ESTIMATION AND FORECASTING METHODOLOGIES FOR
NONPARAMETRIC REGRESSION MODELS VIA DYNAMIC LINEAR
MODELS

A Dissertation in
Statistics
by
Yuejiao Jiang

© 2009 Yuejiao Jiang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2009
The thesis of Yuejiao Jiang was reviewed and approved* by the following:

Runze Li  
Professor of Statistics  
Dissertation Advisor, Chair of Committee

Jingzhi Huang  
Associate Professor of Finance

Murali Haran  
Assistant Professor of Statistics

Aleksandra B. Slavković  
Assistant Professor of Statistics

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

*Signatures are on file in the Graduate School.
Abstract

Nonparametric regression models have been used to explore features of data. Various estimation procedures have been proposed for estimating nonparametric regression models in the literature. The existing procedures may enjoy some statistically optimal properties, but it is difficult to use the estimated regression functions for the purpose of forecasting, i.e., prediction with independent variables beyond the observed range. Motivated by dynamic linear model estimation and forecasting methodologies developed by West and Harrison (1997), our research extends the study in two directions by building connections between nonparametric regression models and dynamic linear models.

First, for nonparametric regression models with independently and identically distributed errors, we develop a regularized regression spline to determine hyper-parameter values for errors and compare estimation performance with a local linear estimator, a commonly used smoothing method for nonparametric regression. Second we study nonparametric regression models with autoregressive errors process since when collected over time, data are likely to be serially correlated. From our simulation studies, the proposed methods outperform the local linear regression under certain statistical settings for both models with i.i.d. and autoregressive error. Last, we establish a connection between varying coefficient models
and dynamic linear models. Both estimation and forecasting performances of the proposed method are examined via simulations. The proposed methods perform well for larger sample sizes. The proposed procedures are demonstrated by analysis of two real data examples.
# Table of Contents

List of Figures viii  
List of Tables x  
Acknowledgments xi  

<table>
<thead>
<tr>
<th>Chapter 1</th>
<th>Introduction</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.1 Nonparametric regression model with independent data</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1.2 Nonparametric regression model with serially correlated errors</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1.3 Varying coefficient model</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>1.4 The outline of the dissertation</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 2</th>
<th>Literature Review</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.1 Bayesian Inference and MCMC</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>2.1.1 Bayesian Inference</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>2.1.2 Markov Chain Monte Carlo</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>2.2 Dynamic Linear Model</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>2.3 Univariate nonparametric regression models</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>2.3.1 Local linear estimation</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>2.3.2 Bayesian nonparametric regression models</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>2.4 Varying coefficient model</td>
<td>25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 3</th>
<th>A Dynamic Linear Model Approach for Nonparametric Regression Model</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.1 Introduction</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>3.2 Regularized linear regression splines and Dynamic Linear Models</td>
<td>33</td>
</tr>
</tbody>
</table>
### Chapter 3

3.3 The Gibbs sampler algorithm .................................................. 36
3.3.1 Fully Bayesian estimation procedure ........................................ 36
3.3.2 Empirical Bayesian estimation procedure .................................... 37
3.4 Simulations ............................................................................ 39
3.5 Discussion ............................................................................. 43

### Chapter 4

#### Nonparametric Regression Function with Serially Correlated Errors

4.1 Introduction ........................................................................... 48
4.2 An estimation and forecasting DLM framework for nonparametric regression model with correlated errors ........................................... 49
4.2.1 A connection between DLM and penalized splines ....................... 50
4.2.2 Dynamic linear models with an AR error ................................... 51
4.3 The Gibbs sampler algorithm .................................................... 55
4.4 Simulations and an application .................................................. 56
4.4.1 Simulation study .................................................................. 57
4.4.2 Interest rate data set .............................................................. 59
4.5 Proof ..................................................................................... 67
4.6 Discussion ............................................................................. 78

### Chapter 5

#### A DLM Approach for Varying Coefficient Models

5.1 Introduction ........................................................................... 79
5.2 An estimation and forecasting DLM framework for varying coefficient models with correlated errors ........................................... 80
5.3 The Gibbs sampler algorithm .................................................... 82
5.4 Simulations and an application .................................................. 84
5.4.1 Simulation study .................................................................. 84
5.4.2 Environmental health data set ................................................ 90

### Chapter 6

#### Conclusion

6.1 Research contributions .............................................................. 97
6.2 Future research interests ........................................................... 98

### Appendix A

#### Dynamic Linear Models .......................................................... 102

### Appendix B

#### Forward Filtering and Backward Sampling ............................... 107
List of Figures

1.1 An illustration of the presence of both low and high frequency signals in the regression function. ................................................................. 3
1.2 Rates on 3-month T-Bills ............................................................... 5
3.1 Plot of a typical estimation for a fixed design with $T = 500$ .............. 44
3.2 Plot of a typical estimation for a fixed design with $T = 1000$ .......... 45
3.3 Plot of a typical estimation for a random design with $T = 500$ ........ 46
3.4 Plot of a typical estimation for a random design with $T = 1000$ ...... 47
4.1 Plot of a typical estimator for a fixed design with AR(1) model error, $\rho = 0.4$ and $T = 500$ ................................................................. 60
4.2 Plot of a typical estimator for a fixed design with AR(1) model error, $\rho = 0.4$ and $T = 1000$ ................................................................. 61
4.3 Plot of a typical estimator for a random design with AR(1) model error, $\rho = 0.4$ and $T = 500$ ................................................................. 62
4.4 Plot of a typical estimator for a random design with AR(1) model error, $\rho = 0.4$ and $T = 1000$ ................................................................. 63
4.5 Trajectory plots for US interests rate from Jan. 1st 1992 to June 11, 2007 65
4.6 Trajectory plots for the US interests rate from August 11 1998 to Dec. 29, 1998 66
4.7 One-step forecasting vs observed value trend plot from 12/02/1991 to 04/17/1992 67
4.8 One-step forecasting error from 12/02/1991 to 4/17/1992 .................. 68
4.9 One-step forecasting vs observed value plot from 12/02/1991 to 4/17/1992 69
4.10 Two-step forecasting vs observed value plot from 12/02/1991 to 04/20/1992 70
5.1 Plot of typical estimated coefficients for a fixed design with sample size 500 . 86
5.2 Plot of typical estimated coefficients for a random design with sample size 500. 87
5.3 Plot of typical estimated coefficients for a fixed design with sample size 1000 88
5.4 Plot of typical estimated coefficients for a random design with sample size 1000. 89
5.5 One-step forecasting plot for 500 $\hat{y}_{496}$’s, sample size = 500
5.6 One-step forecasting plot for 500 $\hat{y}_{996}$’s, sample size = 1000
5.7 Plots for daily environmental data in Hong Kong, Jan. 1, 1994 - Dec. 31, 1995
5.8 Autocorrelation plot and partial autocorrelation plot for varying coefficient model error.
5.9 Estimated coefficient function plot with the 95% HPD credit set
5.10 One-step forecasting for the total number of daily hospital admissions for circulatory and respiratory problems from June 1, 1991 - Dec. 31, 1995
List of Tables

2.1 Pointwise asymptotic bias and variance of local smoothers (Fan, 1992) . . . . 21
3.1 Mean squared Error ($f = 40$) .................................................. 41
3.2 Computation time for 500 simulations ........................................... 42
3.3 Mean squared error ($c = 2$) .................................................. 42
4.1 Mean squared error ($f = 40, \rho = 0.4$) ........................................... 58
4.2 Mean squared error ($c = 2, \rho = 0.4$) ........................................... 59
5.1 Computation time for 500 simulations ........................................... 84
5.2 Mean squared error .............................................................. 85
I would like to express my deepest appreciation and indebtedness to many people who have supported me as I worked to achieve my Ph.D. degree. Without all of their help, I could never have made it to this stage. I owe a great intellectual debt to my Ph.D adviser Dr. Runze Li for his insightful guidance, consistent encouragement, tremendous patience, and summer financial support as we worked through my research. His top-notch and systematic statistical background and sharp research abilities led me through each step and stoked my interests in statistical applications in the real world. I have been impressed by his non-stop hardworking characteristics and have learned quite a few of his methods, such as searching out every avenue to get around apparent dead-ends. His studious attitude, and his ingenious, humble, and optimistic personality will influence me throughout my life. My special thanks also go to my other committee members, Dr. Murali Haran, Dr. Jingzhi Huang, and Dr. Aleksandra B. Slavković for their valuable time spent serving on my committee. It is indeed a privilege to thank Dr. Huang who spent a considerable number of hours discussing and explaining related finance knowledge and providing me with additional financial data. Institutionally, I would thank the Penn State University Department of Statistics which gave me the opportunity to study here and provided a financial stipend for the long-term pursuit of my Ph.D.

On a personal note I would like to thank my MBA adviser Dr. Richard M. Robinson, his long-term trust in me and for valuable corrections on my interest data analysis. Thanks also to my friends Lan Wang and Quiling Wang for continuously encouraging me all the way through. Thanks to Joe Gyekis and Loanne Snavely for helping to polish my writing. Thanks Jean and Jim Brock for the unlimited spiritual support given to me over the past 17 years. I deeply appreciate my parents who raised me and instilled in me the importance of education and provided me many necessities to achieve this goal. Finally, my thanks go to my husband for his endless love, understanding, care, and inspiration.

This research is supported by National Science Foundation grant DMS 0348869.
Introduction

Nonparametric regression models have been widely applied to various scientific research fields. In many situations, the regression function may consist of both low-frequency and high-frequency signals. This makes it difficult for existing nonparametric methods to estimate the nonparametric regression. This motivates us to propose new estimation procedures for nonparametric regression models. Another challenge in the nonparametric regression modeling is to predict the values of independent variables beyond the range of the observed independent variables. For example, when data are collected over time, it is of interest to forecast future values. This dissertation aims to develop new estimation procedures for nonparametric regression models in the presence of both low-frequency and high-frequency signals, and to provide a good forecasting procedure for nonparametric regression, in particular, when data are collected over time. To this end, we propose to estimate the nonparametric regression function via a dynamic linear model approach (West & Harrison, 1997). As a result, we renew a Bayesian framework for nonparametric regression models.
1.1 Nonparametric regression model with independent data

Consider a nonparametric regression model

\[ y = m(u) + \varepsilon, \quad (1.1) \]

where \( m(\cdot) \) is a nonparametric regression function, and \( \varepsilon \) is a random error with mean zero and variance \( \sigma^2 \). Nonparametric regression models are very useful to explore features of the data and provide a powerful diagnostic tool for data analysis. Various estimation procedures have been developed for model (1.1) in the literature (Wahba, 1990, Fan & Gijbels, 1996, Gu, 2002 and Stone, Hansen, Kooperberg, & Troung, 1997).

Unlike linear regression models, \( m(\cdot) \) in model (1.1) does not have a parametric form. Therefore, forecasting, i.e., prediction with values of the independent variable beyond the observed range of \( x \)-variable, is challenging for model (1.1). Most existing estimation procedures were designed for estimation of the regression function rather than for forecasting. Thus, it is desirable to develop an estimation procedure that can be utilized effectively for forecasting.

Based on our limited experience, the commonly used smoothing methods, including local linear regression and spline regression, may perform poorly in the presence of both low and high frequency signals simultaneously occurring in the observed data. For example,

\[ m(u) = 4 \times \cos(2 \times \pi u) + c \times \sin(40 \times \pi u) \]

for \( u \in [0, 1] \). Figure 1.1 depicts the regression function with \( c = 0, 1, 2, 3 \). From our simulation study in Chapter 3, the local linear regression can be dramatically improved.
when $c = 1, 2$, and 3. This motivates us to propose an alternative to the commonly used smoothing methods.

Our goal in Chapter 3 is to develop an estimation procedure for model (1.1) using a dynamic linear model approach. The dynamic linear model (DLM), also known as a state-space model, has been well studied in the literature of Bayesian inference. See, for example, West and Harrison (1997). In Chapter 3, we systematically utilize the DLM approach to develop a new estimation procedure for the nonparametric regression model (1.1). We make the following contributions in Chapter 3.
1. We establish a connection between regularized regression splines and dynamic linear models. This connection allows formulation of the nonparametric regression problem as a dynamic linear model. The connection also provides insights into how to estimate the hyper-parameters in the dynamic linear model. The hyper-parameters control the model complexity, and play the same roles as the regularization parameter in the regularized regression splines.

2. We renew a Bayesian framework for nonparametric regression model. Nonparametric Bayesian methods have been developed for distribution function using the Dirichlet process as the prior (Ferguson, 1973) and for the hazard function in Cox (1972). Catia (2008) introduced a prior on the coefficients in the expansion of the regression function through an orthonormal system. However, as far as we know, there is no Bayesian estimation procedure for mean function in model (1.1), although some nonparametric smoothing methods, such as, smoothing splines (Smith & Robert, 1996), admit Bayesian interpretation.

1.2 Nonparametric regression model with serially correlated errors

Based on results on Chapter 3, we extend the DLM methodology for nonparametric regression with a serially random error. Consider a nonparametric regression model

\[ y_t = m(u_t) + \varepsilon_t \]  

(1.2)

where \( \varepsilon_t \) is a correlated error, \( t = 1, \cdots, T \). Model (1.2) has been studied in the literature. For cases where \( u_t \) has a fixed design, i.e., \( u_t = t/T \) or \( (t - 0.5)/T \), Altman (1990) and Hart (1991) examined how to construct a better estimator for \( m(\cdot) \) by incorporating the
correlated error $\varepsilon_t$’s. Specifically, they assumed that the correlation between $\varepsilon_t$ and $\varepsilon_s$ is of the form $r(|t - s|)$, and discussed how to select a bandwidth adjusting the correlation structure. Opsomer, Wang, and Yang (2001) provided a good review on the topic.

For cases with $u_t$ being a random rather than a fixed design, Xiao, Linton, Carroll, and Mammen (2003) proposed an estimation procedure based on de-correlating the correlated error. Li and Li (2008) further proposed an estimation procedure using profile least squares by observing that model (1.2) can be written as a partially linear model. The work in Chapter 4 was motivated by the desire to analyze US interest rate data. Examination of the annualized continuously-compounded interest rates on newly issued 3-month US Treasury securities (T-Bills) from January 1st 1990 to June 11, 2007 provides interesting results. By “newly issued”, we mean the rates discerned from the prices of T-Bills when they are first auctioned to the public rather than on prices from subsequent sales on the so called “secondary market”. Figure 1.2 depicts the plotting of observed data, which is collected over time. Therefore, it is expected that the error $\varepsilon_t$’s are correlated.

![Rates on 3−month T−Bills, 1/1/1990−6/11/2007](image)

**Figure 1.2.** Rates on 3-month T-Bills from January 1, 1990 to June 11, 2007
We first develop a Bayesian theory and estimation procedure for a general DLM with an autoregressive (AR) error. We further establish a connection between a dynamic linear model with an AR error and a nonparametric regression with an AR error from the point of view of penalized splines. The connections enable us to apply the Bayesian procedure to the DLM for estimating \( m(\cdot) \) in model (1.2). We then conduct extensive Monte Carlo simulations to assess the finite sample performance of the proposed procedure. From our numerical comparisons, we find that the proposed procedure can often outperform the local linear estimator and characterize the situation where this approach is superior.

### 1.3 Varying coefficient model

Since its systematic introduction in Hastie and Tibshirani (1993), the varying coefficient model has become popular in the literature. In Chapter 5, we study the varying coefficient model:

\[
Y = \beta_1(U)X_1 + \cdots + \beta_p(U)X_p + \varepsilon = x^T \beta(U) + \varepsilon \tag{1.3}
\]

for given covariates \((U, X_1, X_2, \cdots, X_p)\). \( \{\beta_1(U), \cdots, \beta_p(U)\} \) are unknown regression coefficient functions. The \( \varepsilon \) is a random error, here we consider both an independent random error and an AR(p) random error

\[
\varepsilon_t = \sum_{j=1}^{d} \rho_j \varepsilon_{t-j} + \gamma_t,
\]

where \( \gamma_t \sim N(0, \sigma^2) \) with finite \( \sigma^2 \).

The class of dynamic linear models can be applied to a variety of models, for example, dynamic linear regression model, ARIMA models and so on. In Chapter 5, we extend its application to varying coefficient model with correlated error AR(p) in the model (1.3). We establish a connection between a dynamic linear model with an AR error and a varying
coefficient model with an AR error.

1.4 The outline of the dissertation

The rest of this dissertation is organized as follows. A literature review is given in chapter 2 which provides a brief review of nonparametric regression models, varying coefficient models, dynamic linear models and Bayesian inference in order to convey useful ideas and methodologies which have been established on these topics. In line with our research objective that is to develop an estimation and forecasting method for nonparametric regression with correlated error through Bayesian inference, chapter 3, provides a first look at the relationship between regularized linear regression splines and dynamic linear models with i.i.d. errors. Chapter 3 also provides a data-driven method to select values of hyper-parameters in Gibbs sampler algorithms. We compare its performance to local linear estimation through intensive simulation examples. Chapter 4 develops the theorem on estimation and forecasting for dynamic linear models with a correlated error, a procedure that, in one form or another, commands a major share of the research. A large number of simulations and a real data example involve either estimation or forecasting assessment. Chapter 5 focuses on varying coefficient models with an auto-correlated error. Lastly Chapter 6 summarizes our research work and the key contributions of the dissertation. Suggestions for future research is also given in Chapter 6.
Chapter 2

Literature Review

The primary goal of this chapter is to present a comprehensive literature review relevant to the dissertation research topics. Since our research objective is to develop an estimation and forecasting method for nonparametric regression with correlated error through Bayesian inference, we narrow the scope of the reviews to materials relevant to the Bayesian approach. The rest of this chapter is organized as follows: we describe Bayesian inference in section 2.1, including selected Markov Chain Monte Carlo algorithms. We introduce the dynamic linear model and its properties in section 2.2. Last, we review related literatures of local modeling estimation with the nonparametric regression and the varying coefficient model in section 2.3 and 2.4, respectively.

2.1 Bayesian Inference and MCMC

The Bayesian approach updates and summarizes current knowledge in terms of probability distributions for estimation and forecasting of current and future values. It permits all uncertainties be measured by probability distributions, both observed values and parameters so as to produce statistical inferences about any quantity of interests. The name “Bayesian”
came from use of Bayes’ theorem that was derived from the work of the Reverend Thomas Bayes in the 18th century. The purpose of this section is to present ingredients required to construct Bayesian analysis. Details are illustrated through several models: the classic linear regression model, the hierarchical model and the dynamic linear model. Our research is relevant to concepts of the dynamic linear model, thus we introduce it in the next section. For more discussion about Bayesian inference, please refer to the books by Berger (1980), Bernardo and Smith (1994) and Gelman, Carlin, Stern, and Rubin (1995). We review Markov Chain Monte Carlo stochastic simulation algorithms at the end of the section.

2.1.1 Bayesian Inference

Assume an uncertain random sample $Y = (y_1, \cdots, y_n)'$ is extracted from a population which is distributed according the density function $f(Y|\theta)$. The observations $y_i$ are independently and identically distributed with parameters $\theta = \{\theta_1, \cdots, \theta_p\}$ which have their own distributions. The Bayesian approach incorporates the information about $\theta$ into model analysis through a prior density $p(\theta)$ even though this information is not observed. Bayesian Inference contains three ingredients:

1. Prior distribution $p(\theta)$. It contains the probability distributions of parameter $\theta$ before the observation of the value $Y$. The prior distributions can be specified with constant parameters that are called the hyper-parameters in that they are the parameters that describe the distributions of the original parameters. In general, the values of hyper-parameters are assumed to be known.

2. Observational distribution $f(Y | \theta)$. The likelihood function $\mathcal{L}(\theta) = f(Y | \theta)$ can be treated as a function of $\theta$ and it provides the chances of each value of $\theta$ having led to that observation value.
3. Posterior distributions $p(\theta \mid Y)$. Once an observation value $y_i$ is observed and becomes part of the available information, it is natural that inferences should be based on the conditional probability distributions of $p(\theta \mid Y)$, which are called the posterior distribution and can be obtained by Bayes’ Theorem:

$$
p(\theta \mid Y) = \frac{f(Y \mid \theta)p(\theta)}{f(Y)}
$$

where $f(Y) = \int f(Y \mid \theta)p(\theta)d(\theta)$. At the left hand side is the conditional posterior probability density function for $\theta$, and the observed random sample $Y$ is simply a constant as well as $f(Y)$. Bayes’ Theorem can be written in a more compact form

$$p(\theta \mid Y) \propto L(\theta)p(\theta).$$

The prior and posterior distributions are called conjugate distributions if both distributions belong to the same family of probability density functions. The concepts of prior and posterior distributions are always relative to the observation considered at a given moment. It is possible that after observing $y_{i-1}$ and obtaining the posterior, we can detect the information of a new observation $y_i$ from the new likelihood function $\theta$ that becomes available. In this case, the posterior, relative to $y_{i-1}$, becomes the prior, relative to $y_i$. After this a new posterior distribution can be obtained by a new application of Bayse’ Theorem, no matter the order of values for $y_{i-1}$ and $y_i$. Once the posterior distribution is available, statistical inferences may be summarized. This concept is discussed further in section 2.2.

