{\theta}_0)\} K_h(x_i - x_0). \tag{3.25}$$

Given a fixed $\tilde{\boldsymbol{\theta}}^{(l)}(x_i), i = 1, \cdots, n$, it is obvious that $\int g\{c|y_i, \tilde{\boldsymbol{\theta}}^{(l)}(x_i)\}dc = 1$. Then local likelihood (3.25) is

$$\begin{aligned}
\ell_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log\{f_Y(y_i|\boldsymbol{\theta})\} \left\{ \int g\{c|y_i, \tilde{\boldsymbol{\theta}}^{(l)}(x_i)\}dc \right\} K_h(x_i - x_0) \\
&= \sum_{i=1}^{n} \left\{ \int \log\{f_Y(y_i|\boldsymbol{\theta})\} g\{c|y_i, \tilde{\boldsymbol{\theta}}^{(l)}(x_i)\}dc \right\} K_h(x_i - x_0).
\end{aligned} \tag{3.26}$$

By (3.24), we also have

$$\log\{f_Y(y_i|\boldsymbol{\theta})\} = \log\{h(y_i, c|\boldsymbol{\theta})\} - \log\{g(c|y_i, \boldsymbol{\theta})\}. \tag{3.27}$$

Thus, we have

$$\begin{aligned}
\ell_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \mathrm{E}\{\log\{h(y_i, \mathcal{C}|\boldsymbol{\theta})\}|\boldsymbol{\theta}^{(l)}(x_i)\} K_h(x_i - x_0) \\
&- \sum_{i=1}^{n} \mathrm{E}\{\log\{g(\mathcal{C}|y_i, \boldsymbol{\theta})\}|\boldsymbol{\theta}^{(l)}(x_i)\} K_h(x_i - x_0),
\end{aligned} \tag{3.28}$$

where $\boldsymbol{\theta}^{(l)}(x_i)$ is the l-th step local estimations at x_i. Taking the expectation is equivalent to calculating (3.3). In M step, we choose $\boldsymbol{\theta}^{(l+1)}(x_0)$ such that

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left\{ \log[h\{y_i, \mathcal{C}|\boldsymbol{\theta}^{(l+1)}(x_0)\}]|\boldsymbol{\theta}^{(l)}(x_i) \right\} K_h(x_i - x_0) \\
\geq & \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left\{ \log[h\{y_i, \mathcal{C}|\boldsymbol{\theta}^{(l)}(x_0)\}]|\boldsymbol{\theta}^{(l)}(x_i) \right\} K_h(x_i - x_0).
\end{aligned}$$

It suffices to show that

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\log\left\{\frac{g\{\mathcal{C}|y_i,\boldsymbol{\theta}^{(l+1)}(x_0)\}}{g\{\mathcal{C}|y_i,\boldsymbol{\theta}^{(l)}(x_0)\}}\right\}\middle|\boldsymbol{\theta}^{(l)}(x_i)\right] K_h(x_i-x_0) \le 0 \qquad (3.29)$$

in probability. Denote

$$L_g = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\log\left\{\frac{g\{\mathcal{C}|y_i,\boldsymbol{\theta}^{(l+1)}(x_0)\}}{g\{\mathcal{C}|y_i,\boldsymbol{\theta}^{(l)}(x_0)\}}\right\}\middle|\boldsymbol{\theta}^{(l)}(x_i)\right] K_h(x_i-x_0),$$

and

$$L_J = \frac{1}{n}\sum_{i=1}^{n} \log\left[\mathrm{E}\left\{\frac{g\{\mathcal{C}|y_i,\boldsymbol{\theta}^{(l+1)}(x_0)\}}{g\{\mathcal{C}|y_i,\boldsymbol{\theta}^{(l)}(x_0)\}}\middle|\boldsymbol{\theta}^{(l)}(x_i)\right\}\right] K_h(x_i-x_0).$$

By Jensen's inequality, $L_g \le L_J$. Next we show that $L_J \to 0$ in probability. To this end, we first calculate the expectation of L_J.

$$E(L_J) = E\left(\log\left[\int_{\mathcal{C}} \frac{g\{c|Y,\boldsymbol{\theta}^{(l+1)}(x_0)\}}{g\{c|Y,\boldsymbol{\theta}^{(l)}(x_0)\}} g\{c|Y,\boldsymbol{\theta}^{(l)}(X)\}dc\right] K_h(X-x_0)\right),$$

which tends to 0 by a standard argument. We next calculate the variance of L_J. Note that the variance of L_J is dominated by the following term

$$\frac{1}{n}E\left(\log\left[\int_{\mathcal{C}} \frac{g\{c|Y,\boldsymbol{\theta}^{(l+1)}(x_0)\}}{g\{c|Y,\boldsymbol{\theta}^{(l)}(x_0)\}} g\{c|Y,\boldsymbol{\theta}^{(l)}(X)\}dc\right] K_h(X-x_0)\right)^2,$$

which can be shown to have the order $O_p\{(nh)^{-1}\}$. Then we have $L_J = o_p(1)$ by Chebyshev inequality. This completes the proof.

Chapter 4

Nonparametric Mixture of Regression Models with Constant Mixing Proportions

4.1 Introduction

In this chapter, we consider a situation in which the mean functions and the variance functions are all nonparametric, but the mixing proportion is constant. Thus, the model studied in this chapter indeed is a semiparametric model. Using a kernel regression technique, we propose an estimation procedure for the unknown functions via local likelihood approach, and further develop backfitting algorithm for in nonparametric mixture of regression model. The sampling properties of the proposed estimation procedure are investigated. We derive the asymptotic bias and variance of the local likelihood estimates, and establish its asymptotic normality.

In Chapter 3, we proposed a modified EM algorithm (Dempster, Laird and Rubin, 1977) to maximize the local likelihood functions for nonparametric mixture of regression models. The EM algorithm is used in the proposed estimation procedure of nonparametric mixture of regression models. Given an estimate of the mixing proportion, we maximize the local likelihood functions and obtain the estimates of nonparametric functions using the modified EM algorithm. Given the estimates of nonparametric functions, we can further derive a more efficient estimate for the mixing proportion. The backfitting algorithm

works well in our simulation and a real data example.

We derive a standard error formula for the resulting estimate by the conventional sandwich formula. A bandwidth selector is proposed for the local likelihood estimate using a multi-fold cross-validation method. A simulation study is conducted to examine the performance of the proposed procedures and test the accuracy of the proposed standard error formula. We further demonstrate the proposed model and estimation procedure by a continued analysis of the US housing price index data.

The rest of this paper is structured as follows. In section 2, we present the nonparametric finite mixture of regression models with constant mixing proportions, and develop an estimation procedure for nonparametric mixture of regression models. Simulation results and an empirical analysis of a real data are presented in section 3. Some discussions are provided in section 4. Technical conditions and proofs are given in section 5.

4.2 A Semiparametric Model

Let $\mathcal{C}$ be a latent class variable with a discrete distribution $P(\mathcal{C} = c) = \pi_c$ for $c = 1, 2, \cdots, C$. Conditioning on $\mathcal{C} = c$, the relationship between X and Y follows a nonparametric regression model,

$$Y = m_c(X) + \sigma_c(X)\epsilon, \tag{4.1}$$

where $\epsilon \sim N(0, 1)$, $m_c(\cdot)$ and $\sigma_c(\cdot)$ are unknown but smooth functions. In other words, conditioning on x,

$$Y \sim \sum_{c=1}^{C} \pi_c N\{m_c(x), \sigma_c^2(x)\}. \tag{4.2}$$

Compared with the nonparametric mixture of regression model (3.1), model (4.2) indeed is a semi-parametric model because π_c is an unknown parameter rather than unknown nonparametric function of x. Thus, we may derive a more efficient estimate for π_c. Suppose that $\{(x_i, y_i), i = 1, \cdots, n\}$ are random samples from the population (X, Y). The likelihood function of the collected data is

$$\ell_n(\pi, \mathbf{m}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \pi_c \phi\{y_i | m_c(x_i), \sigma_c^2(x_i)\} \right\}, \tag{4.3}$$

where $\mathbf{m} = \mathbf{m}(\cdot) = \{m_1(\cdot), \cdots, m_C(\cdot)\}$, and $\boldsymbol{\sigma}^2 = \boldsymbol{\sigma}^2(\cdot) = \{\sigma_1^2(\cdot), \cdots, \sigma_C^2(\cdot)\}$. Since $\mathbf{m}(\cdot), \boldsymbol{\sigma}^2(\cdot)$ are nonparametric functions, (4.3) is not ready for maximization. In our approach, we first use kernel regression techniques (Fan and Gijbels, 1996) to estimate $\mathbf{m}(\cdot)$, and $\boldsymbol{\sigma}^2(\cdot)$, substitute the resulting estimates to (4.3) and then maximize (4.3) with respect to π. Given the estimate of π, we may further derive a more efficient estimate for $m_c(x)$ and $\sigma_c^2(x)$ by the local likelihood method.

4.2.1 An Estimation Procedure

We first propose an approach to obtain good initial values. The idea is as follows. We treat the constant mixing proportion as a function of x, and apply the estimation procedure proposed in Chapter 3 to estimate π, $m_c(\cdot)$ and $\sigma_c^2(\cdot)$. Let $K_h(\cdot) = h^{-1}K(\cdot/h)$ be a rescaled kernel for a kernel function $K(\cdot)$ and a bandwidth h. Further, denote $\phi(y|\mu, \sigma^2)$ to be the density function $N(\mu, \sigma^2)$. The local likelihood method is to maximize the local likelihood function

$$\ell_n(\pi_0, \boldsymbol{\sigma}_0^2, \mathbf{m}_0) = \sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \pi_{c0} \phi(y_i | m_{c0}, \sigma_{c0}^2) \right\} K_h(x_i - x_0), \tag{4.4}$$

where $\mathbf{m}_0 = (m_{10}, \cdots, m_{C0})^T$, $\boldsymbol{\sigma}_0^2 = (\sigma_{10}^2, \cdots, \sigma_{C0}^2)^T$, $\pi_0 = (\pi_{10}, \cdots, \pi_{C-1,0})^T$, and $\pi_{C0} = 1 - \sum_{c=1}^{C-1} \pi_{c0}$. Let $\{\tilde{\pi}, \tilde{\boldsymbol{\sigma}}^2, \tilde{\mathbf{m}}\}$ be the solution of maximizing the local

likelihood function (4.4). Then the estimates for $\pi_c(x_0)$, $\sigma_c^2(x_0)$, and $m_c(x_0)$ are

$$\tilde{\pi}_c(x_0) = \tilde{\pi}_{c0}, \quad \tilde{\sigma}_c^2(x_0) = \tilde{\sigma}_{c0}^2, \quad \textit{and} \quad \tilde{m}_c(x_0) = \tilde{m}_{c0}.$$

The maximization can be achieved by the modified EM algorithm proposed in Chapter 3. We have showed that the resulting estimates are $\sqrt{nh}$ consistent.

To achieve the root consistency of an estimate for π, we should use all data rather than data in a local neighborhood. For given estimates $\hat{m}_c(\cdot)$ and $\hat{\sigma}_c^2(\cdot)$, we maximize

$$\ell_1(\pi) = \sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \pi_c \phi\{y_i | \hat{m}_c(x_i), \hat{\sigma}_c^2(x_i)\} \right\}, \tag{4.5}$$

with respect to π. Let $\hat{\pi} = (\hat{\pi}_1, \cdots, \hat{\pi}_{C-1})^T$ be a maximizer of (4.5). If $\hat{m}_c(\cdot)$ and $\hat{\sigma}_c^2(\cdot)$ are chosen to be $\tilde{m}_c(\cdot)$ and $\tilde{\sigma}_c^2(\cdot)$, the resulting estimate is referred to as a **one step** estimate of π. In section 3.2 we will show that under certain regularity conditions, the one step estimate of π is a root n consistent estimator of π.

To maximize (4.5), define Bernoulli random variables

$$z_{ic} = \begin{cases} 1, & \text{if } (x_i, y_i) \text{ is in the } c^{th} \text{ group,} \\ 0, & \text{otherwise.} \end{cases}$$

and let $\mathbf{z}_i = (z_{i1}, \cdots, z_{iC})^T$. The complete data are $\{(x_i, y_i, \mathbf{z}_i), i = 1, 2, \cdots, n\}$, and the corresponding complete log likelihood function is

$$\sum_{i=1}^{n} \sum_{c=1}^{C} z_{ic} \left\{ \log \pi_c + \log \phi\{y_i | \hat{m}_c(x_i), \hat{\sigma}_c^2(x_i)\} \right\}. \tag{4.6}$$

In E step, we calculate the expectation of z_{ic}, given by

$$r_{ic}^{(l)} = \frac{\pi_c^{(l)} \phi\{y_i | \hat{m}_c(x_i), \hat{\sigma}_c^2(x_i)\}}{\sum_{c=1}^{C} \pi_c^{(l)} \phi\{y_i | \hat{m}_c(x_i), \hat{\sigma}_c^2(x_i)\}}. \tag{4.7}$$

Then in M step, we only need to maximize

$$\sum_{i=1}^{n}\sum_{c=1}^{C} r_{ic}^{(l)} \log \pi_c,$$

which gives solution

$$\pi_c^{(l+1)} = \frac{1}{n}\sum_{i=1}^{n} r_{ic}^{(l)}. \tag{4.8}$$

For a given $\hat{\pi}$, we maximize

$$\ell_2(\boldsymbol{\sigma}_0^2, \mathbf{m}_0) = \sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \hat{\pi}_c \phi(y_i | m_{c0}, \sigma_{c0}^2) \right\} K_h(x_i - x_0), \tag{4.9}$$

with respect to $\mathbf{m}_0$ and $\boldsymbol{\sigma}_0^2$. The EM algorithm proposed in the last chapter can be adapted to maximize (4.9). Note that the complete log-likelihood function is

$$\sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic} \left[\log \hat{\pi}_c + \log \phi\{y_i | m_c(x_i), \sigma_c^2(x_i)\} \right].$$

In the l-th step of the EM algorithm iteration, we have $m_c^{(l)}(\cdot)$, and $\sigma_c^{2(l)}(\cdot)$. In the E-step, the expectation of the latent variable z_{ic} is given by

$$r_{ic}^{(l)} = \frac{\hat{\pi}_c \phi\{y_i | m_c^{(l)}(x_i), \sigma_c^{2(l)}(x_i)\}}{\sum_{c=1}^{C} \hat{\pi}_c \phi\{y_i | m_c^{(l)}(x_i), \sigma_c^{2(l)}(x_i)\}}. \tag{4.10}$$

Let $\{u_1, \cdots, u_N\}$ be a set of grid points at which the estimated functions are evaluated, where N is the number of grid points. In the M-step, we maximize

$$\sum_{i=1}^{n}\sum_{c=1}^{C} r_{ic}^{(l)} \left[\log \phi\{y_i | m_{c0}(x_0), \sigma_{c0}^2(x_0)\} \right] K_h(x_i - x_0), \tag{4.11}$$

for $x_0 = u_i$, $i = 1, \cdots, N$. The closed form solution is, for $x_0 \in \{u_j, j = 1, \cdots, N\}$,

$$m_c^{(l+1)}(x_0) = \frac{\sum_{i=1}^{n} w_{ci}^{(l)}(x_0) y_i}{\sum_{i=1}^{n} w_{ci}^{(l)}(x_0)}, \tag{4.12}$$

$$\sigma_c^{2(l+1)}(x_0) = \frac{\sum_{i=1}^{n} w_{ci}^{(l)}(x_0) \{y_i - m_c^{(l)}(x_i)\}^2}{\sum_{i=1}^{n} w_{ci}^{(l)}(x_0)}, \tag{4.13}$$

where $w_{ci}^{(l)}(x_0) = r_{ic}^{(l)} K_h(x_i - x_0)$. Furthermore, we update $m_c^{(l+1)}(x_i)$ and $\sigma_c^{2(l+1)}(x_i)$, $i = 1, \cdots, n$ by linearly interpolating $m_c^{(l+1)}(u_j)$ and $\sigma_c^{2(l+1)}(u_j)$, $j = 1, \cdots, N$, respectively. Similar to the effective EM algorithm in Chapter 3, the proposed EM algorithm for (4.9) also has a corresponding ascent property. For large n, we have

$$\ell_2\{\mathbf{m}^{(l+1)}(u_j), \boldsymbol{\sigma}^{2(l+1)}(u_j)\} - \ell_2\{\mathbf{m}^{(l)}(u_j), \boldsymbol{\sigma}^{2(l)}(u_j)\} \geq 0.$$

A Backfitting Algorithm:

Step 1: Calculate $\tilde{\pi}(x_i)$, $\tilde{m}_c(x_i)$ and $\tilde{\sigma}_c^2(x_i)$ using the effective EM algorithm in Chapter 3, regarding the semiparametric model as the fully nonparametric model.

Step 2; Set the initial value for maximizing $\ell_1(\pi)$ as follows. $\pi_c^{(0)}$ to be the average of $\tilde{\pi}_c(x_i)$s. Take $\hat{m}_c(x_i)$ and $\hat{\sigma}_c^2(x_i)$ to be $\tilde{m}_c(x_i)$ and $\tilde{\sigma}_c^2(x_i)$, respectively. Maximize $\ell_1(\pi)$ by using an EM algorithm. Denote the resulting estimate by $\hat{\pi}$, which will be referred to as the **one-step estimate** for π.

Step 3: Maximize $\ell_2(\boldsymbol{\sigma}^2, \mathbf{m})$ with $\hat{\pi}$. Obtain $\hat{\mathbf{m}}(x_i)$ and $\hat{\boldsymbol{\sigma}}^2(x_i)$ using the proposed EM algorithm.

Step 4: Maximize $\ell_1(\pi)$ with $\hat{\mathbf{m}}(x_i)$ and $\hat{\boldsymbol{\sigma}}^2(x_i)$ obtained in Step 3.

Because we can construct a good initial value, it may be unnecessary to iterate the EM algorithms in both Steps 3 and 4 until they converge; by avoiding this, computational cost can be reduced. Indeed, for a good initial value, we may iteratively calculate (a) (4.7) and (4.8), and (b) (4.10), (4.12) and (4.13), until the algorithm converges. Even more aggressively, we may iteratively calculate (4.7),

(4.8), (4.12) and (4.13). This may be viewed as the gradient ECM algorithm by combining the idea the gradient EM algorithm in Lange (1995) and the ECM algorithm proposed in Meng and Rubin (1993). The convergence behavior of the gradient ECM algorithm can be studied along the lines in Lange (1995). Further discussion will be given in section 4.4.

