ABSTRACT

WANG, JIANGDIAN. Shape Restricted Nonparametric Regression with Bernstein Polynomials.
(Under the direction of Sujit K. Ghosh.)

There has been increasing interest in estimating a multivariate regression function subject to various shape restrictions, such as nonnegativity, isotonicity, convexity and concavity among many others. The estimation of such shape-restricted regression curves is more challenging for multivariate predictors, especially for functions with compact support. Most of the currently available statistical estimation methods for shape restricted regression functions are generally computationally very intensive. Some of the existing methods have perceptible boundary biases.

This thesis considers a suitable class of univariate and multivariate Bernstein polynomials and proposes sieved estimators obtained from a nested sequence of shape-restricted multivariate Bernstein polynomials. The proposed nonparametric estimators are shown to be: (i) the regression function estimate is shown to be the solution of a quadratic programming problem; making it computationally attractive (ii) the nonparametric estimator is shown to be universally consistent under some mild regularity conditions and (iii) the estimation methodology is flexible in the sense that it can be easily adapted to accommodate many popular multivariate shape restrictions. Numerical results derived from simulated data sets and real data analysis are used to illustrate the superior performance of the proposed estimators compared to a few other existing estimators in terms of various goodness of fit metrics.

The inference is derived through the bootstrap method and the bayesian method, as the standard methods to derive asymptotic distributions may not be applicable directly to the estimator. In addition, we extend our ordinary estimator obtained from shape-restricted Bernstein polynomials and apply it in the linear mixed effects model framework. The work is motivated by a longitudinal data, and the proposed estimator is obtained through iterative methods. We show the effectiveness of our algorithm through both simulations and the real data study. Another extension is to fit the regression function under the assumption of heteroscedasticity by

modifying the ordinary algorithm to an iterative estimation method. The proposed estimator is applied in the analysis of several simulation data and real data examples.

© Copyright 2011 by Jiangdian Wang

All Rights Reserved

Shape Restricted Nonparametric Regression with Bernstein Polynomials

by
Jiangdian Wang

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2012

APPROVED BY:

Barry Goodwin

Huixia (Judy) Wang

Brian Reich

Sujit K. Ghosh
Chair of Advisory Committee

DEDICATION

To my father, my husband
and in memory of my mother

BIOGRAPHY

Jiangdian Wang was born on February 23rd, 1981 in Fuzhou, China. She got her Bachelor of Science degree in Electronic and Computer Engineering from Zhejiang University, Hangzhou, China in July 2003. After that, she worked as a research assistant in Singapore. To seek a graduate education, she came to the United States. She got her Master degree in Statistics at North Carolina State University (NCSU) in August 2009 and she has been to pursue a Ph.D. degree in Statistics since then. Her research focuses on the methodology of nonparametric regression with shape restrictions and its applications.

ACKNOWLEDGEMENTS

I would like to express my deep and sincere appreciation to my advisor Professor Sujit K. Ghosh for his patience, motivation, enthusiasm, and immense knowledge. His great guidance, encouragement and support helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank my committee members: Professor Judy Wang, Professor Brian Reich, and Professor Barry Goodwin for their valuable helps and insight comments.

My appreciations also go to Dr. Pam Arroya, Professor Leonard Stefanski, Professor John Monahan and Professor Sujit K. Ghosh. They are brilliant Directors of Graduate Programs (GDPs). Dr. Arroya and Professor Stefanski approved my application to transfer from the department of ECE, and offered me the Teaching Assistant scholarship. Without their helps, I would still be a suffering student there. Dr. Arroya gave me incredible encouragements and supports when she was in our department. Professor Monahan and Professor Ghosh are our current GDPs. They work very hard to provide more and more opportunities to students (including me), encourage us to attend several academic conferences, and solve my financial support issue.

I am lucky to receive the support and encouragement from all my friends in the department. We have had many great moments together and I would like to thank them for sharing the happy school life with me in the last few years.

Last but not the least I would like to thank my parents and my husband for all their endless love and support. They always have confidence in me and always encourage me to overcome any difficulties I faced during my Ph.D. study.

TABLE OF CONTENTS

List of Tables	viii
List of Figures	x
Chapter 1 Introduction to Shape Restricted Regression	1
1.1 Shape Restricted Regression with a Single Predictor Variable	2
1.2 Shape Restricted Regression with Multiple Predictor Variables	6
1.3 Bayesian Models for Shape-Restricted Regression	8
1.4 Bernstein Polynomials	9
1.4.1 Univariate Bernstein Polynomials	9
1.4.2 Multivariate Bernstein Polynomials	13
1.5 Organization of The Thesis	14
Chapter 2 Shape Restricted Nonparametric Regression with Univariate Bernstein Polynomials	15
2.1 Introduction	15
2.2 Restricted Bernstein Polynomial Estimator	19
2.2.1 Examples of Constraint Bernstein Polynomial Sieves	21
2.2.2 Computation of the Sieved Estimator	23
2.3 Asymptotic Properties	25
2.4 Simulation Studies	29
2.4.1 Monotone Regression	29
2.4.2 Increase-concave Regression	32
2.5 Real Data Examples	34
2.5.1 Fuel Efficiency Study using ASA Car Data	34
2.5.2 Growth Curve Study using Rabbit Data	35
2.6 Conclusion	36
Chapter 3 Shape Restricted Nonparametric Regression with Multivariate Bernstein Polynomials	43
3.1 Introduction	43
3.2 Modelling Shape Restrictions Using Multivariate Bernstein Polynomials	47
3.2.1 Constraint Bernstein Polynomial Sieve ($d = 2$)	49
3.2.2 Computational Details	54
3.3 Asymptotic Properties	56
3.4 Simulation Studies	60
3.4.1 Data Generation and Computational Details	61
3.4.2 Results	62
3.5 Real Data Application	66
3.6 Discussion	68

Chapter 4 Inference using Bootstrap and Bayesian Methods	84
4.1 Introduction	84
4.2 Bootstrap Method	85
4.2.1 Bootstrap Distribution of Estimator $\hat{m}_N(\mathbf{x})$	86
4.2.2 Real Data Example	87
4.3 Bayesian Method	89
4.3.1 Bayesian Approach to Shape Restricted Regression	90
4.3.2 Simulation Studies	97
4.3.3 Real Data Application	103
4.4 Conclusion	104
Chapter 5 Mixed Effects Model Subject to Shape Restrictions	109
5.1 Introduction	109
5.2 Models and Algorithms	111
5.2.1 Linear Mixed Models	114
5.2.2 Iterative Method for Linear Mixed Model with Heteroscedastic Errors	115
5.2.3 Extension to Shape Restricted Mixed Regression Model	118
5.3 A Simulation Study	120
5.4 Real Data Application	121
5.5 Discussion	126
Chapter 6 Shape Restricted Nonparametric Regression with Heteroscedastic Errors	130
6.1 Introduction	130
6.2 Models and Algorithms	132
6.2.1 Estimator	133
6.2.2 Computation of the Estimator	134
6.3 Simulations	137
6.4 Infant Mortality Rate Data Example	139
6.5 Conclusion	140
Chapter 7 Discussions and Future Work	143
7.1 Bayesian Model with Varying Dimensions	144
7.2 Testing for Shape Restrictions	145
References	147
Appendices	153
Appendix A Detailed Proofs of Asymptotic Properties	154
A.1 Proof of Property 3.3.1	154
A.2 Proof of Property 3.3.2	157
A.3 Proof of Lemma 3.3.1	158
A.4 Proof of Lemma 3.3.2	159
A.5 Proof of Theorem 3.3.1	164

Appendix B	Display of Variance Estimation in Section 5.4	170
B.1	Estimated Variance of Subject Level Measurement Errors	170
B.2	Estimated Variance of Random Effects	171

LIST OF TABLES

Table 2.1	Average integrated mean squared errors (IMSE) over 1000 repetitions. The sample size is $n = 50$ and all the noise distributions are centered and scaled to have mean zero and standard deviation $\sigma = 0.4$. Standard error of IMSE is displayed in the parentheses.	37
Table 2.2	Average integrated mean squared prediction errors (IMSPE) over 1000 repetitions. The sample size is $n = 50$ and all the noise distributions are centered and scaled to have mean zero standard deviation $\sigma = 0.4$. Standard error of IMSPE is displayed in the parentheses. P-values are obtained based on two-sample t-tests.	38
Table 2.3	Average integrated mean squared errors (IMSE) over 1000 repetitions. The sample size is $n = 50$ and the noise has standard deviation $\sigma = 0.5$ and 0.1 . Standard error of IMSE is displayed in the parentheses. P-values are obtained by using two-sample t-tests between the BP method and the rest methods.	38
Table 3.1	RMISE and MIAE ($\times 100$) for $m_1(x_1, x_2) = \min(x_1, x_2)$ based on 500 Monte Carlo replications as global measurements of performance. <i>Note</i> : Standard error is displayed in the parentheses. BP stands for the proposed Bernstein polynomial based estimate. MP stands for the monoProc estimate.	71
Table 3.2	RMISE and MIAE ($\times 100$) for $m_2(x_1, x_2) = x_1 x_2 I(x_1 \leq 0.5, x_2 \leq 0.5) + \{(x_1 - 0.5)(x_2 - 0.5) + x_1 x_2\} I(x_1 > 0.5, x_2 > 0.5)$ based on 500 Monte Carlo replications as global measurements of performance. <i>Note</i> : Standard error is displayed in the parentheses. BP stands for the proposed Bernstein polynomial based estimate. MP stands for the monoProc estimate.	71
Table 3.3	RMISE and MIAE ($\times 100$) for $m_3(x_1, x_2) = x_1^{\frac{1}{3}} x_2^{\frac{2}{3}}$ based on 500 Monte Carlo replications as global measurements of performance. <i>Note</i> : Standard error is displayed in the parentheses. BP stands for the proposed Bernstein polynomial based estimate. MP stands for the monoProc estimate.	72
Table 3.4	Prediction biases for $m_1(x_1, x_2) = \min(x_1, x_2)$ based on 500 Monte Carlo replications as local measurement of performance. Column 3 and 4: Mean bias (S.E. of bias); Column 5 and 6: $[2.5^{th} \text{ percentile}, 97.5^{th} \text{ percentile}]$ of the biases.	72
Table 3.5	Prediction biases for $m_2(x_1, x_2) = x_1 x_2 I(x_1 \leq 0.5, x_2 \leq 0.5) + \{(x_1 - 0.5)(x_2 - 0.5) + x_1 x_2\} I(x_1 > 0.5, x_2 > 0.5)$ based on 500 Monte Carlo replications as local measurement of performance. Column 3 and 4: Mean bias (S.E. of bias); Column 5 and 6: $[2.5^{th} \text{ percentile}, 97.5^{th} \text{ percentile}]$ of the biases.	72

Table 3.6	Prediction biases for $m_3(x_1, x_2) = x_1^{1/3} x_2^{2/3}$ based on 500 Monte Carlo replications as local measurement of performance. Column 3 and 4: Mean bias (S.E. of bias); Column 5 and 6: $[2.5^{th}$ percentile, 97.5^{th} percentile] of the biases.	73
Table 3.7	Pearson correlations with Kendall's Tau in parenthesis among three variables.	73
Table 4.1	Simulation study for $m_1(x) = \sin(\frac{\pi}{2}x)$. The results are averaged over 200 Monte Carlo replications. Monte Carlo standard error is displayed in the parentheses.	102
Table 4.2	Simulation study for $m_2(x) = 2x \cdot I(x \in [0, 0.25]) + 0.5 \cdot I(x \in [0.25, 0.75]) + (2x - 1) \cdot I(x \in [0.75, 1])$. The results are averaged over 200 Monte Carlo replications. Monte Carlo standard error is displayed in the parentheses.	102
Table 4.3	Simulation study for $m_3(x) = (16/9)(x - 1/4)^2$. The results are averaged over 200 Monte Carlo replications. Monte Carlo standard error is displayed in the parentheses.	103
Table 4.4	Simulation study for $m_4(x) = (-4x + 1) \cdot I(x \in [0, 0.25]) + 0 \cdot I(x \in [0.25, 0.75]) + (4x - 3) \cdot I(x \in [0.75, 1])$. The results are averaged over 200 Monte Carlo replications. Monte Carlo standard error is displayed in the parentheses.	103
Table 5.1	Summary of $\hat{\beta}$ estimations. The results are based on 100 MC repetitions.	121
Table 6.1	Performance of regression estimator under heteroscedastic and homoscedastic error assumptions based on 401 simulations. Monte Carlo standard errors are displayed in the parentheses.	138

LIST OF FIGURES

Figure 2.1	An increasing-concave curve (solid), and observed data (scatter plot). . .	39
Figure 2.2	Box plots of predicted responses. The sample size is $n = 50$ and the noise has standard deviation $\sigma = 0.5$	40
Figure 2.3	ASA car fuel efficiency as a function of engine output. Top left: Convex and decreasing BP estimator (with $\hat{N} = 12$). Top middle: Convex BP estimator (with $\hat{N} = 17$). Top right: Decreasing BP estimator (with $\hat{N} = 17$). Bottom left: Unconstrained BP estimator (with $\hat{N} = 7$). Bottom middle: MP decreasing estimator.	41
Figure 2.4	Eye lens weight for rabbits as a function of age. Top left: Concave and increasing BP estimator (with $\hat{N} = 9$). Top middle: Concave BP estimator (with $\hat{N} = 9$). Top right: Increasing BP estimator (with $\hat{N} = 10$). Bottom left: Unconstrained BP estimator (with $\hat{N} = 4$). Bottom middle: MP increasing estimator.	42
Figure 3.1	3-D plots of two shape-restricted regression functions: (a) $f(x_1, x_2) = x_1^2 + 2e^{-x_2^2} + \frac{1}{4}x_1(x_2 + 2)$ (left panel); (b) $f(x_1, x_2) = x_1^{1/3}x_2^{2/3}$ (right panel). 74	
Figure 3.2	3-D surface plots and contour plots for $m_1(x_1, x_2) = \min(x_1, x_2)$ when $\sigma^2 = 0.01$ and $n = 400$. <i>Note:</i> Data is randomly generated from one Monto Carlo simulation out of 500 simulations. From top to bottom: surface plots, and contour plots. From left to right: true curve, monoProc (MP) method, and restricted Bernstein Polynomial (BP) method.	75
Figure 3.3	3-D surface plots and contour plots for $m_2(x_1, x_2) = x_1x_2I(x_1 \leq 0.5, x_2 \leq 0.5) + \{(x_1 - 0.5)(x_2 - 0.5) + x_1x_2\}I(x_1 > 0.5, x_2 > 0.5)$ when $\sigma^2 = 0.01$ and $n = 400$. <i>Note:</i> Data is randomly generated from one Monto Carlo simulation out of 500 simulations. From top to bottom: surface plots, and contour plots. From left to right: true curve, monoProc (MP) method, and restricted Bernstein Polynomial (BP) method.	75
Figure 3.4	3-D surface plots and contour plots for $m_3(x_1, x_2) = x_1^{\frac{1}{3}}x_2^{\frac{2}{3}}$ when $\sigma^2 = 0.1$ and $n = 400$. <i>Note:</i> Data is randomly generated from one Monto Carlo simulation out of 500 simulations. From top to bottom: surface plots, and contour plots. From left to right: true curve, monoProc (MP) method, and restricted Bernstein Polynomial (BP) method.	76
Figure 3.5	Plot of average predicted values vs. true values for $m_1(x_1, x_2) = \min(x_1, x_2)$. Solid line is 45° reference line.	76
Figure 3.6	Plot of average predicted values vs. true values for $m_2(x_1, x_2) = x_1x_2I(x_1 \leq 0.5, x_2 \leq 0.5) + \{(x_1 - 0.5)(x_2 - 0.5) + x_1x_2\}I(x_1 > 0.5, x_2 > 0.5)$. Solid line is 45° reference line.	77
Figure 3.7	Plot of average predicted values vs. true values for $m_3(x_1, x_2) = x_1^{\frac{1}{3}}x_2^{\frac{2}{3}}$. Solid line is 45° reference line.	78

Figure 3.8	Box plots of bias at five points using the BP and the MP method with sample size $n=400$ and 500 replications. <i>Note:</i> Simulation scenarios are described as follows. From top to bottom: $\sigma = 0.1$ and $\sigma = 1.0$; From left to right: $m_1(x_1, x_2) = \min(x_1, x_2)$, $m_2(x_1, x_2) = x_1 x_2 I(x_1 \leq 0.5, x_2 \leq 0.5) + \{(x_1 - 0.5)(x_2 - 0.5) + x_1 x_2\} I(x_1 > 0.5, x_2 > 0.5)$, and $m_3(x_1, x_2) = x_1^{\frac{1}{3}} x_2^{\frac{2}{3}}$	79
Figure 3.9	Optional caption for list of figures	80
Figure 3.10	Predicted values of $\hat{m}(x_{1i}, x_{2i})$ vs. predictors x_{1i} and x_{2i} . $i=1, \dots, 176$. <i>Note:</i> X1=ALT, X2=HEPC, Y=IMR. Top Panel: observed y_i vs. x_{1i} and x_{2i} ; Middle Panel: predicted values obtained by Bernstein polynomial method vs. predictors; Bottom: predicted values obtained by monoProc vs. predictors.	81
Figure 3.11	Comparison of observed and predicted values of the response. Solid line represents the 45° reference line, i.e., $y=x$	82
Figure 3.12	Residual vs. predicted values. Residual = predicted value - observed response. Top panel: Bernstein polynomial based estimate. Bottom panel: monoProc estimate.	82
Figure 3.13	Comparison of observed responses and predicted values: Reduced model in (a), Full model in (b), and predicted values from Full model vs. Reduced model in (c). Solid line represents the 45° reference line, i.e., $y=x$	83
Figure 3.14	CV score vs. N for the infant mortality data.	83
Figure 4.1	Histogram of the Bootstrap Distribution of 10000 Resampled IMSPE's.	106
Figure 4.2	Simulation results over 200 simulation data sets under $\sigma = 0.1$. Scatter plot: randomly generated data from one Monte Carlo simulation out of 200 simulations. Solid line: true regression curve. Dashed line: (average) Bayesian regression posterior median curve. Shaded region: (average) 95 percent pointwise credible (confidence) region for the Bayesian fit. Top Panel: increasing functions $m_1(x) = \sin(\frac{\pi}{2}x)$; $m_2(x) = 2x \cdot I(x \in [0, 0.25]) + 0.5 \cdot I(x \in [0.25, 0.75]) + (2x-1) \cdot I(x \in [0.75, 1])$. Bottom Panel: convex functions $m_3(x) = (16/9)(x - 1/4)^2$; $m_4(x) = (-4x + 1) \cdot I(x \in [0, 0.25]) + 0 \cdot I(x \in [0.25, 0.75]) + (4x - 3) \cdot I(x \in [0.75, 1])$	107
Figure 4.3	The Rabbit data alongside the Bayesian increasing Bernstein polynomial regression posterior mean curve (solid line) and 95 percent credible intervals (dotted line).	108
Figure 5.1	Spaghetti plot: each line represents the growth of an individual Sitka spruce tree over time. Left: control group. Right: ozone-enriched group.	127
Figure 5.2	Box plots of β estimation. Reference lines in red represent true β 's.	127
Figure 5.3	5-fold CV at the subject level	128
Figure 5.4	CV score vs. N for Sitka spruce data.	128
Figure 5.5	Thin-plate spline fit by Crainiceanu et al. (2005) for the function for Sitka spruce data (solid) with 95% credible sets.	129

Figure 5.6	Restricted Bernstein polynomial fit for Sitka spruces date: (a) mean growth rate fit by restricted Bernstein polynomial model; (b) ozone effect shown as growth rate difference at different time points. The fixed ozone effect obtained from Crainiceanu et al. (2005) is shown as dashed line. . .	129
Figure 6.1	The heteroscedastic estimation: the estimated regression function (black dashed), the true function (red solid), 95% confidence interval (gray). . .	141
Figure 6.2	The homoscedastic estimation: the estimated regression function (black dashed), the true function (red solid), 95% confidence interval (gray dashed).141	
Figure 6.3	The estimation of conditional variance function (dashed) vs. the true conditional variance function (solid).	142
Figure 6.4	Infant mortality data (a) Raw data and the estimated regression function; (b) Residuals and their estimated volatility function.	142
Figure B.1	The estimated variance $\hat{\sigma}_i^2$ for subject i	170

Chapter 1

Introduction to Shape Restricted Regression

The statistical regression method is often used to explore the inherent relationship between several predictor variables (usually denoted with $X_1, X_2, \dots, X_d$) and a response variable (usually denoted with Y). More specifically, the response variable is often assumed to have a relationship with predictor variables as follows:

$$Y_i = m(x_{1i}, x_{2i}, \dots, x_{di}) + \varepsilon_i, \quad i = 1, 2, \dots, n,$$

where ε_i 's are independently distributed with $E(\varepsilon_i|\mathbf{X}_i) = 0$ and $m(\mathbf{x}) = \mathbf{E}(Y|\mathbf{X} = \mathbf{x})$ when $\mathbf{X} = (X_1, \dots, X_d)$ and $\mathbf{x} = (x_1, \dots, x_d)$ denote the predictor vector and its realized value respectively. This model is called the general regression model, and $m(\cdot)$ is known as the regression function.

In many practical settings, the predictors and the response are known to preserve certain shape restrictions, such as monotonicity, concavity etc. For example, utility functions, cost functions and profit functions in economics are all known to be convex or concave curves (Gallant and Golub 1984; Terrell 1996). The dose response curve in the phase I clinical trial takes the form of a sigmoid curve with increasing trend. The estimation of the hazard rate and the

failure rate in reliability and survival analysis is another typical curve with convexity restrictions (Chang et al. 2007; Molitor and Sun 2002). Over the past decades, efforts have been devoted to search for a smooth and computationally efficient estimator of a shape restricted regression function.

In this thesis, we develop a new nonparametric regression methodology which is based on multivariate Bernstein Polynomials as basis functions. Our proposed method can be easily adopted to fit several shape-restricted regression curves. In the remainder of this chapter, we briefly review the relevant background literature for the estimators which are then fully developed in the remaining chapters of this thesis. We begin by reviewing the existing methods in the literature to estimate shape restricted regression function of a single univariate predictor. We then review the literature on multivariate shape restricted estimators. For completeness, we also provide a brief review of Bernstein Polynomials that are used to develop our proposed estimator, as well as some of their fundamental properties. Finally, we end this chapter with a discussion of the organization of this thesis proposal.

1.1 Shape Restricted Regression with a Single Predictor Variable

In this section we review the literature on shape restricted regression with a single predictor variable. Hildreth (1954) pioneered the research by proposing a maximum likelihood method to estimate a regression function under the restrictions of concavity. The regression model considered by Hildreth is $Y_{ij} = m(x_i) + \varepsilon_{ij}$, $i = 1, \dots, n$, and $j = 1, \dots, J_i$, where x_i 's represent distinct values of predictor variable X , and Y_{ij} 's represent the replicates of observed responses corresponding to x_i . The method is based on estimating a step function subject to restrictions of concavity satisfying $m(x_{i+1}) - m(x_i) \geq m(x_{i+2}) - m(x_{i+1})$, $\forall x_{i+2} \geq x_{i+1} \geq x_i$ where $i = 1, 2, \dots, n-2$. The problem is converted into a quadratic programming problem with the Kuhn-Tucker conditions. However, the estimator does not guarantee concavity for values outside the

range of observed x_i 's. Barlow et al. (1972) proposed the Pool Adjacent Violators Algorithm (PAVA) for solving monotonic regression problems, which is a special (univariate) case of the problem considered by Brunk (1955). Assume the regression model $Y_i = m(x_i) + \varepsilon_i$, the PAVA iteratively finds $\hat{m}_i$ which minimizes $\sum_{i=1}^n (y_i - \hat{m}_i)$ subject to a monotonicity constraint $\hat{m}_{i+1} \geq \hat{m}_i$ for $i = 1, 2, \dots, n-1$. The major shortcoming of this method is that the estimator obtained by the PAVA is not necessarily smooth (Friedman and Tibshirani 1984). A variety of the smoothed PAVA were developed subsequently by Friedman and Tibshirani (1984) and Mukerjee (1988) to obtain smoothed monotonic estimators. Friedman and Tibshirani (1984) combined a running mean smoother with the PAVA algorithm. They first smoothed the data by taking the average of the current prediction values with a certain number of the nearest neighbors and then monotonicizing the fit using PAVA. Mukerjee (1988) on the other hand obtained the isotonic estimator by the PAVA algorithm first, and then smoothed the fit using a kernel smoother.