**Bayesian inference for linear regression models**

Let us start with a classic linear regression model. Assume that the response variables are normally distributed, and a linear regression model is obtained with observations $Y =$
\((y_1, \cdots, y_n)\)' described by

\[ y_i = x_{i1}\theta_1 + \cdots + x_{ip}\theta_p + \varepsilon_i, \quad \varepsilon \sim N(0, \sigma^2), \quad i = 1, \cdots, n \]  

(2.1)

where \((x_{i1}, \cdots, x_{ip})\) are the values of \(p\) explanatory variables for the \(i^{th}\) observation and \(\theta = \{\theta_1, \cdots, \theta_p\}\) are the regression coefficients associated with these variables. Further restricting to a case where classical linear regression has equal conditional variance, \(\text{var}(y_i \mid \theta, X) = \sigma^2\).

For independent observations the model is written in matrix form as:

\[ Y \sim N(X\theta, \sigma^2 I_n) \]  

(2.2)

with \(X\) as \(n \times p\) design matrix, \(I\) is the \(n \times n\) identity matrix and \(\theta\) is a \(p \times 1\) constant vector.

**Posterior Distribution for Conditional Conjugate**

Continuing the model (2.2), it is completed with a prior distribution for parameters \((\theta, \sigma^2)\). In this case assume a conjugate family of conditional prior distribution for \(\theta\) and \(\sigma^2\) are

\[ \theta \mid \sigma^2 \sim N(\mu, \sigma^2 \Delta), \quad \text{and} \quad \phi = (\sigma^2)^{-1} \sim \text{Gamma}(a, b) \]

The joint prior density is:

\[ p(\theta, \phi) \propto p(\theta \mid \phi)p(\phi) = \phi^{p/2+a-1}\exp\left\{-\frac{\phi}{2} [(\theta - \mu)^T \Delta^{-1} (\theta - \mu) + 2b] \right\} \]  

(2.3)

The likelihood function can be written in terms of \(\theta\) and \(\phi\) as:

\[ L(\theta, \phi) = \phi^{n/2}\exp\left\{-\frac{\phi}{2} (Y - X\theta)^T (Y - X\theta) \right\} \phi^{n/2}\exp\left\{-\frac{\phi}{2} [q(\theta) + \text{SSE}] \right\} \]  

(2.4)

where \(q(\theta) = (\theta - \hat{\theta})^T(X^TX)(\theta - \hat{\theta})\), \(\hat{\theta} = (X^TX)^{-1}X^TY\) and \(\text{SSE} = (Y - X\hat{\theta})^T(Y - X\hat{\theta})\).
Combining 2.3 and 2.4 results in joint posterior density:

\[
p(\theta, \phi \mid Y) \propto L(\theta, \phi)p(\theta, \phi)
\]

\[
\propto \phi^{n/2+p/2+a-1}\exp\left\{-\frac{\phi}{2} \left[ (\theta - \mu)^T \Delta^{-1} (\theta - \mu) + 2b + q(\theta) + \text{SSE} \right] \right\}
\]

\[
\propto \phi^{p/2}\exp\left\{-\frac{\phi}{2} (\theta - \mu_1)^T \Delta_1^{-1} (\theta - \mu_1) \right\} \phi^{a_1-1}\exp\{-\phi b_1\}
\]

where \( \mu_1 = (\Delta^{-1} + X'X)^{-1}(\Delta^{-1} \mu + X'Y) \), \( \Delta_1 = (\Delta^{-1} + X'X)^{-1} \), \( a_1 = n/2 + a \) and \( b_1 = \frac{1}{2} \left[ (Y - X\mu_1)^T Y + (\mu - \mu_1)^T \Delta^{-1} \mu + 2b \right] \). The conditional posterior distributions for \( p(\theta \mid \phi) \) and \( p(\phi \mid \theta) \) are:

\[
\theta \mid \phi \sim N(\mu_1, \sigma^2 \Delta_1) \quad \text{and} \quad \phi \mid \theta \sim \text{Gamma}(a_1, b_1)
\]

**Posterior prediction distribution for \( y \)**

Suppose \( \tilde{X} \) is a new value of an explanatory variable data set and will be used to predict \( \tilde{y} \). The posterior predictive distribution of unobserved data, \( p(\tilde{y} \mid Y) \), can be sampled by two steps. First draw \((\theta, \sigma^2)\) from their joint posterior distribution then draw \( \tilde{y} \) from \( N(\tilde{X}\theta, \sigma^2 I) \). The mean and variance for \( \tilde{y} \) are derived by average over \( \theta \).

\[
E(\tilde{y} \mid \sigma^2, Y) = E \left( E(\tilde{y} \mid \theta, \sigma^2, Y) \mid \sigma^2, Y \right) = E(\tilde{X}\theta \mid \sigma^2, Y) = \tilde{X}\theta
\]

where the inner expectation averages over \( \tilde{y} \), conditional on \( \theta \), and the outer expectation averages over \( \theta \). All expressions are taken with respect to \( \sigma^2 \) and \( Y \). Similarly, \( \text{var}(\tilde{y} \mid \sigma^2, Y) \)
can be derived by:

\[
\text{Var}(\hat{y} \mid \sigma^2, \mathbf{Y}) = E(\text{Var}(\hat{y} \mid \theta, \sigma^2, \mathbf{Y}) \mid \sigma^2, \mathbf{Y}) + \text{Var}(E(\hat{y} \mid \theta, \sigma^2, \mathbf{Y}) \mid \sigma^2, \mathbf{Y})
\]

\[
= E(\sigma^2\mathbf{I} \mid \sigma^2, \mathbf{Y}) + \text{Var}(\mathbf{X}\theta \mid \sigma^2, \mathbf{Y}) = (\mathbf{I} + \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\sigma^2
\]

Bayesian inference for hierarchical models

Regression parameters \(\theta\) and \(\sigma^2\) are used to explore the relationship between response variable \(\mathbf{Y}\) and explanatory variable \(\mathbf{X}\). Performance of \(\theta\) and \(\sigma^2\) can also be studied by their distributions through hyper-parameters, especially when statistical applications involve multiple parameters. Hierarchical models solve such a problem hierarchically, with observable outcomes modeled conditionally on certain parameters, which themselves are given a probabilistic specification in terms of hyper-parameters. Such hierarchical thinking helps in understanding multi-parameter problems and also plays an important role in developing computational strategies.

Continuing from regression model (2.2), besides the prior distributions for \(\theta\) and \(\sigma^2\), probability distributions for hyper-parameters are provided:

1. \(1^{st}\) level
   \[\theta \mid \mu, \tau^2 \sim N(\mu_1, \tau^2 I_d)\]

2. \(2^{nd}\) level
   \[\mu \sim N(\delta, \omega)\]

   \[\sigma^2 \sim \text{Inverse Gamma}(\alpha_1, \beta_1)\]

   \[\tau^2 \sim \text{Inverse Gamma}(\alpha_2, \beta_2)\]

Let \(\phi = (\sigma^2)^{-1}\) and \(\xi = (\tau^2)^{-1}\). By Bayesian inference, the full conditional posterior distributions are:

1. \(\theta \mid \mu, \xi, \phi \sim N\left(\left[\phi(\mathbf{X}'\mathbf{X}) + \xi\mathbf{I}_p\right]^{-1} \left[\phi(\mathbf{X}'\mathbf{Y}) + \xi\mu\mathbf{1}_p\right]; \left[\phi(\mathbf{X}'\mathbf{X}) + \xi\mathbf{I}_p\right]^{-1}\right)\);  

2. \(\mu \mid \theta, \xi \sim N\left((p\xi + \omega^{-2})^{-1}(\omega^{-2}\delta + \theta'\mathbf{1}_p); (p\xi + \omega^{-2})^{-1}\right)\)

3. \(\phi \mid \theta \sim \text{Gamma}\left((n/2 + \alpha_1), \left[\frac{1}{2}(\theta - \tilde{\theta})'(\mathbf{X}'\mathbf{X})(\theta - \tilde{\theta}) + \frac{1}{2}\text{SSE} + 2\beta_1\right]\right)\)
4. $\xi \mid \theta, \mu \sim \text{Gamma} \left( (p/2 + \alpha_2), \left[ \frac{1}{2}(\theta - \mu_1, p)'(\theta - \mu_1, p) + 2\beta_2 \right] \right)$

2.1.2 Markov Chain Monte Carlo

The Markov Chain Monte Carlo (MCMC) are simulation-based methods that enable statisticians to apply Bayesian statistical modeling. The Markov chains are defined in such a way that the posterior distribution in the given statistical inference problem is the asymptotic distribution. In this section we summarize basic facts about Markov Chain and Monte Carlo.

Markov Chain

A Markov chain is a mathematical model for stochastic systems where the description of the present state fully captures all the information that could influence the future evolution of the process through a probabilistic process. A stochastic process is a collection of random vectors $\{X(t), t \in T\}$. For MCMC procedures, only the discrete time homogenous Markov chain is needed. A chain is finite if the state space is finite. Given the present state in a Markov Chain, the future and past state are independent, and the current state only depends on the most recent states:

$$p(X_t|X_{t-1}, \cdots, X_1) = p(X_t|X_{t-1})$$

The next state $\theta^t$ with $t \geq 0$ is distributed according to $p(\theta^{t+1} \mid \theta^t)$ which is called the transition kernel. MCMC designs choose transition densities in order to ensure that this stationary distribution is the true target posterior $p(\theta)$. Therefore, simulations run for some number of iterations until the effect of the chosen starting value is forgotten and the process is approximately stationary. Thereafter, successively sampled values may be assumed to be approximately drawn from the posterior distribution. These values can be appropriately summarized to produce approximate posterior inferences.
Monte Carlo integration

Monte Carlo integration is a technique for estimating an integral numerically. In the 1990’s, Bayesian statisticians developed such simulation algorithms based on iterative MCMC methods to address problems when the posterior distribution is difficult to simulate directly due to the complicated mathematical form of posterior distributions and high dimensionality of the data. Let $\theta$ represent a collection of quantities of interest with a prior density distribution and observed data that lead to a posterior distribution $f(\theta)$. Suppose that a simulation study draws a direct sample of Monte Carlo size $N$, $\{\theta_1, \cdots, \theta_N\}$. Then the approximate posterior inference is often based on simple sample summaries such as sample means to approximate posterior means. This is based simply on the law of large numbers, which states that sample averages converge to corresponding exact posterior expectations:

$$\frac{1}{N} \sum_{i=1}^{N} f(\theta_i) \longrightarrow \int f(\theta)p(\theta)d\theta = E[f(\theta)]$$

where the probability distribution follows $N \longrightarrow \infty$, assuming we discard the first $m$ iterations. The amount of data required for burn-in depends upon the application, and current research in MCMC methods continues to evaluate methods for determining the optimal estimators. Literature on MCMC methods, including design and convergence theory are already huge and growing dramatically (Gilks, Richardson, & Spiegelhalter, 1996). Geyer (1992) suggested the burn-in can be between 1% and 2% of iteration, where iteration is large enough to obtain adequate precision in the estimation.

MCMC Algorithms

Two of the most popular MCMC algorithms are discussed, Metropolis-Hastings and Gibbs sampler.

Metropolis-Hastings algorithm is developed by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) and generalized by Hastings (1970). It obtains state of the chain
at $t + 1$ by sampling a candidate point $X$ from a proposal distribution $q(\cdot \mid \theta^t)$ which should be easy to sample from. This depends only on the previous state $\theta^t$ and can have any form, subject to regularity conditions (Robert & Casella, 1999). For example, $q(\cdot \mid \theta^t)$ is the multivariate normal with mean $\theta^t$ and fixed covariance matrix. The proposal distribution $q(\cdot \mid \theta^t)$ should have the properties of irreducibility and aperiodicity. Irreducibility means that there is a positive probability that the Markov chain can reach any non-empty sets from all starting points. Aperiodicity ensures that the chain will not oscillate between different sets of states. These conditions are satisfied if $q(\cdot \mid \theta^t)$ and the full conditional posterior distribution $p(\theta)$ share the same support. The candidate point, $\tilde{\theta}$ is accepted as the next state of the chain with probability given by a Metropolis criterion:

$$\alpha(\tilde{\theta}, \theta^t) = \min \left\{ 1, \frac{p(\tilde{\theta})q(\theta^t \mid \tilde{\theta})}{p(\theta^t)q(\theta \mid \theta^t)} \right\}$$

If the point $\tilde{\theta}$ is not accepted, then the chain does not move and $\theta^{t+1} = \theta^t$, otherwise $\theta^{t+1} = \tilde{\theta}$.

Metropolis-Hastings algorithms share several properties, but one of the most useful properties is that they can be used in applications where $p(\theta)$ is known up to the constant of proportionality. Another property that makes them useful in a lot of applications is that the analyst does not have to know the conditional distributions, which is the case with the Gibbs sampler.

**Gibbs Sampler** (Smith & Robert, 1993) is one of several Markov Chain Monte Carlo methods which have proven useful in Bayesian analysis. It achieves this simulation directly from the full conditional density distribution of the parameter of interest, given other parameters and observations. The Gibbs sampler was originally developed by Tanner (1987) and Geman and Geman (1984). It was brought into mainstream statistics through the articles of Gelfand and Smith (1990). Casella and George (1992) assumed that the joint
density is given by:

\[ f(x, y_1, y_2, \cdots, y_m) \]

The properties of random variable \( X \) from its marginal density distribution \( f(x) \) can be obtained by integration:

\[ f(x) = \int \cdots \int f(x, y_1, y_2, \cdots, y_m) dy_1 dy_2 \cdots dy_m \]

and the Gibbs sampler is a way to get \( f(x) \) by simulation. As with the other MCMC methods, we use the Gibbs sampler to generate a sample from \( f(x) \) and then use the sample to estimate the desired characteristic of \( f(x) \). Casella and George (1992) noted that if the number of iterations was large enough, then any population characteristic can be calculated with the required degree of accuracy.

Gibbs sampler is special case of Metropolis-Hastings algorithms. Two main differences between the Gibbs sampler and Metropolis-Hastings algorithm are that in the Gibbs sampler a candidate point is always accepted and the proposal distribution which is full posterior conditional distribution is known. These properties make the Gibbs sampler less applicable. More properties can be found in Bernardo and Smith (1994) and Gelman et al. (1995) which illustrate many real world applications of Gibbs sampler estimation.

### 2.2 Dynamic Linear Model

Classical linear regression model (2.1) fails to capture dynamic changes when data involve multiple overlapping patterns of periodicity. West and Harrison (1997) extended the model (2.1) in two directions. One models the random error \( \varepsilon \) while the other contains sequences of \( \theta_t \) as “dynamic models”. West and Harrison (1997) modeled the dynamic nature of \( \theta_t \) as
a random walk which is

$$\theta_t = \theta_{t-1} + \omega_t \quad \text{for } \omega_t \sim N(0, W_t)$$

It allowed the routine characterization of $\theta$ to be a slowly evolving parameter. Combining
the above equation and model (2.1) results in a dynamic linear regression model defined by
West and Harrison (1997), at each period $t$,

**Observational equation**

$$y_t = X_t \theta_t + \nu_t, \quad \nu_t \sim N(0, \nu_t)$$

**Evolution equation**

$$\theta_t = \theta_{t-1} + \omega_t, \quad \omega_t \sim N(0, W_t)$$

The dynamic linear regression model assumes that the linear regression form is only locally
appropriate in time, with the regression parameter vector $\theta_t, (p \times 1), t = 1, \cdots, T,$ varying
according to a random walk. The first equation is an observational equation with a known
regression ($p \times 1$) vector $X_t$ and $\nu_t$ as an observational error which describes local random
fluctuations. Second is the evaluation equation, also known as the system equation. The
evolution of error term $\omega_t$ describes the changes in the elements of the parameter vector
between times $t - 1$ and $t$. The zero mean vector reflects that $\theta_t$ is expected to be constant
over a short interval, while evolution variance matrix $W_t$ ($p \times p$) governs the extent of the
movement in $\theta_t$. $\nu_t$ and $W_t$ are two independent sequences and assumed to be mutually
independent. Dynamic linear models are only useful for short-term applications and particu-
larly in cases when the observation variation as measured by $\nu_t$ is greater than the systematic
level of variation measured by $W_t$. A large class of models with these time-varying param-
eters, adequate for the modeling of time series and regression, was presented by Harrison
and Stevens (1976). More properties of dynamic linear models, including some simulation
algorithms are discussed in Appendix A.
2.3 Univariate nonparametric regression models

In this section, a brief summary of local estimation and Bayeisan approaches for univariate nonparametric regression is described. Given observations of data \((x_1, y_1), \cdots, (x_T, y_T)\) and let \((x, y)\) denote a generic member of the sample. Consider the nonparametric regression model:

\[ y = m(x) + \varepsilon, \]

with conditional mean and conditional variance:

\[ E(\varepsilon | X = x) = 0 \quad \text{and} \quad Var(\varepsilon | X = x) = \sigma^2(x), \]

where \(m(x) = E(Y | X = x)\) is a mean function and assumed to be smooth without a parametric form. Estimation of mean function \(m(x)\) and its \(\nu\)th-derivative \(m^{(\nu)}(x)\) are our main interests. The estimation performance of \(\hat{m}_\nu(x)\) for \(m^{(\nu)}(x)\) is accessed via mean squared error (MSE) which is defined as:

\[ MSE(x) = E\left[ (\hat{m}_\nu(x) - m^{(\nu)}(x))^2 | X \right], \]

where \(X = (X_1, \ldots, X_T)\). The main object is to estimate the function \(m^{(\nu)}(.)\) at the point \(x\). We can easily to rewrite MSE as a sum of bias and variance:

\[ MSE(x) = E \left[ (\hat{m}_\nu(x)|X) - m^{(\nu)}(x) \right]^2 + Var\{\hat{m}_\nu(x)|X\}, \]

The first term is referred to the bias and the second is the variance.

With relaxed model assumptions of the model, an intuitive estimator for the mean function \(m(x)\) is the local averaging estimator:

\[ \hat{m}(x) = \sum_{t=1}^{T} \omega_t(x)Y_t, \]
where \( \omega(\cdot) \) is a weighted function depending upon \( x \) and the estimated mean function \( \hat{m}(x) \) is a weighted sum of \( Y_t \). It is natural to assign higher weights to the observation close to \( m(x) \) and lower weights to those that are farther away. The local averaging estimator, first suggested by Nadaraya (1964) and Watson (1964), is called the Nadaraya-Watson kernel regression estimator. It is expressed as:

\[
\hat{m}_h(x) = \frac{\sum_{t=1}^{T} K_h(X_t-x)Y_t}{\sum_{t=1}^{T} K_h(X_t-x)}
\]

Gasser and Müller (1979) proposed another local averaging estimator:

\[
\hat{m}_h(x) = \sum_{t=1}^{T} \int_{s_{t-1}}^{s_t} K_h(u-x) \, du \, Y_t
\]

with \( s_t = \frac{x_t + x_{t+1}}{2}, \, x_0 = -\infty, \, x_{T+1} = +\infty \), where the function \( K_h(.) = K(.) / h \) is called a kernel function which is equivalent to the weighted function \( \omega(.) \). \( K(.) \) is a unimodal function centered at zero (symmetric) and declines in either direction at a rate controlled by bandwidth, \( h \), also called the smoothing parameter. Commonly used kernel functions include the Gaussian kernel \( K(t) = (2\pi)^{-1/2}\exp(-t^2/2) \) and the symmetric Beta family \( K(t) = \text{Beta}(\frac{1}{2}, \nu + 1)^{-1}(1 - t^2)_+^{\gamma}, \gamma = 0, 1, 2, 3, \ldots \). The choices of \( \gamma = 0, 1, 2, 3 \) result in uniform, Epanechnikov, biweight and triweight kernel functions, respectively. The advantage of a local averaging estimator is that as long as the weights are known or can be easily calculated, estimation function \( \hat{m}(x) \) is also easy to calculate. A disadvantage is that both estimators use local constant approximations and thus suffer from large bias at the boundary. The Nadaraya-Watson estimator has zero minimax efficacy. On the opposite local liner fit is asymptotic optimal in an asymptotic minimax sense and automatically corrects boundary bias.
2.3.1 Local linear estimation

Local linear fit is a natural extension of local averaging. Suppose we run a liner regression using only observations that lie in a neighborhood of $x$ such that $X_t \in x \pm h, h$ is the smooth parameter which determines the size of the local neighborhood. If we assume that observations were given equal weight, we can estimate mean function $\hat{m}(x)$ at point $x$ by minimizing:

$$
\min_{a,b} \sum_{X_t \in x \pm h} [y_t - a(x) - b(x)X_t]^2
$$

Then the estimator of $m(x)$ at $x$ would be given by $\hat{m}(x) = \hat{a}(x) + \hat{b}(x)x$. One can also incorporate a weighted function:

$$
\min_{a,b} \sum_{t=1}^{T} [y_t - a(x) - b(x)X_t]^2 K(\frac{X_t - x}{h})
$$

The estimated mean function $m(\cdot)$ at $x$ is the same as before.