4.2.2 Asymptotic Properties

We study the sampling properties of the proposed estimation procedure in Section 3.1. Following the convention, we will show that the one-step estimator $\hat{\pi}$ obtained in Step 2 in the backfitting algorithm is root n consistent and follows an asymptotic normal distribution. We further study the asymptotic property of $\hat{m}(\cdot)$ and $\hat{\boldsymbol{\sigma}}^2(\cdot)$ for any given root n consistent estimate $\hat{\pi}$.

Denote $\boldsymbol{\eta} = \{(\boldsymbol{\sigma}^2)^T, \mathbf{m}^T\}^T$, $\boldsymbol{\eta}(x) = \{\{\boldsymbol{\sigma}^2(x)\}^T, \mathbf{m}(x)^T\}^T$, and

$$\ell(\pi, \boldsymbol{\eta}) = \log\left\{\sum_{c=1}^{C} \pi_c \phi\left\{y|m_c, \sigma_c^2\right\}\right\}.$$

Further define

$$\mathcal{I}_\pi(x) = -\mathrm{E}\left(\frac{\partial^2 \ell(\pi, \boldsymbol{\eta})}{\partial \pi \partial \pi^T}\right)\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}(x)}, \quad \mathcal{I}_{\pi\eta}(x) = -\mathrm{E}\left(\frac{\partial^2 \ell(\pi, \boldsymbol{\eta})}{\partial \pi \partial \boldsymbol{\eta}^T}\right)\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}(x)},$$

where the latter has a block-matrix representation form $\mathcal{I}_{\pi\eta}(x) = (\mathcal{I}^T_{\mathbf{m}\pi}, \mathcal{I}^T_{\sigma\pi})$. Denote the one-step estimate by $\hat{\pi}_{OS}$.

Theorem 3. *Suppose that $nh^4 \to 0$, $nh^2 \log(1/h) \to \infty$, and conditions (A)—(H) in Section 5 hold. Then we have the asymptotic normality*

$$\sqrt{n}(\hat{\pi}_{OS} - \pi) \xrightarrow{D} N\{0, B^{-1}\Sigma B^{-1}\},$$

where $B = \mathrm{E}\{\mathcal{I}_\pi(X)\}$, and

$$\Sigma \;=\; Var\left\{\frac{\partial \ell\left(\pi, \boldsymbol{\eta}(X), Y\right)}{\partial \pi} - \boldsymbol{\omega}(X, Y)\right\},$$

where

$$\boldsymbol{\omega}(x,y) = \mathcal{I}_{\pi\eta}(x)\boldsymbol{\psi}(x,y) + o_p(1),$$

and $\boldsymbol{\psi}(x,y)$ *a* $2C \times 1$ *vector, where the elements are taken from* $[C^{th}, \cdots, (3C-1)^{th}]$ *entries of* $\mathcal{I}^{-1}(x) \times \{\partial\ell(\boldsymbol{\theta}(x), y)/\partial\boldsymbol{\theta}\}$.

Given $\hat{\pi}$, let $\hat{\mathbf{m}}, \hat{\boldsymbol{\sigma}}^2$ be the maximizer of (4.9). Denote

$$\hat{m}_c^* = \sqrt{nh}\{\hat{m}_{c0} - m_c(x_0)\},$$
$$\hat{\sigma}_c^{2*} = \sqrt{nh}\{\hat{\sigma}_c^2 - \sigma_c^2(x_0)\}.$$

Let $\hat{\mathbf{m}}^* = (\hat{m}_1^*, \cdots, \hat{m}_C^*)^T$, $\hat{\boldsymbol{\sigma}}^{2*} = (\hat{\sigma}_1^{2*} \cdots, \hat{\sigma}_C^{2*})^T$. Define $\hat{\boldsymbol{\eta}}^* = \{(\hat{\boldsymbol{\sigma}}^{2*})^T, \hat{\mathbf{m}}^{*T}\}^T$, and

$$\ell(\boldsymbol{\eta}, y) = \log\left\{\sum_{c=1}^{C} \pi_c \phi\left\{y | m_c, \sigma_c^2\right\}\right\},$$

where $\boldsymbol{\eta} = (\boldsymbol{\sigma}^{2T}, \mathbf{m}^T)^T$, and π is true value. Further define

$$\mathcal{I}_\eta(x_0) = -\mathrm{E}\left(\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}\right)\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}(x_0)}.$$

Theorem 4. *Assume that conditions (A)—(H) in Section 5 hold. Then as* $n \to \infty$, $h \to 0$, $nh \to \infty$, *we have the asymptotic normality results for* $\hat{\boldsymbol{\eta}}^*$

$$\sqrt{nh}\{\gamma_n\hat{\boldsymbol{\eta}}^* - \mathcal{B}_\eta(x_0)\} \xrightarrow{D} N\left\{0, f^{-1}(x_0)\mathcal{I}_\eta^{-1}(x_0)\nu_0\right\},$$

where $\mathcal{B}_\eta(x_0)$, *is a* $2C\times 1$ *vector, which the elements are taken from* $[C^{th}, \cdots, (3C-1)^{th}]$ *entries of* $\mathcal{B}(x_0)$ *in Theorem 1.*

4.3 Simulation and Application

In this section, we address some practical implementation issues such as standard error formula and bandwidth selection for nonparametric mixture of

regression model. To assess the performance of the estimates of the unknown regression functions $m_c(x)$, we consider the square root of the average square errors (RASE) for mean functions,

$$\text{RASE}_m^2 = N^{-1} \sum_{c=1}^{C} \sum_{j=1}^{N} \{\hat{m}_c(u_j) - m_c(u_j)\}^2,$$

where $\{u_j, j = 1, \cdots, N\}$ is the grid points at which the unknown functions $m_c(\cdot)$ are evaluated. For simplification, the grid points are taken evenly on the range of the x-variable. In simulation, we set $N = 100$. Similarly, we can define RASE for variance functions $\sigma_c^2(x)$s, denoted by RASE_σ.

4.3.1 Standard Error Formula

Define the fitted value for the i-th observation as a weight sum of the estimated means,

$$\hat{y}_i = \sum_{c=1}^{C} r_{ic} \hat{m}_c(x_i),$$

where r_{ic} are the posterior of the identities when the proposed back-fitting algorithm converges. Then, the residual is $e_i = y_i - \hat{y}_i$. Rewrite the estimate of $m_c(x)$ in the proposed algorithm as

$$\hat{m}_c(x) = (\mathbf{E}^T W_c \mathbf{E})^{-1} \mathbf{E}^T W_c \mathbf{y},$$

where $\mathbf{E}$ is a $n \times 1$ vector with all entries equal to 1; $W_c = \text{diag}\{w_{c1}, \cdots, w_{cn}\}$ with $w_{ci} = r_{ic} K_h(x_i - x_0)$. We consider the following approximate standard error formula for $\hat{m}_c(x)$:

$$\widehat{\text{Var}}\{\hat{m}_c(x)\} = (\mathbf{E}^T W_c \mathbf{E})^{-1} \mathbf{E}^T W_c \widehat{\text{Cov}}(\mathbf{y}) W_c \mathbf{E} (\mathbf{E}^T W_c \mathbf{E})^{-1}, \tag{4.14}$$

where $\widehat{\text{Cov}}(\mathbf{y}) = \text{diag}\{e_1^2, e_2^2, \cdots, e_n^2\}$, a diagonal matrix consisting of the squared residuals e_i^2. Furthermore, (4.14) can be written as

$$\widehat{\text{Var}}\{\hat{m}_c(x)\} = \frac{\sum_{i=1}^n w_{ic}^2 e_i^2}{(\sum_{i=1}^n w_{ic})^2}. \tag{4.15}$$

4.3.2 Bandwidth Selection

Bandwidth selection is fundamental to nonparametric smoothing. In practice, data driven methods can be used to choose the bandwidth, such as cross-validation (CV). Denote by $\mathcal{D}$ as the full data set. We then partition $\mathcal{D}$ into a training set $\mathcal{R}_j$ and test set $\mathcal{T}_j$, $\mathcal{D} = \mathcal{T}_j \cup \mathcal{R}_j$ $j = 1, \cdots, J$. We use the train set $\mathcal{R}_j$ to obtain the estimates $\{\hat{m}_c(\cdot), \hat{\sigma}_c^2(\cdot), \hat{\pi}_c\}$. Then we can estimate $m_c(x)$, $\sigma_c^2(x)$ and π_c for the data points belong to the corresponding test set. For $(x_l, y_l) \in \mathcal{T}_j$,

$$\begin{aligned}
\hat{m}_c(x_l) &= \frac{\sum_{\{i:x_i \in \mathcal{R}_j\}} r_{ic} K_h(x_i - x_l) y_i}{\sum_{\{i:x_i \in \mathcal{R}_j\}} r_{ic} K_h(x_i - x_l)}, \\
\hat{\sigma}_c^2(x_l) &= \frac{\sum_{\{i:x_i \in \mathcal{R}_j\}} r_{ic} K_h(x_i - x_l)(y_i - \hat{m}_c(x_i))^2}{\sum_{\{i:x_i \in \mathcal{R}_j\}} r_{ic} K_h(x_i - x_l)}, \\
\hat{\pi}_c &= \frac{\sum_{\{i:x_i \in \mathcal{R}_j\}} r_{ic}}{\sum I_{\{i:x_i \in \mathcal{R}_j\}}}.
\end{aligned}$$

Based on the estimated $\hat{m}_c(x_l)$ of test set $\mathcal{T}_j$, we again calculate the posterior memberships in test set $\mathcal{T}_j$. For $(x_l, y_l) \in \mathcal{T}_j, c = 1, \cdots, C$,

$$r_{lc} = \frac{\hat{\pi}_c \phi\{y_l | \hat{m}_c(x_l), \hat{\sigma}_c^2(x_l)\}}{\sum_{q=1}^C \hat{\pi}_q \phi\{y_l | \hat{m}_q(x_l), \hat{\sigma}_q^2(x_l)\}}.$$

Now we can implement regular CV criterion in this mixture model

$$CV = \sum_{j=1}^J \sum_{l \in \mathcal{T}_j} (y_l - \hat{y}_l)^2, \tag{4.16}$$

where $\hat{y}_l = \sum_{c=1}^C r_{lc}\, \hat{m}_c(x_l)$ is the predicted value of y_l in the test set $\mathcal{T}_j$.

4.3.3 Simulation Study

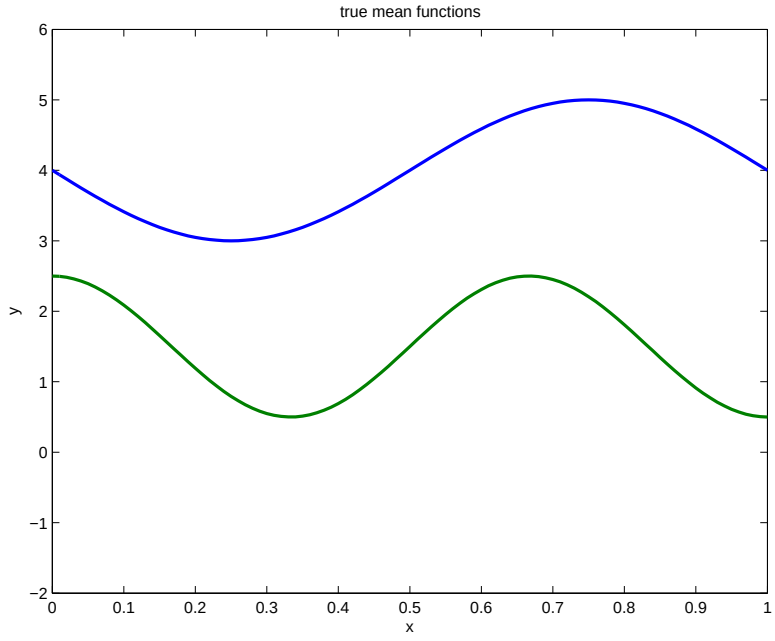
In the following example, we conduct a simulation for a 2-component nonparametric mixture of regressions model with

$$\pi_1 = 0.4, \text{ and } \pi_2 = 0.6,$$
$$m_1(x) = 4 - \sin(2\pi x), \text{ and } m_2(x) = 1.5 + \cos(3\pi x),$$
$$\sigma_1(x) = 0.25\exp(0.5x), \text{ and } \sigma_2(x) = 0.3\exp(-0.2x).$$

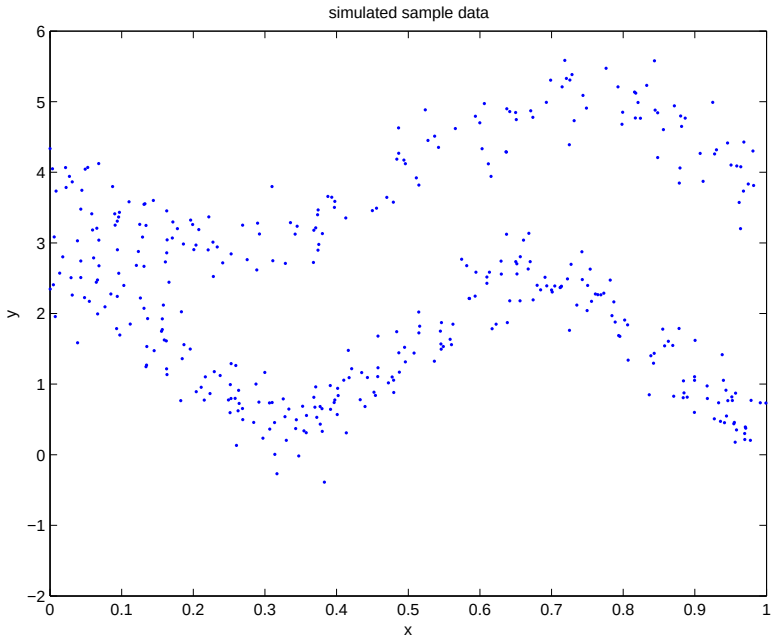
We take the sample size $n = 200, 400, 800$. For each case, 500 simulations were conducted. The predictor x is generated from one dimension uniform distribution in $[0, 1]$. The Epanechnikov kernel is used in our simulation. Figure 4.1 shows the plots of true mean functions with a typical sample data.

To obtain a good initial value, we first fit a mixture of polynomial regression models, which gives the estimates of mean functions $\bar{m}_c(x)$, and parameters $\bar{\sigma}_c^2$, $\bar{\pi}_c$. Then we set the initial values $m_c^{(1)}(x) = \bar{m}_c(x)$, $\sigma^{2(1)}(x) = \bar{\sigma}^2$, and $\pi_c^{(1)} = \bar{\pi}_c$. In our simulation, we first generate several simulation data sets for a given sample size, and then use the CV bandwidth selectors to choose a bandwidth for each data set. This provides us an idea about the optimal bandwidth for a given sample size. To demonstrate that the proposed procedure works quite well over a wide range of bandwidths, we consider three different bandwidths: two-third of the selected bandwidth, the selected bandwidth, and 1.5 times the selected bandwidth, which corresponds to the under-smoothing, optimal smoothing and over-smoothing, respectively. Table 4.1 displays the mean and standard deviation of RASEs over 500 simulations. From Table 4.1, the proposed procedure performs quite well for all three different bandwidths.

We next test the accuracy of the standard error formulas. Table 4.2 summarizes the simulation results for the unknown functions $m_c(x)$ at points



(a)



(b)

Figure 4.1: (a) Plot of true mean functions; (b) A typical sample of simulated data (n=400)

Table 4.1: RASE: Mean and Standard Deviations

		RASE_m	RASE_σ	$\pi_1 = 0.4$
n	h	Mean(Std)	Mean(Std)	Mean(Std)
200	0.12	0.0440(0.0164)	0.0036(0.0030)	0.4038(0.0318)
	0.08	0.0315(0.0127)	0.0052(0.0043)	0.3998(0.0341)
	0.053	0.0383(0.0144)	0.0061(0.0022)	0.3893(0.0397)
400	0.09	0.0193(0.0068)	0.0021(0.0009)	0.4036(0.0253)
	0.06	0.0170(0.0054)	0.0028(0.0010)	0.4007(0.0232)
	0.04	0.0211(0.0062)	0.0043(0.0018)	0.3927(0.0263)
800	0.075	0.0103(0.0038)	0.0012(0.0005)	0.4002(0.0186)
	0.05	0.0092(0.0029)	0.0016(0.0006)	0.3999(0.0187)
	0.033	0.0125(0.0028)	0.0025(0.0007)	0.4007(0.0165)

$0.25, 0.5, 0.75$. The standard deviation of 500 estimates, denoted by SD, can be viewed as the true standard errors. We then calculate the mean and standard deviation of the estimated standard error using the proposed standard error formulas (4.14), denoted by MSD(Std). The result in Table 4.2 shows that the proposed sandwich formula works reasonably well because the difference between the true value and the estimate is less than twice of the standard error of the estimate.