Ramsay (1988) pioneered the use of regression splines for monotone regression functions. The method is based on utilizing the I-spline basis (the integrated M-spline basis) for monotone splines. Assume that the regression function is defined in an interval $[L, U]$ which is subdivided at the points $L = \xi_1 < \dots < \xi_q = U$. Let $\{t_1, \dots, t_{n+k}\}$ be a set of designated positions along the x -axis (or knots) by replacing boundary value ξ_i such that $t_1 \leq \dots \leq t_{n+k}$, $t_1 = \dots = t_k = L$, $t_{n+1} = \dots = t_{n+k} = U$, $t_{k+i} = \xi_i$ and $t_i < t_{i+k}$ for all i . M-spline functions of order k with n free parameters are piece-wise polynomials of degree $k-1$ defined at the knot sequence t . The I-spline basis is the integration of M-spline, and known as a piecewise polynomial of degree k . The monotone regression function is estimated using a linear combination of I-spline basis functions. The isotonic regression estimator by Barlow et al. (1972) is a special case of the I-spline method when $k = 1$. More recently, Meyer (2008) extended the method in Ramsay (1988) and proposed convex C-splines (integration of I-splines) to fit spline regression with convex constraints.

He and Shi (1998) presented another smoothing spline method using B-spline as basis functions. The method finds the solution which minimizes L_1 objective function $\sum_{i=1}^n |y_i - \pi(x_i)^T \alpha|$

subject to $\pi'(x_i)^T \alpha \geq 0$ for $i = 1, \dots, n$. $\pi(x) = (\pi_1(x), \pi_2(x), \dots, \pi_N(x))^T$ denotes normalized B-splines of degree two (quadratic spline), π' denotes the derivative of π , and α is a vector of spline coefficients. The restriction $\pi'(x_i)^T \alpha \geq 0$ ensures the monotonicity of the resulting spline fit. The optimization can be solved using quadratic programming method. In addition, the number of knots k is determined by a criterion similar to Akaike information criterion (AIC). Similarly, Wang and Li (2008) presented an estimator for monotone smoothing spline regression. The regression function is constructed to build upon the natural cubic splines.

Another type of method for regression subject to monotonicity constraints is kernel-type estimators (Hall and Huang 2001; Dette et al. 2006). Hall and Huang (2001) suggested a method to monotonize a general kernel-type estimator of a regression function (e.g. Gasser-Müller, Nadaraya-Watson, Priestley-Chao, local linear estimators, etc.). They generalized the definition of the linear estimator to be a weighted sum of observed responses expressed as $\hat{m}(x) = \sum_{i=1}^n p_i A_i(x) y_i$, where $p = (p_1, \dots, p_n)$ is a set of probability weights supported on the observed x -values $\{x_1, \dots, x_n\}$, and the weight functions A_i depends on a specific kernel function. The proposed monotone kernel regression estimator is obtained by choosing a proper value of p which minimizes a well-defined distance $\mathcal{D}(p)$ from p to the uniform distribution $p_0 = (1/n, \dots, 1/n)$. Dette et al. (2006) combined an unrestricted regression estimate with a density estimate to produce the new monotone estimator. As proposed, the inverse of the monotone estimator is given by $\hat{m}^{-1}(t) = \frac{1}{nh_d} \sum_{i=1}^n \int_{-\infty}^t K_d(\frac{\phi(i/n)-u}{h_d}) dt$, where $\phi(x) = \frac{\sum_{i=1}^n K_r((X_i-x)/h_r) Y_i}{\sum_{i=1}^n K_r((X_i-x)/h_r)}$ is the classical Nadaraya-Watson estimate, K_d and K_r are proper symmetric kernels with finite second moments, and h_d, h_r are corresponding bandwidths for regression estimate and density estimate. The accuracy of the regression results greatly relies on the proper choice of tuning parameters h_d and h_r . Although this method does not require constrained optimization, the selection of tuning parameters can be problematic and make an impediment in the practical implementation. Birke and Dette (2007) extended the method in Dette et al. (2006) to fit a convex regression.

Mammen et al. (2001) observed that many smoothing methods can be considered as a

projection of the data with respect to proper norms, and took one step further to propose a new projection based estimator. Assume a vector space denoted by $\mathcal{V}_s$ containing all possible candidate regression functions space denoted by $\mathcal{V}_s^m$ and a space of data vectors $\mathcal{V}_s^Y$. The space of all possible constrained smooth function is denoted by $\mathcal{C}_s^m$, and obviously $\mathcal{C}_s^m \subset \mathcal{V}_s^m$. Then unconstrained estimator $\hat{m}_s$ minimizes the objective function $\|\underline{Y} - \underline{m}\|^2$ in the $\mathcal{V}_s^m$ space. Similarly, the constrained estimator $\hat{m}_{s,c}$ minimizes $\|\underline{Y} - \underline{m}\|^2$ in the $\mathcal{C}_s^m$ space. Here, $\underline{Y}$ is a vector of constant functions with $f_i(x) = Y_i$, and $\underline{m}$ is a vector of the same function with $f_i(x) = m(x)$ for all i . Using a Pythagorean relationship, $\|\underline{Y} - \underline{m}\|^2$ yields to be the sum of $\|\underline{Y} - \underline{m}_s\|^2$ and $\|\underline{m}_s - \underline{m}\|^2$, where $\hat{m}_s$ is the projection of $\underline{Y}$ on the space $\mathcal{V}_s^m$. Using this decomposition, the constrained estimator turned out to be the projection of the unconstrained estimator on the space $\mathcal{C}_s^m$. Therefore, Mammen et al. (2001) suggested deriving an unconstrained smooth estimator first, and finding the projection of the fit on the constrained space.

Birke and Dette (2007) pointed out that the least squares based as well as the projection based estimators are not smooth enough, in the case that the underlying regression function is known to be smooth. In addition, some of these estimators suffer from the computational inefficiency. For example, the least squares based convex estimators in Dykstra (1983) and Han (1988) are usually calculated by the procedure of iterative loops. Moreover, a few methods are believed to be too complicated to handle the large sample case (e.g., sample size $n > 10,000$). For example, the estimator in Aït-Sahalia and Duarte (2003) requires one to construct restrictions at each individual observed data point, and thus number of constraint conditions could be as large as the sample size. Another drawback is that many of these estimators guarantee to maintain the convexity and/or other shape constraints only at observed data (i.e., at x_i 's only). Hence such shape constraint might be violated at prediction points which are not in the vicinity of the observed data. The monotone estimator by Hildreth (1954) is a typical example.

1.2 Shape Restricted Regression with Multiple Predictor Variables

The literature for shape restricted regression problems with more than one predictor is relatively scarce. The study begins with Brunk (1955). Brunk used the similar maximum likelihood method in Hildreth (1954) and derived an estimator for monotone parameters in the exponential family of distributions. This method has the same drawback as the one of Hildreth (1954), that the estimator can only provides estimation for the values of predictors observed in the data set. In order to interpret the individual effects of each predictor variable easily, Bacchetti (1989) presented an additive isotonic regression model which also has the link function to allow its use in generalized linear models. Bacchetti borrowed the standard PAV algorithm by Barlow et al. (1972) as well as some other algorithms, and defined the cyclic pooled-adjacent-violators algorithm for fitting additive isotonic regression models. The proposed estimator inherits some drawbacks of the PAVA algorithm. For instance, the resulting estimator is not necessary smooth and exhibit the step function like behavior when the observed data seem to be partitioned into groups.

Villalobos and Wahba (1987) presented a smoothing spline estimator using inequality-constrained thin-plate splines. Their proposed estimator minimized the objective function which is defined as the sum of a L_2 norm and the thin-plate penalty function. The tuning parameter which controls the smoothing level is chosen using the generalized cross-validation. The shape constraints are enforced on a fine regular grid in a subset of Euclidean space R^d . Therefore, it may fail to preserve the desired shape restrictions outside this grid. The approach was applied to estimate a bivariate distribution which is strictly concave in one dimension and monotone in the other dimension, and the resulting fit failed to satisfy the concavity in some areas. Mammen et al. (2001) pointed out that the fitted spline smoother by Villalobos and Wahba (1987) does not guarantee to satisfy the shape constraints everywhere on the support of the predictor vector. Other attempts of smoothing spline regression under shape restrictions in-

clude Beresteanu (2004), Bollaerts, Eilers and van Mechelen (2006), and Leitenstorfer and Tutz (2007). Beresteanu (2004) constructed a sieve least squares estimator. The basis function used is the familiar normalized B-splines. Beresteanu restricted the coefficients of B-splines basis to satisfy shape restrictions. The method can be solved using quadratic programming technique. Bollaerts, Eilers and van Mechelen (2006) presented P-splines regression with additional asymmetric discrete penalties to fit a multivariate isotonic regression curve. Leitenstorfer and Tutz (2007) also introduced a B-spline based estimators under monotonicity constraints. These three methods all assume that the observed values of the predictor variables lie on equidistant grids of predictor domain. If the sample size is relatively small, then having a small and discrete number of knots can only allow very limited control over smoothness and flexibility. The fit is especially problematic in higher dimensions.

To allow for the presence of more relaxed spatial arrangement of observed predictors values, Dette and Scheder (2007) presented a kernel-based approach to estimate multivariate non-parametric regression function which is strictly monotone in all or a subset of its arguments and established rigorous asymptotic results of their proposed estimator. The method in Dette and Scheder (2007) is a natural extension of the univariate monotone estimator in Dette et al. (2006) to two or more variables. Therefore, the method cannot escape some of the drawbacks of the univariate version like the choice of the tuning parameters. In addition, this method could be considerably time consuming to implement even with only moderately large sample size (e.g., with two predictors and a sample size of only 400, it takes on average around 286 seconds. For the same scenario, our proposed method takes only 6.36 seconds on average). Moreover, their method is limited to fit only monotonic curves and it is not clear how the method can be extended to fit other popular shape constraints (e.g., convex, concave etc.).

1.3 Bayesian Models for Shape-Restricted Regression

Bayesian methods for a regression problem begins with the joint distribution $p(\theta, \mathcal{D}_n)$ of the parameters θ and the observed data $\mathcal{D}_n$. Following the Bayes' theorem, we have

$$p(\theta|\mathcal{D}_n) \propto p(\mathcal{D}_n|\theta)p(\theta),$$

where $p(\theta|\mathcal{D}_n)$ denotes the posterior distribution of θ (i.e., the distribution of parameters θ obtained by combining the information from the data with the information from the prior distribution), $p(\theta)$ denotes the prior distribution of θ , and $p(\mathcal{D}_n|\theta)$ denotes the likelihood function.

Lavine and Mockus (1995) first applied a nonparametric Bayesian approach to estimate an isotonic regression function. The method used the Dirichlet process by assigning a prior distribution on a space of distribution functions which are nondecreasing. The most notable drawback of using a Dirichlet process is that the probability measures are almost discrete. Therefore, if the true regression function is smooth, the Dirichlet process is not a proper prior distribution. Holmes and Heard (2003) proposed another Bayesian piecewise-constant model for a monotonic regression function. In their approach, they considered a random number of constant pieces and random locations of these knots, and then gave a prior distribution. The approach itself does not add any constraint onto the prior distribution. The monotonicity constraint was guaranteed by discarding all posterior samples that does not satisfy the constraint. The idea is straightforward to implement, but could be very inefficient.

The approach by Neelon and Dunson (2004) adopted a piecewise-linear model for a unknown monotone regression function. To assign the prior distribution of β , the coefficients in the piecewise linear model, Neelon and Dunson (2004) chose a density consisting of the mixture of point masses and truncated normal densities. Then the MCMC algorithm was used to obtain the posterior distributions and the estimation of the regression function. However, it is believed that the piecewise linear function may not be appropriate to model the unknown regression function which satisfies certain differentiability conditions (Curtis and Ghosh 2011). Chang et

al. (2007) used Bernstein polynomial expansions as basis function to fit nonparametric shape-restricted regression functions. They considered monotonic, unimodal and unimodal-concave as shape constraints. Bayesian algorithm was applied by placing prior distributions on the coefficients of Bernstein polynomial expansions. The authors claimed that the prior distributions of model parameters satisfy certain shape-constraints, but the detailed prior distributions were not explicitly given in the literature. The authors used reverse jump MCMC as the sampling scheme from the posterior distribution, but the algorithm was not clearly expressed either in the article.

1.4 Bernstein Polynomials

The Bernstein polynomial is one of the most popular class of polynomials in approximation theory. It was introduced by S.N. Bernstein in 1912 which he used to provide a constructive proof to the famous Weierstrass approximation theorem. In this section, we briefly review both univariate and multivariate Bernstein polynomials, followed by introducing some of their fundamental properties available in the literature of approximation theory.

1.4.1 Univariate Bernstein Polynomials

Let $f : [0, 1] \rightarrow \mathbb{R}$ be a continuous function. The Bernstein polynomial of the degree N of $f(\cdot)$ on the interval $[0, 1]$ is defined by (Lorentz 1986),

$$B_N(f)(x) = \sum_{k=0}^N f\left(\frac{k}{N}\right) \cdot b_k(x, N) = \sum_{k=0}^N f\left(\frac{k}{N}\right) \cdot \binom{N}{k} x^k (1-x)^{N-k}.$$

Note that, if f is defined on the closed interval $[a, b]$, then the corresponding Bernstein polynomial is written as $B_N(f; [a, b])(x) = \frac{1}{(b-a)^N} \sum_{k=0}^N f(a + k\frac{b-a}{N}) \cdot \binom{N}{k} (x-a)^k (b-x)^{N-k}$. It has been shown that as the order N grows, the Bernstein polynomial of a continuous function converges to the function uniformly (see Bernstein (1912) or Lorentz (1986) for more details). That is, $\lim_{N \rightarrow \infty} \sup_{x \in [0, 1]} |B_N(f)(x) - f(x)| = 0$. In this subsection, if not specifically mentioned, we

assume $f : [0, 1] \rightarrow \mathbb{R}$ is a continuous function throughout.

The Bernstein polynomial is appropriate for shape-preserving regression, as it has the optimal shape restriction property among all polynomials (Carnicer and Pena 1993). Here we list a few shape-preserving properties of the Bernstein polynomial. The complete proofs of these properties and many other interesting properties can be found in Chapter 1 of the excellent book by Gal (2008).

- If f is convex (strictly convex) of order $k \in \{0, 1, 2, \dots, N\}$ on $[0, 1]$, then $B_N(f)$ is convex (strictly convex) of order k on $[0, 1]$, for all $N \in \mathbb{N}$; here notice that the usual convex function is of order $k = 1$.
- If f is quasi-convex of order $k \in \{0, 1, 2, \dots, N\}$ on $[0, 1]$, then $B_N(f)$ is quasi-convex of order k on $[0, 1]$, for all $N \in \mathbb{N}$; here notice that a real-valued function $f : [a, b] \rightarrow \mathbb{R}$ defined on $[a, b]$ of a real vector space is quasi convex if whenever $x_1, x_2 \in [a, b]$ and $\lambda \in [0, 1]$ then $f(\lambda x_1 + (1 - \lambda)x_2) \leq \max(f(x_1), f(x_2))$.
- If f is u-monotone, where $u(x) = x^\lambda$, for all $x \in [0, 1]$ and $\lambda \in (0, 1)$ is arbitrary and fixed, then $B_N(f)$ is u-monotone for all $n \in \mathbb{N}$. Here notice that a function $f \in C[a, b]$ is u-monotone if $\exists u \in C[a, b], u(x) > 0$, and $u(x_1)f(x_2) - u(x_2)f(x_1) \geq 0$ for all $a \leq x_1 < x_2 \leq b$.

A (univariate) Bernstein polynomial of order (or degree) N is defined as a linear combination of Bernstein basis polynomials (Lorentz 1986),

$$B_N(x; \beta) = \sum_{k=0}^N \beta_k \cdot b_k(x, N) \text{ where } \beta = (\beta_0, \dots, \beta_N)',$$

where β_k are Bernstein coefficients, and $b_k(x, N)$ are Bernstein basis polynomials. The $N + 1$ Bernstein basis polynomials of order N are defined as

$$b_k(x, N) = \binom{N}{k} x^k (1 - x)^{N-k}, \quad \binom{N}{k} = \frac{N!}{k!(N-k)!},$$

where $k = 0, 1, \dots, N$ and $0 \leq x \leq 1$. From the definition, it is easy to show that all Bernstein basis polynomials are non-negative for all $x \in [0, 1]$. The Bernstein basis polynomials have the following nice properties (Lorentz 1986; Joy 2000).

- The Bernstein basis polynomials of order N can be recursively represented by the sum of two Bernstein polynomials of order $N - 1$. More precisely,

$$b_k(x, N) = (1 - x) \cdot b_k(x, N - 1) + x \cdot b_{k-1}(x, N - 1).$$

Using this fact, we obtain the following property,

$$\sum_{k=0}^N b_k(x, N) = \sum_{k=0}^{N-1} b_k(x, N - 1) = \dots = \sum_{k=0}^1 b_k(x, 1) = (1 - x) + x = 1.$$

- Bernstein basis polynomials of order $N - 1$ can be expressed as a linear combination of Bernstein basis polynomial of order N . That is,

$$b_k(x, N - 1) = \frac{N - k}{N} b_k(x, N) + \frac{k + 1}{N} b_{k+1}(x, N).$$

It is easy to show that any Bernstein basis polynomial with the order less than N can be written as a linear combination of Bernstein basis polynomials of order N .

- The derivatives of Bernstein basis polynomials are polynomials of lower order. That is,

$$b'_k(x, N) = \frac{d}{dx} b_k(x, N) = N \cdot [b_{k-1}(x, N - 1) - b_k(x, N - 1)].$$

Hence, it follows that $B'_N(x, \beta) = N \sum_{k=0}^{N-1} (\beta_{k+1} - \beta_k) b_k(x, N - 1) = N B_{N-1}(x, N \nabla \beta)$ where $\nabla \beta = (\beta_1 - \beta_0, \beta_2 - \beta_1, \dots, \beta_N - \beta_{N-1})'$. In other words the derivative of a Bernstein polynomial is also a Bernstein polynomial (of a lower degree). This property can be easily

extended to higher order derivatives. More precisely,

$$B_N^{(l)}(x, \beta) = \frac{N!}{(N-l)!} \sum_{k=0}^{N-l} \nabla^{(l)} \beta_k b_k(x, N-l) = B_{N-l} \left(x, \frac{N!}{(N-l)!} \nabla^{(l)} \beta \right)$$

- Bernstein basis polynomials can be presented in terms of the power basis, and each power basis element can also be written as a linear combination of Bernstein Polynomials. More precisely,

$$\begin{aligned} b_k(x, N) &= \sum_{i=k}^N (-1)^{i-k} \binom{N}{i} \binom{i}{k} x^i, \\ x^i &= \sum_{k=i-1}^{N-1} \frac{\binom{k}{i}}{\binom{N}{i}} b_k(x, N). \end{aligned}$$

The estimator based on univariate Bernstein polynomial have played an important roles in nonparametric curve estimation. Chak, Madras and Smith (2005), Chang et al. (2007) and Curtis and Ghosh (2011) have applied the Bernstein polynomials to shape-restriction regression. One of the most fundamental advantages of using Bernstein polynomial is that the estimation of the entire function $f(x)$ for $x \in [0, 1]$ can be reduced to the finite dimensional estimation of $\{f(\frac{k}{N}) \text{ for } k = 0, 1, \dots, N\}$ and that the restrictions on $\beta_k = f(\frac{k}{N})$ for $k = 0, 1, \dots, N$ leads to the restriction on not only the function $f(x)$ but also on all derivatives of $f(x)$. The shape constraints can be easily applied by imposing certain linear constraints on the coefficients β_k 's. Chang et al. (2007) enlisted a large collection of constraints on β_k 's for monotonic, unimodal, and unimodal-concave regression functions. They also proposed a Bayesian approach to fit a shape-restricted regression with random Bernstein polynomials. A reversible-jump Markov chain Monte Carlo (RJMCMC) algorithm is used to choose a proper order of the Bernstein polynomial. However, their method was shown to be inefficient in capturing the flat portions of monotone curves and subsequently extended with variable selection priors by Curtis and Ghosh (2011).

1.4.2 Multivariate Bernstein Polynomials

Let f be a continuous function defined on $[0, 1]^d \rightarrow \mathbb{R}$. The multivariate Bernstein polynomials of f is defined as (Lorentz 1986)

$$\begin{aligned} B_{N_1, \dots, N_d}(f)(x_1, \dots, x_d) &= \sum_{k_1=0}^{N_1} \dots \sum_{k_d=0}^{N_d} f\left(\frac{k_1}{N_1}, \dots, \frac{k_d}{N_d}\right) \cdot \prod_{i=1}^d b_{k_i}(x_i, N_i) \\ &= \sum_{0 \leq k_i \leq N_i, i \in \{1, \dots, d\}} f\left(\frac{k_1}{N_1}, \dots, \frac{k_d}{N_d}\right) \cdot \prod_{i=1}^d \binom{N_i}{k_i} x_i^{k_i} (1 - x_i)^{N_i - k_i}. \end{aligned}$$

The multivariate Bernstein polynomials $B_{N_1, \dots, N_d}(f)(x_1, \dots, x_d)$ have been shown to be both pointwise and uniformly convergent to f when $\min(N_1, \dots, N_d)$ goes to infinity (Lorentz 1986). Moreover, the Bernstein polynomials as well as other Bernstein-type polynomials have interesting shape-preserving properties. We recommend Sauer (1999); Gal (2008) for more details about the shape-preserving properties of the multivariate Bernstein-type polynomials.

The bivariate Bernstein polynomial estimator was first investigated by Tenbusch (1994) to estimate two-dimensional density functions. The resulting regression estimator has been shown to be universally consistent and asymptotically normal under a set of mild regularity conditions in the two-dimensional case without any shape restriction (Tenbusch 1997). Here we list some properties of multivariate Bernstein polynomials. For simplicity, we consider only $d = 2$ to illustrate bivariate regression subject to various shape constraint.

- The first order partial derivatives can be presented as

$$\begin{aligned} \frac{\partial B_N}{\partial x_1} &= N_1 \sum_{k_2=0}^{N_2} \sum_{k_1=0}^{N_1-1} (\beta_{k_1+1, k_2} - \beta_{k_1, k_2}) b_{k_1}(x_1, N_1 - 1) b_{k_2}(x_2, N_2) \\ \frac{\partial B_N}{\partial x_2} &= N_2 \sum_{k_1=0}^{N_1} \sum_{k_2=0}^{N_2-1} (\beta_{k_1, k_2+1} - \beta_{k_1, k_2}) b_{k_1}(x_1, N_1) b_{k_2}(x_2, N_2 - 1) \end{aligned}$$

The nondecreasing constrain in x_1 is satisfied if $\beta_{k_1, k_2} \leq \beta_{k_1+1, k_2}$ for $k_1 = 1, \dots, N_1 - 1$.

The nondecreasing constrain in both x_1 and x_2 is satisfied if $\beta_{k_1, k_2} \leq \beta_{k_1+1, k_2}$ for $k_1 =$

$1, \dots, N_1 - 1$ and $\beta_{k_1, k_2} \leq \beta_{k_1, k_2+1}$ for $k_2 = 1, \dots, N_2 - 1$.

- The second order partial derivatives can be presented as

$$\begin{aligned}\frac{\partial^2 B_N}{\partial x_1^2} &= N_1(N_1 - 1) \sum_{k_2=0}^{N_2} \sum_{k_1=0}^{N_1-2} (\beta_{k_1+2, k_2} - 2\beta_{k_1+1, k_2} + \beta_{k_1, k_2}) b_{k_1}(x_1, N_1 - 2) b_{k_2}(x_2, N_2) \\ \frac{\partial^2 B_N}{\partial x_2^2} &= N_2(N_2 - 1) \sum_{k_1=0}^{N_1} \sum_{k_2=0}^{N_2-2} (\beta_{k_1, k_2+2} - 2\beta_{k_1, k_2+1} + \beta_{k_1, k_2}) b_{k_1}(x_1, N_1) b_{k_2}(x_2, N_2 - 2)\end{aligned}$$

Using the above equations, the concavity in x_1 will be enforced by $\beta_{k_1+2, k_2} - 2\beta_{k_1+1, k_2} + \beta_{k_1, k_2} \leq 0$ for $k_1 = 1, \dots, N_1 - 2$. Similarly, the convexity in x_2 can be satisfied by $\beta_{k_1, k_2+2} - 2\beta_{k_1, k_2+1} + \beta_{k_1, k_2} \geq 0$ for $k_1 = 1, \dots, N_1 - 2$. By reversing the signs in the above inequalities, many other shapes can be enforced.

1.5 Organization of The Thesis

The thesis is organized as follows. Chapter 2 introduces the shape restricted estimator with univariate Bernstein polynomials. Chapter 3 presents the multivariate shape restricted estimator with multivariate Bernstein polynomials. Chapter 4 demonstrates two methods of making inferences of the proposed algorithm. In Chapter 5 we extend our algorithm into the linear mixed model framework, and present its application through numerical studies. In Chapter 6, we present how to adjust our method under the assumption of heteroscedasticity. We conclude in Chapter 7 with a brief discussion of our findings, and provide potential research areas for further extensions.