Fan (1992) proved that the Epanechnikov kernel function, $K = \frac{3}{4}(1 - u^2), u \in [-1, 1]$ achieved the optimal efficiency in the class of all regression functions whose second derivative was bounded by a constant in the neighborhood of $x$. We use Epanechnikov kernel function in local linear fitting in the rest of this research.

The asymptotic biases and variances for Nadaraya-Watson estimator, Gasser-Müller estimator and local linear estimator are summarized in Table 2.1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Bias</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nadaraya-Watson</td>
<td>$(m''(x) + \frac{2m(x)f'(x)}{f(x)})b_n$</td>
<td>$V_n$</td>
</tr>
<tr>
<td>Gasser-Muller</td>
<td>$m''(x)b_n$</td>
<td>$1.5V_n$</td>
</tr>
<tr>
<td>Local linear</td>
<td>$m''(x)b_n$</td>
<td>$V_n$</td>
</tr>
</tbody>
</table>

Table 2.1. Pointwise asymptotic bias and variance of local smoothers (Fan, 1992)
The local liner approach ends up with the lowest bias and variance value and bring us the most accurate and precise estimator. Cheng, Fan, and Marron (1997) proved that the local linear fit was more efficient in correcting the boundary bias than the other two local averaging estimators.

Local polynomial fit was well studied. See, for example W. Cleveland (1979) and Fan and Gijbels (1996). Here we briefly introduce the basic ingredient of local polynomial fitting.

Assume that the mean function $m(z)$ can be locally approximate by using Taylor’s expansion in the neighborhood of $x$, thus

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

and $m(x)$ can be estimated by a simple polynomial model through minimizing the locally weighted squared loss function:

$$\sum_{t=1}^{T} \left\{ Y_t - \sum_{j=0}^{p} \beta_j (X_t - x)^j \right\}^2 K_h(X_t - x),$$

with $\hat{m}^{(\nu)}(x) = \nu! \hat{\beta}_\nu$. When $p = 1$, the local polynomial fit is the local linear regression or local linear regression smoother estimator. The estimator $\hat{m}(x)$ can be explicitly expressed as

$$\hat{m}(x) = \frac{\sum_{t=1}^{T} \omega_t Y_t}{\sum_{t=1}^{T} \omega_t}$$

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

It is more straightforward to work with matrix notation for equation (2.6). Let $X$ be a
$T*(p+1)$ design matrix and $y$ be the observed vector, thus they can be written as:

$$X = \begin{pmatrix}
1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\
\vdots & \vdots & \ddots & \vdots \\
1 & (X_T - x_0) & \cdots & (X_T - x_0)^p
\end{pmatrix},
y = \begin{pmatrix}
Y_1 \\
\vdots \\
Y_T
\end{pmatrix}, 
\hat{\beta} = \begin{pmatrix}
\hat{\beta}_0 \\
\vdots
\end{pmatrix}$$

Let $W$ be $T*T$ diagonal matrix of weights such as $W = \text{diag}\{K_h(X_t - x)\}$, then minimizing (2.6) can be written as

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

The weighted least squared estimator for $\beta$ is: $\hat{\beta} = (X^TWX)^{-1}X^TY$.

Local linear estimation provides an excellent smoothing method with high statistical efficiency in an asymptotic minimax (Fan & Gijbels, 1996) and design adaptation (Fan, 1993).

### 2.3.2 Bayesian nonparametric regression models

Ferguson (1973) introduced Bayesian nonparametric regression model analysis through the Dirichlet process in 70's. Recently the analysis involves Markov Chain Monte Carlo methods. There are two main areas for Bayesian analysis on the model. First, researcher assigned prior distributions to the univariate mean function 1.2 and assumed prior distributions follow stochastic processes (Ferguson, 1973). Second, researchers developed Bayesian framework on regression functions through a series parametric distributions. I review both analysis in this section.

Müller and Quintana (2004) reviewed current nonparametric Bayesian inference including density estimation. However Müller and Quintana (2004) did not cover nonparametric
Bayesian inference for time series data. Assuming that a sample with i.i.d. random variable $x$ follows a unknown distribution $F$ which is generated by a Dirichlet process with ay partition, $\{A_1, \cdots, A_k\}$ such as

$$\left( F(A_1), \cdots, F(A_k) \right) \sim D(M \cdot F_0(A_1), \cdots, M \cdot F_0(A_k))$$

where $F \sim D(M, F_0)$. $M$ is the weight parameter that defines variance and $F_0$ is the base measure and defined as the expectation $E(B) = F_0(B)$. We can express $x_i \sim D(M, F_0), i = 1, \cdots, n$. Then the posterior distribution is $F \mid x_1, \cdots, x_n \sim D(M + n, F_1)$ with $F_1 \propto F_0 + \sum_{i=1}^{n} \delta_{x_i}$ and $\delta_{x_i}$ represents a point mass at $x$. The asymptotic discreteness is not appropriate in many case thus Escobar and West (1996) and MacRachern (1994) extend a model DP mixtures (MDP). This extension remove the constraint to discrete measures and add an additional convolution:

$$F(x) = \int f(x \mid \theta) dG(\theta), \quad \text{with } G \sim D(M, G_0)$$

For nonparametric model 1.2, I review smoothing spline method. Others such as wavelet based method can be found in Chipman, Kolaczyk, and McCulloch (1997) and B. (1998). Denison, Mallick, and Smith (1998b) and DiMatteo, Genovese, and Kass (2001) considered some basis $B = \{f_1, f_2, \cdots\}$ and parameterized mean function by a sequence basis coefficients $b = (b_1, b_2, \cdots)$. Thus the model 1.2 can be expressed as

$$y_t = \sum b_h f_h(u_t) = \varepsilon_t$$

The model is then completed with a joint prior distributions $p(\xi, b, \sigma^2)$ on the set of knots $\xi = (\xi_1, \cdots, \xi_k)$ and coefficients $(b_1, b_2, \cdots)$. Consider cubic spline $B = (1, x, x^2, x^3, (x -$
\( (\xi_1)^3, \cdots (x - \xi_k)^3 \) with \((x)_+ = \text{max}(x, 0)\). Prior distribution is assumed to be independent,

\[
p(\xi, b, \sigma^2) = p(\xi)p(\sigma^2)p(b|\sigma^2)
\]

Assume that conditional prior distribution \( b|\sigma^2 \sim N(0, c\sigma^2(B^TB)^{-1}) \) where \( B \) is design matrix. DiMatteo et al. (2001) use Zellen g-prior (Zellner, 1986) for scale \( c \). The conditional posterior mean \( E(b|\xi, \sigma^2) \) is a linear shrinkage of the least squared estimate \( \hat{b} \).

### 2.4 Varying coefficient model

The varying coefficient model is a useful extension of classical linear model that allows the regression coefficients of the independent variables to vary over some explanatory variables. The model was first explored by W. S. Cleveland, Grosse, and Shyu (1991) and by Hastie and Tibshirani (1993). The varying coefficient model follows a linear structure:

\[
Y = \theta_1(U_1)X_1 + \theta_2(U_2)X_2 + \cdots + \theta_p(U_p)X_p + \varepsilon \tag{2.8}
\]

where covariates \( \{U_1, \ldots, U_p\} \) change the explanatory variable vector \( \{X_1, \cdots, X_p\}^T \) via the unspecified functions \( \{\theta_1(.), \ldots, \theta_p(.)\} \). To include a varying intercept function, we may set \( X_1 \equiv 1 \). Unlike traditional regression models, varying coefficient models relax model assumptions and increase model interpretability and applicability. The advantage of the model is that it allows the regression coefficients \( \theta_j(U_j) \) to vary according to \( U_j \) so as to significantly reduce the modeling bias and avoid the “curse of dimensionality”.

Varying coefficient models arise in various forms based upon different data procedures, for example, functional data analysis (Ramsay & Silverman, 1997) and longitudinal data analysis (Hoover, Rice, Wu, & Yang, 1998, and Fan & Zhang, 2000). When \( \theta_j(U_j) = \theta_j \), the
model becomes the ordinary linear regression model such as

\[ Y = \theta_1 X_1 + \cdots + \theta_p X_p + \varepsilon. \]

When \( U_j = U \), such as point in time, the model can be treated as the conditionally parametric model. For a given \( U = u \):

\[ Y = \theta_1(U)X_1 + \cdots + \theta_p(U)X_p + \varepsilon = \sum_{j=1}^{p} \theta_j(U)X_j + \varepsilon \tag{2.9} \]

with

\[ E(\varepsilon|U, X_1, \cdots, X_p) = 0, \]

and

\[ \text{Var}(\varepsilon|U, X_1, \cdots, X_p) = \sigma^2(U). \]

Next let us briefly describe estimation methods developed in the recent literature for both independent and correlated data.

**Independent Data**

Let \( \{(U_t, X_{t1}, \cdots, X_{tp})\} \) be a random sample from model (2.9) with \( \varepsilon_t \) is i.i.d. error and \( X_{t0} \equiv 1 \). Many estimation methods, including kernel estimation and local spline smoothing, are used to estimate the varying coefficients \( \theta_j(U)(j = 1, \cdots, p) \) in the model. Here we concentrate on local linear regression methods.

W. S. Cleveland et al. (1991) introduced an estimation method through local linear regression. For each given point \( u_0 \), approximate the function locally as

\[ \theta_j(u) \approx \theta_j(u_0) + \theta_j'(u_0)(u - u_0) \equiv a_j + b_j(u - u_0), \quad j = 1, \ldots, p \]
for \( u \) in a neighborhood of \( u_0 \). The estimator is achieved by a function to minimize local least-squares:

\[
\sum_{t=1}^{T} \left[ Y_t - \sum_{j=1}^{p} \{ a_j + b_j(U_t - u_0) \} X_{tj} \right]^2 K_h(U_t - u_0)
\]

(2.10)

for a given kernel function \( K \) and bandwidth \( h \); \( K_h(\cdot) = K(\cdot/h)/h \). W. S. Cleveland et al. (1991) assumed that the smoothing function \( \theta_j(\cdot) \) had the same degree of smoothness across the same intervals. The resulting estimator is called a one-step estimator. Fan and W. Zhang (1999) showed that the bias of this one-step estimator was \( O(h_1^2) \) and the variance was \( O((nh_1)^{-1}) \); the optimal rate \( O(n^{-8/9}) \) cannot be achieved. In order to achieve the optimal rate \( O(n^{-8/9}) \) and fit the different degrees of smoothness, Fan and W. Zhang (1999) proposed the following two-step estimation procedure.

**Step 1:** Use local linear regression (2.10) to obtain a preliminary estimator \( \{ \hat{\theta}_{1,0}(\cdot), \hat{\theta}_{2,0}(\cdot) \ldots \hat{\theta}_{p,0}(\cdot) \} \) for \( \{ \theta_1(\cdot), \theta_2(\cdot) \ldots \theta_p(\cdot) \} \) and substitute bandwidth \( h \) to \( h_0 \) as a given initial bandwidth. These initial estimators are usually under-smoothed so the bias is relatively small and variance is large.

**Step 2:** Fan and W. Zhang (1999) substituted \( \{ \hat{\theta}_{1,0}(\cdot), \hat{\theta}_{2,0}(\cdot) \ldots \hat{\theta}_{p-1,0}(\cdot) \} \) and used a local cubic fit to estimate \( \theta_p(u_0) \).

A simple description of local cubic fit assumes that \( \theta_p(\cdot) \) is a smoother function than the other \( \theta_j(\cdot), (j = 1, \ldots, p - 1) \) functions. It assumes that \( \theta_p(\cdot) \) possesses a bounded fourth derivative. In this case, the function can be approximated by a cubic function:

\[
\theta_p(u) \approx a_p + b_p(u - u_0) + c_p(u - u_0)^2 + d_p(u - u_0)^3
\]
Thus the weighted least square estimation follows as

$$\sum_{t=1}^{T} \left[ Y_t - \sum_{j=1}^{p-1} \{a_j + b_j(U_t - u_0)\} X_{tj} \right]$$

$$- \{a_p + b_p(U_t - u_0) - c_p(U_t - u_0)^2 - d_p(U_t - u_0)^3\} X_{tp} \right]^2 \times K_{h_1}(U_t - u_0)$$

(2.11)

After finding \(\{\hat{\theta}_{1,0}(\cdot), \hat{\theta}_{2,0}(\cdot) \ldots \hat{\theta}_{p-1,0}(\cdot)\}\), we can plug them into equation (2.11) to find the estimator \(\theta_p(u_0)\):

$$\sum_{t=1}^{T} \left[ Y_t - \sum_{j=1}^{p-1} \hat{\theta}_{j,0}(U_t) X_{tj} \right]$$

$$- \{a_p + b_p(U_t - u_0) - c_p(U_t - u_0)^2 - d_p(U_t - u_0)^3\} X_{tp} \right]^2 \times K_{h_2}(U_t - u_0)$$

with respect to \(a_p, b_p, c_p, d_p\). Note that \(h_2\) is a different bandwidth in the second step.

Fan and W.Zhang (1999) proved that the bias and variance of the two-step estimator was \(O(h_2^4)\) and \(O\{(nh_2)^{-1}\}\), provided that \(h_0 = o(h_2^2), nh_0/\log h_0 \to \infty\) and \(nh_0^3 \to \infty\). The two-step estimator was not dependent on the first step bandwidth \(h_1\) and the optimal convergence rate \(O(n^{-\frac{8}{9}})\) was fulfilled.

**Correlated Data**

Longitudinal data are repeated measurements collected on a subject through time (Diggle, Heagerty, Liang, & Zeger, 2002). The interest is to study the associations between the covariates and the response variables. The subjects are assumed to be independent of each other and the measurements within each subject are assumed to be correlated. Ignoring these properties could lead to two mistakes. First, if we ignore the correlations within each
subject, variances of time-independent predictors such as gender would be underestimated
and the Type I error rate is inflated for these variables. Second, for time-dependent pre-
dictors such as how the measurements change over time, ignoring their correlations would
inflate Type II error rate (Dunlop, 1994). Numerous data sets for application of these mod-
els are available in biomedical, epidemiological, economic, and many other fields, see Jones
(1993) for more examples.

Let \( \{t_{ij}, j = 1, \cdots, T_i\} \) be the times over which the measurements of the \( i^{th} \) subject
took place. Let \( Y_{ij} \) be the observed response and \( X_{ij} = \{X_{ij1}, \cdots, X_{ijp}\} \) be the \( p \) ob-
served explanatory variables for the \( i^{th} \) subject at time \( j \). This results in data patterned as
\( \{t_{ij}, X_{ij}, y_{ij}\}, j = 1, 2, \cdots, T_i, i = 1, 2, \cdots, n \) and the corresponding time varying coefficient
model is

\[
y_i(t_{ij}) = X_i(t_{ij})^T \theta(t_{ij}) + \varepsilon_i(t_{ij}),
\]

the \( \varepsilon_i \) is a zero-mean stochastic process that cannot be explained by the explanatory variables.

The basic question addressed by longitudinal studies is how the responses vary through time,
and how the patterns of change or stability in the response variables relate to the explanatory
variables. Hoover et al. (1998) and Brumback B (2000) proposed using varying coefficient
model to fit the data. Estimation of varying coefficient functions have been widely studied,
for example, Brumback B (2000) and Hoover et al. (1998) proposed to use kernel, polynomial
and smoothing spline methods. Fan and Zhang (2000) proposed a two step procedure to
estimate the regression coefficient indicating the local degree of correlation between two data
sets. Simply speaking, the first step was to fit a standard linear model to obtain the raw
estimated coefficient function and second step smoothed these raw estimators.

**First step:** Since there are a number of observations collected at time \( t_j, j = 1, \cdots, T \), all
these data can be used to fit a linear model

\[ Y(u) = X(u)^T \theta(u) + \varepsilon(u) \]

to obtain the raw estimates \( \hat{\theta}(t_j) = (\hat{\theta}_1(t_j), \ldots, \hat{\theta}_p(t_j))^T \) for \( \theta(t_j) = (\theta_1(t_j), \ldots, \theta_p(t_j))^T \) at a time point \( t_j \).

**Second step:** Since the raw estimators \( \hat{\theta}(t_j) \) are usually not smooth, the second step smoothes them to get the smooth estimators for the varying coefficient functions. For each given component \( \theta_r(t_j) \), apply a smoothing technique to the data \( \{(t_j, \hat{\theta}_r(t_j)), j = 1, 2, \ldots, T\} \). Thus the smoothing step simplifies to one dimensional estimation.

As Fan and Zhang (2000) reviewed, their method had multiple advantages. A visualization of the raw estimates can assist in choosing a sensible amount of smoothing. The flexible structure allowed different degrees of smoothing to separate components of the coefficients functions, and it was possible to use any of the existing smoothing methods such as spline, kernel, and local polynomial fits. And finally, the two step estimations were easy to implement, program, calculate, and interpret.

This literature review builds the foundation for our research. From the review, it is easy to see that most existing estimation procedures were designed for estimating the mean functions for nonparametric regression models and varying coefficient functions. Based on our knowledge, these estimation procedures were not designed for forecasting. This inspires us to develop an estimation method which also can be applied for prediction of both nonparametric regression and varying coefficient models. As long as data involve time periods, it is hard to avoid correlation among the observations. Therefore we also explore estimation and forecasting performance for models with correlated error terms. In chapter 3, we propose an estimation method through connecting DLM and nonparametric regression models with
independently and identically distributed error.
Chapter 3

A Dynamic Linear Model Approach for Nonparametric Regression Model

3.1 Introduction

Consider a nonparametric regression model

\[ y = m(u) + \varepsilon \]  \hspace{1cm} (3.1)

where \( m(\cdot) \) is a nonparametric smoothing function, and \( \varepsilon \) is an i.i.d. random error with mean zero and variance \( \sigma^2 \). The goal of this chapter is to develop an empirical Bayesian estimation procedure for model (3.1) using a dynamic linear model approach and show how the hyper-parameters can be estimated in a way that is both computationally feasible and statistically valid. This chapter is organized as follows. In Section 3.2, we establish a connection between regularized regression splines and dynamic linear models. In Section 3.3, we develop an estimation procedure for model (3.1) using a Gibbs sampler. Section 3.4 provides numerical comparisons between the local linear regression method and the proposed estimation procedures. Section 3.5 summarizes this chapter and gives discussion for the next
research topic. For self-consistency, some technical materials related to the dynamic linear model are given in the Appendix.

### 3.2 Regularized linear regression splines and Dynamic Linear Models

To get insights into the relationship between regularized linear regression splines and dynamic linear models, let us begin with the canonical nonparametric regression:

\[ y_t = m(u_t) + \varepsilon_t, \]

where \( u_1 < u_2 < \cdots < u_T \) are fixed design points, and \( \varepsilon_t, t = 1, \cdots, T \) are independently and identically distributed random errors with mean zero and variance \( \sigma^2 \). Denote \( m_t = m(u_t) \), and parameterize \( m(u) \) by

\[
m_k(u) = m_{k-1} + \frac{m_k - m_{k-1}}{u_k - u_{k-1}}(u - u_{k-1}), \quad \text{for } u \in (u_{k-1}, u_k]. \tag{3.2}
\]

Then the least squares method is to minimize

\[
\sum_{t=1}^{T} (y_t - m(u_t))^2 \tag{3.3}
\]

with respect to \( m_1, \cdots, m_T \). Note that the least squares function in (3.3) equals

\[
\sum_{t=1}^{T} (y_t - m_t)^2 \tag{3.4}
\]

which implies that the parametrization equation (3.2) is over-parameterized. Therefore, regularization on equation (3.3) is necessary. Typically, it is assumed that \( m(u) \) is continuous
so the difference between \(m_{t-1}\) and \(m_t\) should be small. Thus, we consider a penalized least squares from:

\[
\sum_{t=1}^{T} (y_t - m_t)^2 + \lambda \sum_{t=2}^{T} (m_t - m_{t-1})^2,
\]

(3.5)

where \(\lambda\) is a regularization parameter or a smoothing parameter. We next show that minimization of equation (3.5) with respect to \(m_1, \ldots, m_T\) is equivalent to maximizing the log-likelihood function of a simple dynamic linear model defined below:

\[
\begin{align*}
  y_t &= m_t + \varepsilon_t \\
  m_t &= m_{t-1} + \eta_t,
\end{align*}
\]

(3.6)

where \(\varepsilon_t, t = 1, \ldots, T\) are independently and identically distributed according to \(\mathcal{N}(0, \sigma^2)\), and \(\eta_t, t = 1, \ldots, T\) are independently and identically distributed according to \(\mathcal{N}(0, \sigma^2_{\eta})\).