We now illustrate the performance of the proposed procedure by using a typical simulated sample, which is selected to the one with the median of RASE_m in the 500 simulations. For this data set, we use the cross-validation (CV) criterion proposed in Section 3.2 to select a bandwidth. The cross-validation scores are depicted in Figure 4.2. The CV bandwidth selector yields the bandwidth 0.07

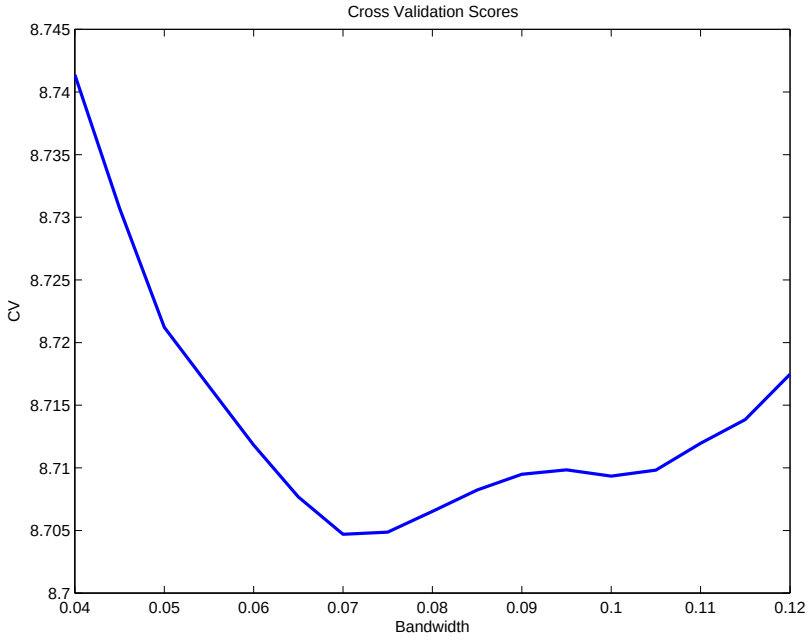
Table 4.2: Standard error of the unknown mean functions

		$m_1(x)$		$m_2(x)$	
n	x	SD	MSD(Std)	SD	MSD(Std)
	0.25	0.0839	0.0788(0.0190)	0.0978	0.0757(0.0169)
200	0.50	0.1186	0.0859(0.0237)	0.1105	0.0754(0.0177)
(h=0.08)	0.75	0.1299	0.1066(0.0316)	0.0893	0.0715(0.0171)
	0.25	0.0732	0.0699(0.0157)	0.0715	0.0656(0.0126)
400	0.50	0.1059	0.0762(0.0180)	0.0866	0.0630(0.0123)
(h=0.06)	0.75	0.0863	0.0824(0.0201)	0.0639	0.0623(0.0116)
	0.25	0.0622	0.0591(0.0114)	0.0593	0.0584(0.0097)
800	0.50	0.0679	0.0677(0.0132)	0.0611	0.0550(0.0088)
(h=0.04)	0.75	0.0814	0.0748(0.0137)	0.0561	0.0513(0.0083)

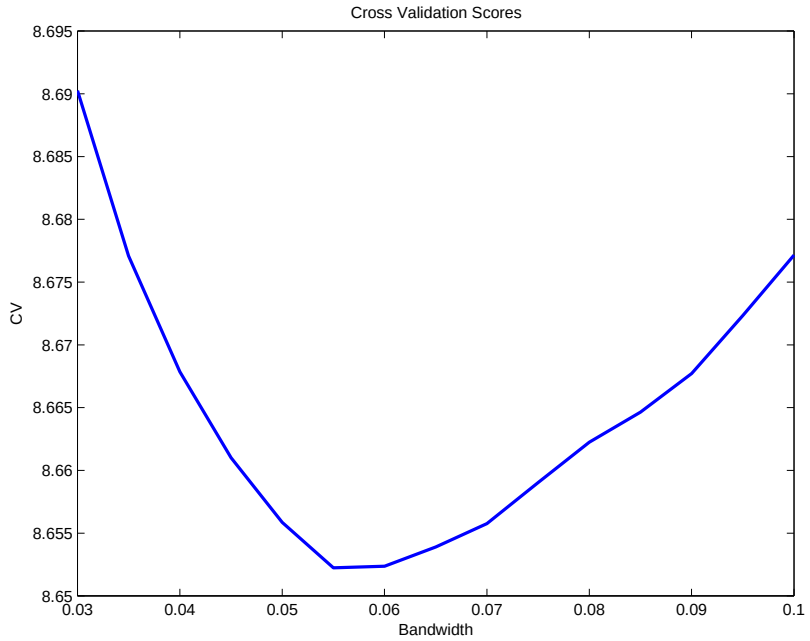
for $n = 400$, and 0.055 for $n = 800$. The resulting estimate with $n = 400$ along with its pointwise confidence interval is depicted in Figure 4.3, from which we can see that the true mean functions lies within the confidence interval. This implies that the proposed estimation procedure performs quite with the moderate sample size.

4.3.4 Analysis of US Housing Index Data (Continued)

We continue the analysis for US housing index data in last chapter. The data set contains the SP-Case Shiller House Price Index (HPI) change and United States GDP growth rate from Jan, 1990 to Dec, 2006. In this analysis, we set HPI change to be the response variable, and the GDP growth rate to be predictor, and limit ourselves to the data with positive GDP growth rate. By an analysis using nonparametric mixture of regression model, we model the impact

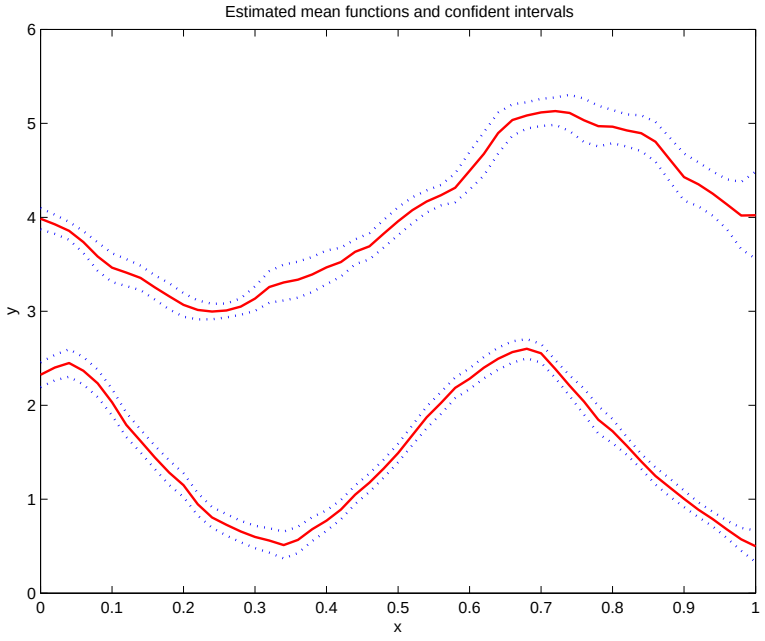


(a)

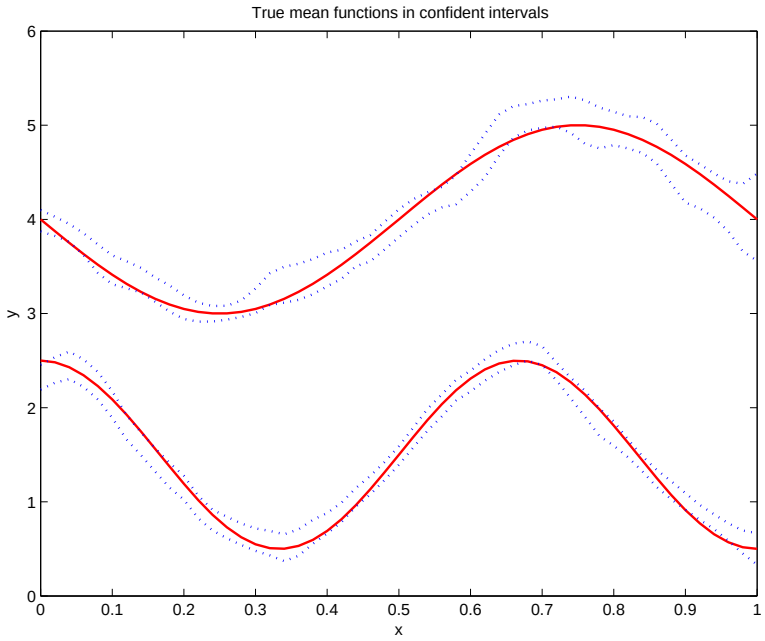


(b)

Figure 4.2: Cross-validation error versus the bandwidth: (a) n=400; (b) n=800.



(a)

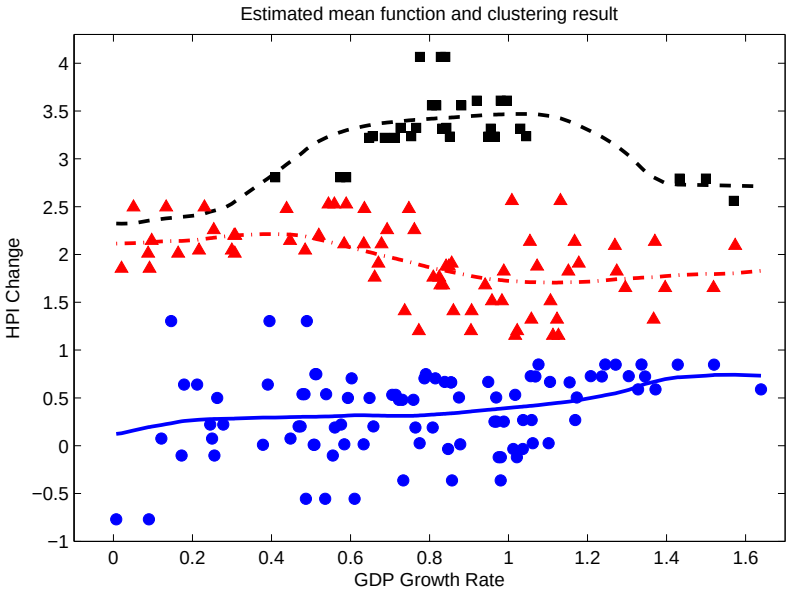


(b)

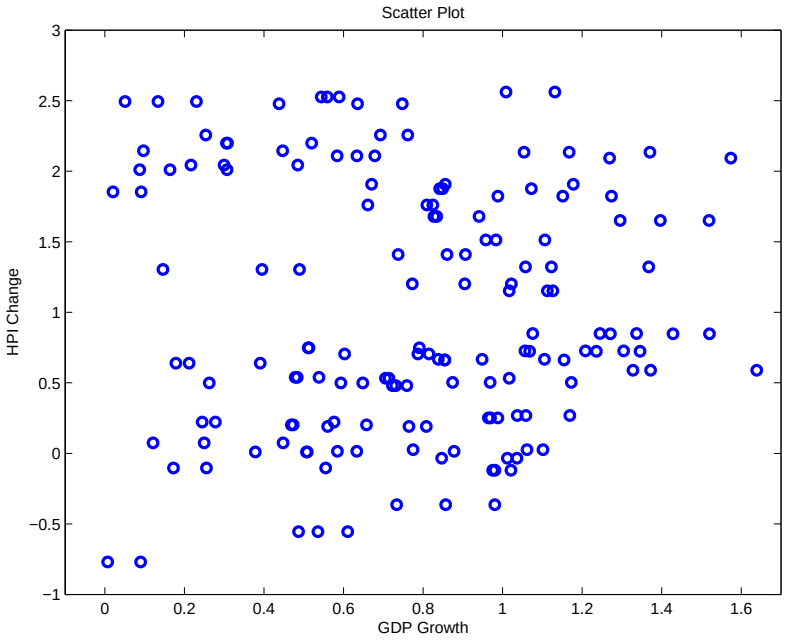
Figure 4.3: (a) The estimated mean functions and 95% pointwise confidence intervals ($n = 400, h = 0.07$); (b) True mean functions and 95% confidence intervals

of GDP growth rate on HPI change in 3 nonparametric curves, corresponding for 3 underlying macroeconomic cycles. We observe that the estimated mixing proportion function of the "housing bubble component" has dramatic changes as GDP growth rate increased. However, the estimated mixing proportion functions of the rest clusters do not change a lot. For further analysis, we remove the "housing bubble component" (upper cluster in Figure 4.4(a)), and apply the proposed semiparametric mixture of regression model and estimation procedure to analyze the data with moderate and low HPI change. The scatter plot of this data set is depicted in Figure 4.4(b). We consider a two component nonparametric mixture of regression models with constant mixing proportion in our analysis.

We first select the bandwidth by a 5-fold CV selector described in (4.16). The selected optimal bandwidth is 0.50, as shown in Figure 4.5(a). With the optimal bandwidth we fit the data in a 2-component semiparametric mixture of regression models. The estimated mean functions and a hard-clustering result based on posterior estimates of component identities are shown in Figure 4.5(b). The dotted points from lower cluster are mainly from January 1990 to September 1997, and from April 2006 to December 2006. The upper cluster are mainly from October 1997 to September 2003, during the internet boom and bust. The estimated variance functions are plotted in Figure 4.6(a). We observe that the variance proportion function for the lower cluster decrease as GDP growth rate increase. The 95% pointwise confidence intervals of each component are plotted in Figure 4.6(b).

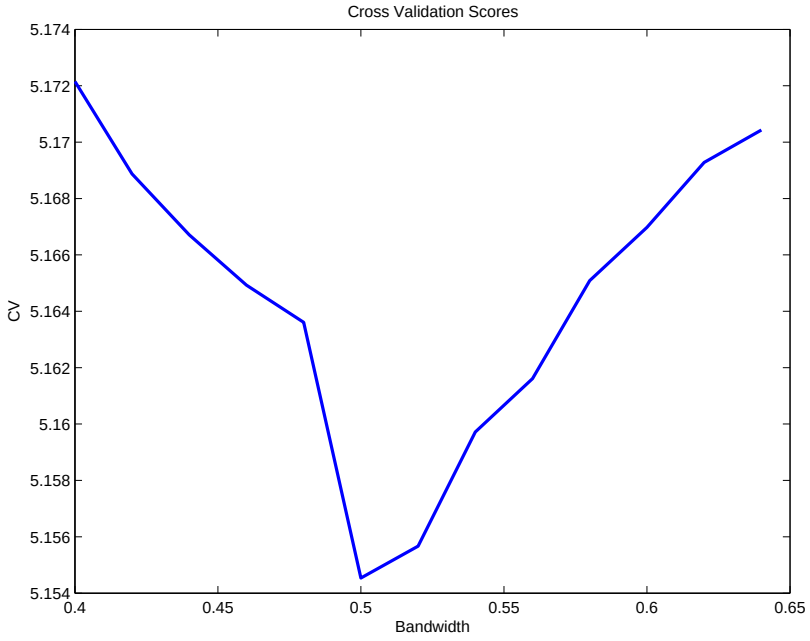


(a)

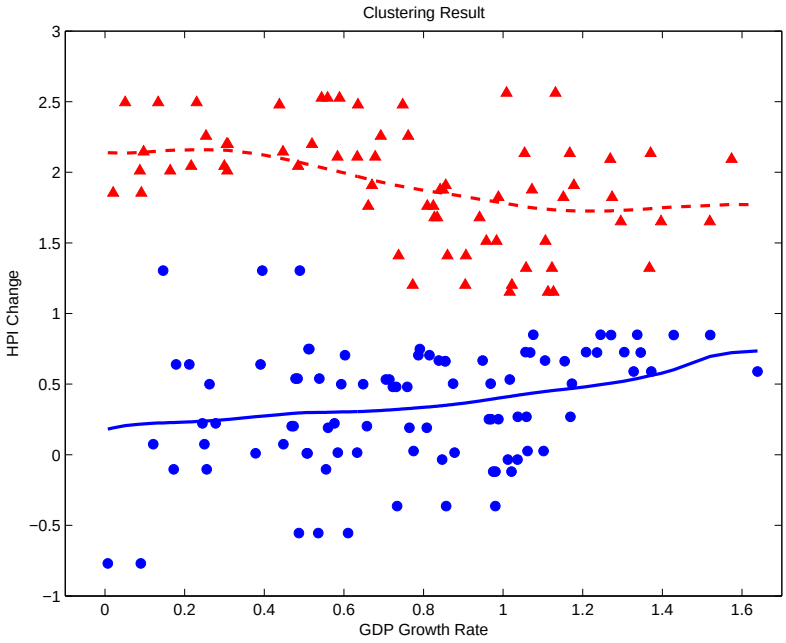


(b)

Figure 4.4: (a) Clustering results from nonparametric mixture of regressions model; (b) Scatterplot.

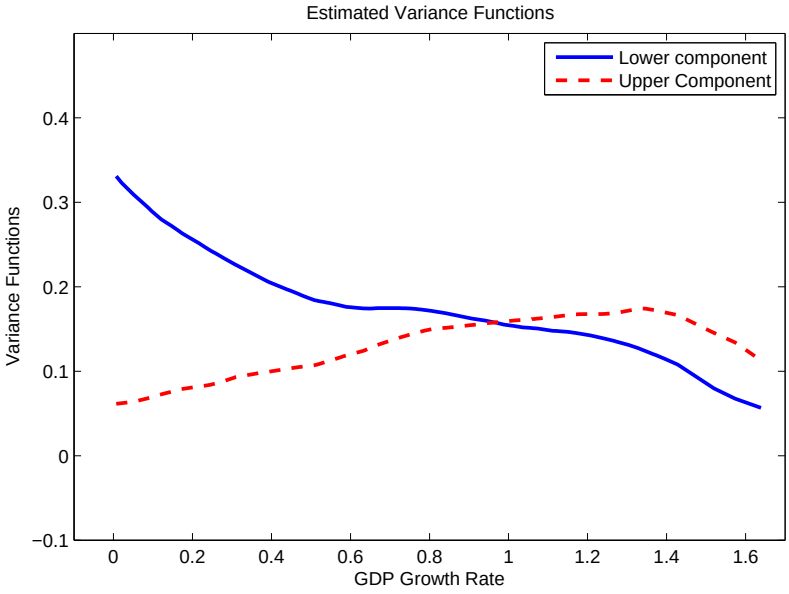


(a)

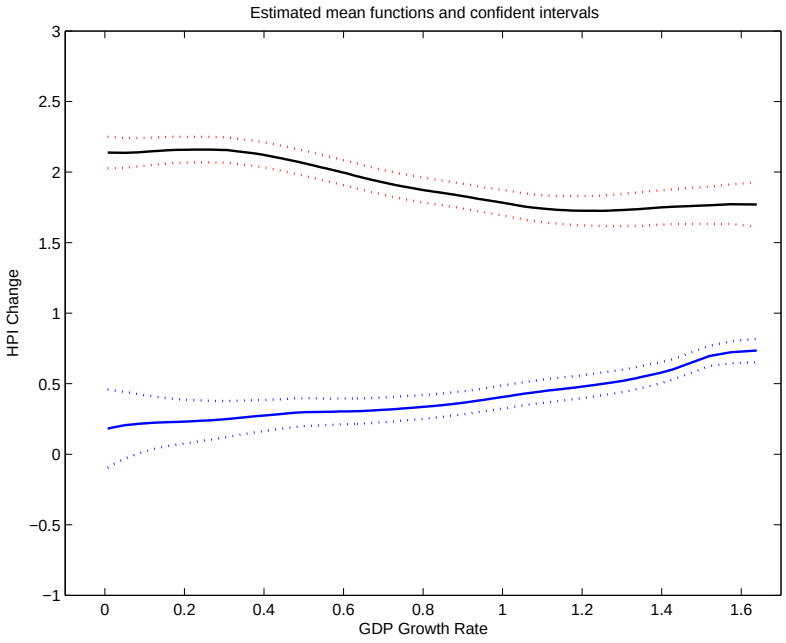


(b)

Figure 4.5: (a)Cross-validation error versus the bandwidth; (b)Clustering result.



(a)



(b)

Figure 4.6: (a)Estimated variance functions; (b)Mean functions and 95% confidence intervals.