Chapter 2

Shape Restricted Nonparametric Regression with Univariate Bernstein Polynomials

2.1 Introduction

In many practical settings, subject-matter information about the relationships between a response and a predictor variable are often available. In econometrics, social sciences, biology and allied areas of applications, the regression function is often known to satisfy various shape constraints such as non-negativity, monotonicity and convexity (or concavity). Some popular examples include the study of utility functions, cost functions, and profit functions in economics (Gallant and Golub 1984; Terrell 1996), the study of dose response curve in the phase I clinical trials, growth curves of animals and plants in ecology and the estimation of the hazard rate and the failure rate in reliability and survival analysis (see e.g., Molitor and Sun (2002); Chang et al. (2007) for many such examples).

The estimation of the regression function subject to a given set of shape restrictions can be formulated as follows. Consider a general regression model where we assume that (X, Y)

is a $\mathbb{R} \times \mathbb{R}$ -valued random vector arising from an arbitrary bivariate distribution such that $\mathbf{E}[Y^2] < \infty$. Let $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ be a sample of observations which are assumed to be independently and identically distributed (i.i.d) as (X, Y) . The regression model in terms of observed data can be expressed as,

$$Y_i = m(X_i) + \varepsilon_i, \quad \text{for } i = 1, \dots, n, \quad (2.1)$$

where ε_i 's are independently distributed with $E(\varepsilon_i|X_i) = 0$ and $m(x) = \mathbf{E}(Y|X = x)$. Assume that the regression function $m(\cdot)$ is known to satisfy a set of certain shape restrictions but not necessarily based on a parametric form. Our goal is to obtain a nonparametric estimator $\hat{m}(\cdot)$ as a function of the observed data $\mathcal{D}_n$ that will satisfy the set of shape restrictions for all values of x in the support of X and for any realization of the data $\mathcal{D}_n$.

To illustrate the benefit of using a constrained estimator, in Figure 2.1, we consider an increasing-concave function $m(x) = 0.5x^2 \log(x) - 1.75x^2 + 3x$ which is displayed as the solid curve, along with a set of observed data simulated from model (2.1) with errors $\epsilon_i \stackrel{iid}{\sim} N(0, 0.5)$. Without any background knowledge about the underlying regression function the scatter plot apparently might suggest a linear increasing regression function, and the least squares estimate of such a line is displayed as the dotted line. Later in Section 2.4.2 we revisit this simulated data example and show the substantial improvements in the fit brought by the subject matter knowledge of an increasing concave function.

[insert Figure 2.1 here.]

Over the past decades, a lot of efforts have been devoted to search for a smooth and computationally efficient nonparametric estimator of a regression function satisfying a desired set of shape restrictions. Brunk (1955) first proposed a method based on estimating a step function subject to restrictions on the steps using the maximum likelihood method. Barlow et al. (1972) proposed the so-called Pool Adjacent Violators Algorithm (PAVA) and a variety of the smoothed versions were attempted in the literature (see e.g., Friedman and Tibshirani (1984);

Mukerjee (1988); Aït-Sahalia and Duarte (2003) among others). In addition, several additional efforts have been made to enrich the literature of estimating a monotone regression function using other approaches, including the smoothing spline methods (Ramsay 1988; Mammen et al. 2001; Wang and Li 2008), projection methods (Mammen et al. 2001), polynomial basis estimators (Chang et al. 2007), and many alike. Hildreth (1954) pioneered the constrained least squares method to estimate a concave function, and Wu (1982) and Fraser and Massam (1989) proposed efficient algorithms to compute the estimator. Aït-Sahalia and Duarte (2003) have also considered a decreasing and convex function with certain bounds on the derivative. Hanson and Pledger (1976), Mammen (1991) and Groeneboom et al. (2001) derived several asymptotic rate of convergence. Alternative approaches which combine shape restriction and smoothing process have also been proposed; for example, kernel-based estimators (Hall and Huang 2001; Dette et al. 2006; Birke and Dette 2007), spline-based estimators (Pal, Woodroffe and Meyer 2007; Meyer 2008), and polynomial-based estimators (Chang et al. 2007; Curtis and Ghosh 2011).

However, Birke and Dette (2007) pointed out the estimators produced by the constrained least squares techniques and other projection based techniques may suffer from a lack of smoothness, even when the underlying regression function is known to be smooth. In particular, those estimators that rely on the PAVA algorithm often exhibit a step function like behavior when the observations seem to be partitioned into groups. In addition, some of these estimators suffer from the computational challenges especially when the sample sizes are large. For example, the least squares based convex estimators in Dykstra (1983) and Han (1988) are usually computed using iterative loops. Moreover, a few methods are believed to be too time consuming when the sample sizes are extraordinarily large (e.g., sample size $n > 10,000$). For instance, the estimator proposed by Aït-Sahalia and Duarte (2003) requires one to construct restrictions at each individual observed data point, and thus the number of constraint conditions could be as large as the sample size. Another major drawback of these methods is that many of these estimators are guaranteed to satisfy the given shape constraint only at observed x -values and not necessarily at unobserved x -values in the support of X . Hence, the desired shape constraint might

be violated for values of the predictor variable which are not in the vicinity of the observed data. The constrained least squares based estimators in Hildreth (1954) and Brunk (1955) are commonly known to suffer from this crucial deficiency. Villalobos and Wahba (1987) proposed a bivariate spline smoother which enforces the constraints on a grid of points. Mammen et al. (2001) pointed out that the fitted spline smoother obtained by this estimation method would not be guaranteed to satisfy the shape constraints everywhere in the domain of the predictor space. We also observed from their numerical results (see Figures 3 and 4 in Villalobos and Wahba (1987)) that some fitted smooth surface areas outside of the grid points violate the concave assumption in the y coordinate.

We adopt the method of sieve and propose an estimator based on restricted Bernstein polynomials. The Bernstein polynomial is appropriate for shape-preserving regression, as it has the optimal shape restriction property among all polynomials (Carnicer and Pena 1993), and all of its derivatives possess the same convergence properties (Lorentz 1986). Moreover, the class of Bernstein polynomials is dense (with respect to sup-norm) in the function space of all continuous functions with compact support. When the true function has sufficient smoothness, Bernstein polynomial and B-splines perform equivalently in terms of asymptotic properties. However, there are at least two advantages of Bernstein polynomial. First, using Bernstein polynomial, we are able to convert a fairly general shape restricted functional regression problem into a (finite-dimensional) least squares problem with (only) linear constraint (for a multitude of different shapes) which can be solved easily using well established quadratic programming method. Secondly, in our theory, higher order smoothness conditions are not required to establish asymptotic properties of the estimator. We only need the underlying true regression function to be continuous (and may not be even be first order differentiable), which is in sharp contrast to the assumptions that are generally used for B-splines theory. In addition, even if the function is discontinuous of finite pointwise type, we can still use Bernstein polynomial to estimate the regression curve (Herzog and Hill 1946).

The Bernstein polynomials type estimators have been applied to nonparametric monotone

curve estimation in Chak, Madras and Smith (2005); Chang et al. (2007); Curtis and Ghosh (2011) and Stadtmuller (1986) among others. We build on these works and extend the methodologies to obtain a shape-restricted Bernstein polynomial estimator. There are several advantages of our new estimator: Our estimator can be easily adapted to accommodate various popular shape constraints (such as nonnegativity, monotonicity, convexity, and increasing-concavity and their variations and combinations). Our proposed estimator is guaranteed to satisfy the required shape constraints everywhere in the domain of the predictor variable, including all prediction points and not just at observed points. The method is computationally efficient, because the proposed estimator can be easily computed as a solution of a quadratic programming problem. Finally, under some mild regularity conditions, our estimator is shown to be universally consistent.

The chapter is organized as follows. Section 2.2 presents the sieve-based Bernstein polynomial estimator subject to a class of popular shape constraints and hence demonstrates how to construct proper sieves corresponding to various shape restrictions, and concludes with computational details. In Section 2.3, we provide the asymptotic properties of the proposed estimator. The corresponding proofs of our main results are provided in the Appendix. Section 2.4 presents simulation studies when the underlying regression curve is subject to various shape constraints such as monotonically increasing, concave or both. The numerical performance of our approach is thoroughly examined and compared with a few currently available nonparametric regression estimators subject to shape restrictions. In Section 2.5, we apply the proposed methodology on real data sets and compare its predictive performance with other available methods. Finally, in Section 2.6, we conclude with a brief discussion of our findings and future research directions.

2.2 Restricted Bernstein Polynomial Estimator

Consider again $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ to be a set of observations which are assumed to be independently and identically distributed (i.i.d) as (X, Y) satisfying the regression model $Y = m(X) + \varepsilon$ as described in Section 2.1. Let $\mathcal{F}$ be a class of continuous functions subject to

a given class of shape restrictions. The regression function $m(\cdot)$ is the one that minimizes the L_2 risk within $\mathcal{F}$ (Gyorfi et al. 2002), i.e.,

$$m(\cdot) = \arg \min_{f \in \mathcal{F}} \mathbf{E}\{(f(X) - Y)^2\}. \quad (2.2)$$

Throughout the chapter, we assume that predictor variables are suitably transformed to lie in the unit $[0, 1]$ (see Section 2.2.2 for more computational details). Adopting the method of sieves, we consider the constrained Bernstein polynomial sieve $\{\mathcal{F}_N\}$ as follows:

$$\mathcal{F}_N = \{B_N(x) = \sum_{k=0}^N \beta_k \cdot b_k(x, N) : \mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{0} \text{ and } \sum_{k=0}^N |\beta_k| \leq L_N\}, \quad (2.3)$$

for $N = 1, 2, \dots$, where $b_k(x, N) = \binom{N}{k} x^k (1-x)^{N-k}$ are the Bernstein basis polynomials. Notice that we can express $B_N(x) = \mathbf{b}_N^\top(x) \boldsymbol{\beta}_N$ where $\mathbf{b}_N^\top(x) = (b_0(x, N), \dots, b_N(x, N))'$ and $\boldsymbol{\beta}_N = (\beta_{N,0}, \dots, \beta_{N,N})'$. For a vector $\mathbf{b}$, we use $\mathbf{b} \geq \mathbf{0}$ to denote the fact that the inequality is satisfied componentwise. We allow the order of the polynomial, N to grow with the sample size n , e.g., $N = o(n^k)$ with a suitably chosen $k > 0$, and the bound $L_N > 0$ is chosen suitably to grow with N (see Main Theorem). $\mathbf{A}_N$ represents a full row rank restriction matrix with dimension $R_N \times (N+1)$ where R_N denotes the rank of $\mathbf{A}_N$. The matrix A_N is chosen such a way that each (function) member in the sieve $\mathcal{F}_N$ preserves the desired shape restrictions (i.e., $\mathcal{F}_N \subseteq \mathcal{F}, \forall N$). Given the data set $\mathcal{D}_n$, the estimator m_N , which is selected among all possible functions in $\mathcal{F}_N$, minimizes the empirical L_2 risk. That is:

$$m_N(\cdot) = \mathbf{b}_N^\top(x) \hat{\boldsymbol{\beta}}_N \text{ where } \hat{\boldsymbol{\beta}}_N = \arg \min_{\boldsymbol{\beta}_N \in \mathcal{B}_N} \frac{1}{n} \sum_{i=1}^n (\mathbf{b}_N^\top(X_i) \boldsymbol{\beta}_N - Y_i)^2, \quad (2.4)$$

where $\mathcal{B}_N = \{\boldsymbol{\beta}_N \in \mathbb{R}^{N+1} : \mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{0}, \sum_0^N |\beta_k| \leq L_N\}$. Note that the existence of m_N follows by the compactness of $\mathcal{F}_N$. In Section 2.3, we show that when $N = o(n^k)$ for some $k > 0$, m_N is weakly universally consistent for m in the L_2 space and with additional mild regularity conditions, m_N is also strongly universally consistent.

2.2.1 Examples of Constraint Bernstein Polynomial Sieves

- Nonnegative Restriction

Let $m(x)$ be a univariate function in the parameter space $\mathcal{F} = \{f \in C[0, 1] : f(x) \geq 0, \forall x \in [0, 1]\}$, where $C[0, 1]$ denotes the space of all real-valued continuous functions defined on $[0, 1]$. We define the restriction on the coefficients as follows:

$$\mathbf{A}_N \boldsymbol{\beta}_N \equiv \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_N \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Clearly in this case $\mathbf{A}_N = I_{N+1}$, the identity matrix of order $(N + 1)$ and its rank $R_N = N + 1$. The above restriction ensures $\beta_k \geq 0$, for all k . Since Bernstein basis polynomials $b_k(x, N) = \binom{N}{k} x^k (1 - x)^{N-k}$ are always nonnegative, $B_N(\mathbf{x}) \equiv \sum_{k=0}^N \beta_k \cdot b_k(x, N)$ are also nonnegative. This implies $\mathcal{F}_N \subset \mathcal{F}$.

- Monotone Restriction

Without loss of any generality, a monotone function in this chapter refers to a non-decreasing function. Non-increasing monotonicity can be simply obtained by reversing the inequalities. Let $m(x)$ be a real-valued monotone function in the space $\mathcal{F} = \{f \in C[0, 1] : f(x_1) \leq f(x_2), \forall 0 \leq x_1 \leq x_2 \leq 1\}$ where $C[0, 1]$ denotes the class of all continuous functions defined on $[0, 1]$. $\mathcal{F}_N$ is the sieve defined in (2.3) and the increasing monotonicity is satisfied by the following conditions:

$$\mathbf{A}_N \boldsymbol{\beta}_N \equiv \begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots \\ & & \ddots & & \\ 0 & \dots & 0 & -1 & 1 \end{pmatrix}_{N \times (N+1)} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_N \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \quad (2.5)$$

Obviously, the constraint shown in (2.5) guarantees $\beta_0 \leq \beta_1 \leq \dots \leq \beta_N$. For any $B_N(x) \in \mathcal{F}_N$, its derivative is given by $B'_N(x) = N \sum_{k=0}^{N-1} (\beta_{k+1} - \beta_k) b_k(x, N-1) \geq 0$ for all $x \in [0, 1]$, which implies $B_N(\cdot) \in \mathcal{F}$ and hence $\mathcal{F}_N \subset \mathcal{F}$. Note that in this case $R_N = N$.

- Convex/Concave Restriction

Let $m(x)$ be a real-valued convex function in $\mathcal{F} = \{f \in C[0, 1] : 2f(\frac{x_1+x_2}{2}) \leq f(x_1) + f(x_2), \forall x_1, x_2 \in [0, 1]\}$. Notice that we are not assuming that the true function $m(x)$ has derivatives of any order. $\mathcal{F}_N$ takes the form in (2.3) with the following restriction matrix:

$$\mathbf{A}_N \beta_N \equiv \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ & & \ddots & & & \\ 0 & \dots & 0 & 1 & -2 & 1 \end{pmatrix}_{(N-1) \times (N+1)} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_N \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \quad (2.6)$$

Since the second derivatives of $B_N(x)$ can be written as $B''_N(x) = N(N-1) \sum_{k=0}^{N-2} (\beta_{k+2} - 2\beta_{k+1} + \beta_k) b_k(x, N-2)$, the above restriction ensures $B''_N(\cdot) \geq 0$ for all N , and therefore $\mathcal{F}_N \subset \mathcal{F}$. Concavity can be simply obtained by taking negative of the restriction matrix $\mathbf{A}_N$ (i.e., $\mathbf{A}_N^{\text{concave}} = -\mathbf{A}_N$, when $\mathbf{A}_N$ takes the form in (2.6). Note that in this case $R_N = N - 1$.

- Monotone and Convex Restriction

Let $m(x)$ be a real-valued increasing and convex function in the function space $\mathcal{F} = \{f \in C[0, 1] : f(x_1) \leq f(x_2), \forall 0 \leq x_1 \leq x_2 \leq 1; \text{ and } 2f(\frac{x_1+x_2}{2}) \leq f(x_1) + f(x_2), \forall x_1, x_2 \in [0, 1]\}$. Again, notice that we are not making any assumptions regarding the existence of the derivatives of the true regression function. This is in sharp contrast to assumptions made when splines are used as approximation space. We define the restriction for increase

convexity on the coefficients as follows:

$$\mathbf{A}_N \boldsymbol{\beta}_N \equiv \begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ & & \ddots & & & \\ 0 & \dots & 0 & 1 & -2 & 1 \end{pmatrix}_{N \times (N+1)} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_N \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \quad (2.7)$$

This restriction guarantees $\beta_N - \beta_{N-1} \geq \dots \geq \beta_1 - \beta_0 \geq 0$, which in turn implies that $B_N''(\cdot) \geq 0$ and $B_N'(\cdot) \geq 0$ for all N . Therefore $\mathcal{F}_N \subset \mathcal{F}$. Other combinations of increasing (or decreasing) monotonicity and convexity (or concavity) can be simply obtained by altering the signs of the element of the $\mathbf{A}_N$ matrix. Note that in this case $R_N = N$.

In all of the examples and their suitable variations (by suitably choosing the elements of $\mathbf{A}_N$) we have shown that the estimation of the regression function $m(x)$ can be reduced to the estimation of the finite dimensional parameter $\boldsymbol{\beta}_N$ subject to linear constraints $\mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{0}$. Next we show how to obtain the estimates of $\boldsymbol{\beta}_N$ and N .

2.2.2 Computation of the Sieved Estimator

The estimator m_N in (2.4) can be shown to be the solution of the following optimization problem:

$$\begin{aligned} & \text{minimize} && \frac{1}{n} \sum_{i=1}^n (\mathbf{b}_N^\top(X_i) \boldsymbol{\beta}_N - Y_i)^2 \text{ with respect to } \boldsymbol{\beta}_N \\ & \text{subject to :} && \mathbf{A}_N \boldsymbol{\beta}_N \geq 0, \sum_k |\beta_k| \leq L_N \end{aligned} \quad (2.8)$$

where $\boldsymbol{\beta}_N$ is a $(N+1) \times 1$ vector of regression coefficients $\{\beta_k\}$, and $\mathbf{b}_N(\mathbf{x})$ is the Bernstein polynomial basis vector as defined in (2.4). $\mathbf{A}_N$ is a suitably chosen matrix of full row rank that is constructed to preserve desired shape restrictions. The above optimization problem can be effectively solved by the general quadratic programming (Goldfarb and Idnani 1982, 1983).

Quadratic programming has been used to impose the necessary shape restrictions based on smoothing-spline methods. In this study, we use the available R package *quadproc* developed by Turlach and Weingessel (2010) to solve quadratic programming problem. It is to be noted that in practice the L_N can be chosen to be a reasonably large number and hence we do not need to choose its value to compute β_N .

In order to choose a proper order of Bernstein polynomials N , we used the popular V -fold cross-validation method. It is one of the most widely used methods to estimate prediction error (Wang and Li 2008). Given V , for each N , the cross validation term $CV(N)$ for our BP estimator takes the following form:

$$CV(N) = \frac{1}{V} \sum_{v=1}^V \sum_{i \in I_{-v}} (y_i - \tilde{B}_N(x_i, v))^2 \quad (2.9)$$

where $\tilde{B}_N(x, v) = \mathbf{b}_N^T(x) \tilde{\beta}_N(v)$ with $\tilde{\beta}_N(v)$'s obtained from the v -th training data set consisting of $\lfloor n \frac{V-1}{V} \rfloor$ observation points and I_{-v} denotes the corresponding validation set consisting of $\lfloor n/V \rfloor$ points. We computed the cross validation function $CV(N)$ defined in (2.9) for a series of N values starting with $N = 2$ to a relatively large integer $N_{max} (< \lfloor n \frac{V-1}{V} - 1 \rfloor)$. The integer N_{max} depends on the data in the sense that it is chosen to be the maximum integer for which the matrix $\sum_{i=1}^n \mathbf{b}_N(x_i) \mathbf{b}_N^T(x_i)$ remains empirically invertible in terms of its condition number (see Section 3.3 of the book by Golub and Van Loan (1996)). The optimal value $\hat{N}$ is chosen to minimize (2.9), i.e., $\hat{N} = \arg \min_{N \in [2, N_{max}]} CV(N)$.

In general, the predictors may not be observed to lie in the domain $[0, 1]$. To satisfy the domain restriction, we use the “linear” transformation defined as follows for all of our empirical applications,

$$\tilde{x}_i = \frac{x_i - (x_{(1)} - s)}{(x_{(n)} + s) - (x_{(1)} - s)}, \quad (2.10)$$

where $x_{(1)} = \min_{1 \leq i \leq n} x_i$ and $x_{(n)} = \max_{1 \leq i \leq n} x_i$ denote the minimum and maximum order statistics of the predictor, and s represents the sample standard deviation of $\{x_i : i = 1, \dots, n\}$.

The above transformation allows us to predict at values outside the observed range of predictor variable. Notice that often a nonlinear transformation can be applied to map an unbounded region to the unit interval $[0, 1]$, however such a nonlinear transformation may no longer satisfy the desired shape constraint of the regression function when viewed as a function of original x -values.

In (2.7), we provide the neat form of $\mathbf{A}_N$ defined for monotone and concave restriction. In practice, we may replace the neat $\mathbf{A}_N$ by its “loose” version as follows:

$$\tilde{\mathbf{A}}_N = \begin{pmatrix} \mathbf{A}_N^{(1)} \\ \mathbf{A}_N^{(2)} \end{pmatrix}.$$

where $\mathbf{A}_N^{(1)}$ has the same form of restriction matrix in (2.5) to ensure monotonicity, and $\mathbf{A}_N^{(2)}$ has the same form of restriction matrix in (2.6) to guarantee convexity. In other words, we can simply combine both restriction matrices by rows which satisfy monotone and convex assumptions respectively to produce a new restriction matrix $\tilde{\mathbf{A}}_N$ for an increasing and convex curve. The alternative matrix $\tilde{A}_N$ is not of full row rank, but it has the same rank as that of its tight-form matrix A_N . The requirement of full row rank basically ensures that set of inequality constraints no longer can be reduced, which results into the tight-form A_N . However, it is to be noted that the alternative (expanded) form of $\tilde{A}_N$ still need to have the same rank as that of the tight form (otherwise some constraints are simply redundant). This alternative matrix $\tilde{\mathbf{A}}_N$ and its neat version $\mathbf{A}_N$ work equally well within the quadratic programming even though the restriction matrix $\tilde{\mathbf{A}}_N$ is no longer of full row rank (see Meyer (1999) on the irreducibility of $\mathbf{A}_N$).

2.3 Asymptotic Properties

In this section, we present the asymptotic properties of our nonparametric regression estimator $m_N(\cdot)$ where $N = o(n^k)$ for a suitable choice of $k > 0$. Consider again the function space $\mathcal{F}$,

and the true regression function m that minimizes the L_2 risk among all possible measurable functions $f \in \mathcal{F}$, i.e., $m(\cdot)$ satisfies (2.2). Recall that in Sections 2.2.1, we provided several illustrations of $\mathcal{F}$ with various shape restrictions and their corresponding sieves $\mathcal{F}_N$. The estimator $m_N(\cdot)$ minimizes the empirical L_2 risk within the sieve $\mathcal{F}_N$ and satisfies (2.4). The sieve $\mathcal{F}_N$ has the following properties:

Property 2.3.1. *The sequence of function spaces $\mathcal{F}_N$ is nested in $\mathcal{F}$, i.e., $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots \subset \mathcal{F}_N \subset \dots \subset \mathcal{F} \subset L_2[0, 1]$.*

Proof. To establish the stated property it is sufficient to show that:

- (1) $\mathcal{F}_N \subset L_2[0, 1]$, $\forall N \in \mathbb{N}$, since Bernstein basis polynomials $b_k(x, N) = \binom{N}{k} x^k (1-x)^{N-k}$ and their linear combinations belong to $L_2[0, 1]$.
- (2) $\mathcal{F}_N \subset \mathcal{F}_{N+1}$, $\forall N \in \mathbb{N}$. Following the iterative property of the Bernstein polynomials (Lorentz 1986): $\forall k, b_k(x, N-1) = \frac{N-k}{N} b_k(x, N) + \frac{k+1}{N} b_{k+1}(x, N)$, we rewrite $B_N(x)$ as:

$$B_N(x) = \sum_{k=0}^{N+1} \tilde{\beta}_k \cdot b_k(x, N+1) = \tilde{B}_{N+1}(x),$$

where $\tilde{\beta}_0 = \beta_0$, $\tilde{\beta}_{N+1} = \beta_N$, and $\tilde{\beta}_k = \beta_k \frac{N+1-k}{N+1} + \beta_{k-1} \frac{k}{N+1}$, $k = 1, \dots, N$. To show $\mathcal{F}_N \subset \mathcal{F}_{N+1}$, it is equivalent to show for any $B_N(x) \in \mathcal{F}_N$, it also satisfies $B_N(x) \in \mathcal{F}_{N+1}$.