The logarithm of the likelihood function for model (3.6) is

\[
\mathcal{L}(\mathbf{m}) = -\frac{T}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^{T} (y_t - m_t)^2 - \frac{T - 1}{2} \log \sigma_{\eta}^2 - \frac{1}{2\sigma_{\eta}^2} \sum_{t=2}^{T} (m_t - m_{t-1})^2
\]

(3.7)

after dropping a constant, where \(\mathbf{m} = (m_1, \ldots, m_T)^T\). Maximizing \(\mathcal{L}(\mathbf{m})\) with respect to \(m_1, \ldots, m_T\) is equivalent to minimizing:

\[
\sum_{t=1}^{T} (y_t - m_t)^2 - \frac{\sigma^2}{\sigma_{\eta}^2} \sum_{t=2}^{T} (m_t - m_{t-1})^2
\]

(3.8)

Denote \(\lambda = \sigma^2/\sigma_{\eta}^2\), then equation (3.8) is exactly the same as equation (3.5). This allows us to establish a Bayesian framework for a nonparametric regression model.

The penalty term in equation (3.5) is designed to penalize large differences between \(m_{t-1}\) and \(m_t\). One may consider to penalizing roughness of the \(k\)-derivative of \(m(\cdot)\). Denote \(m_t^{(0)} = (m_t - m_{t-1})\), and for \(k \geq 1\), define \(m_t^{(k)} = (m_t^{(k-1)} - m_{t-1}^{(k-1)})/(u_t - u_{t-1})\). Thus, it is
natural to consider the following penalized least squares:

$$\sum_{t=1}^{T} (y_t - m_t)^2 + \lambda \sum_{t=k+2}^{T} (m_t^{(k)} - m_{t-1}^{(k)})^2. \tag{3.9}$$

Similar to the arguments in equation (3.7) and equation (3.8), it can be shown by some calculations that minimization of equation (3.9) with respect to $m_1, \ldots, m_T$ is equivalent to maximizing the log-likelihood function of a dynamic linear model defined below:

$$y_t = m_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2),$$

$$m_t = \sum_{j=1}^{k+1} \alpha_j m_{t-j} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma^2),$$

where $\alpha_j$'s are regression coefficients, $\varepsilon_t$, $t = 1, \ldots, T$ are independently and identically distributed according to $\mathcal{N}(0, \sigma^2)$, and $\eta_t$, $t = 1, \ldots, T$ are independently and identically distributed according to $\mathcal{N}(0, \sigma^2)$. The second equation in (3.10) offers us a forecasting for $u \geq u_T$. The choice of $k$ depends on prior knowledge about the regression function. For example, if the signal likely will be a smoothing one, such as the one displayed in Figure 1.1 (a), then it is recommended setting $k = 2$, while if regression function likely contains a high-frequency signal such as the ones displayed in Figure 1.1(b), (c) and (d), then we may take $k = 0$.

The ideas discussed above can be naturally applied to a random design setting. If $(u_t, y_t), t = 1, \cdots, T$, is a random sample from model (3.1) and $u_t$ is independent of $\varepsilon_t$, then we sort the data according to $u_t$, and obtain the sorted data $(u_t(y_t), y_t)$ with $u_t(y_t) < \cdots < u_T(y_T)$. Thus,

$$y_t = m(u_t) + \varepsilon_t.$$ 

Since $u_t$ is independent of $\varepsilon_t$, $\varepsilon_t$'s are independently and identically distributed with mean 0 and variance $\sigma^2$. Therefore, the arguments for equation (3.10) are still valid, and the proposed
procedures in the next section are directly applicable for the nonparametric regression with random design.

3.3 The Gibbs sampler algorithm

In this section, we propose an empirical Bayesian estimation procedure for the dynamic linear model (3.10). Before we proceed further, let us give a brief review of the existing Bayesian estimation procedure for a general dynamic linear model.

3.3.1 Fully Bayesian estimation procedure

Define \( F = (1, 0, \cdots, 0)^T \), a \((k + 1) \times 1\)-dimensional vector, \( G = (g_{ij}) \) is a \((k + 1) \times (k + 1)\) matrix with \( g_{1j} = \alpha_j, j = 1, \cdots, k + 1, g_{j,j-1} = 1 \) for \( j = 2, \cdots, k + 1 \) and \( g_{ij} = 0 \) if \( i > 1 \) and \( i \neq j + 1 \). Furthermore, denote \( \theta_t = (m_t, m_{t-1}, \cdots, m_{t-k})^T \).

Represent equation (3.10) as a constant dynamic linear model:

Observation equation:

\[ y_t = F^T \theta_t + \nu_t, \]

where \( \nu_t \sim N(0, V) \).

System equation:

\[ \theta_t = G\theta_{t-1} + \omega_t, \quad (3.11) \]

where \( \omega_t \sim N(0, W) \) with \( W = \text{diag}\{\sigma_\eta^2, 0, \cdots, 0\} \).

Initial information:

\[ (\theta_0|D_0) \sim N(\mu_0, C_0), \]

where the error sequences \( \nu_t \) and \( \omega_t \) are internally independent, mutually independent,
and independent of \((m_0|D_0)\), with \(D_0\) as the \(\sigma\)-field generated by information at time 0.

Thus, model (3.11) is a special case of the general dynamic linear model with a quadruple \(\{F_t, G_t, V_t, W_t\} = \{F, G, V, W\}\). See the definition of the general dynamic linear model in the Appendix.

Let \(D_t = \{Y_t, D_{t-1}\}\). By directly applying Theorem 4.1 for the above dynamic linear model, we may obtain the one-step forecasting. All posterior distributions for any time \(t > 0\) can be obtained sequentially as follows.

**Step 1.** For some mean \(\mu_{t-1}\) and variance \(C_{t-1}\), the posterior for \(\theta_{t-1}: (\theta_{t-1}|D_{t-1}) \sim \mathcal{N}(\mu_{t-1}, C_{t-1})\).

**Step 2.** Prior for \(\theta_t: (\theta_t|D_{t-1}) \sim \mathcal{N}(a_t, R_t)\), where \(a_t = G\mu_{t-1}\), and \(R_t = GC_{t-1}G^T + W\).

**Step 3.** One-step forecasting: \((Y_t|D_{t-1}) \sim \mathcal{N}(f_t, Q_t)\), where \(f_t = F^Ta_t = F^TG\mu_{t-1}\), and \(Q_t = F^TR_tF + V\).

**Step 4.** Posterior distributions for \(\theta_t: (\theta_t|D_t) \sim \mathcal{N}(\mu_t, C_t)\), where \(\mu_t = a_t + A_t e_t\), \(C_t = R_t - A_tQ_tA_t^T\), \(A_t = R_tF/Q_t\) and \(e_t = y_t - f_t\).

To make a statistical inference using the posterior distributions, one needs to calculate mean \(\mu_t\) and variance \(C_t\) iteratively. In particular, this can be done by using a forward filtering and backward sampling algorithm given in the Appendix A.

### 3.3.2 Empirical Bayesian estimation procedure

In the empirical Bayesian estimation procedure, we use data to estimate nuisance parameters \(\sigma^2\) and \(\sigma^2_\eta\) rather than to specify a prior distribution for the nuisance parameters. This may dramatically reduce computing time, as demonstrated in our simulations.
As in the literature of nonparametric regression, we propose to estimate \( \sigma^2 \) by a difference-based method. Let us consider the case of a fixed design with \( u_1 < \cdots < u_T \). When \( u_1, \cdots, u_T \) are dense, then it is expected that \( m(u_t) - m(u_{t-1}) \approx 0 \). Thus,

\[
d_t = y_t - y_{t-1} \approx \varepsilon_t - \varepsilon_{t-1},
\]

and hence, \( \text{var}(d_t) = 2\sigma^2 \). The difference based estimate for \( \sigma^2 \), denoted by \( \hat{\sigma}^2 \), is half of the sample variance of \( \{d_t, t = 2, \cdots, T\} \). We use GCV to estimate \( \sigma^2_\eta \) by ratio of \( \hat{\sigma}^2 / \hat{\lambda} \), where \( \hat{\lambda} \) is the selected \( \lambda \) as originally proposed by (Craven & Wahba, 1979) for selecting smoothing parameter in the smoothing splines. The fitted values \( \hat{\mathbf{y}} \) can be expressed as

\[
\hat{\mathbf{y}}(\lambda) = H(\lambda)\mathbf{y},
\]

where \( H(\lambda) \) is the \( T \times T \) hat matrix, depending only on \( u \)-variate. Define the effective number of parameters to be \( e(\lambda) = \text{tr}(H(\lambda)) \), then the GCV approach selects \( \lambda \) that minimizes

\[
GCV(\lambda) = \frac{||\mathbf{y} - \hat{\mathbf{y}}(\lambda)||^2}{T \left(1 - e(\lambda)/T\right)^2},
\]

That is,

\[
\hat{\lambda} = \text{argmin}_\lambda GCV(\lambda).
\]

Note that for \( k = 0 \), the fitted values of \( \mathbf{y} \) can be easily expressed as a linear estimator of \( \mathbf{y} \). Thus, \( \hat{\sigma}^2_\eta = \hat{\sigma}^2 / \hat{\lambda} \).

The Gibbs sampler procedures follow:

**Step 1.** Before starting the iteration, we use a GCV data driven method to determine an initial value for smooth parameters \( \hat{\lambda} \);

**Step 2.** Let \( \mathbf{d} = \{d_1, \cdots, d_T\} \) denote the difference of a sequential observed values: \( d_t = \)
\[ y_t - y_{t-1}, t = 2, \ldots, T. \] Setting \( \hat{\sigma}^2 = \frac{1}{2} \text{Var}(d) \), we compute the estimated value for \( \hat{\sigma}^2_n = \hat{\sigma}^2/\hat{\lambda} \).

**Step 3.** Draw initial values for \( m(u_0) \) from \( \mathcal{N}(0, \hat{\sigma}^2) \);

**Step 4.** At iteration \( i = 1 \), denote \( \theta_i = (m_t, m_{t-1}, \ldots, m_{t-k})^T \). Draw \( \{m(u_{t-k}), \ldots, m(u_t)\}^{(1)} \) by the following FFBS (Forward Filtering Backward Sampling) algorithm:

i) Sample \( \theta_t^{(1)} \sim \mathcal{N}(\mu_t, C_t) \)

ii) For each \( n = t - 1, t - 2, \ldots, 1 \), sample \( \{\theta_{t-1}, \theta_{t-2}, \ldots, \theta_1\}^{(1)} \) from

\[
\theta_n^{(1)} \mid D_t \sim N(a_t(-1), R_t(-1))
\]

where values for \( a_t(-1) \) and \( R_t(-1) \) are defined in the Appendix.

**Step 5.** Set \( i = i + 1 \), repeat step 4 until \( i = I \) iterations.

We choose starting value of \( \{\sigma^2, \sigma^2_\eta\} \) by a data driven GCV method and keep the same value through all Gibbs sampler iterations for mean function \( \{m(u_1), \ldots, m(u_T)\} \).

### 3.4 Simulations

In this section, we assess the finite sample performance of the proposed procedure by using a Monte Carlo simulation and compare it with the local linear regression, a commonly used smoothing method. All simulations are conducted using MATLAB code. In our simulation, the size of the Gibbs sampler is 2000 and the first 1000 are set as burn-in samples. We have also varied the size of the Gibbs sampler from 1000 to 10000 and found that the number of the Gibbs sampler does not affect the convergence of the estimator because the values
of MSE indeed only slightly change from iteration 1000 to 10,000. Thus, we set the size of
Gibbs sampler as 2000. The estimation accuracy of the proposed procedure will be assessed
by the Mean Squared Error (MSE) defined as

$$\text{MSE}(\hat{m}(\cdot)) = \frac{1}{T} \sum_{t=1}^{T} \{\hat{m}(u_t) - m(u_t)\}^2,$$

and computing time.

Example 3.4.1. In this example, we generate 500 data sets from

$$y_t = m(u_t) + \varepsilon_t,$$

where $\varepsilon_t$ follows a standard normal distribution $N(0, 1)$ and $m(u) = 4 \times \cos(2\pi u) + c \times \sin(f \times \pi u)$. This regression function is a mixture of the low-frequency signal, $4 \times \cos(2\pi u)$, and the high-frequency signal, $c \times \sin(f \times \pi u)$. To compare the performance of the proposed procedures, we consider several combinations of $c$ and $f$ values. In our simulation, we take the sample size $T$ to be 500 and 1000 and we consider two scenarios for the independent variable $u_t$: one is the fixed design with $u_t = (t - 0.5)/T$ and the other one is the random design with $u_t$ following a uniform distribution over $[0, 1]$.

We first set the high-frequency signal with frequency $f = 40$, and take $c = 0.5, 1, 1.5, 2, 2.5$ and 3. Table 3.1 displays the sample average of MSE’s over the 500 simulations. In Table 3.1, ‘LLE’ stands for the local linear estimator with bandwidth selected by the plug-in method proposed by Ruppert, Sheather and Wand (1995). ‘FB’ stands for the full Bayesian estimation procedure described in Section 3.1, and ‘EB’ stands for the empirical Bayesian estimation procedure proposed in Section 3.2. From Table 3.1, it can be seen that when the high-frequency signal is weak, i.e., $c = 0.5$, the local linear estimator is slightly better than the newly proposed procedure. However, when the high-frequency signal becomes stronger,
Table 3.1. Mean squared Error ($f = 40$)

<table>
<thead>
<tr>
<th>$T$</th>
<th>$c$</th>
<th>Fixed Design</th>
<th>Random Design</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LLE FB EB</td>
<td>LLE FB EB</td>
</tr>
<tr>
<td>500</td>
<td>0.5</td>
<td>0.1423 0.1527 0.1542</td>
<td>0.1394 0.1581 0.1597</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.5109 0.2069 0.2066</td>
<td>0.4979 0.2262 0.2253</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>1.1129 0.2412 0.2370</td>
<td>1.0648 0.2776 0.2717</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.9294 0.2667 0.2577</td>
<td>1.7626 0.3206 0.3087</td>
</tr>
<tr>
<td></td>
<td>2.5</td>
<td>2.9880 0.2874 0.2738</td>
<td>2.5653 0.3600 0.3432</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4.3167 0.3038 0.2881</td>
<td>3.5100 0.3966 0.3779</td>
</tr>
<tr>
<td>1000</td>
<td>0.5</td>
<td>0.1325 0.0929 0.0937</td>
<td>0.1323 0.0952 0.096</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.5140 0.1293 0.1295</td>
<td>0.4957 0.1365 0.1366</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>1.1118 0.1529 0.1516</td>
<td>0.9637 0.1672 0.1656</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.9112 0.1707 0.1674</td>
<td>1.4799 0.1928 0.1889</td>
</tr>
<tr>
<td></td>
<td>2.5</td>
<td>2.9330 0.1848 0.1787</td>
<td>2.0551 0.2157 0.2087</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4.1918 0.1971 0.1880</td>
<td>2.6907 0.2377 0.2273</td>
</tr>
</tbody>
</table>

The proposed procedure outperforms the local linear estimator. For example, when $c = 3$ for fixed design, the MSE of the local linear estimator is more than 20-fold larger than the MSE of the proposed method. The overall patterns for $T = 500$ and 1000 are similar. It seems that in terms of MSE, local linear estimator performs better in random design cases, while the proposed procedure works better in the fixed design cases. It is interesting to observe from Table 3.1 that the full Bayesian procedure and the empirical Bayesian procedure have almost the same performance, but the empirical Bayesian procedure can dramatically reduce computational burden, as evidenced in Table 3.2, which summarizes the computing time of 500 simulations. From Table 3.2, we can see that the local linear estimator needs the least computational time and the empirical Bayesian estimation procedure may save half of the computation time of the full Bayesian estimation procedure.

We next investigate how the proposed procedure performs with different frequencies $f$. Table 3.3 depicts the average of MSE over the simulations with $c = 2$ and $f = 10, 20, \cdots, 60$. From Table 3.3, it can be seen that when $f$ is small, for example, $f = 10$, the local
linear estimator outperforms the dynamic linear model approach. When $f$ increases, the performance of the local linear estimator worsens dramatically, while the dynamic linear model approach has more gain over the local linear estimator. Again, the performance of the full Bayesian procedure and the empirical Bayesian procedure is similar.

**Table 3.3.** Mean squared error ($c = 2$)

<table>
<thead>
<tr>
<th>$T$</th>
<th>$f$</th>
<th>Fixed Design</th>
<th>Random Design</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LLE FB EB LLE FB EB</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>10</td>
<td>0.0598 0.1388 0.1369 0.0582 0.1480 0.1466</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.2759 0.1885 0.1891 0.1018 0.2119 0.2141</td>
<td></td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>1.9267 0.2264 0.2317 0.4856 0.2640 0.2716</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>1.9294 0.2577 0.2667 1.7626 0.3087 0.3206</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>1.9948 0.2873 0.2973 1.9424 0.3504 0.3637</td>
<td></td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>2.0261 0.3217 0.3276 1.9694 0.3932 0.4025</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>10</td>
<td>0.0338 0.0756 0.0752 0.0339 0.0803 0.0797</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.1649 0.1149 0.1138 0.0598 0.1244 0.1232</td>
<td></td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>1.5440 0.1449 0.1438 0.2072 0.1596 0.1582</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>1.9112 0.1674 0.1653 1.4799 0.1889 0.1870</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>1.9225 0.1872 0.1849 1.9225 0.2124 0.2100</td>
<td></td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>1.9960 0.2402 0.2375 2.0147 0.2402 0.2375</td>
<td></td>
</tr>
</tbody>
</table>

To get insights into why the proposed method may outperform the local linear estimator, let us plot the resulting estimators of a typical data set. The mean function $m(u_t) = 4 \times \cos(2\pi u_t) + 2 \times \sin(40 \times \pi u_t)$ and $\epsilon_t \sim \mathcal{N}(0, 1)$ For the dynamic linear model approach we may easily construct a pointwise confidence interval using highest probability density
(HPD) credit set. Thus we provide the 95% HPD credit set for both full Bayesian procedure and the empirical Bayesian procedure in the top panel of Figure 3.1 to 3.4. To have a close look at the figures, we zoom in on the estimated function over \([0.3, 0.4]\). The zoomed-in plot clearly indicates that the local linear estimator with the plug-in bandwidth selector completely misses the high-frequency signal.

### 3.5 Discussion

In this chapter, we propose a dynamic linear model approach for nonparametric regression models and establish a Bayesian framework for nonparametric regression models. We demonstrate the proposed methodology may outperform the local linear regression, a commonly used smoothing method. We address issues of practical implementation and illustrate methods to reduce computational burden. We focus on the situation in which the observations are independent samples. It is of great interest to extend the proposed methodology for settings in which the observations are dependent. We will study how to estimate the nonparametric regression function with serially dependent errors in Chapter 4. It is also of great interest to apply the proposed methodology for other nonparametric regression models, such as varying coefficient models which are investigated in Chapter 5.
Figure 3.1. Plot of a typical estimation for a fixed design with $T = 500$ showing (a) the whole estimated regression function and (b) the partial estimated function over $[0.3, 0.4]$. Solid curves are the estimated regression function, Dashed curves are the true regression function, and dash-dotted curves are the 95% HPD credit set.
Figure 3.2. Plot of a typical estimation for a fixed design with $T = 1000$ showing (a) the whole estimated regression function and (b) the partial estimated function over $[0.3, 0.4]$. Solid curves are the estimated regression function, Dashed curves are the true regression function, and dash-dotted curves are the 95% HPD credit set.
Figure 3.3. Plot of a typical estimation for a random design with $T = 500$ showing (a) the whole estimated regression function and (b) the partial estimated function over $[0.3, 0.4]$. Solid curves are the estimated regression function, Dashed curves are the true regression function, and dash-dotted curves are the 95% HPD credit set.
Figure 3.4. Plot of a typical estimation for a random design with $T = 1000$ showing (a) the whole estimated regression function and (b) the partial estimated function over $[0.3, 0.4]$. Solid curves are the estimated regression function, Dashed curves are the true regression function, and dash-dotted curves are the 95% HPD credit set.
Nonparametric Regression Function with Serially Correlated Errors

4.1 Introduction

Based on results in chapter 3 we know that our proposed methodology outperforms the local linear estimator, a commonly used smoothing method. A natural extension of our research focuses on nonparametric regression model with an auto-correlated error. Consider a nonparametric regression model

\[ y_t = m(u_t) + \varepsilon_t \]  

where \( \varepsilon_t \) is an auto-correlated error and \( u_t \) either follows a fixed design or a random design.