4.4 Discussion

In this chapter we develop an estimation procedure for nonparametric mixture of regression models with constant mixing proportion, and derive the asymptotic properties for the resulting estimates. The proposed methodology is examined by Monte Carlo simulation and illustrated by a real data analysis. In this section we discuss the backfitting algorithm and one simplified version, which can be viewed as ECM algorithm. Note that both step 3 and step 4 in the backfitting algorithm consist of EM iterations. We have pointed out that it may not be necessary to iterate both EM procedures until they converge. We first suggest compute one EM iteration in both backfitting steps. This simplified algorithm is described in the following. In the l-th iteration, we first calculate the expectation of z_{ic}, given by

$$r_{ic}^{(l)} = \frac{\pi_c^{(l)} \phi\{y_i | m_c^{(l)}(x_i), \sigma_c^{2(l)}(x_i)\}}{\sum_{c=1}^{C} \pi_c^{(l)} \phi\{y_i | m_c^{(l)}(x_i), \sigma_c^{2(l)}(x_i)\}}. \tag{4.17}$$

Then we update

$$\pi_c^{(l+1)} = \frac{1}{n} \sum_{i=1}^{n} r_{ic}^{(l)}. \tag{4.18}$$

We complete one EM iteration of step 3. It then follows by one EM iteration of step 4. First, the expectation of z_{ic} is given by

$$r_{ic}^{(l+\frac{1}{2})} = \frac{\pi_c^{(l+1)} \phi\{y_i | m_c^{(l)}(x_i), \sigma_c^{2(l)}(x_i)\}}{\sum_{c=1}^{C} \pi_c^{(l+1)} \phi\{y_i | m_c^{(l)}(x_i), \sigma_c^{2(l)}(x_i)\}}. \tag{4.19}$$

Then we update for $x_0 \in \{u_j, j = 1, \cdots, N\}$,

$$m_c^{(l+1)}(x_0) = \frac{\sum_{i=1}^{n} w_{ci}^{(l+\frac{1}{2})}(x_0) y_i}{\sum_{i=1}^{n} w_{ci}^{(l+\frac{1}{2})}(x_0)}, \tag{4.20}$$

$$\sigma_c^{2(l+1)}(x_0) = \frac{\sum_{i=1}^{n} w_{ci}^{(l+\frac{1}{2})}(x_0) \{y_i - m_c^{(l)}(x_i)\}^2}{\sum_{i=1}^{n} w_{ci}^{(l+\frac{1}{2})}(x_0)}, \tag{4.21}$$

where $w_{ci}^{(l+\frac{1}{2})}(x_0) = r_{ic}^{(l+\frac{1}{2})} K_h(x_i - x_0)$. Furthermore, we update $m_c^{(l+1)}(x_i)$ and $\sigma_c^{2(l+1)}(x_i)$, $i = 1, \cdots, n$ by linearly interpolating $m_c^{(l+1)}(u_j)$ and $\sigma_c^{2(l+1)}(u_j)$, $j = 1, \cdots, N$, respectively. (4.17) and (4.18) may be regarded as maximizing $\ell_1(\pi)$ conditioning on $m_c(\cdot)$ and $\sigma^2(\cdot)$. (4.19), (4.20), and (4.21) may be regarded as maximizing $\ell_2(\boldsymbol{\sigma}_0^2, \mathbf{m}_0)$ conditioning on π. Thus the simplified version can be viewed as an ECM algorithm. Furthermore, one may omit (4.19), and replace $w_{ci}^{(l+\frac{1}{2})}(x_0)$ with $w_{ci}^{(l)}(x_0) = r_{ic}^{(l)} K_h(x_i - x_0)$, which is a acceleration scheme of the ECM algorithm.

4.5 Proofs

In this section we outline the key steps of proofs for Theorems 1 and 2. Note that $\boldsymbol{\theta} = (\pi^T, \boldsymbol{\sigma}^{2T}, \mathbf{m}^T)^T$ is a $(3C-1) \times 1$ vector. Whenever necessary, we rewrite $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_{3C-1})^T$ without changing the order of $\pi, \boldsymbol{\sigma}^2$, and $\mathbf{m}$. Correspondingly, we rewrite $\boldsymbol{\eta} = (\boldsymbol{\sigma}^{2T}, \mathbf{m}^T)^T$ as $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_{2C})^T$ if necessary. Otherwise, we will use the same notations as we defined in Section 2.

Regularity Conditions

A. The sample $\{(x_i, y_i), i = 1, \cdots, n\}$ is independent and identically distributed from the population (x, y). The support for x, denoted by $\mathcal{X}$, is closed and bounded of R^1.

B. The marginal density function $f_X(x)$ is continuous and positive for $x \in \mathcal{X}$.

C. The third derivative $|\partial^3 \ell(\boldsymbol{\theta}, y)/\partial\theta_j \partial\theta_k \partial\theta_l| \leq M_{jkl}(\boldsymbol{\theta}, y)$, where $\mathrm{E}\{M_{jkl}(\boldsymbol{\theta}, Y)\}$ is bounded for all j, k, l, and all $\boldsymbol{\theta}$. The expectation $E\left(|\partial\ell(\boldsymbol{\theta}, Y)/\partial\theta_j|^3\right)$ is finite.

D. The unknown functions $\boldsymbol{\theta}(x)$ have continuous second derivative.

E. The kernel density function $K(\cdot)$ is symmetric, and has a closed and bounded support.

F. For any j, l, $\mathrm{E}|\partial^2 \ell(\boldsymbol{\theta}, y)/\partial\theta_j \partial\theta_l|^2 < \infty$, and

$$\sup_x \int y^2 \frac{\partial^2 \ell(\boldsymbol{\theta}, y)}{\partial\theta_j \partial\theta_l} f(x, y) dy < \infty.$$

G. For $c = 1, \cdots, C$, $\pi_c(x) > 0$ and $\sigma_c^2(x) > 0$ hold for all $x \in \mathcal{X}$.

H. The second derivative matrix $E\partial^2 \ell(\boldsymbol{\theta}(x), y)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$ is positive definite.

Lemma 2. Let $\{(X_i, Y_i), i = 1 \cdots, n\}$ be i.i.d random variables from (X, Y), where X and Y are both scalar random variables. Denote f to be the joint density of (X, Y), and further assume that $\mathrm{E}|Y|^r < \infty$ and $\sup_x \int |y|^r f(x, y) dy < \infty$. Let $K(\cdot)$ be a bounded positive function with bounded support, satisfying a Lipschitz condition. Then

$$\sup_{x \in \mathcal{X}} \left| n^{-1} \sum_{i=1}^{n} [K_h(X_i - x) Y_i - \mathrm{E}\{K_h(X_i - x) Y_i\}] \right| = O_p\{\gamma_n \log^{1/2}(1/h)\},$$

given $n^{2\varepsilon - 1} h \to \infty$, for some $\varepsilon < 1 - r^{-1}$.

Denote $\gamma_n = (nh)^{-1/2}$,

$$\begin{aligned}
\tilde{m}_c^* &= \sqrt{nh}\{\tilde{m}_{c0} - m_c(x_0)\}, \\
\tilde{\sigma}_c^{2*} &= \sqrt{nh}\{\tilde{\sigma}_c^2 - \sigma_c^2(x_0)\}, \\
\tilde{\pi}_c^* &= \sqrt{nh}\{\tilde{\pi}_{c0} - \pi_c(x_0)\}, \\
\tilde{\pi}_C^* &= \sqrt{nh}\{\tilde{\pi}_{C0} - \pi_C(x_0)\} = \sqrt{nh}[1 - \sum_{c=1}^{C-1}\{\tilde{\pi}_{c0} - \pi_c(x_0)\}].
\end{aligned}$$

Let $\tilde{\mathbf{m}}^* = (\tilde{m}_1^*, \cdots, \tilde{m}_C^*)^T$, $\tilde{\boldsymbol{\sigma}}^{2*} = (\tilde{\sigma}_1^{2*} \cdots, \tilde{\sigma}_C^{2*})^T$, and $\tilde{\pi}^* = (\tilde{\pi}_1^*, \cdots, \tilde{\pi}_{C-1}^*)^T$. Define $\tilde{\boldsymbol{\theta}}^* = \{(\tilde{\pi}^*)^T, (\tilde{\boldsymbol{\sigma}}^{2*})^T, (\tilde{\mathbf{m}}^*)^T\}^T$, $\nu_l = \int u^l K^2(u)\, du$. We use the following

notations:

$$Q_1(\boldsymbol{\theta}(x), y) = \frac{\partial \ell(\boldsymbol{\theta}(x), y)}{\partial \boldsymbol{\theta}}, \quad Q_2(\boldsymbol{\theta}(x), y) = \frac{\partial^2 \ell(\boldsymbol{\theta}(x), y)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}.$$

Recall that

$$\ell_n^*(\boldsymbol{\theta}^*) = \Delta_n \boldsymbol{\theta}^* + \frac{1}{2} \boldsymbol{\theta}^{*T} \Gamma_n \boldsymbol{\theta}^* + o_p(1), \tag{4.22}$$

where

$$\begin{aligned} \Delta_n &= \sqrt{\frac{h}{n}} \sum_{i=1}^{n} Q_1(\boldsymbol{\theta}(x_0), y_i) K_h(x_i - x_0), \\ \Gamma_n &= \frac{1}{n} \sum_{i=1}^{n} Q_2(\boldsymbol{\theta}(x_0), y_i) K_h(x_i - x_0). \end{aligned}$$

By the SLLB, it follows that $\Gamma_n = -f_X(x_0)\mathcal{I}(x_0) + o_p(1)$. Therefore,

$$\ell_n^*(\boldsymbol{\theta}^*) = \Delta_n \boldsymbol{\theta}^* - \frac{1}{2} f_X(x_0) \boldsymbol{\theta}^{*T} \mathcal{I}(x_0) \boldsymbol{\theta}^* + o_p(1). \tag{4.23}$$

Lemma 3. Assume that conditions (A)—(H) holds, in addition with $nh \to 0$ as $n \to \infty$, then for all x in the support $\mathcal{X}$, we have

$$\sup_{x \in \mathcal{X}} |\tilde{\boldsymbol{\theta}}^* - f_X^{-1}(x) \mathcal{I}^{-1}(x) \Delta_n| = O_p\{h^2 + \gamma_n \log^{1/2}(1/h)\}.$$

Proof. Since each element in Γ_n is sum of i.i.d. random variables, by condition (F) and Lemma 2, we can show that Γ_n converge to $-f_X(x_0)\mathcal{I}(x_0)$ uniformly for all $x_0 \in \mathcal{X}$. By (4.22) and condition (H), we know $\ell_n^*(\boldsymbol{\theta}^*)$ is a concave function of $\boldsymbol{\theta}^*$ for large n. Then by condition (G), when n is large enough, $-\ell_n^*(\boldsymbol{\theta}^*)$ is a convex function defined on a convex open set. Thus, by the convexity lemma (Pollard, 1991),

$$\sup_{x \in \mathcal{X}} \left| (\Delta_n \boldsymbol{\theta}^* + \frac{1}{2} \boldsymbol{\theta}^{*T} \Gamma_n \boldsymbol{\theta}^*) - (\Delta_n \boldsymbol{\theta}^* - \frac{1}{2} f_X(x) \boldsymbol{\theta}^{*T} \mathcal{I}(x) \boldsymbol{\theta}^*) \right| \xrightarrow{P} 0 \tag{4.24}$$

holds uniformly for all $x_0 \in \mathcal{X}$ and $\boldsymbol{\theta}^*$ in any compact set $\mathcal{C}$. We know that $f_X^{-1}(x)\mathcal{I}^{-1}(x)\Delta_n$ is a unique maximizer of (4.23), and is continue in x; $\tilde{\boldsymbol{\theta}}^*$ is a maximizer of (4.22). Then by Lemma A.1 of Carroll et al. (1997), we have

$$\sup_{x\in\mathcal{X}} |\tilde{\boldsymbol{\theta}}^* - f_X^{-1}(x)\mathcal{I}^{-1}(x)\Delta_n| \xrightarrow{P} 0. \tag{4.25}$$

Then by the definition of $\tilde{\boldsymbol{\theta}}^*$,

$$\left.\frac{\partial \ell_n^*(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*}\right|_{\boldsymbol{\theta}^*=\tilde{\boldsymbol{\theta}}^*} = h\gamma_n \sum_{i=1}^{n} Q_1\{\boldsymbol{\theta}(x), y_i\} K_h(x_i - x). \tag{4.26}$$

By a Taylor expansion, we have

$$\Delta_n + \Gamma_n \tilde{\boldsymbol{\theta}}^* + \frac{h\gamma_n^3}{2} \sum_{i=1}^{n} \sum_{j,l} \frac{\partial^2 Q_1(\boldsymbol{\theta}(x) + \tilde{\xi}_i)}{\partial\theta_j^* \partial\theta_l^*} \tilde{\theta}_j^* \tilde{\theta}_l^{*T} K_h(x_i - x) = 0, \tag{4.27}$$

where $\boldsymbol{\theta}^*$ is rewritten as $\boldsymbol{\theta}^* = (\theta_1^*, \theta_{3C-1}^*)^T$. $\tilde{\xi}_i$ is a vector between 0 and $\gamma_n \boldsymbol{\theta}^*$. The last term of (4.27) is of order $O_p(\gamma_n ||\tilde{\boldsymbol{\theta}}^*||^2)$. Again it can be deduced from Lemma 2, for each element of Γ_n,

$$\sup_{x\in\mathcal{X}} |\Gamma_n(i,j) - \mathrm{E}\{\Gamma_n(i,j)\}| = O_p\{h^2 + \gamma_n \log^{1/2}(1/h)\}. \tag{4.28}$$

By (4.27), $\Gamma_n \tilde{\boldsymbol{\theta}}^* + O_p(\gamma_n ||\tilde{\boldsymbol{\theta}}^*||^2) = -\Delta_n$, then

$$\{\Gamma_n - \mathrm{E}(\Gamma_n)\}\tilde{\boldsymbol{\theta}}^* + O_p(\gamma_n ||\tilde{\boldsymbol{\theta}}^*||^2) = -\Delta_n + f_X(x)\mathcal{I}(x)\tilde{\boldsymbol{\theta}}^*. \tag{4.29}$$

By (4.25), it is obvious that $\sup_{x\in\mathcal{X}} |\tilde{\boldsymbol{\theta}}^*| = O_p(1)$. Thus for the left side of (4.29), we have

$$\sup_{x\in\mathcal{X}} |\{\Gamma_n - \mathrm{E}(\Gamma_n)\}\tilde{\boldsymbol{\theta}}^*| + O_p(\gamma_n) = O_p\{h^2 + \gamma_n \log^{1/2}(1/h)\}.$$

It follows that the order also holds for the right side of (4.29),

$$\sup_{x\in\mathcal{X}} |f_X(x)\mathcal{I}(x)\tilde{\boldsymbol{\theta}}^* - \Delta_n| = O_p\{h^2 + \gamma_n \log^{1/2}(1/h)\}.$$

The proof is completed by the conditions that $f_X(x)$ and $\mathcal{I}(x)$ are bounded and continuous functions in a closed set of $\mathcal{X}$.

Proof of Theorem 3. Denote $\hat{\pi}_c^* = \sqrt{n}\{\hat{\pi}_c - \pi_c\}$, where π_c is the true value of π_c. Let $\hat{\pi}^* = (\hat{\pi}_1^*, \cdots, \hat{\pi}_{C-1}^*)^T$, and $\hat{\pi}_C^* = \sqrt{nh}\{1 - \sum_{c=1}^{C-1}(\hat{\pi}_c - \pi_{c0})\}$. Further, denote $\boldsymbol{\eta} = \{(\boldsymbol{\sigma}^2)^T, \mathbf{m}^T\}^T$, and $\tilde{\boldsymbol{\eta}}(x_i) = \{\{\tilde{\boldsymbol{\sigma}}^2(x_i)\}^T, \tilde{\mathbf{m}}(x_i)^T\}^T$, then

$$\begin{aligned}
\ell(\pi, \tilde{\boldsymbol{\eta}}(x_i), y_i) &= \log\left\{\sum_{c=1}^{C} \pi_c \phi\{y_i|\tilde{m}_c(x_i), \tilde{\sigma}_c^2(x_i)\right\}, \\
\ell(\pi + \pi_c^*/\sqrt{n}, \tilde{\boldsymbol{\eta}}(x_i), y_i) &= \log\left\{\sum_{c=1}^{C} (\pi_c + \pi_c^*/\sqrt{n}) \phi\{y_i|\tilde{m}_c(x_i), \tilde{\sigma}_c^2(x_i)\right\}.
\end{aligned}$$

Then $\hat{\pi}^*$ maximizes

$$\ell_n(\pi^*) = \sum_{i=1}^{n} \{\ell(\pi + \pi_c^*/\sqrt{n}, \tilde{\boldsymbol{\eta}}(x_i), y_i) - \ell(\pi, \tilde{\boldsymbol{\eta}}(x_i), y_i)\}. \tag{4.30}$$

By a Taylor expansion,

$$\ell_n(\pi^*) = A_n \pi^* + \frac{1}{2} \pi^{*T} B_n \pi^* \{1 + o_p(1)\}, \tag{4.31}$$

where

$$\begin{aligned}
A_n &= n^{-1/2} \sum_{i=1}^{n} \frac{\partial \ell(\pi, \tilde{\boldsymbol{\eta}}(x_i), y_i)}{\partial \pi}, \\
B_n &= n^{-1} \sum_{i=1}^{n} \frac{\partial^2 \ell(\pi, \tilde{\boldsymbol{\eta}}(x_i), y_i)}{\partial \pi \partial \pi^T}.
\end{aligned}$$

It can be shown that

$$B_n = -\mathrm{E}\{\mathcal{I}_\pi(X)\} + o_p(1).$$

Then by (4.31)

$$\ell_n(\pi^*) = A_n \pi^* - \frac{1}{2} \pi^{*T} B \, \pi^* + o_p(1). \tag{4.32}$$

Let $\boldsymbol{\eta}(x_i) = (\{\boldsymbol{\sigma}^2(x_i)\}^T)^T, \mathbf{m}(x_i)^T$. We have

$$\begin{aligned} A_n &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ell(\pi, \boldsymbol{\eta}(x_i), y_i)}{\partial \pi} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ell(\pi, \boldsymbol{\eta}(x_i), y_i)}{\partial \pi \partial \boldsymbol{\eta}^T}\{\tilde{\boldsymbol{\eta}}(x_i) - \boldsymbol{\eta}(x_i)\} + O_p(d_{1n}) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ell(\pi, \boldsymbol{\eta}(x_i), y_i)}{\partial \pi} + T_{n1} + O_p(d_{1n}). \end{aligned}$$

where $d_{1n} = n^{-1/2}||\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}||_\infty^2$. By Lemma 3, we have

$$\tilde{\boldsymbol{\theta}}(x_i) - \boldsymbol{\theta}(x_i) = \frac{1}{n} f_X^{-1}(x_i)\mathcal{I}^{-1}(x_i)\sum_{j=1}^{n}\frac{\partial \ell(\boldsymbol{\theta}(x_i), y_j)}{\partial \boldsymbol{\theta}} K_h(x_j - x_i) + O_p(d_{n2}),$$

where $d_{n2} = \gamma_n^2\sqrt{\log(1/h)}$. Let $\boldsymbol{\psi}(x_i, y_j)$ be a $2C \times 1$ vector, which the elements are taken from $[C^{th}, \cdots, (3C-1)^{th}]$ entries of $\mathcal{I}^{-1}(x_i) \times \{\partial \ell(\boldsymbol{\theta}(x_i), y_j)/\partial \boldsymbol{\theta}\}$, then

$$\tilde{\boldsymbol{\eta}}(x_i) - \boldsymbol{\eta}(x_i) = \frac{1}{n} f_X^{-1}(x_i)\sum_{j=1}^{n}\boldsymbol{\psi}(x_i, y_j) K_h(x_j - x_i) + O_p(d_{n2}).$$