(a) *Monotonicity:* For any $B_N(x) \in \mathcal{F}_N$ with $\mathbf{A}_N$ defined in (2.5), we have $\beta_0 \leq \dots \leq \beta_N$. Therefore, $\tilde{\beta}_0 = \beta_0 \leq \tilde{\beta}_1 = \beta_1 - \frac{1}{N+1}\beta_1 + \frac{1}{N+1}\beta_0$, and $\tilde{\beta}_N = \beta_N - \frac{N}{N+1}(\beta_N - \beta_{N-1}) \leq \beta_N = \tilde{\beta}_{N+1}$. Finally, for $k = 2, \dots, N-1$, $\tilde{\beta}_k - \tilde{\beta}_{k-1} = \frac{(N+1-k)}{N+1}(\beta_k - \beta_{k-1}) + \frac{(k-1)}{N+1}(\beta_{k-1} - \beta_{k-2}) \geq 0$. Hence, $\tilde{\beta}_0 \leq \dots \leq \tilde{\beta}_{N+1}$, and $B_N(x) = \tilde{B}_{N+1}(x) \in \mathcal{F}_{N+1}$.

(b) *Convexity:* For any $B_N(x) \in \mathcal{F}_N$ with $\mathbf{A}_N$ defined in (2.6), we have $\beta_{k+2} - 2\beta_{k+1} + \beta_k \geq 0$, $k = 0, \dots, N-2$. Therefore, $\tilde{\beta}_2 - 2\tilde{\beta}_1 + \tilde{\beta}_0 = \frac{N-1}{N+1}(\beta_2 - 2\beta_1 + \beta_0) \geq 0$, and $\tilde{\beta}_{N+1} - 2\tilde{\beta}_N + \tilde{\beta}_{N-1} = \frac{N-1}{N+1}(\beta_N - 2\beta_{N-1} + \beta_{N-2}) \geq 0$. Finally, for $k = 1, \dots, N-2$, $\tilde{\beta}_{k+2} - 2\tilde{\beta}_{k+1} + \tilde{\beta}_k = \frac{N-(k+1)}{N+1}(\beta_{k+2} - 2\beta_{k+1} + \beta_k) + \frac{k}{N+1}(\beta_{k+1} - 2\beta_k + \beta_{k-1}) \geq 0$. Therefore $B_N(x) = \tilde{B}_{N+1}(x) \in \mathcal{F}_{N+1}$. $\square$

Property 2.3.1 can be verified for other shape constraints in a similar manner by using the iterative property of BPs.

Property 2.3.2. $\bigcup_{N=1}^{\infty} \mathcal{F}_N$ is dense in $\mathcal{F}$ with respect to sup-norm.

Proof. If $\mathcal{F} = \{f \in C[0, 1] : f(x_1) \leq f(x_2), \forall 0 \leq x_1 \leq x_2 \leq 1\}$, then $f(x)$ is nondecreasing. We take $\hat{\beta}_k = f(\frac{k}{N})$, for $k = 0, 1, \dots, N$, and have $\hat{\beta}_0 \leq \hat{\beta}_1 \leq \dots \leq \hat{\beta}_N$. If $\mathcal{F} = \{f \in C[0, 1] : 2f(\frac{x_1+x_2}{2}) \leq f(x_1) + f(x_2), \forall x_1, x_2 \in [0, 1]\}$, then $f(x)$ is convex for $x \in [0, 1]$. Taking $\hat{\beta}_k = f(\frac{k}{N})$, it is easy to see that $\hat{\beta}_{k+1} - \hat{\beta}_k \leq \hat{\beta}_{k+2} - \hat{\beta}_{k+1}$ for $k = 0, 1, \dots, N-2$.

Now define $\hat{B}_N(x) = \sum_{k=0}^N \hat{\beta}_k \cdot b_k(x, N) = \sum_{k=0}^N f(\frac{k}{N}) \cdot b_k(x, N) \in \mathcal{F}_N \subset \bigcup_{j=1}^{\infty} \mathcal{F}_j$. By Stone-Weierstrass approximation theorem, $\hat{B}_N(x)$ converges uniformly to $f(x)$ (Lorentz 1986), and this completes the proof. $\square$

In the following lemma, we state the result which provides a bound for the L_2 risk between m and m_N . This result easily follows from Lemma 10.1 in Györfi et al. (2002) and hence the proof is omitted.

Lemma 2.3.1. Let $\mathcal{F}_N = \mathcal{F}_N(\mathcal{D}_n)$ be a class of functions $f_N : \mathbb{R} \rightarrow \mathbb{R}$ depending on the data $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ as defined in (2.3). If m_N satisfies equation (2.4), then

$$\begin{aligned} & \int \{m_N(x) - m(x)\}^2 \mu(dx) \\ & \leq 2 \sup_{B_N \in \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(X_i) - Y_i)^2 - \mathbf{E}\{(B_N(X) - Y)^2\} \right| \\ & \quad + \inf_{B_N \in \mathcal{F}_N} \int (B_N(x) - m(x))^2 \mu(dx), \end{aligned}$$

where μ denotes the distribution of X .

Next, we state the result in Lemma 2.3.2 which provides sufficient conditions to establish the weak and strong universal consistency of regression function estimators. The result is adapted from Theorem 10.2 in Györfi et al. (2002). The complete proof is provided in the Appendix.

Lemma 2.3.2. Let $\mathcal{F}_N$ be the sieve defined in (2.3) and $m_N(\cdot)$ be the estimator defined in (2.4). Let T_L denote the truncation operation $T_L y = y \cdot I(|y| \leq L) + L \cdot \text{sign}(y) \cdot I(|y| > L)$. Then $Y_L = T_L(Y)$ represents the truncated version of Y , and $T_L \mathcal{F}_N = \{T_L f : f \in \mathcal{F}_N\}$ is a class of

truncated functions. When $L_N > 0$ and $L_N \rightarrow \infty$,

$$(a) \text{ If } \lim_{N \rightarrow \infty} \inf_{B_N \in \mathcal{F}_N} \int (B_N(x) - m(x))^2 \mu(dx) = 0 \text{ a.s.}, \quad (2.11)$$

$$\begin{aligned} & \lim_{N \rightarrow \infty} \sup_{B_N \in \mathcal{T}_L \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(X_i) - Y_{i,L})^2 - \mathbf{E}(B_N(X) - Y_L)^2 \right| \\ & = 0 \text{ a.s.}, \end{aligned} \quad (2.12)$$

$$\text{then } \lim_{N \rightarrow \infty} \int (m_N(x) - m(x))^2 \mu(dx) = 0 \text{ a.s.}$$

$$(b) \text{ If } \lim_{N \rightarrow \infty} \mathbf{E} \left\{ \inf_{B_N \in \mathcal{F}_N} \int (B_N(x) - m(x))^2 \mu(dx) \right\} = 0, \quad (2.13)$$

$$\begin{aligned} & \lim_{N \rightarrow \infty} \mathbf{E} \left\{ \sup_{B_N \in \mathcal{T}_L \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(X_i) - Y_{i,L})^2 - \mathbf{E}(B_N(X) - Y_L)^2 \right| \right\} \\ & = 0, \end{aligned} \quad (2.14)$$

$$\text{then } \lim_{N \rightarrow \infty} \mathbf{E} \left\{ \int (m_N(x) - m(x))^2 \mu(dx) \right\} = 0.$$

Using Lemma 2.3.1 and Lemma 2.3.2, we show that if a set of mild regularity conditions are satisfied, the proposed estimator is both weakly and strongly universally consistent. For $\mathcal{F}_N$ defined in (2.3) and the proposed estimator defined in (2.4), the required regularity conditions are described in the following theorem. The complete proof is provided in the Appendix.

Main Theorem. *Let $\mathcal{F}_N$ be the sieve space defined in equation (2.3) satisfying Properties 2.3.1 and 2.3.2, and $m_N(\cdot)$ be the restricted regression function estimator given by equation (2.4). Further, let R_N denotes the rank of the restriction matrix $\mathbf{A}_N$ which is assumed to be of full row-rank.*

(a) *If $\mathbf{E}(Y^2) < \infty$, and R_N and L_N satisfy*

$$R_N \rightarrow \infty, L_N \rightarrow \infty, \text{ and } \frac{R_N L_N^4 \log L_N}{n} \rightarrow 0, \quad (2.15)$$

then $\mathbf{E} \left[\int (m_N(x) - m(x))^2 \mu(dx) \right] \rightarrow 0$ as $n \rightarrow \infty$, i.e., the estimator $m_N(\cdot)$ is weakly universally consistent for $m(\cdot)$.

(b) If $\mathbf{E}(Y^2) < \infty$, R_N and L_N satisfy condition (2.15), and additionally for some $\delta > 0$, L_N satisfies

$$\frac{L_N^4}{N^{1-\delta}} \rightarrow 0, \quad (2.16)$$

then $\int (m_N(x) - m(x))^2 \mu(dx) \rightarrow 0$ a.s. as $n \rightarrow \infty$, i.e., the estimator $m_N(\cdot)$ is strongly universally consistent for $m(\cdot)$.

Notice that for most applications, $R_N = O(N)$ (as illustrated in Section 2.1) and if we choose $L_N = O(N^\alpha)$ and $N = O(n^k)$ then the conditions (a) and (b) of the above theorem are satisfied if $\max\{4\alpha, k + 4k\alpha\} < 1$. In particular, for any $\alpha \in (0, \frac{1}{4})$ if we choose $k \leq \frac{1}{2}$, then the above stated sufficient condition is satisfied. Clearly, these theoretical asymptotic orders for N are not directly usable in practice as these rates involve unknown constants that depend on the underlying true regression function. Following usual practice, we have thus used V -fold cross-validation method to select N and the selected values seem to agree with these theoretical rates.

2.4 Simulation Studies

In this section, we present several scenarios using simulated data to validate the performance of our restricted Bernstein polynomial estimator (restricted-BP) as compared to other available estimation methods. We consider both monotonic regression scenario and concave regression scenario.

2.4.1 Monotone Regression

We compare the performance of monotonic estimators for three different regression functions: the isotonic smoothing spline estimator (IsoSS) proposed by Wang and Li (2008), the strictly monotone estimator¹ (MP) proposed by Dette et al. (2006), and our proposed restricted Bern-

¹The estimator can be obtained through “mono.1d” function in the monoProc package.

stein Polynomial estimator (ResBP). We use the same isotonic examples as used in Wang and Li (2008):

$$\begin{aligned}
\text{Logistic:} \quad m(x) &= \frac{1}{1 + \exp(-x)}, \\
\text{Cubic:} \quad m(x) &= \frac{1}{10^3}x^3, \\
\text{Step function:} \quad m(x) &= \begin{cases} 0 & -10 \leq x \leq -3, \\ 0.2 & -3 < x \leq 0, \\ 0.5 & 0 < x \leq 5, \\ 0.8 & 5 < x \leq 8, \\ 1 & 8 < x \leq 10. \end{cases}
\end{aligned}$$

The first two functions are continuous on the support $[-10, 10]$ while the third one has discontinuities at finite number of points. It is well known that the Bernstein polynomial approximates the function at any point of discontinuity by taking the average of two one-sided limits of the function (see Theorems (5.1) and (5.5) in Herzog and Hill (1946)). Therefore, our proposed Bernstein polynomial estimator can still estimate the above step function quite well. The predictor variable is simulated from the uniform distribution on $[-10, 10]$. The performances of three estimators are compared under various noise distributions: normal, scaled t with 10 degrees of freedom and shifted and scaled beta(3,2). Note that noise distributions are centered and scaled to have mean zero and standard deviation $\sigma = 0.4$. A sample of $n = 50$ observations is generated to estimate the regression function using all three methods, and the data generation and subsequent estimations are repeated 1000 times for each simulation setting. The order of the Res.BP was estimated using 7-fold cross validation method.

The performance of the estimators is evaluated by two criteria: the integrated mean squared error (IMSE) and the integrated mean squared prediction error (IMSPE) (Hansen 2010) defined

as follows:

$$\begin{aligned}\text{IMSE} &= \frac{1}{n} \sum_{i=1}^n (\hat{m}(x_i) - m(x_i))^2, \\ \text{IMSPE} &= \frac{1}{n} \sum_{i=1}^n (\tilde{m}_{-i}(x_i) - y_i)^2,\end{aligned}$$

where m denotes the true function, $\hat{m}$ denotes the estimated function obtained by different methods, and $\tilde{m}_{-i}$ denotes the leave-one-out estimated function. The comparative results averaged over 1000 simulations are summarized in Tables 2.1 and 2.2. The IMSE in this chapter has the same form of definition as the MSE in Wang and Li (2008). Since the programming package of the IsoSS method is not available in public, we directly reproduced IMSE (though they called MSE) values from Wang and Li (2008)’s Table 1, and ignore this method when using the IMSPE criteria.

[insert Table 2.1 here.]

[insert Table 2.2 here.]

Table 2.1 compares estimation accuracy in terms of IMSE. From the results in Table 2.1, it is clearly evident that our proposed ResBP method outperforms the other two methods when the true regression functions are cubic and step functions under each noise distribution in favor of the ResBP estimator. When the true regression function is the logistic function, the IsoSS method performs slightly better than the other two estimators under normally distributed and t-distributed noises. However, when the noise follows the beta distribution, the MP estimator has the minimum average IMSE, and both MP and ResBP perform significantly better than IsoSS. Notice that differences of average IMSE’s among the three methods are not statistically significant for all three regression functions when the noise is not beta-distributed.

Table 2.2 presents the prediction accuracy of our proposed estimator and the MP estimator in terms of IMSPE. When the true regression functions are cubic and step functions, we see that our ResBP estimator performs substantially better in terms of having smaller IMSPE’s

than the MP estimator for all three noise distributions under consideration. We also conducted a two-sample t-test to test the equivalence of the IMSPE's across 1000 simulated data sets, and found that the differences of IMSPE's are statistically significant. When the true regression function is logistic, the IMSPE's of our estimator are slightly larger, but the increases remain statistically insignificant. The last column in Table 2.2 presents the percentage of times the IMSPE of ResBP appears smaller than that of MP over 1000 simulations. It is clearly evident that ResBP produces smaller IMSPE's in majority of the cases.

2.4.2 Increase-concave Regression

In the second part of simulation study, we consider an increasing-concave regression function to illustrate the consequence of various shape restrictions under a misspecified model. We consider five estimators: (i) BP estimator with monotone and concave restrictions, (ii) BP estimator with monotone restriction only, (iii) BP estimator with concave restriction only, (iv) the unconstrained BP estimator, and (v) the least squares estimator using a simple linear regression (SLR) model. Note that the first three are all restricted-BP estimators with different shape assumptions. The true regression function that we used in the simulation study is the same as the example we illustrated in the introduction, given by

$$m(x) = 0.5x^2 \log(x) - 1.75x^2 + 3x, \quad x \in [0, 1]$$

and the predictor variable is randomly generated from a standard uniform distribution $U(0, 1)$. The errors in the regression model are assumed to be normally distributed with mean zero and standard deviation $\sigma = 0.5$ and 0.1 . We use sample size $n = 50$. The data generation and subsequent estimations were repeated 1000 times. Figure 2.1 in the introduction section gives the true function described above as well as one randomly picked sample of observations. The average IMSE's over 1000 simulations were computed and reported in Table 2.3. Under both noise conditions, the performances of the unconstrained BP estimator and the SLR estimator are obviously poorer compared with all three restricted-BP estimators. In terms of the differences

among three restricted-BP estimators, we observe that when $\sigma = 0.5$, the estimator with both monotone and concave restrictions achieved the minimum IMSE among the five estimators. The other two restricted estimators have slightly larger IMSE's, although these increases in IMSE values are statistically significant compared with the minimum one. We conclude that in this study a correct set of shape assumptions improve estimation accuracy when the standard deviation of error is considerably large. On the other hand, the concave+increasing and concave-only estimators show similar behaviors in terms of IMSE under a mild noise condition (i.e., $\sigma = 0.1$).

[insert Table 2.3 here.]

We also investigate box plots of predicted responses obtained from the five estimators in Figure 2.2. Since $\sigma = 0.5$ and 0.1 provide similar patterns, we only present the results when $\sigma = 0.5$. From the plot, we observe that the true regression function $m(x)$ is always bounded by the first and third quartiles of predicted response values obtained from those BP methods. When the predictor variable x approaches zero, the predicted responses have more variability than those in other regions. This might be due to the instability of $\log x$ when x is near 0. The predicted response values by the BP estimator without any restriction show more variability than those from other estimators. We also notice that predicted values from SLR are biased in many regions as expected. The bottom right panel in Figure 2.2 presents the local root mean square error $\text{RMSE} = \sqrt{\sum_{l=1}^{1000} (\hat{m}^{(l)}(x) - m(x))^2 / 1000}$ at every individual x -value to evaluate the local variability. As we expect, the restricted estimator with both increasing and concave constraints achieves minimal RMSE in general. Other two restricted BP estimators produce larger RMSE's. The BP estimator without any restriction has the largest RMSE on average. This finding is consistent with the observation of box plots in Figure 2.2. Thus, we conclude that appropriate background information about the shape of the regression function can provide significantly better estimator than those without such subject-matter information.

[insert Figure 2.2 here.]

2.5 Real Data Examples

2.5.1 Fuel Efficiency Study using ASA Car Data

Our first example is part of the “cars” data from the 1983 ASA Data Exposition. The data is accessible through the StatLib Internet site (<http://lib.stat.cmu.edu/datasets/cars.data>) at Carnegie Mellon University. The data set includes $n = 406$ different cars. The fuel efficiency (y) is studied as a function of the engine’s horsepower (x) among cars. The data is displayed as a scatter plot in Figure 2.3. This data set has been studied by many researchers including Mammen et al. (2001) and Wang and Li (2008) using constrained and unconstrained smoothing methods. Subject matter knowledge assures that more powerful engines consume more fuel and hence a decreasing constraint should be proper to impose.

We now illustrate a possible application of the proposed BP method and construct estimators with different constraints. The order of polynomial N is chosen by 5-fold least squares cross validation. We considered five different estimators: the MP estimator by Dette et al. (2006), one unconstrained BP estimator, three constrained BP estimators with convex and decreasing, decreasing only, as well as convex only restrictions. In Figure 2.3, we observe that three constrained BP estimators perform very similar for this data set. When the engine power is larger than 200, the fitted value obtained from the convex constrained BP estimator starts increasing. As expected, the unconstrained BP estimator does not preserve either decreasing property or convex property. The MP estimator performs comparatively well as the BP estimator with decreasing constraint, although the BP estimator exhibits more smoothness. We also report the coefficient of determination ($R^2 = (\sum_{i=1}^n (y_i - \bar{y})\hat{m}(x_i))^2 / \sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{m}(x_i) - \bar{\hat{m}})^2$) and the IMSPE for each estimator in Figure 2.3. As expected, the unconstrained estimator gives the largest R^2 . It is interesting to find that in this example, the unconstrained estimator also produces the smallest IMSPE.

[insert Figure 2.3 here.]

2.5.2 Growth Curve Study using Rabbit Data

We analyze another data set to explore the relation between eye lens weight and age for rabbits in Australia. The data as well as the detailed description of the experiment is accessible through the website (<http://www.statsci.org/data/oz/rabbit.html>). The sample size in this data set is $n = 71$. The data was analyzed by Birke and Dette (2007) first to fit a nonparametric concave regression curve to estimate eye lens weight for rabbits as a function of age. However, for values of the age < 400 days, their concave nonparametric estimator seems to underestimate the eye lens weight slightly (see Figure 6 in Birke and Dette (2007)).

We present the performance of four ResBP estimators: (i) increasing and concave; (ii) concave only; (iii) increasing only and (iv) unconstrained; by plotting the estimated functions in Figure 2.4. The MP method by Dette et al. (2006) is included in the same figure as well. Three constrained BP estimators show similar performances, and they all fit the data adequately and preserve the desired increasing and concave restrictions. The unconstrained BP estimator suggests non-monotone behavior when the age is greater than 400 days, and this may be due to less frequency of data points in the region of age > 400 . The MP method does not fit well in both boundary areas. We also present R^2 and the IMSPE for four estimating methods. As we expect, the unconstrained BP estimator still has the largest R^2 as it is obtained by minimizing over a larger space of functions containing the constrained spaces. But notice that R^2 obtained from the increasing + concave BP estimator reduces only by 0.051% as compared to that for the unconstrained BP estimator. However, in this data set with only moderately large sample size, the unconstrained BP estimator and the MP method have comparatively very large IMSPEs of 159.13 and 130.62, which indicates poor predictive performances. Other three constrained BP estimators perform similarly, and their IMSPE's are less than 70.6, resulting in a reduction of 55.63% of IMSPE as compared to the unconstrained BP estimator. In this example, the BP estimator with both increasing and concave constraints has the highest prediction accuracy with the minimum *IMSPE* of 68.27, though this gain in efficiency may not be statistically significant from other two restricted BP estimators.

[insert Figure 2.4 here.]

2.6 Conclusion

We proposed a general framework to develop a nonparametric estimator for a regression function subject to a wide variety of shape restrictions. The framework is based on creating a nested sieve of Bernstein polynomials each of which satisfy the desired shape constraints as the degree of the polynomial (slowly) increases with the sample size. The proposed estimator can be easily obtained as the solution of a quadratic programming problem for a fixed degree of the polynomial. We then used a cross validation method to select the optimal order of polynomials. The estimator preserves the required shape restrictions and smoothness simultaneously. The shape restrictions include increasing, decreasing, concave, convex functions and their combinations. The theoretical results provide asymptotic validity by establishing the consistency of the estimator under a mild set of regularity conditions requiring only continuity of the underlying true regression function. Furthermore, the empirical results based on simulated data suggest that our method provides high prediction accuracy especially when the sample size is small (e.g., $n = 50$). Real data applications show that including more (but correct) shape assumptions may not always produce better estimation accuracy. However, the gain in predictive accuracy would be substantial when the true curve indeed satisfies the restriction.

Table 2.1: Average integrated mean squared errors (IMSE) over 1000 repetitions. The sample size is $n = 50$ and all the noise distributions are centered and scaled to have mean zero and standard deviation $\sigma = 0.4$. Standard error of IMSE is displayed in the parentheses.

Simulation functions		Estimators		
True functions	Noise	Iso.-SS	MP	Res.-BP
Logistic	Normal	0.0151	0.0207 (0.0003)	0.0249 (0.0003)
	t_{10}	0.0151	0.0205 (0.0003)	0.0253 (0.0003)
	beta(3,2)	0.1721	0.0174 (0.0003)	0.0220 (0.0002)
Cubic	Normal	0.0160	0.0188 (0.0003)	0.0133 (0.0003)
	t_{10}	0.0156	0.0195 (0.0003)	0.0135 (0.0003)
	beta(3,2)	0.1801	0.0159 (0.0003)	0.0103 (0.0002)
Step function	Normal	0.0175	0.0175 (0.0004)	0.0155 (0.0003)
	t_{10}	0.0174	0.0182 (0.0004)	0.0162 (0.0004)
	beta(3,2)	0.1735	0.0142 (0.0003)	0.0130 (0.0003)

Table 2.2: Average integrated mean squared prediction errors (IMSPE) over 1000 repetitions. The sample size is $n = 50$ and all the noise distributions are centered and scaled to have mean zero standard deviation $\sigma = 0.4$. Standard error of IMSPE is displayed in the parentheses. P-values are obtained based on two-sample t-tests.

Simulation scenarios		IMSPE			
True functions	Noise	MP	ResBP	P-value	ResBP<MP(%)
Logistic	Normal	0.1821	0.1832	0.5535	50.4
		(0.0012)	(0.0012)		
	t_{10}	0.1828	0.1836	0.2091	51.5
		(0.0004)	(0.0005)		
	beta(3,2)	0.1823	0.1832	0.1633	50.7
		(0.0004)	(0.0005)		
Cubic	Normal	0.1805	0.1716	< 0.0001	85.8
		(0.0012)	(0.0011)		
	t_{10}	0.1799	0.1714	< 0.0001	86.2
		(0.0004)	(0.0004)		
	beta(3,2)	0.1805	0.1717	< 0.0001	86.0
		(0.0004)	(0.0003)		
Step function	Normal	0.1779	0.1706	< 0.0001	84.3
		(0.0012)	(0.0012)		
	t_{10}	0.1776	0.1705	< 0.0001	82.3
		(0.0004)	(0.0004)		
	beta(3,2)	0.1823	0.1711	< 0.0001	78.5
		(0.0004)	(0.0004)		

Table 2.3: Average integrated mean squared errors (IMSE) over 1000 repetitions. The sample size is $n = 50$ and the noise has standard deviation $\sigma = 0.5$ and 0.1. Standard error of IMSE is displayed in the parentheses. P-values are obtained by using two-sample t-tests between the BP method and the rest methods.

		concave+increase	concave	increase	unconstraint	
σ		BP	BP	BP	BP	SLR
0.5	MSE	0.0140	0.0168	0.0161	0.0317	0.0195
	S.E.	(0.0004)	(0.0004)	(0.0004)	(0.0007)	(0.0003)
	t-test		-4.8802	-3.7889	-22.9597	-11.2205
	p-value		< 0.0001	0.00015	< 0.0001	< 0.0001
0.1	MSE($\times 10^4$)	7.7452	8.0786	8.7322	99.6084	19.0479
	S.E.($\times 10^4$)	(.1661)	(.1811)	(.1871)	(.2931)	(.1253)
	t-test		-1.3564	-3.945	-460.0796	-26.8255
	p-value		0.175	< 0.0001	< 0.0001	< 0.0001

Figure 2.1: An increasing-concave curve (solid), and observed data (scatter plot).