The chapter is organized as follows. In Section 4.2, we formulate the problem of nonparametric regression function estimation using a dynamic linear model, and further propose Bayesian estimation procedures for the nonparametric regression function. Theoretical properties of dynamic linear models with a correlated autoregressive error are established. In Section 4.3, we investigate the finite sample performance of the proposed estimation procedure via Monte Carlo simulations.
Carlo simulation study. We further illustrate the proposed methodology by an analysis of rates on 3-month T-bill data mentioned previously. Technical proofs are given in Section 4.4. Section 4.5 concludes with a discussion.

4.2 An estimation and forecasting DLM framework for nonparametric regression model with correlated errors

Suppose that \{u_t, y_t\}, \(t = 1, \ldots, T\), is a sample from the model

\[ y_t = m(u_t) + \varepsilon_t, \]

where \(m(\cdot)\) is a nonparametric function to be estimated, and \(\varepsilon_t\) is a random error. In this chapter, we consider the error \(\varepsilon_t\) is correlated. In practice, it is typical to assume that the error process is a stationary time series and to model it by using an autoregressive moving average (ARMA) model. Note that a stationary ARMA model can be represented as an autoregressive (AR) model with an infinite order, i.e. AR(\(\infty\)). In our research, we focus on an AR error with order \(d\) in this paper and we allow \(d\) to be large. That is,

\[ \varepsilon_t = \sum_{j=1}^{d} \rho_j \varepsilon_{t-j} + \gamma_t, \]

where \(\gamma_t \sim \mathcal{N}(0, \sigma^2)\) with a finite \(\sigma^2\).
4.2.1 A connection between DLM and penalized splines

Suppose that \( u_1 < u_2 < \cdots < u_T \) are fixed design points. Denote \( m_t = m(u_t) \), and parameterize \( m(u) \) by

\[
m_k(u) = m_{k-1} + \frac{m_k - m_{k-1}}{u_k - u_{k-1}}(u - u_{k-1}), \quad \text{for } u \in (u_{k-1}, u_k].
\] (4.2)

Thus, if we estimate \( m_k, k = 1, \cdots, T \), we can get an estimator of \( m(\cdot) \) using equation (4.2). However, the equation in (4.2) uses \( T \) parameters to model \( m(\cdot) \), and this leads to an over-parameterized model. Similar to the ideas of smoothing splines, we consider adding a penalty on roughness of the \( k \)-derivative of \( m(\cdot) \). Denote \( m_t^{(0)} = (m_t - m_{t-1}) \), and for \( k \geq 1 \), define \( m_t^{(k)} = (m_t^{(k-1)} - m_{t-1}^{(k-1)}/(u_t - u_{t-1}) \). Thus, it is natural to consider the following penalized least squares

\[
\sum_{t=1}^{T} (y_t - m_t)^2 + \lambda \sum_{t=k+2}^{T} (m_t^{(k)} - m_{t-1}^{(k)})^2.
\] (4.3)

By the definition of \( m_t^{(k)} \), we can represent

\[
m_t^{(k)} - m_{t-1}^{(k)} = \alpha_t \{m_t - \sum_{j=1}^{k+1} \alpha_j m_{t-j}\}
\]

for some constants \( \alpha_{t-k-1}, \ldots, \alpha_{t-1} \) and \( \alpha_t = 1/(u_t - u_{t-1}) \). Thus,

\[
\sum_{t=1}^{T} (y_t - m_t)^2 + \lambda \sum_{t=k+2}^{T} \{m_t - \sum_{j=1}^{k+1} \alpha_j m_{t-j}\}^2.
\] (4.4)

Consider a dynamic linear model
\[
\begin{align*}
y_t &= m_t + \varepsilon_t, \\
m_t &= \sum_{j=1}^{k+1} \alpha_j m_{t-j} + \eta_t,
\end{align*}
\] (4.5)

where \( \alpha_j \)'s are regression coefficients, \( \varepsilon_t \)'s, \( t = 1, \cdots, T \) are independently and identically distributed according to \( N(0, \sigma^2_t) \), and \( \eta_t \), \( t = 1, \cdots, T \) are independently and identically distributed according to \( N(0, \sigma^2_\eta) \). Setting \( \lambda \alpha_t^2 = \sigma^2_t / \sigma^2_\eta \). It can be shown that minimization of equation (4.4) is equivalent to maximizing the logarithm of the likelihood function of model (4.5). In particular, when \( u_t \) is evenly distributed, then \( \alpha_t \) is a constant. Thus, we may consider \( \sigma^2_t \) is constant with respect to \( t \), which is \( \sigma^2_t = \sigma^2 \).

### 4.2.2 Dynamic linear models with an AR error

Let \( F_t = (1, 0, \cdots, 0)^T \), a \( (k+1) \times 1 \)-dimensional vector, \( G_t = (g_{ij}) \) is a \( (k+1) \times (k+1) \) matrix with \( g_{1j} = \alpha_j, \ j = 1, \cdots, k+1 \), \( g_{ij} = 1 \) for \( j = 2, \cdots, k+1 \) and \( g_{ij} = 0 \) if \( i > 1 \) and \( i \neq j+1 \). Furthermore, denote \( \theta_t = (m_t, m_{t-1}, \cdots, m_{t-k})^T \). Then model (4.5) can be written as a special case of the general DLM with an AR error as defined below.

**Definition 4** A dynamic linear model with an AR error is defined as

- **Observation equation:** \( y_t = F_t^T \theta_t + \nu_t \),
- **System equation:** \( \theta_t = G_t \theta_{t-1} + \omega_t \),
- **AR error:** \( \nu_t = \sum_{j=1}^{d} \rho_j \nu_{t-j} + \gamma_t \),
- **Initial information:** \( (\theta_0|D_0) \sim N(\mu_0, C_0) \).

where \( \omega_t \sim N(0, W) \) with \( W = \text{diag}\{\sigma^2_\nu, 0, \cdots, 0\} \) and \( \gamma_t \sim N(0, V) \). \( \omega_t \) and \( \gamma_t \) are independent, and they are also independent of \( (m_0|D_0) \), where \( D_0 \) is the \( \sigma \)-field generated by information at time \( t = 0 \).

The DLM has been studied under various settings. See, for example, West and Harrison (1997). However, to the best of our knowledge, the DLM with an AR error has not yet
been studied. Therefore, we develop the Bayesian theory for the DLM with an AR error. The developed Bayesian theory can be viewed as a natural extension of the DLM with an independent error.

Assume that $\theta_0$, $W$, $V$ and $\rho$ are independent and their prior distributions are given by

Prior distribution for $\theta_0$: $\theta_0 \sim N(\mu_0, C_0)$

Prior Distribution for $W$: $W \sim IW(q, \Delta_0)$

Prior Distribution for $V$: $V \sim IG(\alpha_0, \beta_0)$

Prior Distribution for $\rho$: $\rho \sim U(R^d)$

where $IW(q, \Sigma)$ stands for the inverse-Wishart distribution:

$$f(W) = \frac{1}{2^{\frac{p^2}{2}} \times \Gamma_p\left(\frac{p}{2}\right)} \times \exp\left\{ -\frac{1}{2} \text{trace}\left( W^{-1} \Sigma \right) \right\}$$

where $W$ is a $p \times p$ positive defined random matrix and $\Gamma_p(a) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma(a + \frac{1-j}{2})$.

$IG(\alpha, \beta)$ stands for the inverse of Gamma distribution,

$$f(V) = \frac{\beta^\alpha}{\Gamma(\alpha)} \times (V^{-1})^{\alpha+1} \exp(-\beta V^{-1})$$

and $U(R^d)$ stands for the uniform distribution over $(R^d)$.

Bayesian inferences for observation $y_t$ and evaluation mean function $\theta_t$ are obtained sequentially in Theorem 1.

**Theorem 1.** Denote $D_s$ to be the $\sigma$-field generated by data collected at $s = 0, \cdots , t$. It is assumed that the conditional posterior distribution for $\theta_t$ at state $t - 1$ is

$$\theta_{t-1} \mid D_{t-1} \sim N(\mu_{t-1}, C_{t-1})$$

where $\mu_{t-1}$ and $C_{t-1}$ are the mean and variance for $\theta_t$ at state $t - 1$. For the DLM model
(4.2.2), we have the following conditional posterior distributions.

1. **Conditional prior distribution for** \( \theta_t \) **at** \( t \) **is:**

\[
\theta_t \mid (\theta_{t-1}, W) \sim N(a_t, R_t)
\]

\( a_t = G_t \mu_{t-1} \) and \( R_t = G_t C_{t-1} G_t^T + W \)

2. **One-step forecasting distribution for** \( y_t \) **is**

\[
y_t \mid \left( D_{t-1}, \theta_{t-1}, \cdots, \theta_{t-d}, V, \rho \right) \sim N(f_t, Q_t)
\]

where \( f_t = F_t^T a_t + \sum_{j=1}^d \rho_j (y_{t-j} - F_{t-j}^T \theta_{t-j}) \) and \( Q_t = F_t^T R_t F_t + V \).

3. **Let** \( Y_t = (y_0, \cdots, y_t)^T \). **Conditional posterior distribution for** \( \theta_t \), **given** \( (Y_t, \theta_{t-1}, W, V, \rho) \) **is**

\[
\theta_t \mid (Y_t, \theta_{t-1}, \cdots, \theta_{t-d}, W, V, \rho) \sim N(\mu_t, C_t)
\]

\[
\mu_t = a_t + \frac{R_t F_t^T}{Q_t} (y_t - F_t^T a_t)
\]

and

\[
C_t = R_t - \left( \frac{R_t F_t^T}{Q_t} \right) R_t \left( \frac{C_t F_t^T}{Q_t} \right)^T
\]

4. **The conditional posterior distributions for** \( \{V, W, \rho\} \) **are**

i. **The conditional posterior distributions for** \( V \): \( V \mid (\rho, \{\theta_1, \cdots, \theta_T\}) \sim IG(\alpha, \beta) \)
where 
\[ \alpha = \frac{T}{2} + \alpha_0 \]
\[ \beta = \beta_0 + \frac{1}{2} \sum_{t=1}^{T} \left[ y_t - F_t^T \theta_t - \sum_{j=1}^{d} \rho_j \left( y_{t-j} - F_{t-j}^T \theta_{t-j} \right) \right]^2 \]

ii. The conditional posterior distributions for \( W \)

\[ W \mid (\theta_1, \cdots, \theta_T) \sim IW(q, \Delta) \]

where \( q = T + q_0 \) and 
\[ \Delta^{-1} = \Delta_0 + \sum_{t=1}^{T} (\theta_t - \theta_{t-1}) (\theta_t - \theta_{t-1})^T \]

iii. The conditional posterior distributions for \( \rho \):

\[ \rho \mid (V, \{\theta_1, \cdots, \theta_T\}) \sim N \left( (X^*T X^*)^{-1} X^* Y^* (X^*T X^*)^{-1} \right) \]

where \( Y^* \) and \( X^* \) are defined in equation (4.8).

(5) At point \( t \), the one-step forecasting distribution for \( y_{t+1} \) is

\[ y_{t+1} \mid (Y_t, \theta_t, \cdots, \theta_{t+1-d}, V, W, \rho) \sim N \left( f_t(1), Q_t(1) \right) \]

\[ f_t(1) = F_{t+1}^T G_{t+1} \mu_t + \sum_{j=1}^{d} \rho_j (y_{t+1-j} - F_{t+1-j}^T \theta_{t+1-j}) \]

\[ Q_t(1) = F_{t+1}^T (G_{t+1} C_t G_{t+1}^T + W) F_{t+1} + V \]

The proof of Theorem 1 is given in Section 4.
It is always useful to look back in time in order to get a clearer picture of what may happen in the future and make corresponding decisions accordingly. The use of current data to derive previous inferences of $\theta_t$ is called filtering. The information is filtered back to a previous time point, for example the distribution of $(\theta_{t-1}|\theta_t, D_t)$ is called the one-step filtered distribution for $\theta_{t-1}$ at time $t$. The retrospective estimation of the historical development of a time series mean function $\theta_t$ using the filtered distribution is called smoothing the series. West and Harrison (1997) proved that one-step filtered distribution for $\theta_{t-1}$ is $(\theta_t|\theta_{t+1}, Y_t, V, W, \rho) \sim N(h_t, H_t)$ where $h_t = \mu_t + C_tG_{t+1}^T R_{t+1}^{-1}(\theta_{t+1} - a_{t+1})$ and $H_t = C_t - C_tG_{t+1}^T R_{t+1}^{-1}G_{t+1}C_t$.

4.3 The Gibbs sampler algorithm

A simulation algorithm includes both estimation and forecasting components.

The estimation algorithm

Let $i = 1, \cdots, I$ be the total number of iterations. We sample $\{\theta_1, \cdots, \theta_t\}$ and $\{W, V, \rho\}$ separately at each iteration:

Step 1. Set $a_1 = (0, 0, \cdots, 0)_{(k+1)\times1}$, $R_1 = 0.01 \times I_{(k+1)\times(k+1)}$ and $F_t = (1, 0, \cdots, 0)^T_{(k+1)\times1}$.

Step 2. Pick initial values for $\{y_0, V^{(0)}, W^{(0)}, \rho^{(0)}\} = \{0, 0.1, 0.005 \times I_{(k+1)\times(k+1)}, 0.1 \times (1, 1, \cdots, 1)_{d\times1}\}$.

Step 3. At iteration $i = 1$, sample $\{\theta_1, \cdots, \theta_t\}^{(1)}$ by FFBS algorithm:

i Sample $\theta_t^{(1)} \sim N(\mu_t, C_t)$

ii for each $n = t - 1, t - 2, \cdots, 1$, sample $\{\theta_{t-1}, \theta_{t-2}, \cdots, \theta_1\}^{(1)}$ from

$$\theta_n^{(1)} \left| (\theta_{n+1}^{(1)}, Y_n, V^{(0)}, W^{(0)}, \rho^{(0)}) \right. \sim N(h_n, H_n)$$
Step 4. Sample $\{V, W, \rho\}^{(1)}$ from the conditional posterior distributions:

$$V^{(1)} \mid \left( Y_t, \rho^{(0)}, \{\theta_1, \cdots, \theta_t\}^{(1)} \right) \sim IG(\alpha, \beta)$$

$$W^{(1)} \mid \left( \{\theta_1, \cdots, \theta_n\}^{(1)} \right) \sim IW(q, \Delta)$$

$$\rho^{(1)} \mid \left( Y_n, V^{(1)}, \{\theta_1, \cdots, \theta_t\}^{(1)} \right) \sim N \left( (X^{*T}X^{*})^{-1}X^{*}Y^{*}, V(X^{*T}X^{*})^{-1} \right)$$

where $Y^{*}$ and $X^{*}$ are defined in equation (4.8)

Step 5. Set $i = i + 1$, repeat step 3 and step 4 until $i = I$ iterations.

The sampling algorithm works like a loop, repetitively drawing parameters to read data values and create observations one by one.

The forecasting algorithm

The forecasting distribution for $y_{t+1}$ is given by

$$y_{t+1} \mid \left( Y_t, \{\theta_t, \cdots, \theta_{t+1-d}, \hat{V}, \hat{W}, \rho\}^{(1)} \right) \sim N \left( f_t(1), Q_t(1) \right)$$

where $f_t(1) = F_{t+1}^T a_t + \sum_{j=1}^{d} \rho_j (y_{t-j} - F_{t-j}^T \theta_{t-j})$ and $Q_t(1) = F_{t+1}^T R_q F_{t+1} + V$.

4.4 Simulations and an application

In this section, we illustrate a nonparametric regression model with an auto-correlated error via simulations and an application to a daily interest rate data set.
4.4.1 Simulation study

We conduct Monte Carlo simulation studies to compare the finite sample performance of the proposed DLM estimator with the local linear estimator. All simulations are conducted by using MATLAB code. In our simulation, $k$ is set to be 1, the size of the Gibbs sampler is 2000 and the first 1000 is set as a burn-in sample. We have also tried the size of Gibbs sampler from 1000 to 10,000. We find that the size of Gibbs sampler does not affect the convergence of the estimator because the values of MSE only slightly change from iteration 1000 to 10,000. Thus, we set the size of Gibbs sampler as 2000. The estimation accuracy of the proposed procedure is be assessed by the Mean Squared Error (MSE) defined as

$$\text{MSE}(\hat{m}(\cdot)) = \frac{1}{T} \sum_{t=1}^{T} \{\hat{m}(u_t) - m(u_t)\}^2,$$

and computing time.

In this simulation study, we generate 500 data sets from the following nonparametric regression model:

$$y_t = m(u_t) + \varepsilon_t$$

where mean function $m(u_t) = 4 \times \cos(2\pi u_t) + c \times \sin(f \times \pi u_t)$ and $\varepsilon_t$ follows an AR(1) series $\varepsilon_t = 0.4 \times \varepsilon_{t-1} + \gamma_t$, $\gamma_t \sim \mathcal{N}(0, 0.1)$. This regression function is a mixture of the low-frequency signal, $4 \times \cos(2\pi u_t)$, and the high-frequency signal, $c \times \sin(f \times \pi u_t)$. We select several combinations of $c$ and $f$ values to compare their performances. In our simulation, we take $T$ to be sample sizes of 500 and 1000. We consider two scenarios for the independent variable $u_t$: one is the fixed design with $u_t = (t - 0.5)/T$ and the other is the random design with $u_t$ following a uniform distribution over $[0, 1]$.

First, we set the high-frequency signal with frequency $f = 40$ and take $c \in \{0.5, 1, 1.5, 2, 2.5, 3\}$. Table 4.1 displays sample average of MSEs over 500 simulations. In the Local Linear
Table 4.1. Mean squared error \((f = 40, \rho = 0.4)\)

<table>
<thead>
<tr>
<th>T</th>
<th>c</th>
<th>Fixed Design</th>
<th>Random Design</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LLE</td>
<td>DLM</td>
</tr>
<tr>
<td>500</td>
<td>0.5</td>
<td>0.1946</td>
<td>0.1429</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.5848</td>
<td>0.5035</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>1.2182</td>
<td>1.0577</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.0631</td>
<td>1.7522</td>
</tr>
<tr>
<td></td>
<td>2.5</td>
<td>3.0907</td>
<td>2.57</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4.3374</td>
<td>3.5385</td>
</tr>
<tr>
<td>1000</td>
<td>0.5</td>
<td>0.1683</td>
<td>0.1338</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.5695</td>
<td>0.4923</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>1.2266</td>
<td>0.9662</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.1023</td>
<td>1.477</td>
</tr>
<tr>
<td></td>
<td>2.5</td>
<td>3.1547</td>
<td>2.0443</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4.3839</td>
<td>2.6998</td>
</tr>
</tbody>
</table>

Estimator('LLE'), the bandwidth is selected by the plug-in method proposed by Ruppert, Sheather and Wand (1995). ‘DLM’ stands for the DLM estimator proposed in Section 3.2. Table 4.1 shows that when the high-frequency signal is set at 0.5, the average MSE’s of local linear estimator methods are about half of the newly proposed DLM estimator. As the high-frequency signal increases from 1 to 3, the newly proposed estimator outperforms the local linear estimator. The average MSE’s of the proposed procedures are about 10% of those of the local linear estimator when the high-frequency signal reaches 3. On average, in terms of MSE, the local linear estimator performs better in random design cases, while the proposed DLM estimator works better in the fixed designs. The patterns are applied to both sample sizes (500 and 1000). When setting the high-frequency signal \(c = 2\) with different \(f\) values \(f \in \{10, 20, 30, 40, 50, 60\}\), the simulation outputs are similar to those in Table 4.2.

Figures 4.1 to 4.4 are estimated mean function trajectory plots from typical data set. The mean function \(m(u_t) = 4 \times \cos(2\pi u_t) + 2 \times \sin(40 \times \pi u_t)\) and \(\varepsilon_t\) follows an AR(1) series.
Table 4.2. Mean squared error \((c = 2, \rho = 0.4)\)

<table>
<thead>
<tr>
<th>T</th>
<th>f</th>
<th>Fixed Design</th>
<th>Random Design</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LLE</td>
<td>DLM</td>
</tr>
<tr>
<td>500</td>
<td>10</td>
<td>0.1164</td>
<td>0.4378</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>0.2529</td>
<td>0.4499</td>
</tr>
<tr>
<td>30</td>
<td>10</td>
<td>1.4594</td>
<td>0.4783</td>
</tr>
<tr>
<td>40</td>
<td>10</td>
<td>2.0631</td>
<td>0.5239</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
<td>1.9918</td>
<td>0.5835</td>
</tr>
<tr>
<td>60</td>
<td>10</td>
<td>2.0377</td>
<td>0.6493</td>
</tr>
<tr>
<td>1000</td>
<td>10</td>
<td>0.0661</td>
<td>0.4337</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>0.1429</td>
<td>0.4352</td>
</tr>
<tr>
<td>30</td>
<td>10</td>
<td>1.0118</td>
<td>0.4394</td>
</tr>
<tr>
<td>40</td>
<td>10</td>
<td>2.1023</td>
<td>0.4471</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
<td>1.9229</td>
<td>0.458</td>
</tr>
<tr>
<td>60</td>
<td>10</td>
<td>1.9918</td>
<td>0.4744</td>
</tr>
</tbody>
</table>

\(\varepsilon_t = 0.4 \times \varepsilon_{t-1} + \gamma_t, \quad \gamma_t \sim \mathcal{N}(0, 0.1)\)

They visually illustrate performance of the proposed DLM estimator and the local linear estimator. The plots include 95% HPD credit set for the proposed estimator.