Thus,

$$T_{n1} = n^{-3/2}\sum_{i=1}^{n}\frac{\partial^2 \ell(\pi, \boldsymbol{\eta}(x_i), y_i)}{\partial \pi \partial \boldsymbol{\eta}^T} f_X^{-1}(x_i)\sum_{j=1}^{n}\boldsymbol{\psi}(x_i, y_j) K_h(x_j - x_i) + O_p(n^{1/2} d_{2n}).$$

By condition $nh^2/\log(1/h) \to \infty$, we have $O_p(n^{1/2}d_{2n}) = o_p(1)$. Since $\boldsymbol{\eta}(x_i) - \boldsymbol{\eta}(x_j) = O((x_i - x_j)^2)$, therefore

$$\begin{aligned} T_{n1} &= n^{-3/2}\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{\partial^2 \ell(\pi, \boldsymbol{\eta}(x_i), y_i)}{\partial \pi \partial \boldsymbol{\eta}^T} f_X^{-1}(x_i)\boldsymbol{\psi}(x_i, y_j) K_h(x_i - x_j) + O_p(n^{1/2}h^2) \\ &= T_{n2} + O_p(n^{1/2}h^2). \end{aligned}$$

It can be shown, by calculating the second moment, that

$$T_{n2} - T_{n3} \xrightarrow{P} 0, \tag{4.33}$$

where $T_{n3} = -n^{-1/2}\sum_{j=1}^{n}\boldsymbol{\omega}(x_j, y_j)$, with

$$\begin{aligned} \boldsymbol{\omega}(x_j, y_j) &= \mathrm{E}\left\{\frac{\partial^2 \ell(\pi, \boldsymbol{\eta}(X), Y)}{\partial \pi \partial \boldsymbol{\eta}^T} f_X^{-1}(X)\boldsymbol{\psi}(X, y_j) K_h(X - x_j)\right\} \\ &= \mathcal{I}_{\pi\eta}(x_j)\boldsymbol{\psi}(x_j, y_j). \end{aligned}$$

By condition $nh^4 \to 0$, we know

$$A_n = n^{-1/2} \sum_{i=1}^{n} \left\{ \frac{\partial \ell(\pi, \boldsymbol{\eta}(x_i), y_i)}{\partial \pi} - \boldsymbol{\omega}(x_i, y_i) \right\} + o_p(1).$$

By (4.32) and quadratic approximation lemma,

$$\hat{\pi}^* = B^{-1} A_n + o_p(1).$$

Then we calculate the mean and variance of A_n. It is obvious that $\mathrm{Var}(A_n) = \Sigma$, and

$$\mathrm{E}(A_n) = \sqrt{n}\, \mathrm{E} \left\{ \frac{\partial \ell(\pi, \boldsymbol{\eta}(X), Y)}{\partial \pi} - \boldsymbol{\omega}(X, Y) \right\}.$$

Similar to the proof in Chapter 3, we can show that the elements of $\mathrm{E}(\partial \ell(\pi, \boldsymbol{\eta}(X), Y)/\partial \pi)$ are equal to 0, and

$$\mathrm{E}\left\{ \boldsymbol{\omega}(X, Y) \right\} = -\mathrm{E}\left\{ \mathcal{I}_{\pi\eta}(X) \boldsymbol{\psi}(X, Y) \right\},$$

where $\boldsymbol{\psi}(X, Y)$ are the $[C^{th}, \cdots, (3C-1)^{th}]$ elements of $\mathcal{I}^{-1}(X) \times \{\partial \ell(\boldsymbol{\theta}(X), Y)/\partial \boldsymbol{\theta}\}$. Further calculation shows that $\mathrm{E}\left\{ \boldsymbol{\omega}(X, Y) \right\} = 0$. So we have $\mathrm{E}(A_n) = o_p(\sqrt{n})$. By the Central Limit Theorem we complete the proof of Theorem 3.

Proof of Theorem 4. Using similar arguments in the proof of Theorem 1, we have

$$\hat{\boldsymbol{\theta}}^* = f_X(x_0)^{-1} \mathcal{I}_\eta(x_0)^{-1} \hat{\Delta}_n + o_p(1), \tag{4.34}$$

where

$$\hat{\Delta}_n \;=\; \sqrt{\frac{h}{n}} \sum_{i=1}^{n} \frac{\partial \ell(\hat{\pi}, \boldsymbol{\eta}(x_0), y_i)}{\partial \boldsymbol{\eta}} K_h(x_i - x_0).$$

It can be calculated that

$$\hat{\Delta}_n \;=\; \sqrt{\frac{h}{n}} \sum_{i=1}^{n} \frac{\partial \ell(\pi, \boldsymbol{\eta}(x_0), y_i)}{\partial \boldsymbol{\eta}} K_h(x_i - x_0) + D_n + o_p(1),$$

where

$$\begin{aligned} D_n &= \sqrt{\frac{h}{n}} \sum_{i=1}^{n} \frac{\partial \ell(\pi, \boldsymbol{\eta}(x_0), y_i)}{\partial \boldsymbol{\eta} \partial \pi^T} (\hat{\pi} - \pi) K_h(x_i - x_0) \\ &= \sqrt{h}\, \mathrm{E} \left\{ \frac{\partial \ell(\pi, \boldsymbol{\eta}(x_0), Y)}{\partial \boldsymbol{\eta} \partial \pi^T} K_h(X - x_0) \right\} \times \sqrt{n}(\hat{\pi} - \pi). \end{aligned}$$

Since $\sqrt{n}(\hat{\pi} - \pi) = O_p(1)$, and

$$\mathrm{E} \left\{ \frac{\partial \ell(\pi, \boldsymbol{\eta}(x_0), Y)}{\partial \boldsymbol{\eta} \partial \pi^T} K_h(X - x_0) \right\} = -\mathcal{I}^T_{\pi\eta}(x_0) = O_p(1).$$

Thus, $D_n = o_p(1)$. The rest of the proof follows similar arguments to the proof of Theorem 1.

Chapter 5

Mixture of Gaussian Processes

Since its systematic introduction in Ramsay and Silverman (1997), functional data analysis has become a very active research topic. Various statistical procedures have been developed for functional data. Section 2.3 provides a brief review of the recent development in the topic of functional data analysis. In this chapter, we propose a model for functional data as a mixture of Gaussian processes. In Section 5.2, we propose an estimation procedure for the newly proposed model by assuming independent correlation structure. In Section 5.3, we further propose estimation procedure to incorporate the estimated covariance functions. In Section 5.4, we empirically test the proposed estimation procedures by Monte Carlo simulation study. We further apply the newly proposed model for analysis of the supermarket data introduced in Section 1.2.

5.1 Model Definition and Observed Data

Let $\mathcal{C}$ be a latent class variable with a discrete distribution $P(\mathcal{C} = c) = \pi_c$ for $c = 1, 2, \cdots, C$. Here C is fixed and is assumed to be known. Conditioning on $\mathcal{C} = c$, $\{X(t), t \in T\} = \{X_c(t) : t \in T\}$, which is a Gaussian process with mean $\mu_c(t)$ and covariance function $\text{Cov}\{X_c(s), X_c(t)\} = G_c(s,t)$, which is a positive definite, bivariate smooth function of s and t. We refer to $\{X(t) : t \in T\}$ as a mixture of Gaussian processes.

As a covariance function, $G_c(s,t)$ can be represented as

$$G_c(s,t) = \sum_{q=1}^{\infty} \lambda_{qc} v_{qc}(t) v_{qc}(s),$$

where λ_{qc}'s are eigenvalues, and $v_{qc}(\cdot)$'s are eigenfunctions. Furthermore, $\lambda_{1c} \geq \lambda_{2c} \geq \cdots$ and $\sum_q \lambda_{qc} < \infty$, for $c = 1, \cdots, C$.

By the Karhunen-Loève theorem, $X_c(t)$ can be represented as follows

$$X_c(t) = \mu_c(t) + \sum_{q=1}^{\infty} \xi_{qc} v_{qc}(t),$$

where ξ_{qc}s are considered as independent random variables with $\mathrm{E}\xi_{qc} = 0$, and $\mathrm{Var}(\xi_{qc}) = \lambda_{qc}$.

Since the sample path of $X_c(t)$ is a smooth function of t, $X_c(t)$ is termed a smooth random function (Yao et al., 2003; Yao et al., 2005). In practice, the collected sample of random curves are typically not smooth, and therefore, it is assumed in the literature that the observed curve $\{Y_i(t), t = t_{ij}, j = 1, \cdots, N_i\}$ is

$$Y_i(t) = X_i(t) + \epsilon_i(t),$$

where $\epsilon_i(t)$ is additive measurement error, and it is assumed that $\epsilon_i(t_{ij})$, for all i and j, are independent and identically distributed as $N(0, \sigma^2)$ (See also Rice and Wu, 2000; James and Sugar, 2003).

Denote $y_{ij} = y_i(t_{ij})$ and $\epsilon_{ij} = \epsilon_i(t_{ij})$. Throughout this chapter, it is assumed that conditioning on $\mathcal{C} = c$, the observations y_{ij}, $j = 1, \cdots, N_i$ and $i = 1, \cdots, n$, follows

$$y_{ij} = \mu_c(t_{ij}) + \sum_{q=1}^{\infty} \xi_{iqc} v_{qc}(t_{ij}) + \epsilon_{ij}, \tag{5.1}$$

where ϵ_{ij}s independent and identically distributed according to $N(0, \sigma^2)$. We propose an estimation procedure for π_c, $\mu_c(\cdot)$, $v_{qc}(\cdot)$ and σ^2 in the next two sections.

5.2 An Estimation Procedure with Working Independent Correlations

In this section, we first develop an estimation procedure for $\mu_c(t)$ using working independent correlation. Local estimation procedures have been proposed for estimating the regression function with longitudinal data. As demonstrated in Lin and Carroll (2000), the kernel generalized estimating equations (GEE) method with working independent correlation matrix yields an optimal estimate for the regression function in a certain sense. Another advantage of the working independence kernel GEE is that it is easy to implement. Thus, we propose an estimation procedure for $\mu_c(\cdot)$ under the working independence assumption. The resulting estimate may be refined by incorporating a correct correlation structure when t_{ij}s are dense for each subject i. We will consider a refined estimation procedure.

By working independence, we mean that conditioning on $\mathcal{C} = c$, $G_c(s,t) = 0$ if $s \neq t$. Denote $\sigma_c^{*2}(t) = G_c(t,t) + \sigma^2$, it follows that

$$y_{ij} = \mu_c(t_{ij}) + \epsilon_{ij}^*, \tag{5.2}$$

where ϵ_{ij}^* are independent with $E(\epsilon_{ij}^*) = 0$ and $\mathrm{Var}(\epsilon_{ij}^*) = \sigma_c^{*2}(t_{ij})$. In other words, we assume working independence for the correlation structure, and consider y_{ij}s come from the following distribution

$$y(t) \sim \sum_{c=1}^{C} \pi_c N\{\mu_c(t), \sigma_c^{*2}(t)\}. \tag{5.3}$$

The likelihood function of the collected data is

$$\sum_{i=1}^{n} \log \left[\sum_{c=1}^{C} \pi_c \prod_{j=1}^{N_i} \phi\{y_{ij} | \mu_c(t_{ij}), \sigma_c^{*2}(t_{ij})\} \right]. \tag{5.4}$$

5.2.1 An Effective EM algorithm

The estimation procedure developed in Chapter 4 can be modified for estimating π_c, $\mu_c(\cdot)$ and $\sigma_c^{*2}(\cdot)$. Define the group identity random variables

$$z_{ic} = \begin{cases} 1, & \text{if } \{X_i(t), t \in T\} \text{ is in the } c^{th} \text{ group,} \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the complete likelihood of $\{y_{ij}, j = 1, \cdots, N_i, j = 1, \cdots, n\}$ is

$$\prod_{i=1}^{n} \prod_{c=1}^{C} \left[\pi_c \prod_{j=1}^{N_i} \phi\{y_{ij} | \mu_c(t_{ij}), \sigma_c^{*2}(t_{ij})\} \right]^{z_{ic}}.$$

The kernel regression method and EM algorithm developed in Chapter 4 can be extended for the estimation of model (5.1). In the l-th iteration of the EM algorithm, we have $\pi_c^{(l)}$, $\sigma_c^{*2(l)}(\cdot)$, and $\mu_c^{(l)}(\cdot)$. In the E-step, the expectation of the latent variable z_{ic} is given by

$$r_{ic}^{(l)} = \frac{\pi_c^{(l)} \left[\prod_{j=1}^{N_i} \phi\{y_{ij} | \mu_c^{(l)}(t_{ij}), \sigma_c^{*2(l)}(t_{ij})\} \right]}{\sum_{c=1}^{C} \pi_c^{(l)} \left[\prod_{j=1}^{N_i} \phi\{y_{ij} | \mu_c^{(l)}(t_{ij}), \sigma_c^{*2(l)}(t_{ij})\} \right]}. \tag{5.5}$$

In the M-step of the EM algorithm, we update the estimated curves at a set of grid points for the given label in the E-step. Let $\{u_1, \cdots, u_N\}$ be a set of grid points at which the estimated functions are evaluated, where N is the number of grid points. For $t_0 \in \{u_1, \cdots, u_N\}$,

$$\mu_c^{(l+1)}(t_0) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{N_i} w_{cij}^{(l)} y_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{N_i} w_{cij}^{(l)}}, \tag{5.6}$$

$$\sigma_c^{*2(l+1)}(t_0) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{N_i} w_{cij}^{(l)} \{y_{ij} - \mu_c^{(l)}(t_{ij})\}^2}{\sum_{i=1}^{n} \sum_{j=1}^{N_i} w_{cij}^{(l)}}, \tag{5.7}$$

where $w_{cij}^{(l)} = r_{ic}^{(l)} K_h(t_{ij} - t_0)$. Furthermore, we update $\mu_c(t_{ij})$ and $\sigma_c^{*2}(t_{ij})$, $i = 1, \cdots, n, j = 1, \cdots, N_i$ by linearly interpolating $\mu_c^{(l+1)}(u_k)$ and $\sigma_c^{2(l)}(u_k)$, $k =$

$1, \cdots, N$. We further update mixing proportion π_c as

$$\pi_c^{(l+1)} = \frac{1}{n}\sum_{i=1}^{n} r_{ic}^{(l)}. \tag{5.8}$$

Denote the resulting estimate of π_c and $\mu_c(\cdot)$ to be $\tilde{\pi}_c$ and $\tilde{\mu}_c(\cdot)$, respectively.

5.2.2 A Backfitting Algorithm

In this section, we propose a backfitting algorithm for maximizing likelihood (5.4). We further show the connection of the backfitting algorithm and the effective EM algorithm in section 5.2.2. For given $\hat{\mu}_c(\cdot)$, and $\hat{\sigma}^{*2}(\cdot)$, we maximize

$$\ell_n(\pi) = \sum_{i=1}^{n} \log\left[\sum_{c=1}^{C} \pi_c \prod_{j=1}^{N_i} \phi\{y_{ij}|\hat{\mu}_c(t_{ij}), \hat{\sigma}_c^{*2}(t_{ij})\}\right], \tag{5.9}$$

with respect to π. The maximization can be achieved by an EM algorithm. In E step, we calculate the expectation of z_{ic}, given by

$$r_{ic}^{(l)} = \frac{\pi_c^{(l)}\left[\prod_{j=1}^{N_i} \phi\{y_{ij}|\hat{\mu}_c(t_{ij}), \hat{\sigma}_c^{*2}(t_{ij})\}\right]}{\sum_{c=1}^{C} \pi_c^{(l)}\left[\prod_{j=1}^{N_i} \phi\{y_{ij}|\hat{\mu}_c(t_{ij}), \hat{\sigma}_c^{*2}(t_{ij})\}\right]}. \tag{5.10}$$

Then in M step, we only need to maximize

$$\sum_{i=1}^{n}\sum_{c=1}^{C} r_{ic}^{(l)} \log \pi_c,$$

which gives the solution

$$\pi_c^{(l+1)} = \frac{1}{n}\sum_{i=1}^{n} r_{ic}^{(l)}. \tag{5.11}$$

Given $\hat{\pi}_c$, we maximize the local likelihood function

$$\sum_{i=1}^{n} \log\left[\sum_{c=1}^{C} \frac{\hat{\pi}_c}{(\rho\,\sigma_{c0}^{*2})^{\frac{N_i}{2}}} \exp\Big\{-\frac{1}{2\sigma_{c0}^{*2}}\sum_{j=1}^{N_i}(y_{ij}-\mu_{c0})^2 K_h(t_{ij}-t_0)\Big\}\right], \tag{5.12}$$

with respect to $\boldsymbol{\mu}_0$ and $\boldsymbol{\sigma}_0^{*2}$, where $\boldsymbol{\mu}_0 = (\mu_{10}, \cdots, \mu_{C0})^T$, and $\boldsymbol{\sigma}_0^{*2} = (\sigma_{10}^{*2}, \cdots, \sigma_{C0}^{*2})^T$. Note that π has been used for mixing proportion, here we use the notation ρ to represent 2 times the circular constant to avoid confusion. Since we are interest in maximization of (5.12) at a set of grid points, the effective EM algorithm can be extended for estimation. In the l-th step of the EM iteration, we have $\sigma_c^{*2(l)}(\cdot)$, and $\mu_c^{(l)}(\cdot)$. In the E-step, the expectation of the latent variable z_{ic} is given by

$$r_{ic}^{(l)} = \frac{\hat{\pi}_c \left[\prod_{j=1}^{N_i} \phi\{y_{ij}|\mu_c^{(l)}(t_{ij}), \sigma_c^{*2(l)}(t_{ij})\}\right]}{\sum_{c=1}^{C} \hat{\pi}_c \left[\prod_{j=1}^{N_i} \phi\{y_{ij}|\mu_c^{(l)}(t_{ij}), \sigma_c^{*2(l)}(t_{ij})\}\right]}. \tag{5.13}$$

In the M-step of the EM algorithm, we update the estimated curves at a set of grid points for the given label in the E-step. Let $\{u_1, \cdots, u_N\}$ be a set of grid points at which the estimated functions are evaluated, where N is the number of grid points. For $t_0 \in \{u_1, \cdots, u_N\}$,

$$\mu_c^{(l+1)}(t_0) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{N_i} w_{cij}^{(l)} y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{N_i} w_{cij}^{(l)}}, \tag{5.14}$$

$$\sigma_c^{*2(l+1)}(t_0) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{N_i} w_{cij}^{(l)} \{y_{ij} - \mu_c^{(l)}(t_{ij})\}^2}{\sum_{i=1}^{n}\sum_{j=1}^{N_i} w_{cij}^{(l)}}, \tag{5.15}$$

where $w_{cij}^{(l)} = r_{ic}^{(l)} K_h(t_{ij} - t_0)$. Furthermore, we update $\mu_c(t_{ij})$ and $\sigma_c^{*2}(t_{ij})$, $i = 1, \cdots, n, j = 1, \cdots, N_i$ by linearly interpolating $\mu_c^{(l+1)}(u_k)$ and $\sigma_c^{2(l)}(u_k)$, $k = 1, \cdots, N$.