Figure 2.2: Box plots of predicted responses. The sample size is $n = 50$ and the noise has standard deviation $\sigma = 0.5$.

Figure 2.3: ASA car fuel efficiency as a function of engine output. Top left: Convex and decreasing BP estimator (with $\hat{N} = 12$). Top middle: Convex BP estimator (with $\hat{N} = 17$). Top right: Decreasing BP estimator (with $\hat{N} = 17$). Bottom left: Unconstrained BP estimator (with $\hat{N} = 7$). Bottom middle: MP decreasing estimator.

Figure 2.4: Eye lens weight for rabbits as a function of age. Top left: Concave and increasing BP estimator (with $\hat{N} = 9$). Top middle: Concave BP estimator (with $\hat{N} = 9$). Top right: Increasing BP estimator (with $\hat{N} = 10$). Bottom left: Unconstrained BP estimator (with $\hat{N} = 4$). Bottom middle: MP increasing estimator.

Chapter 3

Shape Restricted Nonparametric Regression with Multivariate Bernstein Polynomials

3.1 Introduction

Statistical regression methods are often used to explore the inherent relationship between several predictor (or explanatory) variables and a response (or dependent) variable. Once the regression function or the curve has been estimated, the prediction of future responses is straightforward. In many practical settings, the predictors and the response are known to preserve certain shape restrictions (e.g., monotonicity, concavity etc.) but not necessarily based on a parametric form. Some popular examples include the study of utility functions, cost functions, and profit functions in economics (Gallant and Golub 1984; Terrell 1996), the analysis of temperature as a function of various environmental factors, the study of dose response curve in the phase I clinical trials, the estimation of the hazard rate and the failure rate in reliability and survival analysis, among others (Chang et al. 2007; Molitor and Sun 2002). Here we present two examples in Figure 3.1 to illustrate regression functions with certain shape restrictions. The first one, depicted in

Figure 3.1(a) corresponds to a regression function of two predictors which is constrained to be increasing in one dimension while it is constrained to be concave in the other dimension. The second one in Figure 3.1(b) shows another example of a regression function which is restricted to follow a nondecreasing (and concave) trend in both dimensions.

[insert Figure 3.1 here.]

Over the past decades, efforts have been devoted to constructing a smooth and computationally efficient estimator of a shape restricted regression function. In the case of a single predictor variable, Hildreth (1954) pioneered a method to estimate a regression function under the restriction of concavity. This well known method is based on estimating a step function subject to restrictions on the steps using the maximum likelihood method, which resulted into the so-called Pool Adjacent Violators algorithm (PAVA) (Barlow et al. 1972). A variety of the smoothed PAVAs were developed subsequently by many researchers (Friedman and Tibshirani 1984; Mukerjee 1988) to obtain smoothed estimators. Other approaches include the shape constrained smoothing spline methods (Ramsay 1988; He and Shi 1998; Meyer 2008; Wang and Li 2008), kernel methods subject to shape constraint (Hall and Huang 2001; Aït-Sahalia and Duarte 2003; Dette et al. 2006; Birke and Dette 2007), projection methods (Mammen et al. 2001), polynomial basis estimators (Chang et al. 2007; Curtis and Ghosh 2011), and many alike. Most of these methods have been well studied to estimate only a monotone regression function with a single predictor. Extensions of these methods to other shape constraints (e.g., convexity) may not be straightforward except for few attempts (Birke and Dette 2007; Pal, Woodroffe and Meyer 2007; Curtis and Ghosh 2011).

The literature for shape restricted regression problems with more than one predictor is comparatively scarce possibly due to the computational difficulties. The literature begins with Brunk (1955) who derives maximum likelihood estimator for monotone parameters in the exponential family of distributions. The estimator by Brunk (1955) however fails to guarantee monotonicity when the values of predictor variables are not observed in the original data set. Villalobos and Wahba (1987) proposed inequality-constrained thin-plate splines, and

applied the method to estimate a bivariate distribution which is strictly concave in one dimension and monotone in the other dimension. As pointed out by Mammen et al. (2001), the above spline smoother is not guaranteed to satisfy the shape constraints everywhere in the support of predictor variables. Borrowing the standard PAVA algorithm as well as some other algorithms, Bacchetti (1989) defined the cyclic PAVA algorithm for fitting additive isotonic regression models. Although the proposed method extends its application in generalized linear models, it inherits some drawbacks of the PAVA algorithm. For instance, the resulting estimator is not necessarily smooth and often step-function type features can be observed with the fitted curve. More recently, Bollaerts, Eilers and van Mechelen (2006) presented P-splines regression with additional asymmetric discrete penalties to fit a multivariate isotonic regression curve. Beresteanu (2004) and Leitenstorfer and Tutz (2007) introduced B-spline based estimators under monotonicity constraints. Both methods appear to rely on the fact that predictors be observed an approximately equidistant grid of values. To allow for the presence of random spatial alignment of predictors values, Dette and Scheder (2007) suggested a two-stage approach to estimate multivariate regression functions which are strictly monotone in all or a subset of its arguments, and established rigorous asymptotic results of their proposed estimators. However, this method could be considerably time consuming to implement even with only moderately large sample size (e.g., with two predictors and a sample size of 400, it takes on average around 286 seconds. For the same scenario, our proposed method takes only 6.36 seconds on average and achieves similar or even better accuracy). Moreover, the method was proposed to fit only monotonic curves and it may not be straightforward to extend their method to address other popular shape constraints (e.g., concavity, convexity etc.).

In this chapter, we adopt the method of sieves, and propose a Bernstein polynomial based estimator to estimate a multivariate regression function subject to various shape restrictions (e.g., nonnegativity, monotonicity, and convexity) in all or a subset of the arguments. The suggested sieve in our method is constructed as a nested sequence of multivariate Bernstein polynomials of all degrees. Univariate Bernstein polynomial estimators have played important roles in non-

parametric curve estimation (Chang et al. 2007; Chak, Madras and Smith 2005; Stadtmuller 1986). Bivariate Bernstein polynomials estimator was first investigated by Tenbusch (1994) to estimate two-dimensional density functions. The resulting regression estimator has been shown to be universally consistent and asymptotically normal under a set of mild regularity conditions in the two-dimensional case without any shape restriction (Tenbusch 1997). Chak, Madras and Smith (2001) used bivariate Bernstein polynomials to approximate the shape-restricted regression function. The Bernstein polynomial is appropriate for shape-preserving regression, as it has the optimal shape restriction property among all polynomials (Carnicer and Pena 1993), and all of the derivatives possess the same convergence properties (Lorentz 1986). More importantly, the multivariate Bernstein polynomials can be used to approximate a function which is not required to be smooth beyond being simply continuous. With all these nice properties of Bernstein polynomials, we develop an algorithm to derive a shape-restricted multivariate Bernstein polynomial estimator and investigate its asymptotic properties. Monotone regression and convex (or concave) regression are explored in some details. In particular, the numerical performance of our approach is thoroughly examined and compared with the two-dimensional monotone estimator proposed by Dette and Scheder (2007).

The chapter is organized as follows. Section 3.2 introduces the multivariate Bernstein polynomial based model, gives a few examples of sieves under different shape restrictions, and provides some properties of the sieve constructed by varying the order of the polynomials. The asymptotic properties of the proposed nonparametric regression estimator are presented in Section 3.3 and the technical details of the proofs are deferred to the Appendix. Section 3.4 illustrates a design to simulate data for several shape constraint scenarios, and the results are presented and compared with a currently available nonparametric regression estimator subject to shape restrictions. In Section 3.5, we applied our method as well as the competing method to the infant mortality rate data obtained from World Health Organization and United Nations Development Programme Annual Report. Finally, in Section 3.6, we conclude with a brief discussion of our findings and present ideas for further extensions.

3.2 Modelling Shape Restrictions Using Multivariate Bernstein Polynomials

Consider a general regression model where we assume that $(\mathbf{X}, Y)$ is a $\mathbb{R}^d \times \mathbb{R}$ -valued random vector arising from an arbitrary distribution. Let $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ be the set of observations which are assumed to be independently and identically distributed (i.i.d) as $(\mathbf{X}, Y)$. The regression model is given by

$$Y_i = m(\mathbf{X}_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where ε_i 's are independently distributed with $E(\varepsilon_i|\mathbf{X}_i) = 0$ and $m(\mathbf{x}) = \mathbf{E}(Y|\mathbf{X} = \mathbf{x})$. Our goal is to estimate the regression function $m(\cdot)$, subject to some known shape restrictions. Let $\mathcal{F}$ be a class of smooth functions subject to a given set of shape restrictions, such as monotonicity, convexity and concavity. For example, when $d = 1$, we may have subject matter information about $m(x)$ to be a convex function. Similarly when $d = 2$, we may have scientific information which restricts $m(x_1, x_2)$ to be monotone in x_1 while convex in x_2 . Among all possible measurable functions within the class $\mathcal{F}$, the regression function obtains the minimal L_2 risk (Gyorfi et al. 2002), i.e.,

$$m(\cdot) = \arg \min_{f \in \mathcal{F}} \mathbf{E}\{(f(\mathbf{X}) - Y)^2\}. \quad (3.1)$$

In most applications $m(\cdot)$ is unknown due to the unknown distribution of $(\mathbf{X}, Y)$. Our goal is to obtain an estimator of $m(\cdot)$ based on the data $\mathcal{D}_n$.

In this chapter, we adopt the method of sieves (Grenander 1981; Geman and Hwang 1982) and propose an estimator based on multivariate Bernstein polynomials (BPs). First, we construct a sequence of sieves $\mathcal{F}_N$, which is a nested sequence of function spaces dense in the space of functions $\mathcal{F}$ with respect to a suitable metric. Additionally, each member in the function space $\mathcal{F}_N$ preserves the desired shape restriction (i.e., $\mathcal{F}_N \subseteq \mathcal{F}$, $\forall N$). The estimator m_N is

chosen from such an approximating space $\mathcal{F}_N$ having the minimum empirical L_2 risk. That is,

$$m_N \in \mathcal{F}_N \text{ and } m_N(\cdot) = \arg \min_{f_N \in \mathcal{F}_N} \frac{1}{n} \sum_{i=1}^n (f_N(\mathbf{X}_i) - Y_i)^2.$$

We will show in Section 3 that m_N is weakly universally consistent for m in the L_2 space where $N = o(n^k)$ for some $k > 0$, and with additional mild set of regularity conditions, we also establish the strong universal consistency (see Theorem 3.3.1 in Section 3.3).

The following notations are used throughout the paper. Without any loss of generality, for any d -dimensional vector $\mathbf{x} = (x_1, \dots, x_d)$, consider a continuous function $h(\mathbf{x}) = h(x_1, x_2, \dots, x_d)$ on the space $[0, 1]^d \rightarrow \mathbb{R}$. Let a d -dimensional integer index of the multivariate BP be denoted by $\mathbf{k} = (k_1, k_2, \dots, k_d)$ and a d -dimensional integer order of the multivariate BP be denoted by $\mathbf{N} = (N_1, N_2, \dots, N_d)$. For convenience, we use $\mathbf{k}/\mathbf{N}$ to denote the d -dimensional fraction $(\frac{k_1}{N_1}, \frac{k_2}{N_2}, \dots, \frac{k_d}{N_d})$. Let the index space be denoted by $\mathbb{M}_d(\mathbf{N}) = \mathbb{M}_1 \times \mathbb{M}_2 \times \dots \times \mathbb{M}_d$, where $\mathbb{M}_j = \{0, 1, \dots, N_j\}$ for $j = 1, \dots, d$. Using the above notations, $\mathbf{k} \in \mathbb{M}_d(\mathbf{N})$ represents the indices $k_j \in \{0, 1, \dots, N_j\}$ for $j = 1, \dots, d$.

We will assume throughout the paper that the predictor variables are suitably transformed to lie in the unit hypercube $[0, 1]^d$ (see Section 3.2.2 for more details). Adopting the method of sieves, we first construct the sieve $\{\mathcal{F}_N\}$ based on a class of multivariate BPs that satisfy the desired shape restrictions. Then the N^{th} sieve is described as linear combinations of multivariate BPs with a linear-combinational restriction on the coefficients. Often we can write the general form of $\mathcal{F}_N$ as follows:

$$\mathcal{F}_N = \{B_N(\mathbf{x}) \equiv \sum_{\mathbf{k} \in \mathbb{M}_d(\mathbf{N})} \beta_{\mathbf{k}} \cdot b_{\mathbf{k}}(\mathbf{x}, \mathbf{N}) : \mathbf{A}_N \beta_N \geq \mathbf{0}, \text{ and } \sum_{\mathbf{k}} |\beta_{\mathbf{k}}| \leq L_N\}, \quad (3.2)$$

where $b_{\mathbf{k}}(\mathbf{x}, \mathbf{N}) = \prod_{j=1}^d b_{k_j}(x_j, N_j)$ and $b_k(x, N) = \binom{N}{k} x^k (1-x)^{N-k}$ are the univariate Bernstein polynomials. The order of the polynomial, N_j , will be chosen as a function of the sample size n , e.g., $N_j = o(n^{k_j})$ with suitably chosen $k_j > 0$ for $j = 1, \dots, d$. To simplify the notation, we assume henceforth that $N_j = N$ for $j = 1, 2, \dots, d$ and $N = o(n^k)$ for some $k > 0$. The

bound $L_N > 0$ is chosen suitably to grow with N (see Theorem 3.3.1) to infinity. We use $\mathbf{A}_N$ to represent a full row rank restriction matrix with dimension $R_N \times (N+1)^d$ where R_N denotes the rank of $\mathbf{A}_N$. The necessity of full row rank follows from the tightness conditions described in Silvapulle and Sen (2006), but also see Meyer (1999) on the irreducibility of $\mathbf{A}_N$.

Let $\boldsymbol{\beta}_N = \{\beta_{\mathbf{k}}\}$ be the $(N+1)^d \times 1$ coefficient vector, and $\mathbf{b}_N(\mathbf{x}) = \{b_{\mathbf{k}}(\mathbf{x}, \mathbf{N}) : N_j = N, \forall j\}$ be the $(N+1)^d \times 1$ Bernstein polynomial basis vector. The expression $B_N(\mathbf{x})$ in (3.2) can therefore be written as $B_N(\mathbf{x}) = \mathbf{b}_N^\top(\mathbf{x})\boldsymbol{\beta}_N$. We will give examples of two-dimensional cases to illustrate how to construct $\mathcal{F}_N$ and $\mathbf{A}_N$ in Sections 3.2.1. For an arbitrary d -dimensional case, our examples of $\mathcal{F}_N$ can be easily extended by adopting similar but somewhat cumbersome additional notations.

Given the data set $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$, the estimator m_N is selected from the Bernstein polynomial sieve $\mathcal{F}_N$ defined in (3.2) that minimizes the empirical L_2 risk. That is:

$$m_N(\cdot) = \arg \min_{\boldsymbol{\beta}_N : \mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{0}, \sum |\beta_{\mathbf{k}}| \leq L_N} \frac{1}{n} \sum_{i=1}^n (\mathbf{b}_N^\top(\mathbf{X}_i) \boldsymbol{\beta}_N - Y_i)^2. \quad (3.3)$$

Note that the existence of m_N follows by the compactness of $\mathcal{F}_N$ and requires only a finite-dimensional optimization (see Section 3.2.2 for computational details).

3.2.1 Constraint Bernstein Polynomial Sieve ($d = 2$)

For notational simplicity we illustrate various shape constraints only for $d = 2$. However, the example developed in this section can easily be extended to higher dimensions with additional but somewhat cumbersome notations. Let $m(\mathbf{x}) = m(x_1, x_2)$ be a real-valued function of a

bivariate predictor $\mathbf{x}$ in the function space $\mathcal{F}$. The N^{th} sieve defined in (3.2) reduces to:

$$\begin{aligned}\mathcal{F}_N &= \{B_N(\mathbf{x}) \equiv \sum_{\mathbf{k} \in \mathbb{M}_d(\mathbf{N})} \beta_{\mathbf{k}} \cdot b_{\mathbf{k}}(\mathbf{x}, \mathbf{N}) : \mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{0}, \text{ and } \sum_{\mathbf{k}} |\beta_{\mathbf{k}}| \leq L_N\}, \\ &= \{B_N(\mathbf{x}) \equiv \sum_{k_1=0}^N \sum_{k_2=0}^N \beta_{k_1, k_2} \cdot b_{k_1}(x_1, N) b_{k_2}(x_2, N) : \\ &\quad \mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{0} \text{ and } \sum_{k_1, k_2} |\beta_{k_1, k_2}| \leq L_N\},\end{aligned}$$

where

$$\boldsymbol{\beta}_N = (\beta_{00}, \beta_{01}, \dots, \beta_{0N}, \beta_{10}, \beta_{11}, \dots, \beta_{1N}, \dots, \beta_{N0}, \beta_{N1}, \dots, \beta_{NN})^\top.$$

(i) *Nonnegativity*

Let $m(\mathbf{x})$ be a function in the parameter space $\mathcal{F} = \{f \in C[0, 1]^2 : f(x_1, x_2) \geq 0, \forall (x_1, x_2) \in [0, 1]^2\}$, where $C[0, 1]^2$ denotes the class of all continuous functions defined on $[0, 1] \times [0, 1] \equiv [0, 1]^2$. We define the restriction matrix with the rank $R_N = (N + 1)^2$ on the coefficients as follows:

$$\mathbf{A}_N = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}_{(N+1)^2 \times (N+1)^2} = I_{(N+1)^2}.$$

The restriction $\mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{0}$ guarantees all coefficients are nonnegative. As a result $B_N(\cdot)$ is nonnegative for any N and $\mathcal{F}_N \subset \mathcal{F}$.

(ii) *Increasing in Both Coordinates*

Let $m(\mathbf{x}) = m(x_1, x_2)$ be a real-valued function with monotonicity on both predictor variables, i.e., $\mathcal{F} = \{f \in C[0, 1]^2 : f(u_1, v_1) \leq f(u_2, v_1), f(u_1, v_1) \leq f(u_1, v_2), \forall 0 \leq u_1 \leq u_2 \leq 1, 0 \leq v_1 \leq v_2 \leq 1\}$. Notice that we have not made any additional smoothness

assumptions about the true regression functions besides continuity. Then the restriction matrix satisfying $\mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{0}$ of the sieve is represented as follows:

$$\mathbf{A}_N = \begin{pmatrix} \mathbf{A}_N^{(1)} \\ \mathbf{A}_N^{(2)} \end{pmatrix}.$$

The sub-matrix $\mathbf{A}_N^{(1)}$ ensures the monotonicity of the function with respect to the first predictor x_1 , and $\mathbf{A}_N^{(2)}$ is used similarly for the second variable x_2 . To be more specific, we write $\mathbf{A}_N^{(1)}$ as follows:

$$\mathbf{A}_N^{(1)} = \begin{pmatrix} -1 & 0 & \dots & 0 & 1 & & \\ & -1 & 0 & \dots & 0 & 1 & \\ & & & & \ddots & & \\ & & & & & & -1 & 0 & \dots & 0 & 1 \end{pmatrix}_{N(N+1) \times (N+1)^2}, \quad (3.4)$$

and write $\mathbf{A}_N^{(2)}$ as follows:

$$\mathbf{A}_N^{(2)} = \begin{pmatrix} \mathbf{B} & & & \\ & \mathbf{B} & & \\ & & \ddots & \\ & & & \mathbf{B} \end{pmatrix}_{N(N+1) \times (N+1)^2} \quad \text{with } \mathbf{B} = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots \\ & & \ddots & & \\ 0 & \dots & 0 & -1 & 1 \end{pmatrix}_{N \times (N+1)} \quad (3.5)$$

where in each row of $\mathbf{A}_N^{(1)}$, there are N 0's between -1 and 1. Submatrix $\mathbf{A}_N^{(1)}$ ensures $\beta_{i,k_2} \leq \beta_{j,k_2}$ and $\mathbf{A}_N^{(2)}$ ensures $\beta_{k_1,i} \leq \beta_{k_1,j}$ when $i \leq j$. Notice that $\mathbf{A}_N$ is a $2N(N+1) \times (N+1)^2$ matrix with rank $R_N = 2N(N+1)$. Further, notice that

$$\frac{\partial B_N}{\partial x_1} = N \sum_{k_1=0}^{N-1} \sum_{k_2=0}^N (\beta_{k_1+1,k_2} - \beta_{k_1,k_2}) b_{k_1}(x_1, N-1) b_{k_2}(x_2, N) \geq 0$$

and

$$\frac{\partial B_N}{\partial x_2} = N \sum_{k_1=0}^N \sum_{k_2=0}^{N-1} (\beta_{k_1, k_2+1} - \beta_{k_1, k_2}) b_{k_1}(x_1, N) b_{k_2}(x_2, N-1) \geq 0,$$

and hence $\mathcal{F}_N \subseteq \mathcal{F}$.

(iii) *Increasing in One Coordinate*

If $m(\mathbf{x}) = m(x_1, x_2)$ is a function with monotonicity in x_1 only, then the restriction matrix $\mathbf{A}_N$ is taken to be $\mathbf{A}_N = \mathbf{A}_N^{(1)}$, where $\mathbf{A}_N^{(1)}$ is defined in (3.4). Similarly, if $m(\mathbf{x}) = m(x_1, x_2)$ is a function with monotonicity in x_2 only, we set $\mathbf{A}_N = \mathbf{A}_N^{(2)}$, where $\mathbf{A}_N^{(2)}$ is defined in (3.5). In both scenarios, $\mathbf{A}_N$ is an $N(N+1) \times (N+1)^2$ matrix of rank $R_N = N(N+1)$. Hence, it easily follows that, $\mathcal{F}_N \subseteq \mathcal{F}$. Notice that we have not assumed any form of additivity structure of the true regression function and all types of interactions between x_1 and x_2 are allowed.