In conclusion, the proposed DLM estimator procedure is superior to local linear estimator in certain circumstance where data are volatile and correlated. Both estimators offer certain trends for change of data.

### 4.4.2 Interest rate data set

Analysis and forecasting of interest rates are of paramount economic interest. Analysis of T-Bill rates directly help with forecasting rates on other fixed-income securities such as long-term corporate bonds or mortgage rates. In these cases, the long-term corporate bond rates are frequently treated as linear functions of the T-Bill rates. Also, increases in T-Bill rates tend to depress longer-term bond and equity markets. Indeed, the Federal Funds interest
Figure 4.1. Plot of a typical estimator for a fixed design with AR(1) model error, $\rho = 0.4$ and $T = 500$ showing (a) the whole estimated regression function and (b) the partial estimated function over $[0.3, 0.4]$. Solid curves are the estimated regression function, dashed curves are the true regression function, and dash-dotted curves are the 95% HPD credit set.
Figure 4.2. Plot of a typical estimator for a fixed design with AR(1) model error, $\rho = 0.4$ and $T = 1000$ showing (a) the whole estimated regression function and (b) the partial estimated function over $[0.3, 0.4]$. Solid curves are the estimated regression function, dashed curves are the true regression function, and dash-dotted curves are the 95% HPD credit set.
Figure 4.3. Plot of a typical estimator for a random design with AR(1) model error, $\rho = 0.4$ and $T = 500$ showing (a) the whole estimated regression function and (b) the partial estimated function over [0.3, 0.4]. Solid curves are the estimated regression function, dashed curves are the true regression function, and dash-dotted curves are the 95% HPD credit set.
Figure 4.4. Plot of a typical estimator for a random design with AR(1) model error, $\rho = 0.4$ and $T = 1000$ showing (a) the whole estimated regression function and (b) the partial estimated function over $[0.3, 0.4]$. Solid curves are the estimated regression function, dashed curves are the true regression function, and dash-dotted curves are the 95% HPD credit set.
rates, i.e., the rate on overnight loans from one bank to another, are the target of Federal Reserve monetary policy aimed at affecting macroeconomic conditions.

As an illustration of the proposed methodology, we analyze the 3-month T-Bill interest rate data, which were collected as 3-month T-Bill prices from January 1, 1990 through June 11, 2007 from Reuters. There are a total of 4,451 daily price observations. Set the interest rate (%) to be the response variable \( y_t \). T-Bill market prices are conventionally fitted to the continuously compounded formula of:

\[
P_0 = \mathcal{M} \times \exp(y_t),
\]

where \( P_0 \) is the market price, \( \mathcal{M} \) is the maturity value, \( y_t \) is the continuously compounded annualized rate of return, and \( t \) measures years including fractions thereof. We compute the annualized continuously-compounded interest rates as:

\[
4 \times \log \left( \frac{\text{3-month T-Bill bond maturity value}}{\text{3-month T-Bill bond market value}} \right).
\]

For instance, assume that a 3-month T-Bill bond with par value 10,000 is sold today at price 9,800, then the annualized continuously-compounded interest rates is

\[
4 \times \log \left( \frac{10,000}{9,800} \right) \approx 0.0808.
\]

We fit the annualized continuously-compounded interest rates by using the nonparametric regression model:

\[
y_t = m(u_t) + \varepsilon_t, \quad t = 1, \ldots, T \tag{4.7}
\]

with \( \varepsilon_t = \rho \varepsilon_{t-1} + \gamma_t \). In our analysis, we set \( k = 1 \). That is, the mean function follows \( m(u_t) = m(u_{t-1}) + \eta_t \). Size of Gibbs sampler is 10,000.

Figure 4.5 depicts the resulting estimators of the mean function by two estimation methods. In order to have a closer look at the resulting estimate, Figure 4.6 depicts a portion of
the resulting estimators for August 11, 1998 through December 29, 1998. It is apparent that the local linear estimator catches the smooth, overall trend, but results in a large bias. MSE (MSE($\hat{y}($) = \frac{1}{T} \sum_{t=1}^{T} (\hat{y}_t - y_t)^2$) for local linear estimator and proposed DLM estimator is 0.0070 and 0.0060, respectively. The proposed DLM approach yields a reasonable estimation.

In the rest of this example, we demonstrate one-step and two-step forecasting performance of the proposed model, a dynamic linear model with correlated error. We compare one-step forecasting output with the dynamic linear model developed by West and Harrison (1997). For one-step forecasting, we use first $1 - 400$ observations, from day January 1st, 1990 to November 29, 1991 to forecast the next 100 days’ interest rate. For example, we use the $\{y_1, \cdots, y_{400}\}$ to forecast $\hat{y}_{401}$, $\{y_1, \cdots, y_{401}\}$ to forecast $\hat{y}_{402}$, $\cdots$, $\{y_1, \cdots, y_{t-1}\}$ to forecast $\hat{y}_t$ and $\{y_1, \cdots, y_{499}\}$ to forecast $\hat{y}_{500}$. The sampling algorithm used is stated in section 2.3 and it ends up with 100 one-step forecasting values $\{\hat{y}_{401}, \cdots, \hat{y}_{500}\}$ for day December.
Figure 4.6. Trajectory plots for the US interest rate from August 11, 1998 to Dec. 29, 1998 showing (a) local linear estimation and (b) Dynamic Linear Model with AR(1) error estimation. Solid curves are the estimated regression function, dashed curves are observed values, and dash-dotted curves are the 95% HPD credit set.

2, 1991 to April 17, 1992. A series of plots show the current method of dynamic linear model estimation alongside the approach of West and Harrison (1997). Figure 4.7 shows the difference in convergence with one step forecasting, with the current model tightly following the observed values. The error plot in Figure 4.8 shows that the current method’s error is consistently close to zero, and the scatter plot of observed vs. forecasting values in Figure 4.9 shows a tight 1 : 1 ratio for forecasting and observed values in the proposed model, but not for West and Harrison (1997). This demonstrates that not only is this new method superior, it is sufficient to utilize in one-step forecasting.

Next, for two-step forecasting, we use first 1 – 400 observations, from day January 1st, 1990 to November 29, 1991 to forecast next 100 days’ interest rate. For example, we use \( \{y_1, \cdots, y_{400}, \tilde{y}_{401}\} \) to forecast \( \tilde{y}_{402} \), \( \{y_1, \cdots, y_{401}, \tilde{y}_{402}\} \) to forecast \( \tilde{y}_{403} \), \cdots , \( \{y_1, \cdots, y_{n-1}, \tilde{y}_t\} \) to forecast \( \tilde{y}_{t+1} \) and \( \{y_1, \cdots, y_{498}, \tilde{y}_{499}\} \) to forecast \( \tilde{y}_{501} \). Thus we forecast 100 two-step fore-
casting values \{\tilde{y}_{402}, \cdots, \tilde{y}_{501}\} from day December 2^{nd}, 1991 to April 20, 1992.

4.5 Proof

This section provides a detailed proof of Theorem 1. The proof is done by induction and the multivariate normal distribution theory.

Proof of (1). We know at state \( t - 1 \) conditional posterior distribution for \( \theta_t \) is \( \mathcal{N}(\mu_{t-1}, C_{t-1}) \) and \( \theta_t \) is the sum of two independent normal quantities, \( \theta_t = G_t \theta_{t-1} + \omega_t \). According to conjugate properties, the posterior distribution of \( \theta_t \) itself is also a normal quantity. The updated mean and variance are obtained by adding means and variances of the summands.

\[
E\left( \theta_t \mid \theta_{t-1}, W \right) = E\left( G_t \theta_{t-1} + \omega_t \mid \theta_{t-1}, W \right)
\]
Figure 4.8. One-step forecasting error from 12/02/1991 to 4/17/1992 showing (a) the Dynamic Linear Model with AR(1) error and (b) the Dynamic Linear Model with i.i.d. error.

\[ = G_t E(\theta_{t-1}) + E(\omega_t) \]
\[ = G_t \mu_{t-1} \]
\[ = a_t \]

since \( E(\omega_t) = 0 \). Furthermore, it follows that

\[ \text{Var}(\theta_t | \theta_{t-1}, W) = \text{Var}(G_t \theta_{t-1} + \omega_t | \theta_{t-1}, W) \]
\[ = G_t \text{Var}(\theta_{t-1}) G_t^T + \text{Var}(\omega_t) \]
\[ = G_t C_{t-1} G_t^T + W \]
\[ = R_t. \]

Thus the conditional prior distribution for \( \theta_t \) at point \( t \) is

\[ \theta_t | (\theta_{t-1}, W) \sim N(a_t, R_t) \]
Figure 4.9. One-step forecasting vs observed value plot from 12/02/1991 to 4/17/1992 showing (a) the Dynamic Linear Model with AR(1) error and (b) the Dynamic Linear Model with i.i.d. error.

This completes the proof of part (1).

**Proof of Part (2).** At point $t$, let $Y_t = \{y_1, y_2, \cdots, y_t\}$ thus $Y_{t-1} = \{y_1, y_2, \cdots, y_{t-1}\}$ and prior information of $\theta_t$, $y_t$ is the sum of two independent random quantities, $G_t\theta_t$ and $\nu_t$

\[
\begin{align*}
\mathbb{E}(y_t \mid Y_{t-1}, \theta_{t-1}, \cdots, \theta_{t-d}, V, W, \rho) &= \mathbb{E}(F_t^T\theta_t + \nu_t \mid Y_{t-1}, \theta_{t-1}, \cdots, \theta_{t-d}, V, W) \\
&= F_t^T\mathbb{E}(\theta_t) + \mathbb{E}\left(\sum_{j=1}^{d} \rho_j(y_t-j - F_{t-j}^T\theta_{t-j}) + \gamma_t \mid Y_{t-1}, \theta_{t-1}, \cdots, \theta_{t-d}, V, W, \rho\right)
\end{align*}
\]
Figure 4.10. Two-step forecasting vs observed value trend plot from 12/02/1991 to 04/20/1992 showing (a) the two-step forecasting vs observed value trend, solid curves are the two-step forecasting trend and dots are the observed value, (b) the two-step forecasting error and (c) the two-step forecasting vs observed value.

\[ f_t = F_t^T a_t + \sum_{j=1}^{d} p_j (y_{t-j} - F_{t-j}^T \theta_{t-j}) \]
and

\[
\begin{align*}
\text{Var}(y_t \mid Y_{t-1}, \theta_{t-1}, \ldots, \theta_{t-d}, V, W, \rho) &= \text{Var}(F_t^\top \theta_t + \nu_t \mid Y_{t-1}, \theta_{t-1}, V, W) \\
&= \text{Var}(F_t^\top \theta_t + \sum_{j=1}^d \rho_j \nu_{t-j} + \gamma_t \mid Y_{t-1}, \theta_{t-1}, V, W) \\
&= \text{Var}(F_t^\top \theta_t + \sum_{j=1}^d \rho_j (y_{t-j} - F_t^\top \theta_{t-j}) + \gamma_t \mid Y_{t-1}, \theta_{t-1}, V, W) \\
&= F_t^\top \text{Var}(\theta_t) F_t + \text{Var}(\gamma_t) \\
&= F_t^\top R_t F_t + V \\
&= Q_t
\end{align*}
\]

Therefore, the one-step forecasting distribution for \(y_t\) at \(t\) is

\[
y_t \mid (Y_{t-1}, \theta_{t-1}, \ldots, \theta_{t-d}, V, \rho) \sim \mathcal{N}(f_t, Q_t).
\]

**Proof of Part (3).** At the point “t”, the density function for \(y_t\) is

\[
f(y_t \mid \{\theta_t, \ldots, \theta_{T-d}\}, \{y_t, \ldots, y_{T-d}\}, V, \rho) = \frac{1}{\sqrt{2\pi} V^{1/2}} \exp\left\{-\frac{1}{2V} \left[(y_t - F_t^\top \theta_t) - \sum_{j=1}^d \rho_j (y_{t-j} - F_{t-j}^\top \theta_{t-j})\right]^2\right\}
\]

The conditional prior distribution for \(\theta_t\) is \(\mathcal{N}(a_t, R_t)\)

\[
f(\theta_t \mid \theta_{t-1}, W) = \frac{|R_t|^{-\frac{1}{2}}}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} (\theta_t - a_t)^T R_t^{-1} (\theta_t - a_t)\right\}
\]
The conditional posterior distribution for $\theta_t$

$$f\left(\theta_t \mid \theta_{t-1}, \cdots, \theta_{t-d}, W, V, \rho, y_t\right) \propto \text{Likelihood function} \times \text{Prior Distribution}$$

Thus, $\theta_t$ has conditional posterior distribution:

$$f\left(\theta_t \mid \theta_{t-1}, \cdots, \theta_{t-d}, W, V, \rho, y_t\right)$$

$$\propto \exp\left\{ -\frac{1}{2} (\theta_t - a_t)^T R_t^{-1} (\theta_t - a_t) \right\}$$

$$\times \exp\left\{ -\frac{1}{2V} \left[ y_t - F_t^T \theta_t - \sum_{j=1}^d \rho_j (y_{t-j} - F_{t-j}^T \theta_{t-j}) \right]^2 \right\}$$

$$\propto \exp\left\{ -\frac{1}{2} (\theta_t - \mu_t)^T C_t^{-1} (\theta_t - \mu_t) \right\}$$

where

$$\mu_t = a_t + \frac{R_t F_t^T}{Q_t} \left( y_t - F_t^T a_t \right)$$

and

$$C_t = R_t - \left( \frac{R_t F_t^T}{Q_t} \right) R_t \left( \frac{C_t F_t^T}{Q_t} \right)^T$$

The conditional posterior distribution for $\theta_t$ is:

$$\theta_t \mid (\theta_{t-1}, W, V, \rho, Y_t) \sim N(\mu_t, C_t)$$

**Proof of Part (4).** By the property of independence for prior distribution, the joint distribution of the model parameters can be expressed as the product of the marginal prior
distribution:

\[ f(\{\theta_1, \cdots, \theta_T\}, W, V, \rho) = f(\theta_1, \cdots, \theta_T) \times f(W) \times f(V) \times f(\rho) \]

The density function is derived from all observations:

\[
f(Y_t|\{\theta_1, \cdots, \theta_T\}, V, \rho) \propto (V)^{-\frac{T}{2}} \exp \left\{ -\frac{1}{2V} \sum_{t=1}^{T} \left[ y_t - F_t^T \theta_t - \sum_{j=1}^{d} \rho_j (y_{t-j} - F_{t-j}^T \theta_{t-j}) \right]^2 \right\}
\]

The joint posterior distribution therefore is

\[
f(\{\theta_1, \cdots, \theta_T\}, W, V, \rho \mid Y_t)
\]

\[
\propto f(Y_t \mid \{\theta_1, \cdots, \theta_T\}, V, \rho) \times f(\{\theta_1, \cdots, \theta_T\}, W, V, \rho)
\]

\[
= (V)^{-\frac{T}{2}} \exp \left\{ -\frac{1}{2V} \sum_{t=1}^{T} \left[ y_t - F_t^T \theta_t - \sum_{j=1}^{d} \rho_j (y_{t-j} - F_{t-j}^T \theta_{t-j}) \right]^2 \right\}
\]

\[
\times |W|^{-\frac{T}{2}} \exp \left\{ -\frac{1}{2} \sum_{t=1}^{T} \left[ (\theta_t - \theta_{t-1})^T W^{-1} (\theta_t - \theta_{t-1}) \right] \right\}
\]

\[
\times (V^{-1})^{n_0-1} \exp(-V^{-1} \beta_0) \times |W|^{-\frac{1}{2}(g_0+p+1)} \exp \left\{ -\frac{1}{2} \text{trace}(W^{-1} \Delta_0) \right\} \times 1^d
\]

This implies that the conditional posterior for the three parameters factor into three independent components. Hence sampling \( \{V, W, \rho\} \) reduces to three independent draws from \( f(V \mid Y_T, \rho, \{\theta_1, \cdots, \theta_T\}) \), \( f(W \mid \{\theta_1, \cdots, \theta_T\}) \) and \( f(\rho \mid Y_T, V, \{\theta_1, \cdots, \theta_T\}) \).

i. The posterior distribution of \( V \) given \( (Y_T, \rho, \{\theta_1, \cdots, \theta_T\}) \) is:
\[
f(V \mid Y_T, \rho, \{\theta_1, \ldots, \theta_T\})
\]

\[
\propto (V^{-1})^{\frac{T}{2}} \exp \left\{ -\frac{V^{-1}}{2} \sum_{t=1}^{T} \left[ y_t - F_t^T \theta_t - \sum_{j=1}^{d} \rho_j \left( y_{t-j} - F_t^T \theta_{t-j} \right) \right]^2 \right\}
\]

\[
\times (V^{-1})^{\alpha_0 - 1} \exp(-V^{-1} \beta_0)
\]

\[
\propto (V^{-1})^{\frac{T+\alpha_0}{2} - 1} \exp \left\{ -V^{-1} \left[ \frac{1}{2} \sum_{t=1}^{T} \left( y_t - F_t^T \theta_t - \sum_{j=1}^{d} \rho_j \left( y_{t-j} - F_t^T \theta_{t-j} \right) \right)^2 + \beta_0 \right] \right\}
\]

\[
f(V \mid Y_T, \rho, \{\theta_1, \ldots, \theta_T\})
\]

\[
\propto (V^{-1})^{\frac{T}{2}} \exp \left\{ -\frac{V^{-1}}{2} \sum_{t=1}^{T} \left[ y_t - F_t^T \theta_t - \sum_{j=1}^{d} \rho_j \left( y_{t-j} - F_t^T \theta_{t-j} \right) \right]^2 \right\}
\]

\[
\times (V^{-1})^{\alpha_0 - 1} \exp(-V^{-1} \beta_0)
\]

\[
= (V^{-1})^{\alpha_0 - 1} \exp(-V^{-1} \beta_0) \times |W|^{-\frac{1}{2}(q_0+p+1)} \exp \left\{ -\frac{1}{2} \text{trace}(W^{-1} \Delta_0) \right\}
\]

\[
\propto (V^{-1})^{\frac{T+\alpha_0}{2} - 1} \exp \left\{ -V^{-1} \left[ \frac{1}{2} \sum_{t=1}^{T} \left( y_t - F_t^T \theta_t - \sum_{j=1}^{d} \rho_j \left( y_{t-j} - F_t^T \theta_{t-j} \right) \right)^2 + \beta_0 \right] \right\}
\]

Thus, conditional posterior distribution for \( V \) is

\[
V \mid (Y_T, \rho, \{\theta_1, \ldots, \theta_T\}) \sim IG(\alpha, \beta)
\]

where \( \alpha = \frac{T}{2} + \alpha_0 \) and

\[
\beta = \beta_0 + \frac{1}{2} \sum_{t=1}^{T} \left[ y_t - F_t^T \theta_t - \sum_{j=1}^{d} \rho_j \left( y_{t-j} - F_t^T \theta_{t-j} \right) \right]^2
\]
ii. The posterior distribution of $W$ given $\{\theta_1, \cdots, \theta_T\}$ is:

$$f(W \mid \{\theta_1, \cdots, \theta_T\})$$

$$\propto |W|^{-\frac{T}{2}} \exp \left\{ -\frac{1}{2} \sum_{t=1}^{T} (\theta_t - \theta_{t-1})^T W^{-1} (\theta_t - \theta_{t-1}) \right\}$$

$$\times |W|^{-\frac{1}{2}(q_0 + p + 1)} \exp \left\{ -\frac{1}{2} \text{trace}(W^{-1} \Delta_0) \right\}$$

$$= |W|^{-\frac{1}{2}(T+q_0+p+1)} \exp \left\{ -\frac{1}{2} \text{trace} \left[ W^{-1} \left( \Delta_0 + \sum_{t=1}^{T} (\theta_t - \theta_{t-1})(\theta_t - \theta_{t-1})^T \right) \right] \right\}.$$ 

Therefore, the conditional posterior distribution for $W$, given $\{\theta_1, \cdots, \theta_T\}$, is

$$W \mid \{\{\theta_1, \cdots, \theta_T\}\} \sim IW(q, \Delta)$$

where $q = T + q_0$ and

$$\Delta = \Delta_0 + \sum_{t=1}^{T} (\theta_t - \theta_{t-1})(\theta_t - \theta_{t-1})^T$$

iii. The full conditional posterior distributions for $\rho$.

For the observed function $y_t = F_t^T \theta_t + \nu_t$ is equivalent to $\nu_t = y_t - F_t^T \theta_t, \forall t = 1, \cdots, T$.