Similar to the backfitting algorithm in Chapter 4, we can reduce the computational cost by iteratively calculating (a) (5.10) and (5.11), and (b) (5.12), (5.14) and (5.15), until the algorithm converges. More aggressively, given a good initial value, we may iteratively calculate (5.10), (5.11), (5.14) and (5.15), which is essentially the same as the effective EM algorithm in section 5.2.1.

5.3 Estimation Procedure with Correlation Structure

From our own limited experience, the estimate of π_c's from Section 5.2.1 can be improved by incorporating the correlation structure. Functional principal analysis provides a convenient way to incorporating the information of the estimated covariance functions. With the estimated $\mu_c(t)$, we calculate the residuals, which are raw material to estimate the covariance function $G_c(\cdot,\cdot)$. We then proposed an estimation procedure to σ^2 and π in Section 5.3.2. With the updated π_c and posteriors r_{ic}, we may further improve the estimation of $\mu_c(t)$.

5.3.1 Estimation of Covariances

Denote

$$\bar{G}_{ic}(t_{ij},t_{il}) = (y_{ij} - \hat{\mu}_c(t_{ij}))(y_{il} - \hat{\mu}_c(t_{il})).$$

Note that $\mathrm{Cov}\{Y(t),Y(t)\} = G_c(t,t) + \sigma^2$ and $\mathrm{Cov}\{Y(s),Y(t)\} = G_c(s,t)$ for $s \neq t$. If z_{ic} were observable, then the covariance function $G(s,t)$ could be estimated by a two-dimensional kernel smoother, which is to minimize

$$\sum_{i=1}^{n} z_{ic} \sum_{1\le j\neq l\le N} [\bar{G}_{ic}(t_{ij},t_{il}) - \beta_0]^2 K_h(t_{ij}-s)K_h(t_{il}-t), \tag{5.16}$$

with respect to β_0. In practice, z_{ic} is a latent variable. Following the idea of the EM algorithm, we replace z_{ic} by its expectation r_{ic}, which was obtained in the estimation procedure for $\mu_c(\cdot)$ described in Section 5.2.1. Thus, we minimize

$$\sum_{i=1}^{n} r_{ic} \sum_{1\le j\neq l\le N} [\bar{G}_{ic}(t_{ij},t_{il}) - \beta_0]^2 K_h(t_{ij}-s)K_h(t_{il}-t), \tag{5.17}$$

with respect to β_0. The minimizer $\hat{\beta}_0$ of (5.17) has a closed form given by

$$\hat{G}_c(s,t) = \frac{\sum_{i=1}^{n} r_{ic} \sum\limits_{1\le j\neq l\le N_i} \bar{G}_c(t_{ij},t_{il})K_h(t_{ij}-s)K_h(t_{iq}-t)}{\sum\limits_{i=1}^{n} r_{ic} \sum\limits_{1\le j\neq l\le N_i} K_h(t_{ij}-s)K_h(t_{iq}-t)}. \tag{5.18}$$

Following Rice and Silverman (1991), the estimation of eigenvalues and eigenfunctions are based on discretizing the covariance $\hat{G}_c(s,t)$. The estimates of eigenvalues $\hat{\lambda}_{qc}$ and eigenfunctions $\hat{v}_{qc}(\cdot)$ are determined by eigenfunctions

$$\int_T \hat{G}_c(s,t)\hat{v}_{qc}(s)ds = \hat{\lambda}_{qc}\hat{v}_{qc}(t), \tag{5.19}$$

where $\hat{v}_{qc}(t)$ satisfies $\int_T \hat{v}_{qc}^2(t)dt = 1$, and $\int_T \hat{v}_{pc}(t)\hat{v}_{qc}(t)dt = 0$ if $p \neq q$. Then, in order for the resulting estimate of $G(s,t)$ to be positive definite, we set

$$\hat{G}(s,t) = \sum_q \hat{\lambda}_{qc} I(\hat{\lambda}_{qc} > 0)\hat{v}_{qc}(s)\hat{v}_{qc}(t).$$

5.3.2 Estimation of σ^2 and π_cs

Given $\hat{\mu}_c(t)$, and $\hat{v}_{qc}(t)$, define

$$\hat{\xi}_{iqc} = \int_T \{y_i(t) - \hat{\mu}_c(t)\}\, \hat{v}_{qc}(t)dt, \tag{5.20}$$

which is an estimate of principal component scores ξ_{iqc}. Further, for $j = 1, \cdots, N_i$ and $j = 1, \cdots, n$, define

$$\hat{X}_{ic}(t_{ij}) = \hat{\mu}_c(t_{ij}) + \sum_q \hat{\xi}_{iqc} I(\hat{\lambda}_{qc} > 0)\hat{v}_{qc}(t_{ij}). \tag{5.21}$$

To estimate π and σ^2, we maximize

$$\ell_n(\pi, \sigma^2) = \sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \pi_c \prod_{j=1}^{N_i} \phi\left(y_{ij} | \hat{X}_{ic}(t_{ij}), \sigma^2\right) \right\}, \tag{5.22}$$

with respect to π_c and σ^2. The EM algorithm can be used to maximize (5.22). In the E-step, the expectation of the latent variable z_{ic} is given by

$$r_{ic}^{(l)} = \frac{\pi_c^{(l)} \left[\prod_{j=1}^{N_i} \phi\{y_{ij} | \hat{X}_{ic}(t_{ij}), \sigma^{2(l)}\}\right]}{\sum_{c=1}^{C} \pi_c^{(l)} \left[\prod_{j=1}^{N_i} \phi\{y_{ij} | \hat{X}_{ic}(t_{ij}), \sigma^{2(l)}\}\right]}. \tag{5.23}$$

In the M-step of the EM algorithm, we update π_c and σ^2 as

$$\pi_c^{(l+1)} = \frac{1}{n}\sum_{i=1}^{n} r_{ic}^{(l)}, \tag{5.24}$$

$$\sigma^{2(l+1)} = \frac{1}{\sum_{i=1}^{n} N_i}\sum_{i=1}^{n}\sum_{c=1}^{C}\sum_{j=1}^{N_i} r_{ic}^{(l)}\{y_{ij} - \hat{X}_{ic}(t_{ij})\}^2. \tag{5.25}$$

5.3.3 An Iterative Estimation Procedure

When maximizing (5.22), we update the posterior component identities r_{ic} by adapting the estimated covariance $\hat{G}_c(\cdot,\cdot)$. This procedure may provide a better estimate of the posterior identities r_{ic} compared to r_{ic} yielded by working independent correlation in section 5.2.1. Given r_{ic}s yielded by maximizing (5.22), we may further improve the estimation of $\mu_c(\cdot)$. For $t_0 \in \{u_1, \cdots, u_N\}$,

$$\hat{\mu}_c(t_0) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{N_i} r_{ic}K_h(t_{ij}-t_0)y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{N_i} r_{ic}K_h(t_{ij}-t_0)}. \tag{5.26}$$

Given the newly updated $\hat{\mu}_c(\cdot)$, and r_{ic}s from the EM algorithm, we may further update $G_c(s,t)$ in section 5.3.1, and improve the estimation of π_c and σ^2 again using the proposed procedure in section 5.3.2. Motivated by the backfitting algorithm in Chapter 4, we propose the following iterative estimation procedure for mixture of Gaussian processes.

An Iterative Estimation Procedure:

Step 1: Calculate $\tilde{\pi}_c$, $\tilde{\mu}_c(\cdot)$ and $\tilde{\sigma}_c^{*2}(\cdot)$ using the proposed EM algorithm for model (5.3). Retain r_{ic}s and set $\hat{\mu}_c(\cdot) = \tilde{\mu}_c(\cdot)$.

Step 2: Given $\hat{\mu}_c(\cdot)$, and r_{ic}s, obtain $\hat{G}_c(s,t)$ using (5.18) in section 5.3.1.

Step 3: Given $\hat{\mu}_c(\cdot)$ and $\hat{G}_c(\cdot,\cdot)$, calculate $\hat{X}_{ic}(\cdot)$ using (5.19), (5.20), and (5.21). Further obtain $\hat{\pi}_c$, and r_{ic}s by maximizing (5.4) by the proposed EM algorithm, and update $\hat{\mu}_c(\cdot)$ using (5.26).

Iteratively update Step 2 and Step 3 until convergence.

5.4 Simulation and Application

In this section, we conduct numerical simulation to demonstrate the performance of the estimation procedure. To assess the performance of the estimates of the unknown regression functions $\mu_c(x)$, we consider the square root of the average square errors (RASE) for mean functions,

$$\text{RASE}_{\mu}^2 = n_{grid}^{-1} \sum_{c=1}^{C} \sum_{j=1}^{n_{grid}} \{\hat{\mu}_c(u_j) - \mu_c(u_j)\}^2,$$

where $\{u_j, j = 1, \cdots, n_{grid}\}$ are the grid points at which the unknown functions $\mu_c(\cdot)$ are evaluated. For simplification, the grid points are taken evenly on the range of the t_{ij}s. In the simulation, we set $n_{grid} = 50$. Similarly, we can define the RASE of the eigenfunctions for the c-th component, which is

$$\text{RASE}_{v_c}^2 = n_{grid}^{-1} \sum_{q=1}^{Q_c} \sum_{j=1}^{n_{grid}} \{\hat{v}_{qc}(u_j) - v_{qc}(u_j)\}^2.$$

We are also interest in the average of mean square of predicted error, given by

$$\text{MSE} = (\sum_{i=1}^{n} N_i)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{N_i} \left\{ y_{ij} - \sum_{c=1}^{C} r_{ic} \hat{X}_{ic}(t_{ij}) \right\}^2,$$

where $\hat{X}_{ic}(t_{ij})$ is defined in (5.21). MSE is considered as a natural estimate of σ^2.

5.4.1 Simulation Study

In the following example, we generate data from a two component mixture of Gaussian processes with

$$\pi_1 = 0.45, \quad \pi_2 = 0.55, \quad \text{and} \quad \sigma^2 = 0.01,$$
$$\mu_1(t) = \delta + 1.5\sin(\pi t), \text{ and } \mu_2(t) = \sin(\pi t),$$
$$\phi_{11}(t) = \sqrt{2}\sin(4\pi t), \text{ and } \phi_{12}(t) = \sqrt{2}\cos(4\pi t),$$
$$\phi_{21}(t) = \sqrt{2}\sin(\pi t), \text{ and } \phi_{22}(t) = \sqrt{2}\cos(\pi t).$$

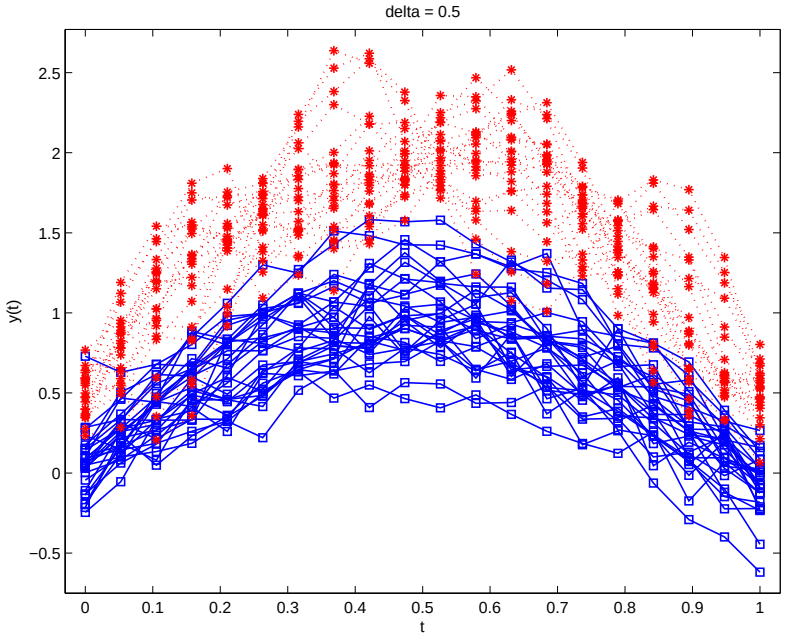
Suppose t is evenly distributed in $[0, 1]$. The data are observed at grid points $t = (1/N, \cdots, 1)$, where N is set to be 20 and 40. Let the eigenvalues for both components be $\lambda_{11} = 0.04$, $\lambda_{12} = 0.01$, $\lambda_{21} = 0.04$, $\lambda_{22} = 0.01$, and $\lambda_{qc} = 0$, for $q > 2$, $c = 1, 2$, and let the principal component scores ξ_{iqc} be generated from $N(0, \lambda_{qc})$, $q = 1, 2$, and $c = 1, 2$.

We consider two classes of simulation data sets from the above generation scheme. In the first class of simulations, we set $\delta = 0.5$, and thus the subjects of the two components are well separated. In this setting, the difference of the two components are mainly from the difference in their mean functions, and thus the estimation procedure with working independent correlation structure is expected to work well. In the second class we set $\delta = 0$, and the mean functions of the two components are close to each other. Thus, the subjects of the two components are heavily overlapping. In this setting, estimation with working independent correlation structure may not work well. We expect that the difference in correlation structures can be used to improve the estimation, and expect that models which incorporate correlations will yield better results than models with working independent correlation structure. Figure 5.1 shows a typical sample data set for the two classes. In the following simulation, we compare

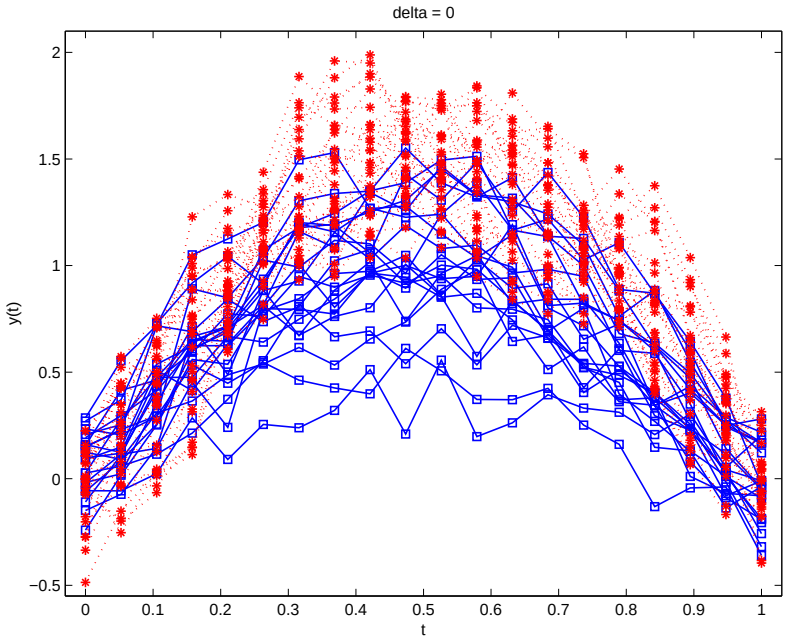
the performance of the two models in both the well-separated setting, and the heavy-overlap setting. For the heavy-overlap setting, we further investigate the performance of eigenfunction estimation using model (5.22).

We first fit a multivariate normal mixture model, which gives the estimates of mean $\bar{\mu}_c(\cdot)$, covariance matrixes $\bar{\Sigma}_c$, and $\bar{\pi}_c$. This would provide us a good initial value. To avoid intensive computation, the smoothing parameter for covariance functions should be predetermined. We use one-curve-leave-out cross validation to choose this smoothing parameter. The selection of the bandwidth for covariance functions is based on $\bar{\Sigma}_c$. In our simulation, we first generate several simulation data sets for a given sample size, and then use the CV bandwidth selectors to choose a bandwidth for each data set. This provides us an idea about the optimal bandwidth for a given sample size. For a typical sample from the overlap setting with $n = 50$, $N = 40$, and given a bandwidth of covariance (0.06), we use 5-fold cross-validation method to select the bandwidth for mean functions. As shown in Figure 5.2, the CV bandwidth selector yields the bandwidth 0.07.

To demonstrate the proposed procedure working quite well in a wide range of bandwidth, we consider three different bandwidths: two-thirds of the selected bandwidth, the selected bandwidth, and 1.5 times the selected bandwidth, which corresponds to the under-smoothing, appropriate smoothing and overs-moothing, respectively. Table 5.1 displays the simulation results of the case that $\delta = 0.5$. The mean and standard deviation of RASE_μ, and the estimate of π over 100 simulations are recorded for both models. From Table 5.1, the proposed procedures performs quite well for all three different bandwidths in the two models. Table 5.2 displays the simulation results of the case that $\delta = 0$. The mean and standard deviation of RASE_μ, and the estimate of π over



(a)



(b)

Figure 5.1: (a) A typical sample data of the well-separated setting, $\delta = 0.5$; (b) A typical sample data of the heavy-overlap setting $\delta = 0$.

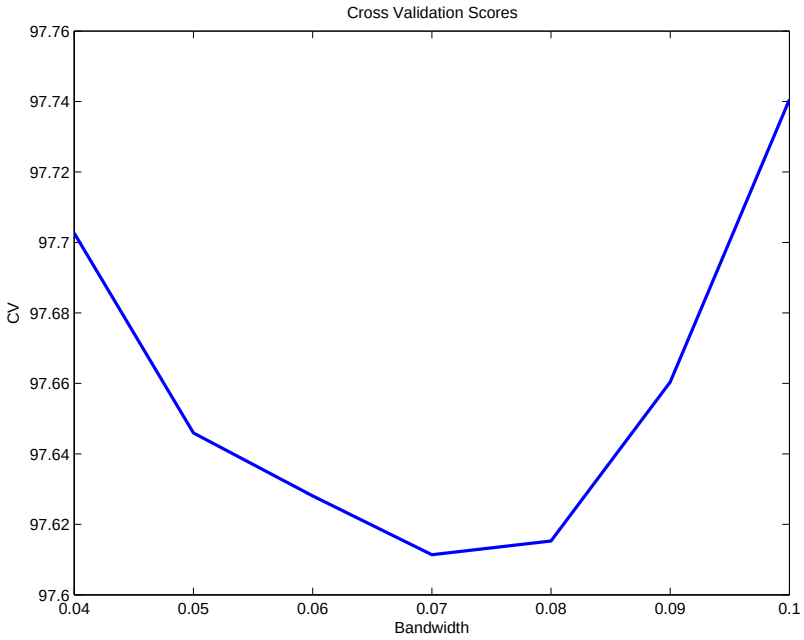


Figure 5.2: Cross-validation error versus the bandwidth. Settings: n=50, N=40, $\delta = 0$.