(iv) *Convexity*

Let $m(\mathbf{x}) = m(x_1, x_2)$ be a real-valued function which is convex as a function of x_1 (for fixed x_2) and as function of x_2 (for fixed x_1). In other words, the function space is given by $\mathcal{F} = \{f \in C[0, 1]^2 : 2f(\frac{u_1+u_2}{2}, v_1) \leq f(u_1, v_1) + f(u_2, v_1) \text{ and } 2f(u_2, \frac{v_1+v_2}{2}) \leq f(u_2, v_1) + f(u_2, v_2), \forall u_1, u_2, v_1, v_2 \in [0, 1]\}$. Again, we have not assumed any smoothness conditions on $m(x_1, x_2)$ in terms of its partial derivatives. Then the restriction matrix $\mathbf{A}_N$ satisfying $\mathbf{A}_N \beta_N \geq \mathbf{0}$ of the sieve is represented as follows:

$$\mathbf{A}_N = \begin{pmatrix} \mathbf{A}_N^{(1)} \\ \mathbf{A}_N^{(2)} \end{pmatrix}.$$

The sub-matrix $\mathbf{A}_N^{(1)}$ ensures that the function is convex with respect to the variable x_1 , and $\mathbf{A}_N^{(2)}$ is used similarly for the second variable x_2 . To be more specific, we write $\mathbf{A}_N^{(1)}$

as follows:

$$\mathbf{A}_N^{(1)} = \begin{pmatrix} 1 & 0 & \dots & 0 & -2 & 0 & \dots & 0 & 1 \\ & 1 & 0 & \dots & 0 & -2 & 0 & \dots & 0 & 1 \\ & & & & & \ddots & & & & \\ & & & & & & & & & \end{pmatrix}_{(N^2-1) \times (N+1)^2},$$

and write $\mathbf{A}_N^{(2)}$ as follows:

$$\mathbf{A}_N^{(2)} = \begin{pmatrix} \mathbf{B} & & & & \\ & \mathbf{B} & & & \\ & & \ddots & & \\ & & & \mathbf{B} & \end{pmatrix}_{(N^2-1) \times (N+1)^2} \quad \text{with } \mathbf{B} = \begin{pmatrix} 1 & -2 & 1 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots \\ & & \ddots & & \\ 0 & \dots & 1 & -2 & 1 \end{pmatrix}_{(N-1) \times (N+1)},$$

where in each row of $\mathbf{A}_N^{(1)}$, there are N 0's between -1 and 2. Notice that the rank R_N of the above matrix $\mathbf{A}_N$ is $2(N^2-1)$. The submatrix $\mathbf{A}_N^{(1)}$ ensures $\beta_{i,k_2} - \beta_{i+1,k_2} \leq \beta_{i+1,k_2} - \beta_{i+2,k_2}$ and $\mathbf{A}_N^{(2)}$ ensures $\beta_{k_1,i} - \beta_{k_1,i+1} \leq \beta_{k_1,i+1} - \beta_{k_1,i+2}$ for all i . Consequently, it guarantees the following, and hence $\mathcal{F}_N \subseteq \mathcal{F}$.

$$\begin{aligned} \frac{\partial^2 B_N}{\partial x_1^2} &= N(N-1) \sum_{k_1=0}^{N-2} \sum_{k_2=0}^N (\beta_{k_1+2,k_2} - 2\beta_{k_1+1,k_2} + \beta_{k_1,k_2}) b_{k_1}(x_1, N-2) b_{k_2}(x_2, N) \geq 0, \\ \frac{\partial^2 B_N}{\partial x_2^2} &= N(N-1) \sum_{k_1=0}^N \sum_{k_2=0}^{N-2} (\beta_{k_1,k_2+2} - 2\beta_{k_1,k_2+1} + \beta_{k_1,k_2}) b_{k_1}(x_1, N) b_{k_2}(x_2, N-2) \geq 0. \end{aligned}$$

Now consider the case when $m(x_1, x_2)$ is a convex function in $[0, 1]^2$, i.e., the regression function belongs to the function space $\mathcal{F} = \{f \in C[0, 1]^2 : f(\lambda(u_1, u_2) + (1-\lambda)(v_1, v_2)) \leq \lambda f(u_1, u_2) + (1-\lambda)f(v_1, v_2), \forall \lambda \in [0, 1], (u_1, u_2), (v_1, v_2) \in [0, 1]^2\}$. In this case, we can no longer express the restriction as $\mathbf{A}_N \boldsymbol{\beta}_N \geq 0$, and we need quadratic restrictions on $\boldsymbol{\beta}_N$ to satisfy $\frac{\partial^2 B_N}{\partial x_1^2} \cdot \frac{\partial^2 B_N}{\partial x_2^2} - (\frac{\partial^2 B_N}{\partial x_1 \partial x_2})^2 \geq 0$, where $\frac{\partial^2 B_N}{\partial x_1 \partial x_2} = N^2 \sum_{k_1=0}^{N-1} \sum_{k_2=0}^{N-1} (\beta_{k_1+1,k_2} - \beta_{k_1,k_2})(\beta_{k_1,k_2+1} - \beta_{k_1,k_2}) b_{k_1}(x_1, N-1) b_{k_2}(x_2, N-1)$, in addition to the restrictions stated above.

(v) *Other Shape Restrictions*

Notice that various other shape restrictions (e.g., $m(x_1, x_2)$ is monotone in x_1 but concave in x_2 etc.) can similarly be handled by constricting appropriate sieve and the corresponding $\mathbf{A}_N$ matrix. Also extensions to higher dimensions is straightforward with additional notations.

3.2.2 Computational Details

The estimator m_N can be shown to be the solution of the following optimization problem:

$$\begin{aligned} & \text{minimize} && \frac{1}{n} \sum_{i=1}^n (\mathbf{b}_N^\top(\mathbf{X}_i) \boldsymbol{\beta}_N - Y_i)^2 \text{ w.r.t. } \boldsymbol{\beta}_N \\ & \text{subject to :} && \mathbf{A}_N \boldsymbol{\beta}_N \geq 0, \sum_{\mathbf{k}} |\beta_{\mathbf{k}}| \leq L_N \end{aligned} \quad (3.6)$$

where $\boldsymbol{\beta}_N$ is a $(N + 1)^d \times 1$ vector of regression coefficients $\{\beta_{k_1, k_2, \dots, k_d}\}$, and $\mathbf{b}_N(\mathbf{x})$ is the Bernstein polynomial basis vector as defined in (3.3). $\mathbf{A}_N$ is a suitably chosen matrix to preserve desired shape restrictions and L_N can be chosen to be a very large number and be practically ignored when solving the optimization problem.. The above optimization problem can be effectively solved by the general quadratic programming ¹ (Goldfarb and Idnani 1982, 1983). The available R package *quadprog* developed by Turlach and Weingessel (2010) is used in this study to solve the quadratic programming problem. The formulation (3.6) essentially reduces the problem to the case of a linear model with linear restrictions on the regression coefficients. Various attempts to solve this problem include Dykstra (1983); Fraser and Massam (1989); Gouriéroux, Holly and Monfrot (1982); Judge and Takayama (1966); Liew (1976) among others with a book length treatment by van Eeden (2006). One distinctive feature of our method to all of these approaches is that we allow the dimension of $\boldsymbol{\beta}_N$ (and hence that of $\mathbf{A}_N$) to vary with sample size n .

We used the well-known and widely used V -fold cross-validation method (Efron and Tibshi-

¹Note that, estimation of convex function with $d \geq 2$ variables may require quadratic programming with quadratic constraints.

rani 1997; Picard and Cook 1984; Stone 1977) to choose the order of Bernstein basis polynomials denoted by N . The main idea of the V -fold cross-validation is to partition the original sample with n observations into V non-overlapping subsets, each including $\lfloor n/V \rfloor$ observations. Among these V subsets, one subset is considered as a validation set, and the remaining $V - 1$ subsets are used as a training set. For a given value of N , the estimators obtained from each training set I_v will be validated by the corresponding validation set I_{-v} using an objective function denoted by $CV(N)$. The whole process is repeated V times (i.e., V -fold) such that each subset is used exactly once as the validation set. As our goal is to solve the optimization problem (3.6) with varying N , the $CV(N)$ for our BP estimator takes the following form:

$$CV(N) = \frac{1}{V} \sum_{v=1}^V \sum_{i \in I_{-v}} (Y_i - \hat{B}_N(\mathbf{X}_i))^2 \quad (3.7)$$

where $\hat{B}_N(\mathbf{x}) = \mathbf{b}_N^T(\mathbf{x})\hat{\boldsymbol{\beta}}_N$ with $\hat{\boldsymbol{\beta}}_N$'s obtained based on the training data set by solving the quadratic programming defined in (3.6).

We computed the cross validation function $CV(N)$ defined in (5.17) on a series of N values starting with $N = 2$ to a relatively large integer N_{max} . The integer N_{max} depends on the data in the sense that it is chosen to be the maximum integer for which the matrix $\sum_{i=1}^n \mathbf{b}_N(\mathbf{X}_i)\mathbf{b}_N^T(\mathbf{X}_i)$ remains empirically invertible. The optimal value $\hat{N}$ is chosen to minimize (5.17), i.e., $\hat{N} = N(n) = \arg \min_{N \in [2, N_{max}]} CV(N)$. In general, if we do not have any additional assumption about the class of regression function, $\mathcal{F}$, we require $(N + 1)^d < \frac{V-1}{V}n$ and $(\frac{V-1}{V}n)^{1/d} - 1$ is the operational upper bound we get for N_{max} . However, if we assume some structural relationships (e.g., additive models) among the predictor variables, this upper bound can be allowed to be much larger. The increase relies on assumption on the order of interactions between predictor variables. The extreme case is when the regression function is assumed to be additive, and the function $m(\mathbf{x})$ can be written as $m(\mathbf{x}) = m_1(x_1) + \cdots + m_d(x_d)$. In this case, we estimate the functions $m_j(x_j)$ each by a univariate Bernstein polynomial estimator then we have $m(\mathbf{x}) = \sum_{j=1}^d m_j(x_j) \approx \sum_{j=1}^d B_N(x_j)$. The shape restriction of each predictor variable x_j

is assigned individually by $\mathbf{A}_N^{(j)}\boldsymbol{\beta}_N^{(j)} \geq 0$. In this case, $(N+1)d < \frac{V-1}{V}n$ needs to be satisfied, and thus an operational upper bound for N is given by $N_{max} < \frac{V-1}{dV}n - 1$. If regression function is partially additive and includes a subset of interaction terms, the operational upper bound of N_{max} will be greater than $(\frac{V-1}{V}n)^{\frac{1}{d}} - 1$ but smaller than $\frac{V-1}{dV}n - 1$. For example, when $d = 2$, $n = 100$ and $V = 7$, the N_{max} can be as large as 41 under additivity assumption compared to only 8 without such assumptions.

Finally, notice that in general the predictors may not be observed to lie in the domain $[0, 1]^d$. To satisfy the domain restriction, we use the following “linear” transformation for our empirical applications,

$$\tilde{x}_{ij} = \frac{x_{ij} - x_{(1)j} + \delta_j}{x_{(n)j} - x_{(1)j} + 2\delta_j},$$

where $x_{(1)j} = \min_{1 \leq i \leq n} x_{ij}$ and $x_{(n)j} = \max_{1 \leq i \leq n} x_{ij}$ denote the minimum and maximum order statistics of the j^{th} predictor for $j = 1, 2, \dots, d$. Finally, by denoting s_j to be the sample standard deviation of $\{x_{ij} : i = 1, \dots, n\}$, we choose $\delta_j = s_j$ for $j = 1, \dots, d$ to allow the estimated function to predict outside of range of the observed x_{ij} ’s. The transformations are linear in the x -values and hence keep essential feature of the x -shape, however, if other (possibly nonlinear) transformations are considered, the shape restrictions may not necessarily be satisfied by a simple set of linear restrictions expressed by $\mathbf{A}_N\boldsymbol{\beta}_N \geq 0$ and hence we caution against nonlinear transformations.

3.3 Asymptotic Properties

In this section, we discuss the asymptotic properties of the nonparametric regression estimator $m_N(\cdot)$ where $N = o(n^k)$ for a suitable choice of $k > 0$. Consider again the function space $\mathcal{F}$, and the true regression function m that minimizes the L_2 risk among all possible measurable functions $f \in \mathcal{F}$, i.e., $m(\cdot)$ satisfies (3.1). Recall that in Section 3.2.1, we provided several illustrations of $\mathcal{F}$ and the corresponding sieve $\mathcal{F}_N$ with various shape restrictions. The estimator

m_N is the function that minimizes the empirical L_2 risk within the sieve $\mathcal{F}_N$ and satisfies (3.3).

In order to establish asymptotic properties of the regression function estimator $m_N(\cdot)$, we verify the following properties of the sieve. The proofs for a few selected cases in one-dimension (i.e., for $d = 1$ only) are outlined in the Appendix. The arguments used in our proofs can be easily extended to other sieves of any arbitrary dimension $d > 1$ with additional notations.

Property 3.3.1. *The sequence of function spaces $\mathcal{F}_N$ is nested in L_2 space, i.e., $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots \subset \mathcal{F}_N \subset \dots \subset L_2[0, 1]^d$.*

Property 3.3.2. $\bigcup_{N=1}^{\infty} \mathcal{F}_N$ is dense in $\mathcal{F}$.

We first state the following result which provides a bound for the L_2 risk between m and m_N . This result easily follows from Lemma 10.1 in Györfi et al. (2002) and hence the proof is omitted.

Lemma 3.3.1. *Let $\mathcal{F}_N = \mathcal{F}_N(\mathcal{D}_n)$ be a class of functions $f_N : \mathcal{R}^d \rightarrow \mathcal{R}$ depending on the data $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ as defined in (3.2). If m_N satisfies equation (3.3), then*

$$\begin{aligned} & \int \{m_N(\mathbf{x}) - m(\mathbf{x})\}^2 \mu(d\mathbf{x}) \\ & \leq 2 \sup_{B_N \in \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_i)^2 - \mathbf{E}\{(B_N(\mathbf{X}) - Y)^2\} \right| \\ & \quad + \inf_{B_N \in \mathcal{F}_N} \int (B_N(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}), \end{aligned}$$

where μ denotes the distribution of $\mathbf{X}$.

Next, we state the result in Lemma 3.3.2 which provides sufficient conditions to establish the weak and strong universal consistency of regression function estimators. The result is adapted from Theorem 10.2 in Györfi et al. (2002). The complete proof is provided in the Appendix.

Lemma 3.3.2. *Let $\mathcal{F}_N$ be the N^{th} sieve defined in equation (3.2) and $m_N(\cdot)$ be the estimator defined in equation (3.3). Let T_L denote the truncation operation $T_L y = y \cdot I(|y| \leq L) + L \cdot$*

$\text{sign}(y) \cdot I(|y| > L)$. Then $Y_L = T_L(Y)$ represents the truncated version of Y , and $T_L\mathcal{F}_N = \{T_L f : f \in \mathcal{F}_N\}$ is a class of truncated functions.

(a) If $L_N \rightarrow \infty$, and

$$\lim_{N \rightarrow \infty} \inf_{B_N \in \mathcal{F}_N} \int (B_N(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) = 0 \text{ a.s.}, \quad (3.8)$$

$$\lim_{N \rightarrow \infty} \sup_{B_N \in T_L \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 - \mathbf{E}(B_N(\mathbf{X}) - Y_L)^2 \right| = 0 \text{ a.s. for all } L > 0, \quad (3.9)$$

then

$$\lim_{N \rightarrow \infty} \int (m_N(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) = 0 \text{ a.s.}$$

(b) If $L_N \rightarrow \infty$, and

$$\lim_{N \rightarrow \infty} \mathbf{E} \left\{ \inf_{B_N \in \mathcal{F}_N} \int (B_N(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) \right\} = 0, \quad (3.10)$$

$$\lim_{N \rightarrow \infty} \mathbf{E} \left\{ \sup_{B_N \in T_L \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 - \mathbf{E}(B_N(\mathbf{X}) - Y_L)^2 \right| \right\} = 0 \text{ for all } L > 0, \quad (3.11)$$

then

$$\lim_{N \rightarrow \infty} \mathbf{E} \left\{ \int (m_N(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) \right\} = 0.$$

Using Lemma 3.3.1 and Lemma 3.3.2, we show that if a set of mild regularity conditions are satisfied, the proposed estimator is both weakly and strongly universally consistent. The required regularity conditions for the sieve space $\mathcal{F}_N$ defined in (3.2) and the regression function estimator (3.3) are described in the following theorem.

Theorem 3.3.1. *Let $\mathcal{F}_N$ be the sieve space defined in equation (3.2) satisfying Properties 3.3.1*

and 3.3.2, and $m_N(\cdot)$ be the restricted regression function estimator given by equation (3.3). Further, let R_N denotes the rank of the restriction matrix $\mathbf{A}_N$ which is assumed to be of full row-rank.

(a) If $\mathbf{E}(Y^2) < \infty$, and R_N and L_N satisfy

$$R_N \rightarrow \infty, L_N \rightarrow \infty, \text{ and } \frac{R_N L_N^4 \log L_N}{n} \rightarrow 0, \quad (3.12)$$

then

$$\mathbf{E} \left[\int (m_N(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) \right] \rightarrow 0 \text{ as } n \rightarrow \infty,$$

i.e., the estimator $m_N(\cdot)$ is weakly universally consistent for $m(\cdot)$.

(b) If $\mathbf{E}(Y^2) < \infty$, R_N and L_N satisfy condition (3.12), and additionally for some $\delta > 0$, L_N satisfies

$$\frac{L_N^4}{N^{1-\delta}} \rightarrow 0, \quad (3.13)$$

then

$$\int (m_N(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) \rightarrow 0 \text{ a.s. as } n \rightarrow \infty,$$

i.e., the estimator $m_N(\cdot)$ is strongly universally consistent for $m(\cdot)$.

Proof. : The proof of the theorem is based on verifying the sufficient conditions stated in Lemma 3.3.2. In other words, using conditions (3.12) and (3.13), we will show conditions (3.8) - (3.11) in Lemma 3.3.2 are satisfied, and thus the estimate is both weakly and strongly universally consistent. The Appendix contains more details. $\square$

Next, we provide some choices for R_N and L_N that satisfy the conditions (3.12) and (3.13) stated above. Assume that $R_N = O(N^r)$, $L_N = O(N^{dl})$ and $N = O(n^k)$, where $r \leq d$, $0 \leq l \leq 1$

and $0 \leq k \leq 1$. Notice that the examples stated in Section 3.2.1 all satisfy these choices. Thus, when n approaches infinity, both R_N and L_N will tend to infinity. To satisfy condition (3.12) in Theorem 3.3.1, we need to show

$$\frac{R_N L_N^4 \log L_N}{n} = C \cdot \frac{n^{kr} n^{4dlk} \log n}{n} = C \cdot \frac{\log n}{n^{1-kr-4dlk}} \rightarrow 0, \quad (3.14)$$

where $C > 0$ is a universal constant. The limit in (3.14) holds if $kr + 4dlk < 1$ is satisfied. For example, when $d = 2$, one may choose $r = 2, k = \frac{1}{10}$ for any $l < 1$. In general, since $r \leq d$ and $0 \leq l < 1$, in order to satisfy $k(d + 4dl) < 1$ we can choose $k = \frac{1}{5d}$ and $l < 1$ because $r + 4dl \leq d + 4dl < 5d$. However, as asymptotic orders typically depend on the unknown regression function, these choices can not be used in practice and hence we have used V-fold cross validation methods to select the optimal order N (see Section 3.2.2).

3.4 Simulation Studies

We evaluated the performance of our multivariate Bernstein polynomial (BP) estimator using several simulated scenarios and compared our approach with a competing shape-restricted regression estimator, the so-called *monoProc* (MP) method¹ proposed by Dette and Scheder (2007). Notice that the MP estimator is proposed specifically for estimating only a bivariate monotone regression function. Although our method can be easily applied to estimate a regression function of any fixed dimensional predictor subject to many popular shape restrictions, in order to make the comparison more convincing and comparable to the MP estimator we only present the monotonicity scenario by studying three different monotone functions of bivariate predictors, which are motivated by a similar study performed by Beresteanu (2004).

¹This method is conveniently available online as a R package “monoProc”

3.4.1 Data Generation and Computational Details

We generated $n = 400$ observations for each of the scenarios $y_i = m_s(x_{1i}, x_{2i}) + \epsilon_i$ (for $s = 1, 2, 3$ and $i = 1, \dots, 400$), where $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$ and $x_{1i}, x_{2i} \stackrel{iid}{\sim} \text{Unif}(0, 1)$. For our study, the regression functions $m_s(\cdot)$ each defined on $[0, 1]^2$ take the following forms:

$$\begin{aligned} \text{(a)} \quad & m_1(x_1, x_2) = \min(x_1, x_2), \\ \text{(b)} \quad & m_2(x_1, x_2) = \begin{cases} (x_1 - \frac{1}{2})(x_2 - \frac{1}{2}) + x_1 x_2 & \text{if } x_1 \geq \frac{1}{2}, x_2 \geq \frac{1}{2} \text{ and} \\ x_1 x_2 & \text{otherwise,} \end{cases} \\ \text{(c)} \quad & m_3(x_1, x_2) = x_1^{1/3} x_2^{2/3}. \end{aligned}$$

All three functions are continuous on the support $[0, 1]^2$ and monotone in each of the coordinates x_1 and x_2 . Notice that, the first function is not differentiable on the 45° line $\{(x_1, x_2) : x_1 = x_2\}$; the second function is differentiable everywhere but the derivative is not continuous at $(\frac{1}{2}, \frac{1}{2})$; and the third function has infinitely many derivatives on $(0, 1)^2$ but the derivatives are unbounded at the boundaries. The error (noise) standard deviation is set to be $\sigma = 0.1$ or 1 associated with each of the three regression functions, and results into six test scenarios. The data generation and subsequent estimation are repeated 500 times in each scenario and the same 500 data sets are used to compute the BP and the MP estimators.

We use $V = 7$ -fold cross-validation method to select N (as described in Section 3.2.2). Other criteria (e.g., leave-one-out cross validation, generalized cross validation, Akaike's information criterion (AIC), Bayesian information criterion (BIC), empirical L_2 loss and other user-defined target functions) can also be used to determine the tuning parameter N . In our simulation study, we found that the results are relatively insensitive to the criteria being used, which is consistent with the findings by Stone (1974). The MP method is based on kernel functions and hence requires the choice of a proper bandwidth as tuning parameter. We first used the default bandwidth choice suggested by Dette and Scheder (2007), but after observing relatively sub-optimal predictive performance we then used a series of bandwidth values includ-

ing the suggested default value of the bandwidth. The one within this series that minimizes the empirical L_2 loss was chosen to be used as the bandwidth for subsequent analysis.

3.4.2 Results

The global performance of the proposed method is measured by the root mean integrated squared errors (RMISE) and also by the mean integrated absolute errors (MIAE) defined as follows:

$$\begin{aligned}
RMISE &= \sqrt{\mathbf{E} \left[\int (\hat{m}(\mathbf{X}) - m(\mathbf{X}))^2 dF(\mathbf{X}) \right]} \\
&\approx \frac{1}{K} \sum_{k=1}^K \sqrt{\frac{1}{J} \sum_{j=1}^J (\hat{m}^{(k)}(x_{1j}, x_{2j}) - m(x_{1j}, x_{2j}))^2}, \\
MIAE &= \mathbf{E} \int |\hat{m}(\mathbf{X}) - m(\mathbf{X})| d\mathbf{X} \\
&\approx \frac{1}{K} \sum_{k=1}^K \left[\frac{1}{J} \sum_{j=1}^J \left| \hat{m}^{(k)}(x_{1j}, x_{2j}) - m(x_{1j}, x_{2j}) \right| \right],
\end{aligned}$$

where $\hat{m}$ denotes an estimated regression function obtained by different estimation methods, and the Monte Carlo simulation repeats K times. The pairs (x_{1j}, x_{2j}) 's do not necessarily indicate the observed predictor variables values from the original sample data set (details are presented later). To compare the local performance, we define the prediction biases of the estimated functions at a given point (x_1, x_2) as

$$BIAS(x_1, x_2) = \hat{m}(x_1, x_2) - m(x_1, x_2),$$

where the first term on the right hand side is the predicted function value, and the second term is the true function value.

In our simulation studies, we used RMISE and MIAE to evaluate the overall global performance of the proposed estimators, and also the prediction biases to assess the local performances

at five selected evaluation points given by

$$(x_1, x_2) \in \{(0, 0), (0, 0.33), (0.2, 0.5), (0.5, 0.5), (0.9, 0.8)\}.$$

Note that all of these global and local measures are obtained for each data set, and then averaged over 500 Monte Carlo (MC) repetitions.

Global Measure of Performances

For each of the three regression function scenarios described above, we obtained 3-D surface plots for each of the true regression functions, and also displayed the corresponding surface plots based on two estimated regression functions (see Figures 3.2 - 3.4). It was observed that $\sigma = 0.1$ or 1 provided similar results, so we only presented the results under the moderate noise condition $\sigma = 0.1$ in Figures 3.2 - 3.4 which is probably more realistic given the fact that the predictors are assumed to lie in $[0, 1]^2$. The plots demonstrate that our proposed BP method performs better than the MP method in terms of capturing the increasing trend. For the regression function in example (a), predicted responses seem to slightly decrease as predictors increase in some areas when using the MP method. The proposed BP method performs better in estimating the function near the boundaries at $x_1 = 0$ and $x_2 = 0$ for the test functions in examples (b) and (c). Nevertheless, both methods appear to over smooth the function near the area where the true function lacks differentiability.

[insert Figures 3.2 - 3.4 here.]

Figures 3.5-3.7 depict the scatter plots of average predicted values $\hat{m}_s(x_1, x_2)$ over 500 MC repetitions against the true function values $m_s(x_1, x_2)$ in a sequence of 22×22 pairs (x_1, x_2) (i.e., 484 points) for $s = 1, 2, 3$ respectively. These (x_1, x_2) points are located in an equidistant grid, from which $x_1, x_2 \in \{0, \frac{1}{21}, \frac{2}{21}, \dots, \frac{20}{21}, 1\}$. The global performances, RMISE and MIAE, are approximated through these $J = 484$ points. We then averaged the prediction values $\hat{m}_s(x_1, x_2)$ obtain from the BP and the MP methods respectively at the 484 evaluation points

mentioned above over 500 Monte Carlo repetitions. Each estimation method exhibits similar performances regardless of noise levels, although with larger noise (i.e., when $\sigma = 1$) more variability is observed as expected. For the first test function (a) in Figure 3.5, underestimation of the MP based predicted values become evident when the true response value is near zero, while overestimation is evident for the test cases (b) and (c) in Figures 3.6 and 3.7, respectively. Compared to MP, our proposed BP method significantly reduces the prediction bias near the boundary, and provides more accurate prediction across the overall surface (further numerical evidence is presented in Tables 3.1-3.3).

[insert Figures 3.5-3.7 here.]