We can rewrite the function as:

$$y_t - F_t^T \theta_t = \sum_{j=1}^{d} \rho_j (y_{t-j} - F_{t-j}^T \theta_{t-j}) + \gamma_t$$

which can be treated as a linear model of response variable $Y^*$ and design matrix $X^*$ with $\rho$ as the model coefficients:

$$Y^* = X^* \rho + \gamma$$  (4.8)
where:

\[
Y^* = \begin{pmatrix}
y_{d+1} - F_{d+1}^T \theta_{d+1} \\
\vdots \\
y_T - F_T^T \theta_T
\end{pmatrix}_{(T-d) \times 1}
\]

\[
X^* = \begin{pmatrix}
(y_1 - F_1^T \theta_1) & \cdots & (y_d - F_d^T \theta_d) \\
\vdots & \ddots & \vdots \\
(y_{T-d} - F_{T-d}^T \theta_{T-d}) & \cdots & (y_{T-1} - F_{T-1}^T \theta_{T-1})
\end{pmatrix}_{(T-d) \times d}
\]

and

\[
\rho = \begin{pmatrix}
\rho_1 \\
\vdots \\
\rho_d
\end{pmatrix}_{d \times 1}
\]

\[
\gamma = \begin{pmatrix}
\gamma_{d+1} \\
\vdots \\
\gamma_T
\end{pmatrix}_{(T-d) \times 1}
\]

where \( \gamma \sim N(0, V I) \) and \( I \) are \((T-d) \times (T-d)\) identical matrices.

The full conditional posterior distributions for \( \rho \) can be factored as:

\[
f(\rho \mid Y_T, V, \{\theta_1, \cdots, \theta_T\}, W) \\
= \frac{f(W, V, \rho \mid \{\theta_1, \cdots, \theta_T\}, Y_T)}{f(W, V \mid \{\theta_1, \cdots, \theta_T\}, Y_T)} \\
= \frac{f(W, V, \rho \mid \{\theta_1, \cdots, \theta_T\}, Y_T)}{f(V \mid \rho, Y_T, \{\theta_1, \cdots, \theta_T\}) \times f(W \mid \{\theta_1, \cdots, \theta_T\})} \\
\propto \exp \left\{ - \frac{1}{2V} \sum_{t=1}^{T} \left[ (y_t - F_t^T \theta_t) - \sum_{j=1}^{d} \rho_j (y_{t-j} - F_{t-j}^T \theta_{t-j}) \right]^2 \right\}
\]

According to the linear model in (4.8), the function can be written as:

\[
f(\rho \mid Y_T, V, \{\theta_1, \cdots, \theta_T\})
\]
\[
\alpha \exp \left\{ -\frac{1}{2V} \sum_{t=1}^{T} \left[ (y_t - F_t^T \theta_t) - \sum_{j=1}^{d} \rho_j (y_{t-j} - F_{t-j}^T \theta_{t-j}) \right]^2 \right\} \\
= \exp \left\{ \frac{1}{2V} \left( Y^* - X^* \rho \right)^T (Y^* - X^* \rho) \right\}
\]

By the ordinal least squared estimation method, the conditional posterior distributions for \( \rho \) is:

\[
\rho \mid (Y_T, V, \{\theta_1, \cdots, \theta_T\}) \sim \mathcal{N}\left((X^*X^*)^{-1}X^*Y^*, V(X^*X^*)^{-1}\right)
\]

where \( Y^* \) and \( X^* \) are defined in equation (4.8)

**Proof of Part (5)** By property of observation and system equations, we obtain \( y_{t+1} = F_{t+1}^T \theta_{t+1} + \nu_{t+1} \) and \( \theta_{t+1} = G_{t+1} \theta_t + \omega_{t+1} \), given conditional posterior distribution for \( \theta_t \mid (\theta_{t-1}, W, V, \rho, Y_t) \sim \mathcal{N}(\mu_t, C_t) \),

\[
E(y_{t+1} \mid Y_t, \theta_t, \cdots, \theta_{t+1-d}, V, W, \rho)
= E(F_{t+1}^T \theta_{t+1} + \nu_{t+1} \mid Y_t, \theta_t, \cdots, \theta_{t+1-d}, V, W, \rho)
= F_{t+1}^T E(G_{t+1} \theta_t) + E\left( \sum_{j=1}^{d} \rho_j (y_{t+1-j} - F_{t+1-j}^T \theta_{t+1-j}) \right) + \gamma_{t+1} \mid Y_t, \theta_t, \cdots, \theta_{t+1-d}, V, W, \rho)
= F_{t+1}^T G_{t+1} \mu_t + \sum_{j=1}^{d} \rho_j (y_{t+1-j} - F_{t+1-j}^T \theta_{t+1-j})
= f_t(1)
\]

\[
\text{Var}(y_{t+1} \mid Y_t, \theta_t, \cdots, \theta_{t+1-d}, V, W, \rho)
\]
\[ \text{Var}(F_{t+1}^T \theta_{t+1} + \nu_{t+1} | Y_t, \theta_t, \cdots, \theta_{t+1-d}, V, W, \rho) \]
\[ = \text{Var}(F_{t+1}^T \theta_{t+1} + \sum_{j=1}^{d} \rho_j (y_{t+1-j} - F_{t+1-j}^T \theta_{t+1-j}) + \gamma_{t+1} | Y_t, \theta_t, \cdots, \theta_{t+1-d}, V, W, \rho) \]
\[ = F_{t+1}^T \text{Var}(G_{t+1} \theta_t + \omega_{t+1}) F_{t+1} + \text{Var}(\gamma_{t+1}) \]
\[ = F_{t+1}^T (G_{t+1} \text{Var}(\theta_t) G_{t+1} + \text{Var}(\omega_{t+1})) F_{t+1} + V \]
\[ = F_{t+1}^T (G_{t+1} \text{C}_{t} G_{t+l}^T + W) F_{t+1} + V \]
\[ = Q_t(1) \]

At point \( t \), one-step forecasting distribution for \( y_{t+k} \) is

\[
y_{t+1} \bigg| (Y_t, \theta_t, \cdots, \theta_{t+1-d}, V, W, \rho) \sim N\left(f_t(1), Q_t(1)\right)
\]

\[
f_t(1) = F_{t+1}^T G_{t+1} \mu_t + \sum_{j=1}^{d} \rho_j (y_{t+1-j} - F_{t+1-j}^T \theta_{t+1-j})
\]

\[
Q_t(1) = F_{t+1}^T (G_{t+1} \text{C}_{t} G_{t+l}^T + W) F_{t+1} + V
\]

### 4.6 Discussion

In this chapter we develop a Bayesian approach to estimation and forecasting nonparametric regression models with an auto-correlated error through a dynamic linear model. Numerous simulation outputs and an application on T-Bills from January 1, 1990 to June 11, 2007 demonstrate that the proposed DLM estimations on correlated error outperform the local linear estimation as data become more dynamic. Thus we decide to apply our methodology on varying coefficient models which is a very important approach to capturing dynamic change in longitudinal observations.
A DLM Approach for Varying Coefficient Models

5.1 Introduction

In Chapter 3 and 4, we were successful in estimation of the regression function in the nonparametric regression models by using the DLM approach. This chapter further advances the DLM procedure for the varying coefficient models, a useful extension of nonparametric regression models and linear regression models. Let $Y$ be a response variable, and $(U, X_1, X_2, \cdots, X_p)$ be a covariate vector. The varying coefficient model is defined as

$$Y = \beta_1(U)X_1 + \cdots + \beta_p(U)X_p + \varepsilon = x^T\beta(U) + \varepsilon,$$ \hspace{1cm} (5.1)

where $\varepsilon$ is a random error with mean zero and finite variance, and $\{\beta_1(U), \cdots, \beta_p(U)\}$ are nonparametric regression coefficient functions. See references in Chapter 2 for a brief literature review on varying coefficient models. In this chapter, we consider both independent
error $\varepsilon_t$ and an AR error process. We represent these two error structures by

$$
\varepsilon_t = \sum_{j=1}^{d} \rho_j \varepsilon_{t-j} + \gamma_t,
$$

where $\gamma_t \sim \mathcal{N}(0, \sigma^2)$ with finite variance. The independent error corresponds to AR error with $d = 0$. In Section 5.2, a connection between varying coefficient model and dynamic linear regression is described. An MCMC sampling algorithm is introduced for the corresponding DLM. The extensive Monte Carlo simulations are conducted to examine the finite sample performance of the proposed procedure in Section 5.3. A real data example is used to illustrate the proposed methodology.

### 5.2 An estimation and forecasting DLM framework for varying coefficient models with correlated errors

Let us assume that $u_1 < \cdots < u_T$. We model $\beta_l(\cdot)$ by a piecewise linear function

$$
\beta_l(u) = \beta_{l,k-1} + \frac{\beta_{l,k} - \beta_{l,k-1}}{u_k - u_{k-1}} (u - u_{k-1})
$$

for $u \in (u_{k-1}, u_k]$.

Suppose that we have a sample $\{u_t, x_t, y_t\}$. Then we consider a penalized least squares method

$$
\sum_{t=1}^{T} \{y_t - x_t^T \beta(u_t)\}^2 + \sum_{l=1}^{p} \lambda_l \sum_{t=2}^{T} \left( \frac{\beta_{l,k} - \beta_{l,k-1}}{u_t - u_{t-1}} \right)^2. \tag{5.2}
$$

The penalty is introduced to avoid over-fitting. Here we allow different coefficient functions to have different regularization parameter $\lambda_l$. 
Denote \( \theta_t = (\beta_{1,t}, \cdots, \beta_{p,t})^T = \beta(u_t) \). Then minimization of the penalized least squares in equation (5.2) is equivalent to considering the following DLM:

\[
y_t = x_t^T \theta_t + \varepsilon_t \\
\theta_t = \theta_{t-1} + \eta_t,
\]

where \( \varepsilon_t \sim N(0, \sigma^2) \), and \( \eta_{lt} \sim N_p(0, \sigma^2/(\lambda \alpha_t^2)) \). In this case \( \alpha_t = u_t - u_{t-1} \), for \( l = 1, \cdots, p \). Moreover, \( \eta_{lt} \)'s are mutually independent. That is \( \eta_t \sim N_p(0, (\sigma^2/\alpha_t^2)E) \), where \( E = \text{diag}\{\lambda_1^{-1}, \cdots, \lambda_p^{-1}\} \). In particular, we set all \( \lambda_i \)'s to be the same, and denoted by \( \lambda \). Then \( E = \lambda^{-1}I_p \).

To build a connection between the varying-coefficient model and a DLM with the auto-correlated error, let us assume that \( F_t = x_t \), and \( G_t = I_p \), the \( p \times p \) identity matrix. Then equation (5.2) can be written as a special case of the general DLM with an AR error

\[
\begin{align*}
\text{Varying coefficient model:} & \quad y_t = F_t^T \theta_t + \nu_t, \\
\text{Varying regression coefficients:} & \quad \theta_t = G_t \theta_{t-1} + \omega_t, \\
\text{AR error:} & \quad \nu_t = \sum_{j=1}^{d} \rho_j \nu_{t-j} + \gamma_t \\
\text{Initial information:} & \quad (\theta_0|D_0) \sim N(\mu_0, C_0).
\end{align*}
\]

where \( \omega_t \sim N(0, W) \) with \( W = (\sigma^2/\alpha_t^2)E \) and \( \gamma_t \sim N(0, V) \). \( \omega_t \) and \( \gamma_t \) are independent, and they are also independent of \((m_0|D_0)\), with \( D_0 \) is the \( \sigma \)-field generated by information at time \( t = 0 \).

Assume that \( \theta_0, W, V \) and \( \rho \) are independent and their prior distributions are given as follows:

\[
\begin{align*}
\text{Prior distribution for } \theta_0 : & \quad \theta_0 \sim N(\mu_0, C_0) \\
\text{Prior Distribution for } W : & \quad W \sim TW(q_0, \Delta_0)
\end{align*}
\]
Prior Distribution for $V$ : $V \sim \mathcal{IG}(\alpha_0, \beta_0)$

Prior Distribution for $\rho$ : $\rho \sim \mathcal{U}(\mathcal{R}^d)$

Here $\mathcal{IW}(q, \Sigma)$ stands for the inverse-Wishart distribution with

$$f(W) = \frac{\mid \Sigma \mid^{\frac{q}{2}} \mid W \mid^{-(q+p+1)/2}}{2^{\frac{q+p+d}{2}} \times \Gamma_p\left(\frac{q}{2}\right)} \times \exp\left\{ -\frac{1}{2} \text{trace}\left( W^{-1} \Sigma \right) \right\},$$

for a positive definite random matrix $W$, and $\Gamma_p(a) = \pi^{p(p-1)/4} \Pi_{j=1}^p \Gamma(a + \frac{1-j}{2})$. $\mathcal{IG}(\alpha, \beta)$ stands for the inverse of Gamma distribution,

$$f(V) = \frac{\beta^\alpha}{\Gamma(\alpha)} \times (V^{-1})^{\alpha+1} \exp(-\beta V^{-1})$$

and $\mathcal{U}(\mathcal{R}^d)$ stands for the uniform distribution over $(\mathcal{R}^d)$.

Bayesian inferences for the observation of $y_t$ and the evaluation of mean function $\theta_t$ can be found in Theorem 1.

### 5.3 The Gibbs sampler algorithm

For the simulation discussed in the next section, we use a “fully Bayesian estimation procedure” which is similar to the sampling algorithm described in chapter 4. The distinctions are the definitions for matrix $F_t$ and $G_t$. $F_t$ is not a $k+1$ vector anymore but equal to $x_t$ which is $(x_{1t}, x_{2t}, \cdots, x_{pt})^T$ vector. $G_t$ here is a $p \times p$ identical matrix instead of a $(k+1) \times (k+1)$ matrix with $G_t = (g_{ij})$

The forecasting algorithm

Sampling regression varying coefficient functions is obtained through the FFBS algorithm.

Let $i = 1, \cdots, I$ be the total number of iterations. Then follow these sampling procedures:

**Step 1.** Give the conditional prior distribution for $\theta_1$ at time period 1: $\theta_1 \mid (\theta_0, W) \sim$
\[ N(\mathbf{a}_1, \mathbf{R}_1), \text{ set } \mathbf{a}_1 = (0, 0, \cdots, 0)_{(p \times 1)}, \mathbf{R}_1 = 0.01 \times \mathbf{I}_{p \times p} \text{ and } \mathbf{F}_t = \mathbf{x}_{p \times 1}. \]

**Step 2.** Pick initial values for \( \{y_0, \mathbf{V}^{(0)}, \mathbf{W}^{(0)}, \mathbf{\rho}^{(0)}\} = \{0, 1, 0.005 \times \mathbf{I}_{p \times p}, 0.1 \times (1, 1, \cdots, 1)_{d \times 1}\}. \)

**Step 3.** At iteration \( i = 1 \), use the FFBS algorithm to sample \( \{\theta_1, \cdots, \theta_t\}^{(1)} : \)

\begin{enumerate}
\item[i] \text{by sampling } \theta_t^{(1)} \sim N(\mu_t, \mathbf{C}_t) \\
\item[ii] \text{for each } n = t - 1, t - 2, \cdots, 1, 0, \text{sample } \{\theta_{t-1}, \theta_{t-2}, \cdots, \theta_1, \theta_0\}^{(1)} \text{ from } \\
\theta_n^{(1)} \mid (\theta_{n+1}^{(1)}, \mathbf{Y}_n, \hat{\mathbf{V}}, \hat{\mathbf{W}}, \mathbf{\rho}^{(0)}) \sim N(h_n, H_n) \\
\text{where } h_t = \mu_t + \mathbf{C}_t \mathbf{G}_t^T \mathbf{R}_{t+1}^{-1} (\theta_{t+1} - \mathbf{a}_{t+1}) \text{ and } H_t = \mathbf{C}_t - \mathbf{C}_t \mathbf{G}_t^T \mathbf{R}_{t+1}^{-1} \mathbf{G}_{t+1} \mathbf{C}_t. \]
\end{enumerate}

**Step 4.** Sample \( \{\mathbf{\rho}\}^{(1)} \) from the conditional posterior distributions:

\[ \rho^{(1)} \bigg| (\mathbf{Y}_n, \hat{\mathbf{V}}, \{\theta_1, \cdots, \theta_t\}^{(1)}) \sim N \left( (X^T X^*)^{-1} X^* Y^*, V(X^T X^*)^{-1} \right) \]

where \( Y^* \) and \( X^* \) are defined in equation .

**Step 5.** Set \( i = i + 1 \), repeat step 3 and step 4 for \( i = I \) iterations.

The sampling algorithm works as a loop, repetitively drawing parameters to read data values and create observations one by one.

**The forecasting algorithm**

The forecasting distribution for \( y_{t+1} \) is given by

\[ y_{t+1} \bigg| (\mathbf{Y}_t, \{\theta_t, \cdots, \theta_{t+1-d}\}, \hat{\mathbf{V}}, \hat{\mathbf{W}}, \mathbf{\rho})^{(1)} \sim N(f_t(1), Q_t(1)) \]

where \( f_t(1) = \mathbf{F}_t^T \mathbf{a}_t + \sum_{j=1}^d \mathbf{\rho}_j (y_{t-j} - \mathbf{F}_t^T \theta_{t-j}) \) and \( Q_t(1) = \mathbf{F}_t^T \mathbf{R}_t \mathbf{F}_t + V.\)
5.4 Simulations and an application

In this section, we investigate the performance of the proposed DLM estimation procedure by a Monte Carlo simulation study. We further illustrate the proposed procedure by an analysis of air pollutant data.

5.4.1 Simulation study

Table 5.1. Computation time for 500 simulations (in hours)

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Fixed</th>
<th>Random</th>
<th>Fixed</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>43</td>
<td>41</td>
<td>37</td>
<td>34</td>
</tr>
<tr>
<td>1000</td>
<td>80</td>
<td>78</td>
<td>72</td>
<td>66</td>
</tr>
</tbody>
</table>

In our simulation, data are generated from the following varying coefficient model:

\[ y_t = 2u_t^2 + 2 \cos(2\pi u_t)x_{1t} + 3 \sin(\pi u_t)x_{2t} + \varepsilon_t, \]

where \( \varepsilon_t \) is an AR error, \( \varepsilon_t = \rho \varepsilon_{t-1} + \eta \), with \( \rho = 0, 0.4 \) and \( \eta \sim \mathcal{N}(0, \sigma_n^2) \) with \( \sigma_n^2 = 0.1 \). When \( \rho = 0 \), the error is independent. We include two scenarios for covariate \( u_t \): the uniform random variables over \([0, 1]\) and evenly spaced on \([0, 1]\) corresponding to random and fixed designs. Both \( x_{1t} \) and \( x_{2t} \) follow standard normal distributions. We generate 500 simulated data sets with sample size \( T = 500 \) and \( T = 1000 \), respectively. Each chain runs for a total of 10,000 iterations and we set the first 2000 as burn-in samples. In order to illustrate the length of time required to do the simulations, we gave computing times in Table 5.1. Average times spent on models with correlated errors are roughly 20% longer than the time required for models with normal errors. Simulations of models with normal error are much faster.
Table 5.2 summarizes the mean squared errors (MSE) of the estimated coefficients over the 500 simulations. From Table 5.2, it can be seen that the DLM procedures perform quite well. It is interesting to note that the performance of the proposed procedures for random designs is similar to that for fixed designs. For the intercept function, the MSE for AR(1) seems to be much larger than that for independent error. This may be due to the variance of AR error $\varepsilon_t$ which is equal to $\sigma_n^2/(1 - \rho^2)$, being larger than the variance of the independent error which is equal to $\sigma_n^2$.