Table 5.1: Comparisons: the well-separated setting

		Working independent		Incorporating correlation	
$\delta = 0.5$		RASE_μ	$\pi_1 = 0.45$	RASE_μ	$\pi_1 = 0.45$
(n, N)	h	Mean(Std)	Mean(Std)	Mean(Std)	Mean(Std)
(50, 20)	0.15	0.2157(0.0104)	0.4382(0.0688)	0.2200(0.0075)	0.4458(0.0679)
	0.10	0.2061(0.0108)	0.4384(0.0687)	0.2105(0.0082)	0.4458(0.0679)
	0.067	0.2063(0.0113)	0.4488(0.0698)	0.2105(0.0087)	0.4562(0.0694)
(50, 40)	0.09	0.2119(0.0115)	0.4538(0.0707)	0.2164(0.0085)	0.4614(0.0702)
	0.06	0.2089(0.0118)	0.4540(0.0705)	0.2135(0.0090)	0.4614(0.0702)
	0.04	0.2086(0.0116)	0.4520(0.0713)	0.2110(0.0090)	0.4558(0.0711)

Table 5.2: Comparisons: the heavy-overlap setting

		Working independent		Incorporating correlation	
$n = 50$		RASE_μ	$\pi_1 = 0.45$	RASE_μ	$\pi_1 = 0.45$
N	h	Mean(Std)	Mean(Std)	Mean(Std)	Mean(Std)
20	0.15	0.3533(0.0263)	0.3049(0.0791)	0.3140(0.0073)	0.4828(0.0838)
	0.10	0.3436(0.0232)	0.3155(0.0788)	0.3103(0.0075)	0.4618(0.0787)
	0.067	0.3508(0.0303)	0.3140(0.0645)	0.3097(0.0079)	0.4746(0.0725)
40	0.09	0.3392(0.0222)	0.3235(0.0713)	0.3065(0.0065)	0.4693(0.0668)
	0.06	0.3358(0.0211)	0.3304(0.0715)	0.3067(0.0068)	0.4657(0.0677)
	0.04	0.3355(0.0246)	0.3344(0.0644)	0.3069(0.0064)	0.4533(0.0691)

100 simulations, are recorded for both models. From Table 5.2, the estimation procedure with working independent correlation does not perform well since the estimation of π_1 has large bias. However, as shown in Table 5.2, model the estimation procedure incorporating correlations does give better results. For the case of $\delta = 0$, we further summarize the estimation of σ^2, MSE, and the RASE of the eigenfunctions for each component in Table 5.2. The result show that both the $\hat{\sigma}^2$ yielded by the iterative procedure and the MSE are good estimates of σ^2. Furthermore, the iterative procedure also provides a good estimate of the eigenfunctions in the heavy overlap setting.

5.4.2 Analysis of Supermarket Data

We apply the proposed mixture of Gaussian processes and estimation procedure to analyze a real dataset, which contains the number of customers visiting a supermarket on each of 139 days. For each day, the number of cus-

Table 5.3: Estimation of eigenfunctions and measurement error

$n = 50$		RASE_{v_1}	RASE_{v_2}	MSE	$\hat{\sigma}^2 = 0.01$
N	h	Mean(Std)	Mean(Std)	Mean(Std)	Mean(Std)
20	0.15	0.3974(0.3934)	0.3749(0.2800)	0.0112(0.0011)	0.0113(0.0011)
	0.10	0.3862(0.3568)	0.3357(0.2796)	0.0093(0.0008)	0.0093(0.0008)
	0.067	0.3912(0.3582)	0.3341(0.2798)	0.0087(0.0007)	0.0087(0.0007)
40	0.09	0.2136(0.0986)	0.2473(0.1469)	0.0100(0.0003)	0.0100(0.0003)
	0.06	0.2342(0.1165)	0.2302(0.1410)	0.0094(0.0004)	0.0094(0.004)
	0.04	0.2182(0.1038)	0.2420(0.1449)	0.0093(0.0003)	0.0093(0.0003)

tomers in the supermarket is recorded every half hour from 7:00am to 5:30pm. In the analysis, we regard each day as one subject. Thus, we have 139 subjects in total. Figure 5.3 depicts the plot of this data set.

Results of two component model

We first analyze the data using a working independent correlation model with two components. The smoothing parameter is chosen to be 0.063. The estimated proportions of the two components are 0.2138 and 0.7862. The estimated mean functions and a hard-clustering result are shown in Figure 5.4. The hard-clustering is is obtained by assigning component identities according to the largest $r_{ic}, c = 1, \cdots, C$. From the hard-clustering result and the original data, we found that the days in the upper component are mainly from those which are 1-3 days before weekends and holidays. The estimated mean functions can be viewed as estimated average customer flows of the two classes. We observed that there are two peaks of customer flows for both components. The first peak occurs around 9:00 am in both components. The second peak occurs around

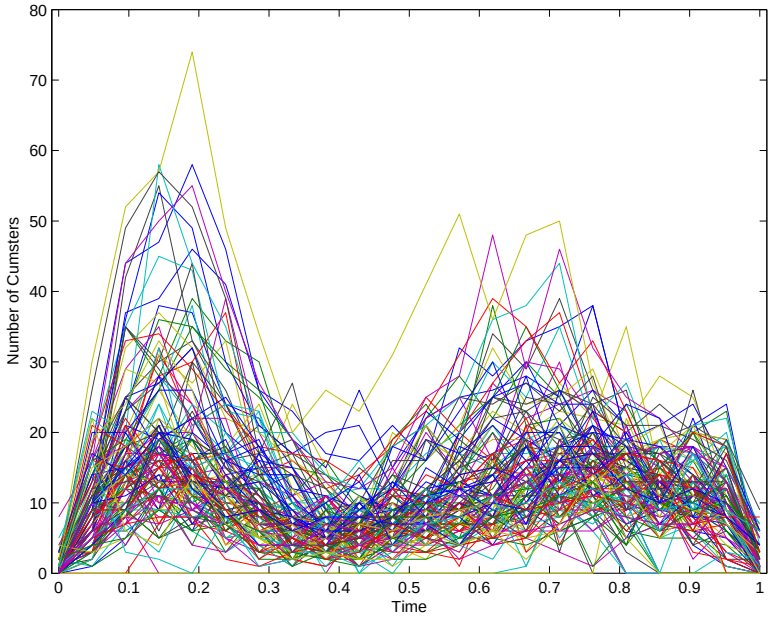


Figure 5.3: Plot of supermarket data

2:00 pm for the first component, and 3:00 pm for the second component. This pattern may indicate that people tend to buy earlier in the afternoon if they are preparing for holidays and weekends. We further plot the estimated variance functions of the two component in Figure 5.5. Combining Figure 5.4 and 5.5, we observed that the variance functions incerese along with the mean functions in both components, in that the higher the mean, the higher the associated variance.

The next step is to analyze the data by incorporating the estimated correlations. Based on the estimated posterior, we estimate the covariance functions and obtain estimates of the eigenfunctions of both components. We plot the first two eigenfunctions of both components in Figure 5.6. For the first component (upper class), the first eigenfunction explains 59.42% of the total variation, and has a negative value along its time interval from 9:00 am to 5:30 pm. It means

that a subject (i.e., a day) with a positive (negative) loading score on this direction tends to have larger (smaller) customer flows than the population average in a whole observed time interval. We also observe that there are two negative peaks in the first eigenfunction, which occurs around 9:00 am and 2:00 pm. It means that the variations of the costumor flows are large in the two peaks, especially for the peak at 9:00 am. Note that these peaks are also observed in the first estimated variance function, that the results agree with each other as we expected. The second eigenfunction, which explains 13.57% of the total variation, has negative values in the morning time and positive values in the afternoon. This means that a subject with a positive loading score on this direction tends to have smaller costumer flow in the morning and a higher costumer flow in the afternoon. The variation characterized by the second eigenfunction has a minor magnitude compared to the variation in the first eigenfunction, where the magnitude is determined by the eigenvalues. The third eigenfunction explains 5.67% of the total variation, and is of little interest. Similarly we can interpret the eigenfunctions of the second component. Further analysis shows that incorporating the estimated covariances does not lead to significant improvement, and thus those results are not reported.

Results of three component model

We next analyze the data using model (5.3) with three components. The smoothing parameter is chosen to be 0.063. The estimated proportions of the three components are 0.1632, 0.4308, and 0.4060. The estimated mean functions and a hard-clustering result based on posterior estimate of identities are shown in Figure 5.7. The estimated variance functions of the three component are plotted in Figure 5.8. The MSE of the two component model is 13.08, while the MSE of the three component model is 12.89. Compared to the two component

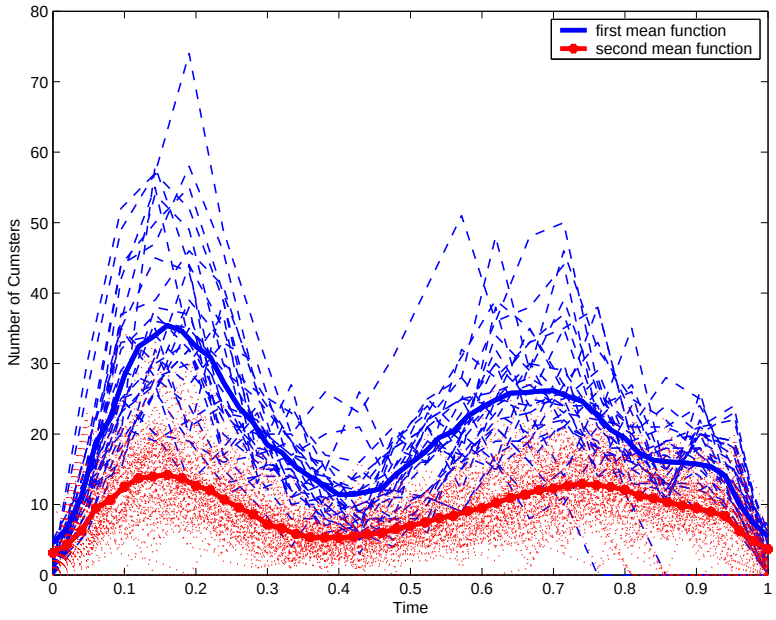


Figure 5.4: Estimated mean functions and clustering results based on posteriors.

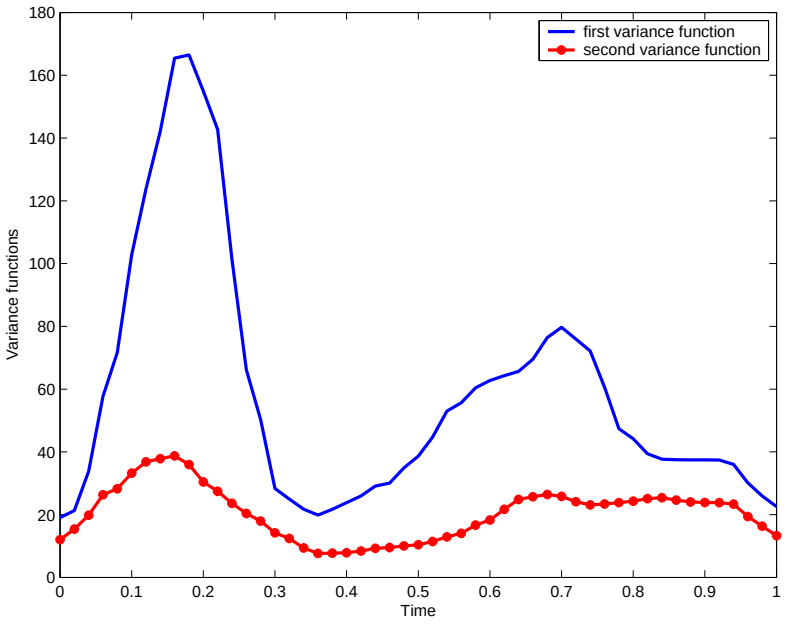
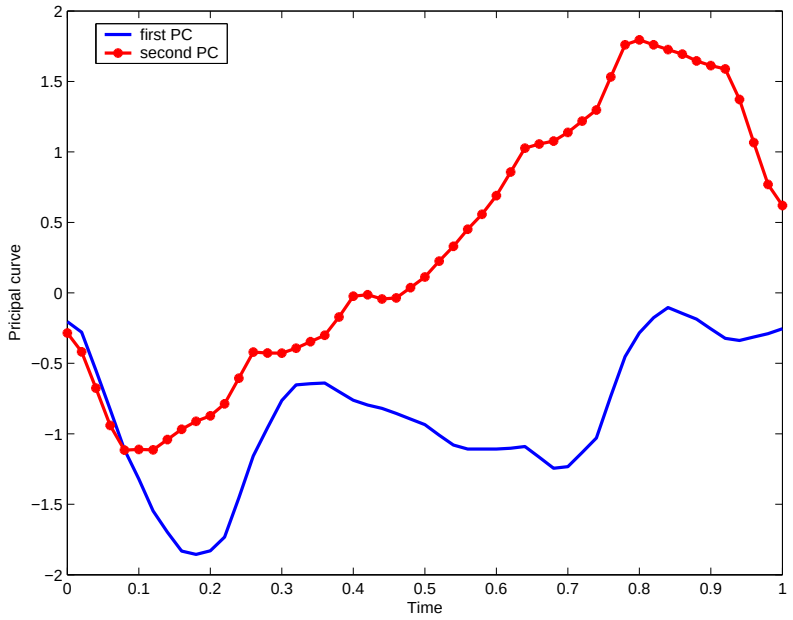
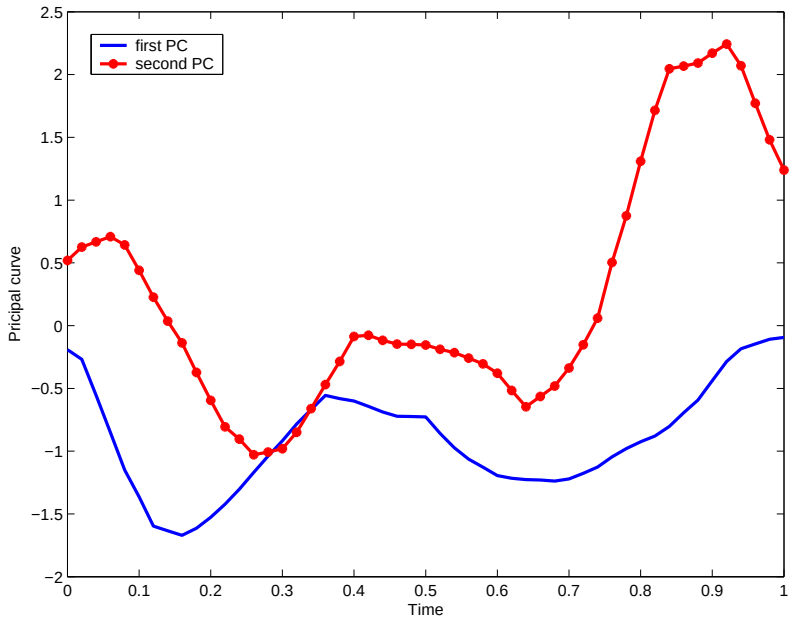


Figure 5.5: Estimated variance functions.



(a)



(b)

Figure 5.6: (a) First two eigenfunctions of the upper cluster; (b) First two eigenfunctions of the lower cluster.

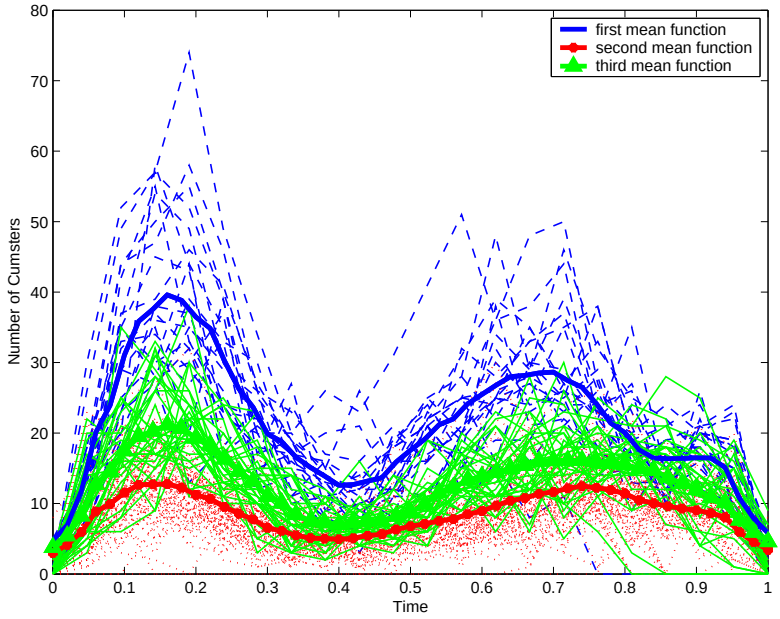


Figure 5.7: Estimated mean functions and clustering results based on posteriors.

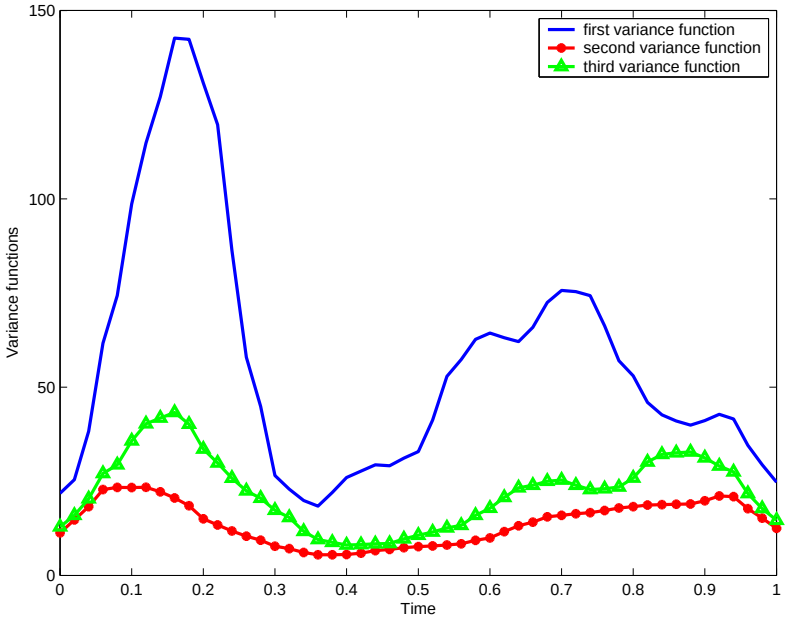


Figure 5.8: Estimated variance functions.

model, the MSE of the three component model is not significantly reduced. This result may suggest that a two component model is enough for the analysis of the supermarket dataset.