The entries in columns 3 and 4 in Tables 3.1-3.3 represent the approximate *RMISE*'s and *MIAE*'s based on 500 MC repetitions as well as their MC standard errors (shown in parentheses). These numerical results clearly suggest that our BP method performs significantly better than the MP method in both noise settings ($\sigma = 0.1$ and 1.0). In particular, the BP method shows an impressive gain over the MP method when the noise is moderate. That is, when the true $\sigma = 0.1$, *RMISE_{bp}*'s are reduced by 53.6%, 70.4% and 57.8%, respectively, for three test functions defined in (a)-(c) compared to *RMISE_{mp}*. *MIAE* provides similar conclusions in favor of BP. In fact, for $\sigma = 0.1$, the *MIAE_{bp}*'s are reduced by 47.2%, 62.1% and 53.9%, respectively, for the three regression functions compared to *MIAE_{mp}*. Finally, in the last columns of Tables 3.1-3.3, we give the percentage of the times (out of 500 MC replications) in which the *RMISE* and the *MIAE* of BP were less than that of MP. It is clearly evident that in majority of the scenarios (e.g., over 84% across all six scenarios), BP provides a better overall fit (in terms of minimizing *RMISE* and *MIAE*) than the MP. In particular, nearly in 100% of the cases, BP is better than MP across all three functions when $\sigma = 0.1$.

[insert Tables 3.1-3.3 here.]

Local Measure of Performances

In addition to overall global performance of the two estimators, we also explored the local prediction accuracy at selected evaluation points. We examined the biases evaluated at five chosen critical points, namely at $(x_1, x_2) \in \{(0, 0), (0, .33), (.2, .5), (.5, .5), (.9, .8)\}$. The first two points are on the boundary, and the remaining three are on the smooth surface. Notice that the test functions (a) and (b) are not differentiable at the point $(.5, .5)$. Tables 3.4-3.6 presents the comparison of the biases of predicted values obtained by the BP and the MP methods. For each critical point, we also report in Tables 3.4-3.6 the 2.5th to 97.5th percentiles as an interval. The intervals that fail to cover the target value zero are marked in bold. The bold interval implies the presence of significant estimation bias at a particular critical point, and we considered them as estimation failure. The length of those intervals covering zero shows the accuracy of estimation (i.e., the shorter the interval, the greater its estimation accuracy).

[insert Tables 3.4-3.6 here.]

Based on this rule, the proposed BP method gives high local estimation accuracy in general. It gives unbiased estimations at all five points when $\sigma = 1$. In the cases of $\sigma = 0.1$, the BP method shows considerable biases only at the point $(.5, .5)$ when the regression function is $m_1(x_1, x_2) = \min(x_1, x_2)$, and at the point $(0, .33)$ when the regression function is $m_3(x_1, x_2) = x_1^{1/3} x_2^{2/3}$. Whereas the MP method had at least one significantly biased local estimation in each scenario, and even underestimated values at four critical points when the first regression $m_1(x_1, x_2) = \min(x_1, x_2)$ is considered.

Similar results can be observed from the comparison of box plots in Figure 3.8. Recall that the bottom side and the top side of the box represent the first and third quartiles respectively. We focus on top three panels which are corresponding to the scenarios when $\sigma = 0.1$, as bottom three panels illustrate similar results with slightly large deviations. From the box plots, we observe that our proposed BP method estimated values too low at the point $(.5, .5)$ for the first regression function (a), and slightly overestimated values at the point $(0, .33)$ for the

third function (c). There is no significant bias shown for the second function (b). However, the estimations by the MP method are significantly biased at all five critical points for the first regression function (a), and three out of five points for the second and third regression functions (b) and (c). Especially on the boundary points, the MP method failed to make proper estimations.

[insert Figure 3.8 here.]

Finally, our proposed BP method is computationally not as intensive as the MP. The total time required to perform all simulations studies (for all six scenarios) is about 2.1 hours for BP compared to a total of 253.3 hours for MP.

3.5 Real Data Application

To demonstrate our approach on a real data set, we applied our proposed BP method on a data set which consists of $n = 176$ countries with three economic variables for the year 2006: (1) Health Expenditure Per Capita (HEPC), (2) Infant Mortality Rate (IMR), and (3) Adult Literacy Rate (ALR). HEPC and IMR data were obtained from the World Health Organization (WHO) website (www.who.org), while the ALR data were obtained from the United Nations (UN) Development Programme Annual Report (www.undp.org). We considered ALR and HEPC as predictors, and IMR as the response variable. The economic theory suggests that the IMR should be a decreasing function of HEPC and ALR. Empirically we found this evidence by computing their Pearson correlations and Kendall's τ rank correlations (see Table 3.7, Figures 3.9 and 3.10). The goal of our study is to estimate the underlying decreasing trend of the regression function relating the level of HEPC and ALR to the response variable IMR using a framework that allows us to model also the interaction terms. Notice that to begin with we are not necessarily using an additive nonparametric regression model.

[insert Table 3.7 here.]

To obtain our proposed estimator, we rescaled both predictors (HEPC and ALR) to lie in the interval $[0, 1]$ by using the transformation described in the last paragraph of Section 3.2.2. We also used 7-fold cross validation as described in Section 3.2.2 to choose N , and the optimal order $\hat{N} = 4$ produced the minimum cross validation score (see Figure 3.14). All subsequent results are based on the BP estimate with the optimal order $\hat{N} = 4$. For comparative purpose, we also conducted the regression analysis using the MP method using the same bandwidth selection procedure as used in our simulation study. Root mean squared prediction error (i.e., $RMSPE = \sqrt{\frac{1}{n} \sum_i (\hat{y}_i - y_i)^2}$) is used to numerically measure the prediction accuracy. We observed that $RMSPE_{mp} = 0.0867$, and $RMSPE_{bp} = 0.0224$. Thus, our method improved the prediction by reducing $RMSPE$ by around 75% relative to that of the MP estimate.

[insert Figure 3.14 here.]

We present 3-D scatter plots of the real data as well as the estimates from two approaches in Figure 3.9. In addition, the 2-D scatter plots relating the estimated response to two predictors respectively, are presented in Figure 3.10. The figures show that both estimates capture the decreasing trend. Our BP based estimate provided smoother curves than the MP estimate. Moreover, all of our predicted responses are within the observed range $[0, 0.165]$ of the response variable. On the other hand, a few predicted values of the response obtained by the MP method fell outside of the range $[0, 0.165]$. To further illustrate this issue, in Figure 3.11, we plot the estimated values against observed values to compare predictive performances. Figure 3.11 shows clear evidence that our BP method is able to well estimate majority of the observed values compared to that of the MP method. In particular, notice that our predictions are closer to the 45° reference line when observed values are close to zero. We also present the residual plot (i.e., residuals against the fitted values) in Figure 3.12. The plot depicts a fan-shaped trend which in turn indicates the possible lack of fit at larger values of the response variable, and further indicates possibly the lack of homoscedasticity of the variance.

[insert Figures 3.9 - 3.12 here.]

Since the correlation between two predictor variables is not very strong (e.g., the Pearson correlation $\rho(\text{ALR}, \text{HEPC}) = 0.45$), we next fitted a reduced additive model given by

$$\begin{aligned} Y_i &= m_1(x_{1i}) + m_2(x_{2i}) + \varepsilon_i, \\ &\approx B_N(x_{1i}) + B_N(x_{2i}) + \varepsilon_i, \\ &= \sum_{i=0}^N b_{k_1}(x_{1i}, N) \cdot \beta_{k_1} + \sum_{i=0}^N b_{k_2}(x_{2i}, N) \cdot \beta_{k_2} + \varepsilon_i. \end{aligned}$$

The reduced model does not include any interaction term between the two predictor variables, and therefore requires only $2(N + 1)$ parameters instead of $(N + 1)^2$ parameters in the full model. We used the same optimal order of polynomials ($\hat{N} = 4$) as in the full model. In Figure 3.13, we compared the observed responses to the predicted responses using both the reduced and full BP methods. The performance of the reduced BP model is almost the same as that of the full BP model (with multiple $R^2 = 0.9964$), even though the reduced model used $N^2 - 1 = (N + 1)^2 - 2(N + 1) = 15$ less number of parameters. For this data set, the reduced model provided an acceptable regression estimate compared to the full model (see Figure 3.13).

[insert Figure 3.13 here.]

3.6 Discussion

In this chapter, we proposed a multivariate nonparametric shape-restricted regression function estimator based on multivariate Bernstein polynomials. Under some mild regularity conditions, the proposed estimator is shown to provide universally consistent estimates of the unknown shape-restricted regression function. The estimator has several advantages over currently available methods for the estimation of multivariate functions with shape restrictions (e.g., non-negativity, monotonicity, and convexity). First, our estimator is generic since it can be easily adapted to estimate regression functions subject to several well-known shape restrictions. Since all derivatives of the Bernstein polynomial estimator possess the same convergence

properties, the estimator has desirable asymptotic properties, and this enables our estimator to accommodate different shape restrictions by using a sample-size dependent finite-dimensional restriction matrix to achieve the desired shape constraint. Secondly, the numerical algorithm to compute the BP estimator is computationally efficient, fast and numerically stable, since the computational complexity is never beyond the scope of a quadratic programming problem covering many different shape restrictions. Furthermore, the empirical results based on both simulated data and real data applications suggest that our method not only provides significant reduction in bias at local points but also improves prediction accuracy globally. Thus, our method provides a better estimate of the regression curve both locally and globally over the observed range of predictors.

Although we have demonstrated the effectiveness of the frequentist approach of the multivariate restricted Bernstein polynomial estimator, the (asymptotic) variance function for the estimator and the rate of convergence have not been established yet. As stated in Section 3.2.2, the calculation of the proposed estimator will involve quadratic programming with varying dimensions, which indicates that the sampling distribution of the estimator depends on the rank of $\mathbf{A}_N$ and hence remains complicated. The standard methods to derive asymptotic distributions may not be applicable directly to the estimator. One simple idea would be to obtain Bootstrap distribution of $m_N(\cdot)$, as the estimator can be computed easily and fast with many Bootstrap replicated data.

Another alternative would be to use a Bayesian framework to obtain posterior inference for the estimator. Other possible extensions include developing estimation methods under non-standard model assumptions (e.g., binary response, heteroscedasticity, autocorrelated errors etc.). In this article, we proposed the estimator under the assumption of homoscedasticity, yet the real data application in Section 3.5 demonstrates that this assumption may not always hold in practice. This motivates us to allow the conditional variance $Var(\varepsilon|\mathbf{x})$ to vary with the predictor vector $\mathbf{x}$ and use a weighted least squares method. For example, for the data set that we analyzed, one may expect a larger variation in the observed responses when the true

value of the regression function is larger. The above issues and many other extensions of our methodology are currently being explored by the authors and will be reported elsewhere.

Table 3.1: RMISE and MIAE ($\times 100$) for $m_1(x_1, x_2) = \min(x_1, x_2)$ based on 500 Monte Carlo replications as global measurements of performance. *Note:* Standard error is displayed in the parentheses. BP stands for the proposed Bernstein polynomial based estimate. MP stands for the monoProc estimate.

σ	$ \hat{m} - m $	BP	MP	BP $\leq$ MP(%)
0.1	RMISE	2.65 (0.01)	5.72 (0.02)	100
	MIAE	2.08 (0.01)	3.94 (0.02)	99.2
1.0	RMISE	11.4 (0.14)	12.72 (0.15)	88.2
	MIAE	9.23 (0.11)	9.68 (0.12)	84.4

Table 3.2: RMISE and MIAE ($\times 100$) for $m_2(x_1, x_2) = x_1x_2I(x_1 \leq 0.5, x_2 \leq 0.5) + \{(x_1 - 0.5)(x_2 - 0.5) + x_1x_2\}I(x_1 > 0.5, x_2 > 0.5)$ based on 500 Monte Carlo replications as global measurements of performance. *Note:* Standard error is displayed in the parentheses. BP stands for the proposed Bernstein polynomial based estimate. MP stands for the monoProc estimate.

σ	$ \hat{m} - m $	BP	MP	BP $\leq$ MP(%)
0.1	RMISE	1.49 (0.02)	5.04 (0.02)	100
	MIAE	1.14 (0.02)	3.01(0.02)	100
1.0	RMISE	10.05 (0.16)	13.35 (0.15)	96
	MIAE	7.96 (0.13)	10.00 (0.12)	93.2

Table 3.3: RMISE and MIAE ($\times 100$) for $m_3(x_1, x_2) = x_1^{\frac{1}{3}}x_2^{\frac{2}{3}}$ based on 500 Monte Carlo replications as global measurements of performance. *Note:* Standard error is displayed in the parentheses. BP stands for the proposed Bernstein polynomial based estimate. MP stands for the monoProc estimate.

σ	$ \hat{m} - m $	BP	MP	BP $\leq$ MP(%)
0.1	RMISE	1.98 (0.01)	4.69 (0.02)	100
	MIAE	1.46 (0.01)	3.17 (0.02)	100
1.0	RMISE	10.69 (0.14)	12.59 (0.17)	85.4
	MIAE	8.38 (0.12)	9.49 (0.13)	85

Table 3.4: Prediction biases for $m_1(x_1, x_2) = \min(x_1, x_2)$ based on 500 Monte Carlo replications as local measurement of performance. Column 3 and 4: Mean bias (S.E. of bias); Column 5 and 6: $[2.5^{th}$ percentile, 97.5^{th} percentile] of the biases.

σ	Point	BP	monoProc	BP	monoProc
0.1	(0, 0)	-0.0291 (0.0010)	-0.4186 (0.002)	[-0.1521, 0.0307]	[-0.5267, -0.3122]
	(0, .33)	-0.0019 (0.0009)	-0.1574 (0.0012)	[-0.0609, 0.0546]	[-0.2143, -0.1112]
	(.2, .5)	0.0057 (0.0004)	-0.0236 (0.0004)	[-0.0181, 0.0286]	[-0.0409, -0.0049]
	(.5, .5)	-0.0690 (0.0005)	-0.0762 (0.0005)	[-0.0958, -0.0437]	[-0.1000, -0.0553]
	(.9, .8)	-0.0038 (0.0009)	-0.0214 (0.0008)	[-0.0432, 0.0441]	[-0.0536, 0.0125]
1.0	(0, 0)	-0.0458 (0.0071)	-0.4450 (0.0090)	[-0.3910, 0.1669]	[-0.8645, -0.1051]
	(0, .33)	0.0157 (0.0054)	-0.0575 (0.0063)	[-0.2492, 0.2173]	[-0.3826, 0.1880]
	(.2, .5)	-0.0201 (0.0030)	-0.0554 (0.0035)	[-0.1541, 0.1099]	[-0.2145, 0.0874]
	(.5, .5)	-0.1408 (0.0029)	-0.0224 (0.0033)	[-0.2586, 0.0083]	[-0.1628, 0.1164]
	(.9, .8)	-0.0010 (0.0055)	0.0582 (0.0054)	[-0.2390, 0.2496]	[-0.1494, 0.3261]

Table 3.5: Prediction biases for $m_2(x_1, x_2) = x_1x_2I(x_1 \leq 0.5, x_2 \leq 0.5) + \{(x_1 - 0.5)(x_2 - 0.5) + x_1x_2\}I(x_1 > 0.5, x_2 > 0.5)$ based on 500 Monte Carlo replications as local measurement of performance. Column 3 and 4: Mean bias (S.E. of bias); Column 5 and 6: $[2.5^{th}$ percentile, 97.5^{th} percentile] of the biases.

σ	Point	BP	monoProc	BP	monoProc
0.1	(0, 0)	-0.0176 (0.0009)	-0.3295 (0.0018)	[-0.0693, 0.0171]	[-0.4233, -0.2564]
	(0, .33)	-0.0041 (0.0007)	-0.1455 (0.0009)	[-0.0411, 0.0271]	[-0.1859, -0.1055]
	(.2, .5)	-0.0005 (0.0004)	-0.0053 (0.0004)	[-0.0198, 0.0184]	[-0.0226, 0.0115]
	(.5, .5)	0.0016 (0.0005)	0.0034 (0.0004)	[-0.0232, 0.0215]	[-0.0164, 0.0218]
	(.9, .8)	-0.0065 (0.0008)	0.0131 (0.0008)	[-0.0427, 0.0386]	[-0.0209, 0.0466]
1.0	(0, 0)	-0.1186 (0.0069)	-0.5517 (0.0079)	[-0.4829, 0.1129]	[-0.9605, -0.2365]
	(0, .33)	-0.0512 (0.0048)	-0.2392 (0.0059)	[-0.2758, 0.1459]	[-0.5444, -0.0235]
	(.2, .5)	-0.0098 (0.0031)	-0.0222 (0.0033)	[-0.1421, 0.1213]	[-0.1706, 0.1171]
	(.5, .5)	0.0086 (0.0026)	0.0096 (0.0032)	[-0.1209, 0.1198]	[-0.1297, 0.1431]
	(.9, .8)	-0.0302 (0.0059)	0.0218 (0.0067)	[-0.2945, 0.2378]	[-0.2428, 0.3229]

Table 3.6: Prediction biases for $m_3(x_1, x_2) = x_1^{1/3} x_2^{2/3}$ based on 500 Monte Carlo replications as local measurement of performance. Column 3 and 4: Mean bias (S.E. of bias); Column 5 and 6: $[2.5^{th} \text{ percentile}, 97.5^{th} \text{ percentile}]$ of the biases.

σ	Point	BP	monoProc	BP	monoProc
0.1	(0, 0)	0.0006 (0.0017)	-0.2845 (0.0028)	[-0.1150, 0.0709]	[-0.4402, -0.1852]
	(0, .33)	0.1213 (0.0012)	-0.0411 (0.0014)	[0.0418, 0.1789]	[-0.1035, 0.0223]
	(.2, .5)	-0.0056 (0.0005)	-0.0318 (0.0005)	[-0.0296, 0.0228]	[-0.0515, -0.0102]
	(.5, .5)	0.0042 (0.0005)	-0.0090 (0.0004)	[-0.0226, 0.0277]	[-0.0279, 0.0088]
	(.9, .8)	-0.0012 (0.0008)	0.0102 (0.0007)	[-0.0358, 0.0367]	[-0.0172, 0.0395]
1.0	(0, 0)	-0.0327 (0.0086)	-0.4371 (0.0086)	[-0.3910, 0.1669]	[-0.8451, -0.1189]
	(0, .33)	0.1385 (0.0056)	-0.0487 (0.0050)	[-0.2492, 0.2173]	[-0.2327, 0.1525]
	(.2, .5)	-0.0356 (0.0031)	-0.0472 (0.0031)	[-0.1541, 0.1099]	[-0.1717, 0.0631]
	(.5, .5)	-0.0273 (0.0029)	-0.0425 (0.0028)	[-0.2586, 0.0083]	[-0.1582, 0.0736]
	(.9, .8)	0.0343 (0.0053)	0.0726 (0.0071)	[-0.2390, 0.2496]	[-0.1279, 0.4922]

Table 3.7: Pearson correlations with Kendall's Tau in parenthesis among three variables.

	IMR	HEPC	ALR
IMR	1	-.48 (-.63)	-.79 (-.62)
HEPC		1	.45 (.55)
ALR			1

(a) Increasing in x_1 , concave in x_2

(b) Increasing and concave in x_1 and x_2

Figure 3.1: 3-D plots of two shape-restricted regression functions: (a) $f(x_1, x_2) = x_1^2 + 2e^{-x_2^2} + \frac{1}{4}x_1(x_2 + 2)$ (left panel); (b) $f(x_1, x_2) = x_1^{1/3} x_2^{2/3}$ (right panel).

Figure 3.2: 3-D surface plots and contour plots for $m_1(x_1, x_2) = \min(x_1, x_2)$ when $\sigma^2 = 0.01$ and $n = 400$. *Note:* Data is randomly generated from one Monto Carlo simulation out of 500 simulations. From top to bottom: surface plots, and contour plots. From left to right: true curve, monoProc (MP) method, and restricted Bernstein Polynomial (BP) method.

Figure 3.3: 3-D surface plots and contour plots for $m_2(x_1, x_2) = x_1x_2I(x_1 \leq 0.5, x_2 \leq 0.5) + \{(x_1 - 0.5)(x_2 - 0.5) + x_1x_2\}I(x_1 > 0.5, x_2 > 0.5)$ when $\sigma^2 = 0.01$ and $n = 400$. *Note:* Data is randomly generated from one Monto Carlo simulation out of 500 simulations. From top to bottom: surface plots, and contour plots. From left to right: true curve, monoProc (MP) method, and restricted Bernstein Polynomial (BP) method.

Figure 3.4: 3-D surface plots and contour plots for $m_3(x_1, x_2) = x_1^{\frac{1}{3}} x_2^{\frac{2}{3}}$ when $\sigma^2 = 0.1$ and $n = 400$. *Note:* Data is randomly generated from one Monto Carlo simulation out of 500 simulations. From top to bottom: surface plots, and contour plots. From left to right: true curve, monoProc (MP) method, and restricted Bernstein Polynomial (BP) method.

Figure 3.5: Plot of average predicted values vs. true values for $m_1(x_1, x_2) = \min(x_1, x_2)$. Solid line is 45° reference line.

Figure 3.6: Plot of average predicted values vs. true values for $m_2(x_1, x_2) = x_1x_2I(x_1 \leq 0.5, x_2 \leq 0.5) + \{(x_1 - 0.5)(x_2 - 0.5) + x_1x_2\}I(x_1 > 0.5, x_2 > 0.5)$. Solid line is 45° reference line.

Figure 3.7: Plot of average predicted values vs. true values for $m_3(x_1, x_2) = x_1^{\frac{1}{3}} x_2^{\frac{2}{3}}$. Solid line is 45° reference line.

Figure 3.8: Box plots of bias at five points using the BP and the MP method with sample size $n=400$ and 500 replications. *Note:* Simulation scenarios are described as follows. From top to bottom: $\sigma = 0.1$ and $\sigma = 1.0$; From left to right: $m_1(x_1, x_2) = \min(x_1, x_2)$, $m_2(x_1, x_2) = x_1 x_2 I(x_1 \leq 0.5, x_2 \leq 0.5) + \{(x_1 - 0.5)(x_2 - 0.5) + x_1 x_2\} I(x_1 > 0.5, x_2 > 0.5)$, and $m_3(x_1, x_2) = x_1^{\frac{1}{3}} x_2^{\frac{2}{3}}$.

(a) Observed data

(b) BP estimate

(c) monoProc estimate

Figure 3.9: 3-D scatter plots of the infant mortality data. *Predictors*: ALR, HEPC; *Response*: IMR .

Figure 3.10: Predicted values of $\hat{m}(x_{1i}, x_{2i})$ vs. predictors x_{1i} and x_{2i} , $i=1, \dots, 176$. *Note:* $X_1=ALT$, $X_2=HEPC$, $Y=IMR$. Top Panel: observed y_i vs. x_{1i} and x_{2i} ; Middle Panel: predicted values obtained by Bernstein polynomial method vs. predictors; Bottom: predicted values obtained by monoProc vs. predictors.

Figure 3.11: Comparison of observed and predicted values of the response. Solid line represents the 45° reference line, i.e., $y=x$.

Figure 3.12: Residual vs. predicted values. Residual = predicted value - observed response. Top panel: Bernstein polynomial based estimate. Bottom panel: monoProc estimate.

Figure 3.13: Comparison of observed responses and predicted values: Reduced model in (a), Full model in (b), and predicted values from Full model vs. Reduced model in (c). Solid line represents the 45° reference line, i.e., $y=x$.

Figure 3.14: CV score vs. N for the infant mortality data.

Chapter 4

Inference using Bootstrap and Bayesian Methods

4.1 Introduction

In the previous chapters, we have demonstrated the methodologies to obtain the restricted Bernstein polynomial estimator along with degree of the polynomial determination using cross validation approaches. However, the inference of the proposed estimators (e.g., the asymptotic and the finite sampled based sampling distribution of the estimators) has remained unresolved. As stated in Section 3.2.2 of Chapter 3, the calculation of the proposed estimator will involve quadratic programming with varying dimensions, which indicates that the sampling distribution of the estimator depends on the rank of $\mathbf{A}_N$ and hence remains complicated. The standard asymptotic methods (e.g., those based on Donsker's Theorems) to derive large sample distribution of the regression function estimator may not be feasible here due to the implicitness of the regression function estimator.

An alternative would be to develop Bootstrap methods (Efron 1979) and obtain the Bootstrapped distribution of $\hat{m}_N(\cdot)$, as the estimators can be computed easily and fast with many Bootstrap replicated data. One drawback of bootstrapping is that the method does not provide

general finite sample properties (Efron and Tibshirani 1997). Another alternative would be to develop a fully hierarchical Bayesian approach to the estimation and inference of the regression function based on Monte Carlo methods.