Table 5.2. Mean Squared Errors

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Coefficient</th>
<th>Fixed Design</th>
<th>Random Design</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AR(1)</td>
<td>AR(0)</td>
<td>AR(1)</td>
</tr>
<tr>
<td>500</td>
<td>$\beta_1(\cdot)$</td>
<td>0.0063</td>
<td>0.0031</td>
</tr>
<tr>
<td></td>
<td>$\beta_2(\cdot)$</td>
<td>0.0141</td>
<td>0.0142</td>
</tr>
<tr>
<td></td>
<td>$\beta_3(\cdot)$</td>
<td>0.0041</td>
<td>0.0044</td>
</tr>
<tr>
<td>1000</td>
<td>$\beta_1(\cdot)$</td>
<td>0.0029</td>
<td>0.0020</td>
</tr>
<tr>
<td></td>
<td>$\beta_2(\cdot)$</td>
<td>0.0053</td>
<td>0.0051</td>
</tr>
<tr>
<td></td>
<td>$\beta_3(\cdot)$</td>
<td>0.0026</td>
<td>0.0024</td>
</tr>
</tbody>
</table>

Figures 5.1 to 5.4 illustrate typical estimated varying coefficient functions along with their 95% HPD credit set. All estimated coefficients follow the trend of true coefficient functions and fall into the 95% HPD credit set.
Figure 5.1. Plot of typical estimated coefficients for a fixed design with sample size 500 showing (a) a model with AR(1) error and $\rho = 0.4$ and (b) a model with normal error $\sigma^2 = 0.1$. Solid curves are estimated coefficient trajectories, dashed curves are the true coefficient trends, and dash-dotted curves are the 95% HPD credit set.
Figure 5.2. Plot of typical estimated coefficients for a random design with sample size 500 showing (a) a model with AR(1) error and $\rho = 0.4$ and (b) a model with normal error $\sigma^2 = 0.1$. Solid curves are estimated coefficient trajectories, dashed curves are the true coefficient trends, and dash-dotted curves are the 95% HPD credit set.
Figure 5.3. Plot of typical estimated coefficients for a fixed design with sample size 1000 showing (a) a model with AR(1) error and $\rho = 0.4$ and (b) a model with normal error $\sigma^2 = 0.1$. Solid curves are estimated coefficient trajectories, dashed curves are the true coefficient trends, and dash-dotted curves are the 95% HPD credit set.
Figure 5.4. Plot of typical estimated coefficients for a random design with sample size 1000 showing (a) a model with AR(1) error and $\rho = 0.4$ and (b) a model with normal error $\sigma^2 = 0.1$. Solid curves are estimated coefficient trajectories, dashed curves are the true coefficient trends, and dash-dotted curves are the 95% HPD credit set.
Figure 5.5. One-step forecasting plot for 500 \( \hat{y}_{496} \)'s, sample size = 500 showing (a) a fixed design with AR(1) model error and \( \rho = 0.4 \); (b) a fixed design with normal model error \( \sigma^2 = 0.1 \); (c) a random design with AR(1) model error and \( \rho = 0.4 \); and (d) a random design with normal model error \( \sigma^2 = 0.1 \).

5.4.2 Environmental health data set

In this section, we illustrate the proposed estimation procedure by an analysis of a real data set. In this example, of interest is the relationship between levels of pollution in the environment and cardiovascular and respiratory hospital visits among people in the area. Many air quality factors affect human circulatory and respiratory system function, and increased pollution levels can cause a physiological stress that leads to adverse health events in susceptible individuals (Thaw Sint & Ghio, 2008). The relationship between ambient
pollution measures and rate of hospital visits may vary across days, weeks or seasons, perhaps due to differences in human behavior or changes in other factors influencing susceptibility to respiratory and cardiovascular events. Here we directly test the association between the level of the pollutants and the number of daily total hospital admissions for circulatory and respiratory problems while concurrently examining how the extent to which the association varies over time. The data set was collected in Hong Kong from January 1, 1994 to December 31, 1995 (courtesy of Professor T.S. Lau). The measure of health outcomes are the number of
hospitalizations registered for combined respiratory and cardiovascular events. The measures of levels of pollutants were sulfur dioxide \( X_2 \) (in \( \mu g/m^3 \)), nitrogen dioxide \( X_3 \) (in \( \mu g/m^3 \)) and dust \( X_4 \) (in \( \mu g/m^3 \)). Figure 5.7 contains daily measurements of pollutants and these three environmental factors. Correlation between nitrogen and dust is the highest at \( r = 0.7820 \) and the correlation of pollutants sulfur dioxide and nitrogen dioxide is 0.4025. The correlation between pollutants sulfur dioxide and dust is the lowest, at 0.2810. All three variables are significantly correlated.

Figure 5.7. Plots for daily environmental data in Hong Kong, Jan. 1, 1994 - Dec. 31, 1995 showing (a) the total number of daily hospital admission for circulatory and respiratory problems, (b) the average level of sulfur dioxide (in \( \mu g/m^3 \)), (c) the average level of nitrogen dioxide (in \( \mu g/m^3 \)), and (d) the average levels of dust (in \( \mu g/m^3 \))
Take $u$ to be the day of data collection and the response variable $y$ to be the logarithm of total hospital admissions, and consider a varying coefficient model for the data:

$$Y_t = \beta_1(u) + \beta_2(u)x_2 + \beta_3(u)x_3 + \beta_4(u)x_4 + \varepsilon_t.$$ 

To determine the AR order of error process $\varepsilon_t$, we first fit the data by the varying coefficient models pretending the error is independent. This provides an estimate of the varying coefficient functions for the 4 regression functions. We then find the fitted value for $\hat{y}_t$, and calculate residuals by $\hat{\varepsilon}_t = y_t - \hat{y}_t$. Figure 5.8 shows the autocorrelation plot and partial autocorrelation plot for the error $\nu_t$ and it indicates that the AR(1) model for $\varepsilon_t$ is appropriate for the model.

The size of Gibbs sampling is 10,000. Figure 5.9 shows changes of varying coefficient functions over all time periods and describes the extent to which the coefficients vary with time. The figure shows that there is a strong time effect on the coefficient functions.

Figure 5.10 shows one-step forecasting and observed values side by side across 230 time points for the total number of daily hospital admissions for circulatory and respiratory problems. We use first 500 data points, from Jan. 1, 1994 to May 31, 1995, to forecast the next 230 response values, June 1, 1995 - Dec. 31, 1995. For example, we use \{$y_1, \cdots, y_{500}$\} to forecast $\tilde{y}_{501}$, use \{$y_1, \cdots, y_{501}$\} to forecast $\tilde{y}_{502}$, $\cdots$, \{$y_1, \cdots, y_{t-1}$\} to forecast $\tilde{y}_t$ and \{$y_1, \cdots, y_{729}$\} to forecast $\tilde{y}_{730}$. It demonstrates that the proposed forecasting method accurately predicts future trends for the total number of daily hospital admission for circulatory and respiratory problems.
Figure 5.8. Autocorrelation plot and partial autocorrelation plot for varying coefficient model error showing (a) the autocorrelation plot and (b) the partial autocorrelation plot for the total number of daily hospital admission for circulatory and respiratory problems.
Figure 5.9. Estimated coefficient function plot with the 95% HPD credit set
Figure 5.10. One-step forecasting for the total number of daily hospital admissions for circulatory and respiratory problems from June 1, 1991 - Dec. 31, 1995. The blue curve is the observed value and the red curve is the one-step forecasting value.
Conclusion

In this dissertation, we proposed an extension of the DLM (West & Harrison, 1997) approach from an independent error to an auto-correlated error through connections between the DLM and a nonparametric regression model. This made it possible to conduct analysis of nonparametric regression models and varying coefficients models with auto-correlated error data for both estimation and forecast. The proposed methods are designed to handle response variables with auto-correlated error.

6.1 Research contributions

The main contributions of our research are:

1. After connecting DLM and nonparametric regression models with an independent error, we develop a method to determine values for model error in equation (3.1) in the Gibbs sampler through the general cross validation criteria. We compare the performance of estimations between the proposed DLM approach and local estimation, a commonly used method for nonparametric regression models. The advantages of our method are that it brings more accurate estimations than local linear linear estimations as high-
frequency signals of data become stronger and it ameliorates the computational burden of the DLM method used by West and Harrison (1997).

2. Based on analysis from nonparametric regression models with independent error, we continue to explore estimation and forecasting performance for the models with autocorrelated error, e.g. AR(1), and compare this to local linear estimation. Again, our methods are built based on DLM. The simulation examples show that the proposed method outperforms local linear method since the proposed method takes into account correlated errors from the data and end up with a much smaller MSE than local linear estimation. We use a real data example to compare the one-step forecasting performance between our proposed method and DLM and demonstrate that our method brings improved forecasting accuracy since forecasting dots fall into a straight line in $1 - 1$ observation and forecasting in Figure 4.10.

3. We further develop a connection between varying coefficient models and a dynamic linear model with an autoregressive error process and compare the performance with the model with i.i.d. errors. The accuracy of the proposed method is examined via simulations which include both estimation and forecasting performance, see Figure 5.5 and Figure 5.5. It performs well for larger sample sizes. As a practical example, we apply the method to study the association between the weekly measurements of pollutants and other environmental factors, collected in Hong Kong from January 1, 1994 to December 31, 1995. See the estimation and forecasting plot in Figure 5.9 and Figure 5.10.

6.2 Future research interests

In the course of this research, we identified several areas of potential development that were beyond the scope of this dissertation. The most promising are listed below.
1. Here we use Gibbs sampler algorithms for all simulations and applications. The Gibbs sampler is the simplest of the Markov Chain simulation algorithms, and it is the first choice for conditionally conjugate models, where we can directly sample from each conditional posterior distribution. However when the varying coefficient model involves AR(d) error $\nu_t$, where $\nu_t = \sum_{j=1}^{d} \rho_j \nu_{t-j} + \gamma_t$ with $\gamma_t \sim \mathcal{N}(0, V)$, we need to use prior distributions for the $d$ initial values $\nu_{t-j}, \nu_{t-j}, \cdots, \nu_{t-j}$ of AR process to derive corresponding posterior distributions. This could be challenging since distribution for $y_t$ which could theoretically depend on parameters $\rho$ and $V$. In such a case, future work should consider using simulation steps other than the direct Gibbs sampler developed here. The Metropolis-Hastings algorithm does not require the conditional posterior distributions. Therefore, one possible research direction is to use Metropolis-Hastings algorithms in the models.

2. Interests in varying coefficient model have increased along with the popularity of longitudinal data modelling. Longitudinal data are repeated measurements collected on a subject through time. The interest is to study the associations between the covariates and the response variables. The subjects are assumed to be independent of each other and the measurements within each subject are assumed to be correlated. Numerous data sets for application of these models are available in biomedical, epidemiological, economic, and many other fields. In our research, we extend the proposed methods to varying coefficient models. The varying coefficient model is a popular method to fit the longitudinal data. It would be meaningful to see performance of our proposed methods for longitudinal data analysis.

3. For some varying coefficient model (5.1), the coefficients $\lambda_i, i \leq 1, \cdots, p_1$ do not depend
on $U$ and they could be constants, thus we can rewrite the model (5.1) as:

\[
Y = \lambda_1X_{11} + \lambda_2X_{12} + \cdots + \lambda_{p_1}X_{1p_1} + \beta_1(U)X_{21} + \beta_2(U)X_{22} + \cdots + \beta_{p_2}(U)X_{2p_2} + \varepsilon
\]

\[
= x_{p_1}^T\lambda + x_{p_2}^T\beta(U) + \varepsilon,
\]

where $(x_{p_1}^T, x_{p_2}^T) = X$ and $(x_{p_1}^T, x_{p_2}^T)$ are $p_1$ and $p_2$ dimensional covariates, and $p_1 + p_2 = p$. $x_{p_1}^T$ are U-independent effects and $x_{p_2}^T$ consist of U-dependent effects. The error $\varepsilon$ is an i.i.d. random variable. This is the semi-varying coefficient model. Zhang, Lee, and Song (2002) proposed a two-step estimation procedure and proved that the $\lambda$ estimator converges at a rate of $O_p(n)^{-1/2}$. Fan and Huang (2005) proposed a generalized likelihood ratio to test whether the coefficients $\lambda$ are independent from $U$. Since our researches already cover estimation and forecasting methodologies for varying coefficient models, we also can apply them on the semi-varying coefficient model to investigate variation of $\lambda$. 
Appendix
Appendix A

Dynamic Linear Models

For self-consistency, this appendix contains some introduction and sampling algorithm for dynamic linear model. The material in this section is adapted from Chapter 4 of West and Harrison (1997).

Consider the external information at times $t \geq 1$, so that given initial prior information $D_0$ at $t = 0$, at any future time the available information set is simply

$$D_t = \{Y_t, D_{t-1}\}$$

$Y_t$ is the observed value of the series at time $t$.

**Definition** For each $t$, the general univariate LDM, is characterized by a quadruple $\{F_t, G_t, V_t, W_t\}$ and is defined by

- **Observation equation:** $Y_t = F_t^T \theta_t + \nu_t$, $\nu_t \sim N(0, V_t)$
- **System equation:** $\theta_t = G_t \theta_{t-1} + \omega_t$, $\omega_t \sim N(0, W_t)$
- **Initial information:** $(\theta_0|D_0) \sim N(\xi_0, C_0)$

for some prior moments $\xi_0$ and $C_0$. The observational and evolution error sequences are assumed to be internally and mutually independent, and are independent of $(\theta_0|D_0)$. 
Theorem A.1 (Theorem 4.1 of West and Harrison (1997)) One-step forecasting and level posterior distributions for any time \( t > 0 \) can be obtained sequentially as follows;

(a) Posterior at \( t - 1 \):
   
   For some mean \( \xi_{t-1} \) and variance matrix \( C_{t-1} \)
   
   \( (\theta_{t-1}|D_{t-1}) \sim N(\xi_{t-1}, C_{t-1}) \)

(b) Prior at \( t \):

   \( (\theta_{t}|D_{t-1}) \sim N(a_t, R_t) \)

   where

   \[ a_t = G_t \xi_{t-1} \quad \text{and} \quad R_t = G_t C_{t-1} G_t^T + W_t \]

(c) One-step forecast:

   \( (Y_{t}|D_{t-1}) \sim N(f_t, Q_t) \)

   where

   \[ f_t = F_t^T a_t \quad \text{and} \quad Q_t = F_t^T R_t F_t + V_t \]

(d) Posterior at \( t \):

   \( (\theta_{t}|D_{t}) \sim N(m_t, C_t) \)

   with

   \[ m_t = a_t + A_t e_t \quad \text{and} \quad C_t = R_t - A_t Q_t A_t^T \]

   where

   \[ A_t = R_t F_t Q_t^{-1} \quad \text{and} \quad e_t = Y_t - f_t. \]

□

Proof can be found in West and Harrison (1997). Besides forecasting and focus study in the future, researchers are often interested in looking back over time in order to get a clearer picture of what happened in order to improve model interpretability and help with future decision making. The process to use recent data to revise inferences about previous state
vector values, the information being filtered back to previous time points, is called filtering. The distribution of \( \theta_{t-1} \mid D_t \) for \( k \geq 1 \) is called the \( k \)-step filtered distribution for the state vector at time \( t \). Another concept is that of smoothing a time series. The retrospective estimation of the historical development of a time series mean response function using the filtered distributions for \( k \geq 1 \) is called smoothing the series.

**Theorem A.2** (Theorem 4.4 of West and Harrison (1997)) *In the univariate DLM, for all \( t \), define*

\[
B_t = C_t G_{t+1}^T R_t
\]

*Then for all \( k, (1 \leq k \leq t) \), the filtered marginal distribution are*

\[
(\theta_{t-k}|D_t) \sim N(a_t(-k), R_t(-k))
\]

*where*

\[
a_t(-k) = \xi_{t-k} + B_{t-k}(a_{-k+1} - a_{t-k+1})
\]

*and*

\[
R_t(-k) = C_{t-k} + B_{t-k}(R_{-k+1} - R_{t-k+1})B_{t-k}^T
\]

*with starting values*

\[
a_t(0) = \xi_t \quad \text{and} \quad R_t(0) = C_t
\]

*and where, as usual*

\[
a_{t-k}(1) = a_{t-k+1} \quad \text{and} \quad R_{t-k}(1) = R_{t-k+1}
\]
Assume $F_t$ and $G_t$ are known and $V_t = V$ and $W_t = W$. They discussed the sampling algorithms in Chapter 15.2. The algorithm samples the state parameters $\theta_t$ jointly. $V$ and $W$ are respectively sampled from inverse Gamma and inverse Wishart conditional posterior distribution. Carter and Kohn (1994) and Frühwirth-Schnatter (1994) noted conditional independence between Markovian structure of the state parameters and the latter article proposed an efficient MCMC simulation method in normal DLM named as **Forward filtering, backward sampling** (FFBS). The sampling procedures are

**Step 1.** Sampling state parameters $\theta_t$ by FFBS method:

1. Sample $\theta_T$ from $(\theta_T|D_T) \sim N(\xi_T, C_T)$ where $\xi_T$ and $C_T$ can be find from Theorem 4.1. They are routinely obtained via the Kalman filter (Jazwinski, 1970);

2. For each $t = T - 1, T - 2, \ldots, 1, 0$, sampling $\theta_t$ from $p(\theta_t|\theta_{t+1}, D_t) \sim N(h_t, H_t)$ where,

   $$h_t = \xi_t + B_t(\theta_{t+1} - a_{t+1}) \quad \text{and} \quad H_t = C_t - B_tR_{t+1}B^T_t$$

   with

   $$B_t = C_tG^T_tR_t^{-1}$$

   This filtered marginal distribution is explained in Theorem 4.5

   Thus FFBS is performed sequentially starting by sampling $\theta_T$ from $N(\xi_T, C_T)$ and recursively working backward one-step at a time by sampling each $\theta_t$ from filtered marginal distribution in Theorem 4.5.

**Step 2.** Sampling $\{V, W\}$

Assuming that prior for $V$ and $W$ are respectively inverse Gamma and inverse Wishart,
their full conditional posterior distributions are also be inverse Gamma and inverse Wishart distribution. The parameters are updated from \((n_V, n_V S_V) \& (n_W, n_W S_W)\) to \((n_V + T, n_V S_V + \sum_{t=1}^{T} (Y_t - F_t^T \theta_t)^2) \& (n_W + T - 1, n_W S_W + \sum_{t=1}^{T} (\theta_t - G_t \theta_{t-1})^T (\theta_t - G_t \theta_{t-1}))\)
Appendix B

Forward Filtering and Backward Sampling

This section raises attention to the technical issues of sampling \( \{\theta_1, \ldots, \theta_n\} \) from a sequence of full multivariate normal posterior conditional distribution \( p(\{\theta_1, \ldots, \theta_n\} \mid D_n) \).

Carter and Kohn (1994) and Frühwirth-Schnatter (1994) independently proposed an iterative simulation algorithm based on Gibbs sampling with the name of *forward filtering and backward sampling*. FFBS procedure applies to general linear Gaussian state-space models. The structure includes two parts.

**Forward Filtering:** Sampling \( \theta_n \sim N(m_n, C_n) \).

starts by running the standard *forward* updating, or *filtering*, from \( t = 0 \) to \( t = n \), saving all concomitant quantities, \( m_t, C_t, a_t, R_t, B_t \) at each stage. Thus at \( t = n \), the parameters for joint multivariate normal distribution of \( \theta_n \) are derived, \( \theta_n \sim N(m_n, C_n) \).

**Backward Sampling:** At each time period \( t \) from \( n - 1 \) to \( 0 \), applying Gibbs sampling directly and apparently sampling \( \theta_t \) from \( (\theta_n \mid \theta_n, D_t) \sim N(h_t, H_t) \) state by state, where \( \theta_{t+1} \) is the value just sampled.
The conditional distribution here $p(\theta_n | \theta_n, D_t)$ can be obtained from the filtering marginal distribution Theorem A.2.:

$$(\theta_t | \theta_{t+1}, D_t) \sim N(h_t, H_t) \quad \text{(A.1)}$$

where $h_t = m_t + B_t(\theta_{t+1} - a_{t+1})$, and $H_t = C_t - B_t B_t' B_t' + B_t C_t G_{t+1} R_{t+1}^{-1}$. Sequence backward through time. Computer $h_t$ and $H_t$ at each time $t$ and sampling from $N(h_t, H_t)$.

Thus, forward filtering and backward sampling is achieved.
References


Vita
Yuejiao Jiang

EDUCATION

08/09  PhD Candidate in Statistics (GPA: 3.82)
       The Pennsylvania State University, University Park, PA

08/03  M.S. in Statistics (GPA: 3.82)
       The Pennsylvania State University, University Park, PA

05/01  MBA in Finance (GPA: 3.80)
       Eastern Kentucky University, Richmond, KY

WORK EXPERIENCE

02/09 - present  Senior quantitative analyst
       The Hartford, 4 Walnut Grove Drive, Horsham PA 19044
       ✓ Perform business and statistical analysis using SAS for customer satisfaction
         surveys and AARP membership markets, including factor analysis, ANOVA,
         chi-square test, and so on
       ✓ Create summary reports and presentations to share internally

08/07 – 01/09  Statistical Consultant
       Statistical Consulting Center, PSU, University Park, PA
       ✓ Determine statistical issues, define data requirements, identify statistical models and develop program codes, mainly SAS and Minitab
       ✓ Complete statistical analysis reports, including explanations and recommendations; make connections between project questions and statistical solutions
       ✓ Hold weekly office hours to solve clients’ emerging statistical questions by phone or face to face; questions cover experimental design and statistical inference analysis

06/03 – 06/07  Statistics Courses Instructor/Assistant
       Department of Statistics, PSU, University Park, PA;
       ✓ Taught Experimental Design (Stat 401) and Elementary Probability (Stat 318) for junior and senior students and led laboratory sections. The courses cover probability concepts and statistical inferences, etc.
       ✓ Assisted students on homework; designed assignments, exams and provided solutions; administered and determined the final grades

LANGUAGES

English and Chinese