5.5 Discussion

In this chapter we introduced mixtures of Gaussian processes for functional data analysis, and developed estimation procedures for model estimation. We conducted a simulation study to evaluate the performance of the proposed model, and illustrated the methodologies by an application in the analysis of supermarket data. Yet there are many issues needing further research. The asymptotic properties of the resulting estimates have not been investigated. The selection of the number of components, the number of eigenfunctions in each component, and the bandwidth selection for the covariance functions needs further research and discussion.

In the simulation and the analysis of supermarket data, we see that the proposed procedures perform quite well for balanced data. The computations are very efficient given a good initial value and a reasonably well-chosen bandwidth. We do not have an accurate idea of the minimum requirements for the dataset, i.e., how large (and dense) the dataset need to be for plausible inference. For balanced data, the model can handle high dimensional data where multivariate Gaussian mixtures fail to work, since our approach perform dimension reduction during the estimation, and maintain the main feature of the original data in a relatively small model. For instance, in the simulation example in section 5.4, the model can handle simulated data up to $n = 100$, $N = 400$, without changing of other settings. Further research in the application of unbalanced data is of interest.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis, we study nonparametric techniques in two newly proposed models: nonparametric mixture of regression models, and mixture of Gaussian processes.

In Chapter 3, we develop an effective estimation procedure for nonparametric mixture of regression model via local likelihood approach and EM algorithm. The modified EM algorithm automatically avoids an issue similar to label switching problem, and enable us to simultaneously maximize the local likelihood functions for the proposed nonparametric mixture of regression model at set of grid points. We demonstrate that the proposed EM algorithm preserves the ascent property in an asymptotic sense. We further established the asymptotic normality of the resulting estimate, conduct Monte Carlo simulation studies, and apply the newly proposed model to real data analysis.

In chapter 4, we study the nonparametric finite mixture of regression models with constant mixing proportion, which is indeed a semiparametric model. We develop an estimation procedure by using back-fitting algorithm, and further suggest one-step back-fitting algorithm, which behaves similar to the gradient ECM algorithm. We studied the asymptotic properties of the resulting estimate,

and demonstrate that the mixing proportion parameter is root n consistent, and follows an asymptotic normal distribution. The asymptotic properties for the resulting estimate of the regression function and variance function are given. Monte Carlo simulation is conducted to assess the performance of the proposed procedure. We further apply the proposed procedure to real data analysis.

In Chapter 5, we propose mixture Gaussian processes, and the estimation procedure for the newly proposed model by using back-fitting algorithm. We further introduce functional principal component analysis as a convenient way to effectively incorporate covariance in the estimation procedure. We examine the finite sample performance of the proposed procedure by Monte Carlo simulation studies, and further apply the proposed procedure to analyze a supermarket dataset.

6.2 Future Work

6.2.1 Mixture of Varying Coefficient Models

The proposed nonparametric mixture of regression models can be naturally extended to mixture of varying coefficient models. Varying coefficient model is a useful extension of classical linear models. Among several nonparametric regression models, the varying coefficient model can be used to explore some functional features in high dimensional data. The varying coefficient model has the following structure:

$$Y = \sum_{j=1}^{J} a_j(U)X_j + \epsilon,$$

where $E(\epsilon|U, X_1, \cdots, X_J) = 0$, and $\mathrm{Var}(\epsilon|U, X_1, \cdots, X_J) = \sigma^2(U)$. By allowing the coefficients $a_j(\cdot)$ depend on U, we can study how the response variable depends on predictors $X_1, \cdots, X_J$ over a range of covariate U, such as time and

temperature and so on. The idea of the varying coefficient model is first introduced by Cleveland *et al.*(1992) and extended in Hastie and Tibshirani (1993). For reference of estimation procedure and inference, see Fan and Zhang (2000) and Cai, Fan and Li (2000). We will further study the estimation procedure, asymptotic properties, and the applications of mixtures of varying coefficient models.

6.2.2 Testing Hypothesis

Future work should include testing hypothesis for the proposed nonparametric mixture of regression model and mixture of Gaussian processes. In practice, it is of interest to know whether the unknown function has a closed parametric form or not. This consideration leads to the following hypothesis test:

$$H_0 : m(x) = \beta_0 + \beta_1 x \quad \text{versus} \quad H_a : m(x) \neq \beta_0 + \beta_1 x.$$

Under the null hypothesis, it is a parametric model, and therefore we can estimate the parameters and calculate the likelihood. Under the alternative hypothesis, we can also calculate the likelihood based on the local likelihood estimators. Denote the log-likelihood under H_0 by $\ell(H_0)$, and the log-likelihood under H_a by $\ell(H_a)$. A generalized likelihood ratio statistic is defined by

$$T = \ell(H_a) - \ell(H_0).$$

The likelihood ratio statistic asymptotically follows a chi-square distribution when H_a is a parametric model. The degrees of freedom is the difference of the effective number of parameters in the two models. Several papers deal with testing problems in varying coefficient models. Theoretical asymptotic results when H_a is a nonparametric model are given by Cai, Fan and Li (2000) and Fan, Zhang and Zhang (2001).

6.2.3 High Dimensional Gaussian Mixtures

The methodologies developed in Chapter 5 may provide some efficient approaches for high dimensional Gaussian mixture models. Consider a multivariate mixture model

$$\mathbf{y} \sim \sum_{c=1}^{C} \pi_c N(\mu_c, \Sigma_c),$$

where $\mathbf{y}$ is a high dimensional random vector. The log-likelihood function of data $\{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$ is

$$\sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \pi_c \phi\Big(\mathbf{y}_i | \mu_c, \Sigma_c\Big) \right\}.$$

The multivariate mixture model may suffer the problem of the "curse of dimensionality", and the computation of the inverse of $\hat{\Sigma}_c$ may not be stable. We suggest an alternative approach for these issues. Consider

$$\mathbf{y} \sim \sum_{c=1}^{C} \pi_c N(\mu_c + \sum_{q=1}^{Q_c} \xi_{iqc} \mathbf{v}_{qc}, \sigma_c^2), \tag{6.1}$$

where $\xi_{iqc} = (\mathbf{y}_i - \mu_c)^T \mathbf{v}_{qc}$, and $\mathbf{v}_{qc}$s are orthogonal with norm 1. That is, for $c = 1, \cdots, C$, we constrain $\mathbf{v}_{qc}^T \mathbf{v}_{qc} = 1$ and $\mathbf{v}_{qc}^T \mathbf{v}_{q'c} = 0$ if $q \neq q'$. Thus, the log-likelihood function is

$$\sum_{i=1}^{n} \log \left\{ \sum_{c=1}^{C} \pi_c \phi\Big(\mathbf{y}_i | \mu_c + \sum_{q=1}^{Q_c} \xi_{iqc} \mathbf{v}_{qc}, \sigma_c^2 I\Big) \right\}. \tag{6.2}$$

From our experience in simulation, the EM algorithm for model (6.2) performs quite well for high dimensional data with correlation structure. Bayesian approaches for model (6.1) may provide comprehensive solutions, including choosing the number of components. Future work on these models is of interest.

6.2.4 Other Issues

In Chapter 3, we focus on estimation of the newly proposed nonparametric finite mixture of regression models when x_0 is an interior point in the range of the covariate. It is certainly also of interest to study the boundary performance of the proposed procedure. The boundary effect has been studied in Cheng, Fan and Marron (1997) for the nonparametric regression model. We may apply local polynomial regression for the mean function, and kernel regression for variance function and mixing proportion functions. It is of interest to test hypotheses such as whether the mixing proportions are constants, whether some of mean functions are constant or of specific parametric forms, and whether the variance functions are parametric. It is of interest to use different bandwidths for different functions, which has not been investigated in this thesis. These can be for further research.

The proposed models focus on mixtures of normal distributions. We may extend the developed methodologies for mixture of other distributions. For example, we could further study the nonparametric mixture of Poisson regressions, the nonparametric mixture of generalized linear models, or mixture of Poisson processes.

Bibliography

[1] Akaike, H. (1974). A New Look at the Statistical Model Identication. *IEEE Trans. on Automatic Control*. 19, 716-723.

[2] Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized Partially Linear Single-Index Models. *Journal of American and Statisical Association*. 92, 477-489.

[3] Carroll, R. J., Ruppert, D., and Welsh, A. H. (1998). Local Estimating Equations. *Journal of American and Statisical Association*. 93, 214-227.

[4] Cai, Z, Fan, J. and Li, R. (2000). Efficient Estimation and Inferences for Varying-coefficient Models. *Journal of American and Statisical Association*. 95, 888-902.

[5] Chen, J. and Kalbfleisch, J. D. (1996). Penalized Minimum-distance Estimates in Finite Mixture Models. *The Canadian Journal of Statistics*. 24, 167- 175.

[6] Cheng, M. Y., Fan, J. and Marron, J. S. (1997). On Automatic Boundary Corrections. *Annals of Statistics*. 25, 1691-1708.

[7] Cleveland, W. S., Grosse, E. and Shyu, W. M. (1992). Local Regression Models. *Statistical Model in S*. 309-376, Pacific Grove, California: Wadsworth & Brooks.

[8] Dempster, A. P., Laird, N. M. and Rubin, B. D. (1977). Maximum Likelihood from Incomplete Data via EM Algorithm. *Journal of American and Statisical Association*. 39, 1-38.

[9] Fan, J. (1993). Local Linear Regression Smoothers and Their Minimax Efficiencies. *Annals of Statistics*. 21, 196-216.

[10] Fan, J and Gijbels, I. (1992). Variable Bandwidth and Local Linear Regression Smoothers. *Annals of Statistics*. 20, 2008-2036.

[11] Fan, J., Heckman, N. E., and Wand, M. P. (1995). Local Plynomial Kernel Regression For Generalized Linear Model and Quasilikelihood Functions. *Journal of American and Statisical Association*. 90, 141-150.

[12] Fan, J., Gijbles, I., and King, M. (1997). Local Likelihood and Local Partial Likelihood in Hazard Regression. *Annals of Statistics*. 25, 1661-1690.

[13] Fan, J., Farmen, M., and Gijbels,I. (1998). Local Maximum Likelihood Estimation and Inference. *Journal of the Royal Statistical Society, Ser. B*. 60, 591-608.

[14] Fan, J. and Gijbles, I. (1996). Local Polynomial Modelling and Its Applications. Chapmam and Hall, London.

[15] Fan, J. and Yao, Q. (1998). Efficient Estimation of Conditional Variance Functions in Stochastic Regression. *Biometrika*. 85, 645-660.

[16] Fan, J. and Zhang, J. (2000). Functional Linear Models for Longitudinal Data. *Journal of the Royal Statistical Society, Ser. B*. 62, 303-332.

[17] Fan, J., Zhang, C., and Zhang, J. (2001). Generalized Likelihood Ratio Statistics and Wilks Phenomenon. *Annals of Statistics*. 29, 153-193.

[18] Frühwirth-Schnatter, S. (2001). Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models. *Journal of American and Statisical Association*. 96, 194-209.

[19] Frühwirth-Schnatter, S. (2005). Finite Mixture and Markov Switching Models. Springer.

[20] Goldfeld, S. M. and Quandt, R. E. (1976). A Markov Model for Switching Regression. *Journal of Econometrics*. 1, 3-16

[21] Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*. 82, 711-732.

[22] Green, P. J. and Richardson, S. (2002). Hidden Markov Models and Disease Mapping. *Journal of American and Statisical Association*. 97, 1055-1070.

[23] Hastie, T.J. and Tibshirani, R.J. (1993). Varying Coefficient Models (with discussion). *Journal of the Royal Statistical Society, Ser. B*. 55, 757-796.

[24] Hathaway, R. J. (1985). A Constrained Formulation of Maximum-likelihood Estimation for Normal Mixture Distributions. *Annals of Statistics*. 13, 795-800.

[25] Hennig, C. (2000). Identifiability of Models for Clusterwise Linear Regression. *Journal of Classification*. 17, 273-296.

[26] Heard, N. A., Holmes, C. C., and Stephens, D. A. (2006). A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves. *Journal of American and Statisical Association*. 101, 18-29.

[27] Hall, P., Müller, H. G., and Wang, J. L. (2006). Properties of Principal Component Methods for Functional and Longitudinal Data Analysis. *Annals of Statistics*. 100, 577-590.

[28] James, G. M. and Sugar, C. A. (2003). Clustering Sparsely Sampled Functional Data. *Journal of American and Statisical Association*. 98, 397-408.

[29] James, G. M., Hastie, T. J., and Sugar, C. A. (2000). Principal Component Models for Sparse Functional Data. *Biometrika*. 87, 587-602.

[30] Lange, K. (1995). A Gradient Algorithm Locally Equivalent to the EM Algorithm. *Journal of the Royal Statistical Society, Ser. B*. 57, 425-437.

[31] Leroux, B. G. (1992). Consistent Estimation of a Mixing Distribution. *Annals of Statistics*. 20, 1350-1360.

[32] 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.

[33] Lindsay, G. B. (1983a). The Geometry of Mixture Likelihoods: A General Theory. *Annals of Statistics*. 11, 86-94.

[34] Lindsay, G. B. (1983b). The Geometry of Mixture Likelihoods: Part 2, the exponential family. *Annals of Statistics*. 11, 783-792.

[35] Lindsay, B. G. (1995). Mixture Models: Theory, Geometry, and Applications. Hayward, Institute of Mathematical Statistics.

[36] Luan, Y. and H. Li (2003). Clustering of Time-course Gene Expression Data Using a Mixed-effects Models With B-spline. *Bioinformatics*. 19, 474-282.

[37] Luan, Y. and H. Li (2004). Model-based Methods for Identifying Periodically Regulated Genes Based on the Time Course Microarray Gene Expression Data. *Bioinformatics*. 20, 332-339.

[38] Ma, P. and Zhong, W. (2008). Penalized Clustering of Large Scale Functional Data with Multiple Covariates. *Journal of the American Statistical Association*. 103, 625-636.

[39] Marron, J. S., and Nolan, D. (1988) Canonical Kernels for Density Estimation. *Stat. Prob. Lett.*. 7, 195-199.

[40] McLachlan, G. J. and Krishnan, T. (1997). The EM algorithm and Extensions. Wiley, New York.

[41] McLachlan, G. J. and Peel, D. (2000). Finite Mixture Models. Wiley, New York.

[42] Meng, X.-L. and Rubin, D. B. (1993). Maximum Likelihood Estimation Via The ECM Algorithm: A General Framework. *Biometrica*. 80, 267-278.

[43] Nadaraya, E. A. (1964). On Estimating Regression. *Theory Prob. Appl.*. 10, 186-190.

[44] Pollard, D. (1991). Asymptotics for Least Absolute Deviation Regression Estimators. *Econometric Theory*. 7, 186-199.

[45] Ramsay, J. O. and Silverman, B. W. (1997). Functional data analysis. Springer-Verlag, New York.

[46] Rice, J., and Wu, C. (2000). Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves. *Biometrics*. 57, 253-259.

[47] Rice, J., and Silverman, B. (1991). Estimating the Mean and Covariance Structure Nonparametrically When the Data Are Curves. *Journal of the Royal Statistical Society, Ser. B.* 53, 233-243.

[48] Rossi, P. E., G. M. Allenby, and R. McCulloch (2005). Bayesian Statistics and Marketing. Chichester: Wiley.

[49] Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of American and Statisical Association*. 90, 1057-1270.

[50] Stephens, M. (2000). Bayesian Methods for Mixture of Normal Distribution - An Alternative to Reversible Jump Methods. *Annals of Statistics*. 28, 40-74.

[51] Staniswalis, J. G., and Lee, J. J. (1998). Nonparametric Regression Analysis of Longitudinal Data. *Journal of the American Statistical Association*. 93, 1403-1418.

[52] Smith, A. F. M. and Roberts, G. O. (1993). Baysian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society, Ser. B*. 55, 3-23.

[53] Simonoff, S. J. (1998). Smoothing Methods in Statistics. Springer, New York.

[54] Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*. 19, 461-464.

[55] Sugar, C., and James, G. (2003). Finding the Number of Clusters in a Data Set : An Information Theoretic Approach. *Journal of the American Statistical Association*. 98, 750-763.

[56] Tibshirani, R. and Hastie, T.(1987). Local Likelihood Estimation. *Journal of American and Statisical Association*. 82, 559-567.

[57] Wang, P., M. L. Puterman, I. Cockburn, and N. Le (1996). Mixed Poisson Regression Models with Covariate Dependent Rates. *Biometrics*. 52, 381-400.

[58] Watson, G. S. (1964). Smooth Regression Analysis. *Sankhya*. Ser. A, 26, 359-372.

[59] Wedel, M. and W. S. DeSarbo (1993). A Latent Class Binomial Logit Methodology for the Analysis of Paired Comparison Data. *Decision Sciences*. 24, 1157-1170.

[60] Wu, C. F. J. (1983). On the Convergence Properties of the EM Algorithm. *Annals of Statistics*. 11, 95-103.

[61] Yao, F., Mller, H. G., Clifford, A. J., Dueker, S. R., Follett, J., Lin, Y., Buchholz, B. A., and Vogel, J. S. (2003). Shrinkage Estimation for Functional Principal Component Scores With Application to the Population Kinetics of Plasma Folate. *Biometrics*. 59, 676-685.

[62] Yao, F., Müller, H. G., and Wang, J. L. (2005). Functional Data Analysis for Sparse Longitudnal Data. *Journal of American and Statisical Association*. 100, 577-590.

[63] Young, D. S. (2007). A study of Mixtures of Regressions. Ph.D. Dissertation, The Pennsylvania State University. Unpublished.

Vita

Mian Huang

Education

- Ph.D. in Statistics (August, 2009), Pennsylvania State University, University Park, PA
- M.S. in Statistics (December, 2006), Pennsylvania State University, University Park, PA
- B.S. in Statistics (July, 2004), University of Science and Technology of China, Hefei

Professional Experience

- 2008-2009, Research Assistant, Department of Statistics, The Pennsylvania State University, University Park, PA
- 2005-2006, Student Consultant, Statistical Consulting Center, The Pennsylvania State University, University Park, PA
- 2004-2007, Teaching Assistant, Department of Statistics, The Pennsylvania State University, University Park, PA