Both methods are computer-based methods to measure the accuracy of sample estimates, and obtain inferences for the estimators of interest (Efron 2011). In this chapter, we develop both inferential methods. In Section 4.2, we present the nonparametric bootstrap method used in our estimation problem. Section 4.2.1 introduces the bootstrap procedure to estimate the distribution of the regression estimator. Section 4.2.2 presents the application of the bootstrap method in a real data example. In Section 4.3, we present the Bayesian method used in our estimation problem. Section 4.3.1 describes the Bayesian regression models. In particular, we show that through a simple re-parametrization over the coefficients in the Bernstein polynomial expansion, we can construct a general class of prior densities with proper support ensuring the shape constraints. Section 4.3.1 also presents the details of posterior sampling scheme. The application of our methodology on four simulated data sets is presented and compared with the method in Chang et al. (2007) in Section 4.3.2. A real data study is described in Section 4.3.3. Finally, the chapter ends with some discussion in Section 4.4.

4.2 Bootstrap Method

The bootstrap method is developed by Efron (1979) and others to make statistical inference by estimating the distributions of statistics based on independent observations. The method derives sampling distribution of the chosen estimator based on the fact that the sampling distribution is a good estimate of the population distribution. The basic bootstrap approach treats the original sample as the population from which repeated samples are drawn, and a Monte-Carlo style sampling procedure is conducted on the original sample. A large number of re-samples with replacement from the original sample is randomly generated. In general, there are two forms of the bootstrap: parametric bootstrap and nonparametric bootstrap. In this chapter, we focus on the nonparametric bootstrap to estimate the sampling distribution

of the regression function of interest empirically, when the population distribution function of the random variables is unknown and we do not make any assumption about the form of the population.

4.2.1 Bootstrap Distribution of Estimator $\hat{m}_N(\mathbf{x})$

Consider a general regression model where we assume that $(\mathbf{X}, Y)$ is a $\mathbb{R}^d \times \mathbb{R}$ -valued random vector arising from a unknown distribution F . Let $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ be the set of random vectors which are assumed to be independently and identically distributed (i.i.d) as $(\mathbf{X}, Y)$. Let $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ be the observations of $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$. The regression model is given by

$$Y_i = m(\mathbf{X}_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where ϵ_i 's are independently distributed with $E(\epsilon_i|\mathbf{X}_i) = 0$, and $m(\mathbf{x}) = \mathbf{E}(Y|\mathbf{X} = \mathbf{x})$ is subject to certain shape constraints such as monotonicity, convexity and concavity. Based on the data set $\mathcal{D}_n$, we obtain the estimator $\hat{m}_N(\mathbf{x})$ by solving the quadratic programming problem as described in Chapter 3. Since $\hat{m}_N(\mathbf{x})$, depending on the various order of polynomials N , is implicitly expressed, its large sample distribution is not easy to derive using the standard asymptotic methods.

We consider nonparametric bootstrap methods. The most commonly used sampling scheme, simple random sampling, is adopted to generate bootstrap distribution. Supposed we have an empirical discrete probability distribution P_n , which gives probability $1/n$ to each observation $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$. Using bootstrap method, we create B samples from P_n . In other words, We draw $(\mathbf{X}_i^*, Y_i^*)$ with replacement from the pool $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ n times to consist of a sample of size n $\{(\mathbf{X}_i^*, Y_i^*)\}_{i=1, \dots, n}$ from P_n . Each bootstrap sample of size n gives us an estimator $\hat{m}_{N^*}^*(\mathbf{x})$, obtained by solving the quadratic programming problem as defined in Chapter 3. This process is conducted B times, and we obtain B samples as well as B estimators

$\hat{m}_{N^*}^*(\mathbf{x})_1, \dots, \hat{m}_{N^*}^*(\mathbf{x})_B$. The variance of $\hat{m}_N(\cdot)$ can be estimated by

$$\text{Var}(\hat{m}_N) = \frac{1}{B} \sum_{i=1}^B (\hat{m}_{N^*}^*(\mathbf{x})_i - \overline{\hat{m}_{N^*}^*}(\mathbf{x}))^2, \quad (4.1)$$

where $\overline{\hat{m}_{N^*}^*}(\mathbf{x}) = \frac{1}{B} \sum_{i=1}^B \hat{m}_{N^*}^*(\mathbf{x})_i$. Usually, B is a large number to guarantee the computing power and estimation accuracy.

Following Efron and Tibshirani (1993), we rewrite the above procedure into the following steps with more details to make the nonparametric bootstrap method clear.

Step 0. Let $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ be the observation of the original sample of size n .

Step 1. Build an empirical probability distribution by placing a probability $P_n\{(\mathbf{X}_i^*, Y_i^*) = (\mathbf{x}_1, y_1)\} = \dots = P_n\{(\mathbf{X}_i^*, Y_i^*) = (\mathbf{x}_n, y_n)\} = \frac{1}{n}$.

Step 2. Based on the empirical probability distribution defined in Step 1, draw a random sample $\{(\mathbf{X}_i^*, Y_i^*)\}_{i=1, \dots, n}$ with replacement, as a bootstrap sample.

Step 3. Obtain $\hat{m}_{N^*}^*(\mathbf{x})$ for this bootstrap sample by solving the quadratic programming problem along with V-fold cross-validation to solve optimal N^* as described in Chapter 3.

Step 4. Repeat Steps 2 and 3 B times to generate B bootstrap samples, where B is a large number. The $\hat{m}_{N^*}^*(\mathbf{x})$ is calculated individually from each bootstrap sample.

Step 5. We estimate the parameters of interest, such as variance of the estimator, confidence intervals of the estimator, from the bootstrap statistics.

4.2.2 Real Data Example

In this section, we would like to apply the bootstrap method in one real data study. We took the rabbit data as an example and generated Bootstrap samples. The rabbit data has been investigated in Section 2.5.2. The data consists of a random sample of size $n = 71$ drawn

from an unknown distribution. We tried to estimate the regression function which shows the relationship between the growth of eyes and the age of rabbits. Following the proposed method as illustrated in Chapter 2, we obtained the estimators $\hat{m}_N(\cdot)$'s according to different shape constraint assumptions. In this example, we are interested in the distribution of integrated mean squared prediction errors (IMSPE) which can be written as a function of $\hat{m}_N(\cdot)$. Our goal is to construct the frequency histogram of the estimated IMSPE's obtained from the bootstrap samples. The resulting histogram based on B bootstrap samples is an approximation to the sampling distribution of IMSPE. Recall that IMSPE is defined as

$$\text{IMSPE} = \frac{1}{n} \sum_{i=1}^n (\tilde{m}_{-i}(x_i) - y_i)^2,$$

where $\tilde{m}_{-i}$ denotes the leave-one-out estimated function.

In this study, we only consider the bootstrapped estimator under the concave-increasing shape constraint and compute the corresponding IMSPE's. We followed the bootstrap procedure described in the last subsection, and generated 10,000 bootstrap samples. The frequency distribution of IMSPE is displayed in Figure 4.1. From the figure, we observe that the bootstrapped distribution gives a bell-shaped distribution with mean 63.16 and standard deviation 11.54. We approximate the 0.05 α -level confidence interval for IMSPE by the values of IMSPE* at the 2.5th and 97.5th percentile of the bootstrap sampling distribution. As a result, the 95% confidence interval is given by [41.21, 88.72]. In the histogram, we marked 95% of probability in shade.

[insert Figure 4.1 here.]

We also marked the IMSPE's obtained from the original data in the bootstrap distribution of IMSPE based on the concave-increasing shape constraint. These four IMSPE's are corresponding to the regression estimators under four different shape-constraint assumptions, namely the concave-increasing estimator, the concave estimator, the increasing estimator, and the unconstrained estimator. Recall that, the IMSPE's of three constrained estimators (i.e.,

the concave-increasing estimator, the concave estimator, the increasing estimator) and the unconstrained estimator are equal to 68.27, 70.57, 69.91, and 159.13 respectively. From the figure, we easily observe that any of the three shape-constraint assumptions used in this data improve the estimation and prediction dramatically. The figure demonstrates that the unconstrained estimator performs poorly compared to three shape-restricted estimators. Among the three different shape constraints implemented in this data, the resulting IMSPE's did not show significant differences. In other words, one can use any of these three constraints to estimate the regression curve in this example. Although we have not demonstrated here, the bootstrap samples $\{\hat{m}_{N*}^*(\mathbf{x})_i, i = 1, \dots, B\}$ can be used to construct various other inferences for $m(\cdot)$.

4.3 Bayesian Method

In previous chapters we studied a nonparametric regression model based on Bernstein polynomials from a frequentist perspective. A challenging issue that remains unresolved is that the (asymptotic) variance function for the estimator does not exist in closed-form solutions. In Section 4.2, we proposed to apply the nonparametric bootstrap distribution method to estimate the sampling distribution of $\hat{m}_N(\cdot)$ through a large number of Bootstrap replicated samples. However, in this case the inference is based on large sample theory. We consider an alternative method by implementing the Bayesian method, which provides finite sample distribution of $m_N(\cdot)$.

The study of nonparametric regression problems with shape restriction in a Bayesian framework has received much attention in recent literature. A nonparametric Bayesian approach to shape restricted regression is first proposed by Lavine and Mockus (1995). They applied the Dirichlet process to give a Bayesian estimator of a monotone regression function. Holmes and Heard (2003) proposed another Bayesian approach by using piecewise-constant functions with random knots. In their method, a very large number of samples from the unconstrained model space were generated but only those for which the monotonic constraint holds were retained. Neelon and Dunson (2004) developed an approach to model the regression curve

using piecewise linear function and apply a proper prior on coefficients which restricts monotonicity. Chak, Madras and Smith (2005) and Chang et al. (2007) used restricted Bernstein polynomial to estimate the regression models. Chang et al. (2007) also used a reversible jump Markov chain Monte Carlo (RJMCMC) to obtain the posterior distribution of the order of polynomials. However most of these developments were restricted to univariate predictors or restricted to monotonicity only. In this section, we develop Bayesian methodology for multivariate predictors using Bayesian Bernstein polynomials. Moreover, our methodology applies to a much broader class of shape constraint ranging from monotonicity to convexity and various other shape constraints discussed in Chapter 3.

4.3.1 Bayesian Approach to Shape Restricted Regression

Assume that $(\mathbf{X}, Y)$ is a $\mathbb{R}^d \times \mathbb{R}$ -valued random vector arising from an arbitrary distribution. Let a sample $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ be assumed to be independently and identically distributed (i.i.d) as $(\mathbf{X}, Y)$. The regression model is given by

$$Y_i = m(\mathbf{X}_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where ϵ_i 's are independently distributed with $\epsilon_i \sim N(0, \sigma^2)$. Without loss of generality¹ the observed value of $\mathbf{X} = (X_1, \dots, X_d)$ is assumed to lie in the unit hypercube $[0, 1]^d$, and $m(\mathbf{x}) = \mathbf{E}[Y|\mathbf{X} = \mathbf{x}]$ preserves some shape properties. We are interested in the estimation of the regression function $m(\cdot)$, subject to shape restrictions.

Consider the multivariate Bernstein polynomial approximation to the above unknown shape restricted regression function

$$m(\mathbf{x}) \approx B_N(\mathbf{x}) = \sum_{\mathbf{k}} \beta_{\mathbf{k}} \prod_{j=1}^d \binom{N_j}{k_j} \mathbf{x}_j^{k_j} (1 - \mathbf{x}_j)^{N_j - k_j},$$

¹If predictor variable X_i is observed to lie outside $[0, 1]$, we can carefully transform the values of X_i to lie in the unit interval $[0, 1]$. More contents on transformation can be found in the computational details.

where the $d \times 1$ vector $\mathbf{k} = (k_1, \dots, k_d)$ denotes the index vector, and the $M \times 1$ column vector $\boldsymbol{\beta}_N = (\beta_{\mathbf{k}}; \mathbf{k} \in \{0, 1, \dots, N_j\}^d)$ where $M = \prod_{j=1}^d (N_j + 1)$ denotes the coefficient of Bernstein polynomial expansions. For simplicity, we assume $N_j = N$ for all j , and $M = (N + 1)^d$. We use $\mathbf{A}_N$ to represent a full row rank restriction matrix with dimension $R_N \times M$ where R_N denotes the rank of $\mathbf{A}_N$, and the shape constraint is guaranteed by the linear inequality $\mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{c}$, defined componentwise.

Notice that we can write the regression function in the vector form as an inner product $B_N(\mathbf{x}_i) = \mathbf{w}_i \boldsymbol{\beta}_N$ where the $1 \times M$ row vector $\mathbf{w}_i$ consists of the product of Bernstein polynomial expansions at $\mathbf{x}_i$. Thus the problem reduces to a constraint linear model as follows:

$$\mathbf{Y} = \mathbf{W} \boldsymbol{\beta}_N + \boldsymbol{\epsilon}, \quad \text{subject to } \mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{c}, \quad (4.2)$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ and $\mathbf{W} = (\mathbf{w}_1^\top, \dots, \mathbf{w}_n^\top)^\top$ is the $n \times M$ “design” matrix of a linear model.

In the Bayesian framework of the above regression problem, the parameters are the regression coefficients vector $\boldsymbol{\beta}_N$ and the error variance of the fitted model σ^2 . We write the model (4.2) that relates observations $\mathcal{D}_n$ and parameters as:

$$\mathbf{Y} | \boldsymbol{\beta}_N, \sigma^2, \mathbf{X} \sim N_n(\mathbf{W} \boldsymbol{\beta}_N, \sigma^2 \mathbf{I}_n),$$

where $\boldsymbol{\beta}_N$ satisfies the requisite shape restrictions. In this regression problem, we are more interested in exploring the predictive distribution of the regression model’s estimator $\hat{m}_N = \hat{B}_N$.

Prior on Bernstein Coefficients

In the following description within this subsection, the order of polynomials N is assumed fixed. The procedure of choosing an appropriate N will be given in the computational details. For a Bayesian approach, the prior distributions of parameters are crucial. Now we specify the prior distributions for the vector of parameters $\boldsymbol{\beta}_N$, which is subject to shape restriction expressed by $\mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{c}$. One simple idea would be generating an unrestricted prior for $\boldsymbol{\beta}_N$ and then

using rejection sampling to obtain samples from the posterior distribution of β_N by accepting only those samples that satisfy $\mathbf{A}_N \beta_N \geq \mathbf{c}$. This idea is similar to the methodology proposed by Holmes and Heard (2003). However when N gets larger, it would lead to very inefficient sampling due to curse of dimensionality.

Instead, we develop a prior distribution for β_N with support $\{\beta_N \in \mathbb{R}^M : \mathbf{A}_N \beta_N \geq \mathbf{c}\}$ for a given N , $\mathbf{A}_N$ and $\mathbf{c}$. As N is fixed, for a moment we ignore the dependence of β 's and $\mathbf{A}$'s on N and write $\beta_N \equiv \beta$ and $\mathbf{A}_N \equiv \mathbf{A}$ for simplicity. Letting $\eta = \mathbf{A}\beta - \mathbf{c}$, we have

$$\beta = \mathbf{A}^-(\eta + \mathbf{c}),$$

where $\mathbf{A}^- = \mathbf{B}\mathbf{A}^\top(\mathbf{A}\mathbf{B}\mathbf{A}^\top)^{-1}$ represents the right inverse of $\mathbf{A}$ satisfying $\mathbf{A}\mathbf{A}^- = \mathbf{I}_R$ where $R \equiv R_N$. Here $\mathbf{B}$ can be chosen to be any symmetric non-singular square matrix. In our applications, we choose $\mathbf{B} = (\mathbf{W}^\top \mathbf{W})^{-1}$. With this re-parametrization, we obtain

$$\begin{aligned} \mathbf{Y} &= \mathbf{W}\beta_N + \epsilon \\ &= \mathbf{W}\mathbf{A}^-(\eta + \mathbf{c}) + \epsilon \\ &= \mathbf{W}\mathbf{A}^-\eta + \mathbf{W}\mathbf{A}^-\mathbf{c} + \epsilon. \end{aligned}$$

By moving $\mathbf{W}\mathbf{A}^-\mathbf{c}$ to the left hand side, we have

$$\mathbf{Y} - \mathbf{W}\mathbf{A}^-\mathbf{c} = \mathbf{W}\mathbf{A}^-\eta + \epsilon.$$

Letting $\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{W}\mathbf{A}^-\mathbf{c}$ and $\tilde{\mathbf{W}} = \mathbf{W}\mathbf{A}^-$, the regression model (4.2) is equivalent to

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{W}}\eta + \epsilon \quad \text{subject to } \eta \geq \mathbf{0}, \quad (4.3)$$

with the parameters σ^2 and $\eta = (\eta_1, \dots, \eta_{R_N})^\top$. Notice that $\tilde{\mathbf{W}}$ is a $n \times R_N$ matrix, and η is a $R_N \times 1$ vector. The restriction on the parameters is $\eta_j \geq 0$ for $j = 1, \dots, R_N$. Instead of

assigning the prior distribution of β_N in (4.2), we can consider any prior on $\boldsymbol{\eta}$ with support $[0, \infty)^{R_N}$. For our illustration, we consider multivariate log-normal distribution with location $\boldsymbol{\mu}_0$ and scale matrix Σ_0 as a possible prior distribution $\pi(\boldsymbol{\eta})$ for $\boldsymbol{\eta}$:

$$\pi(\boldsymbol{\eta}) = \frac{1}{(2\pi)^{R_N/2} \prod_{j=1}^{R_N} \eta_j |\Sigma_0|^{1/2}} \exp \left(-\frac{1}{2} (\log \boldsymbol{\eta} - \boldsymbol{\mu}_0)^\top \Sigma_0 (\log \boldsymbol{\eta} - \boldsymbol{\mu}_0) \right), \quad (4.4)$$

where $\log \boldsymbol{\eta} = (\log \eta_1, \log \eta_2, \dots, \log \eta_{R_N})^\top$ is defined componentwise. Notice that when log-normal distribution is assumed as the prior distribution, we do not specifically capture the boundary values of $\eta_j = 0$ (i.e., corresponding to the flat portion of the curve). To overcome this drawback, one may consider a two-component mixture prior for η_j by assigning a positive probability for $\eta_j = 0$. Alternatively, other nonnegative prior distributions, such as truncated normal distribution or uniform distribution defined in nonnegative intervals should also work as prior distribution for $\boldsymbol{\eta}$.

We next construct the prior distribution of variance parameters σ^2 in model (4.2). There are several noninformative prior distributions suggested for variance parameters in hierarchical models. These noninformative prior distribution of variance parameter usually are not conditionally conjugate, as the marginal likelihood depends in a complex way on the data (Gelman 2006). However, the inverse-gamma distribution for σ^2 is conditionally conjugate with normal model (i.e., if σ^2 has an inverse-gamma prior distribution, then the posterior distribution $p(\sigma^2 | \boldsymbol{\eta}, \mathbf{y})$ is also an inverse-gamma distribution). Because of the conditional conjugacy, one can perform Gibbs sampling scheme to draw samples. In our Bayesian model, we assign the prior distribution of σ^2 to be the inverse-gamma distribution $\text{IG}(a, b)$ with mean $\frac{b}{a-1}$ and variance $\frac{b^2}{(a-1)^2(a-2)}$. In other words, the density of $\text{IG}(a, b)$ is given by

$$\pi(\sigma^2) = \frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} \exp \left(-\frac{b}{\sigma^2} \right), \quad (4.5)$$

The hierarchical Bayesian model expressed in (4.3) is given as follows. For a given order of

polynomials N ,

$$\begin{aligned}\tilde{\mathbf{Y}}|\boldsymbol{\eta}, \sigma^2, \mathbf{X} &\sim \text{N}_n(\tilde{\mathbf{W}}\boldsymbol{\eta}, \sigma^2\mathbf{I}_n) \\ \boldsymbol{\eta}|\sigma^2 &\sim \text{LN}_{R_N}(\boldsymbol{\mu}_0, \Sigma_0) \\ \text{and } \sigma^2 &\sim \text{IG}(a, b),\end{aligned}\tag{4.6}$$

where $a > 0, b > 0, \boldsymbol{\mu}_0 \in \mathbb{R}^{R_N}$ and Σ_0 (a positive definite matrix of order R_N) are known constants to be specified in the model by the user. In most applications, we take $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\Sigma_0 = \gamma_0^2 \cdot \mathbf{I}_{R_N}$.

Posterior Sampling Scheme

In this subsection, we describe the posterior sampling scheme to draw samples of $\boldsymbol{\eta}$ and σ^2 in the scenario where the order of polynomials N is assumed fixed. To make posterior inference, we draw samples from the following full conditional distributions:

$$\begin{aligned}p(\boldsymbol{\eta}|\sigma^2, \tilde{\mathbf{Y}}) &\propto L(\tilde{\mathbf{Y}}|\sigma^2, \boldsymbol{\eta})\pi(\boldsymbol{\eta}) \\ &\propto \frac{1}{\sigma^n} \exp\left[-\frac{1}{2\sigma^2}(\tilde{\mathbf{Y}} - \tilde{\mathbf{W}}\boldsymbol{\eta})^\top(\tilde{\mathbf{Y}} - \tilde{\mathbf{W}}\boldsymbol{\eta})\right] \cdot \pi(\boldsymbol{\eta}) = K(\boldsymbol{\eta}; \sigma^2, \tilde{\mathbf{Y}}),\end{aligned}\tag{4.7}$$

$$\begin{aligned}p(\sigma^2|\boldsymbol{\eta}, \tilde{\mathbf{Y}}) &\propto L(\tilde{\mathbf{Y}}|\sigma^2, \boldsymbol{\eta})\pi(\sigma^2) \\ &\propto (\sigma^2)^{-(a+\frac{n}{2})-1} \exp\left(-\frac{b + \frac{1}{2}(\tilde{\mathbf{Y}} - \tilde{\mathbf{W}}\boldsymbol{\eta})^\top(\tilde{\mathbf{Y}} - \tilde{\mathbf{W}}\boldsymbol{\eta})}{\sigma^2}\right),\end{aligned}\tag{4.8}$$

where $K(\cdot)$ is the kernel of $\boldsymbol{\eta}$'s conditional posterior distribution, $L(\tilde{\mathbf{Y}}|\sigma^2, \boldsymbol{\eta})$ denotes the likelihood function, and $\tilde{\mathbf{W}} = \mathbf{W}\mathbf{A}^-$. Under our model assumption, the conditional posterior distribution of $\sigma^2|\boldsymbol{\eta}, \tilde{\mathbf{Y}}$ shown in (4.8) is the inverse gamma distribution $\text{IG}(a + \frac{n}{2}, b + \frac{1}{2}\|\tilde{\mathbf{Y}} - \tilde{\mathbf{W}}\boldsymbol{\eta}\|_2^2)$, when the prior distribution of σ^2 is $\text{IG}(a, b)$. Therefore, we can use Gibbs sampler (Geman and Geman 1984) to draw $(\boldsymbol{\eta}, \sigma^2)$ from the posterior distribution. However, the conditional distribution of $\boldsymbol{\eta}$ given σ^2 and $\tilde{\mathbf{Y}}$ is non-standard as $\boldsymbol{\eta} \in [0, \infty)^{R_N}$ and hence we use Metropolis sampling within the Gibbs sampler.

Next we describe the algorithm to sample $\boldsymbol{\eta}$ given σ^2 and $\tilde{\mathbf{Y}}$. Recall that multivariate log-normal distribution $\text{LN}_{R_N}(\boldsymbol{\mu}_0, \Sigma_0 = \gamma_0^2 \mathbf{I})$ is used as prior distribution for $\boldsymbol{\eta}$, where $\mathbf{I}$ denotes a $R_N \times R_N$ identity matrix. Since the multivariate log-normal prior and the likelihood function in our regression model do not lead to simple conditional conjugate distributions (see (4.7) for $\boldsymbol{\eta}$'s posterior distribution), we apply random-walk Metropolis-Hastings algorithm (Chib and Greenberg 1995) to generate a sample of $\boldsymbol{\eta}$ from a proposal distribution and accept it with certain probability. In our Bayesian approach, the proposal density at time t is a multivariate log-normal distribution $q(\cdot; \log(\boldsymbol{\eta}^{(t)}), c_0^2 \mathbf{I})$ with mean $\log(\boldsymbol{\eta}^{(t)})$ and variance $c_0^2 \mathbf{I}$ where $c_0 > 0$ is chosen suitably to achieve a desired value of acceptance rate. The next value $\boldsymbol{\eta}^{(t+1)}$ is generated in two steps. The proposal $\boldsymbol{\eta}^*$ is first generated from the proposal density

$$\boldsymbol{\eta}^* \sim q(\cdot; \log(\boldsymbol{\eta}^{(t)}), c_0^2 \mathbf{I}) = \text{LN}_{R_N}(\log(\boldsymbol{\eta}^{(t)}), c_0^2 \mathbf{I}).$$

This proposal $\boldsymbol{\eta}^*$ is accepted with a probability

$$\pi(\boldsymbol{\eta}^* | \boldsymbol{\eta}^{(t)}) = \min \left(1, \frac{K(\boldsymbol{\eta}^*; \sigma^{2(t)}, \tilde{\mathbf{Y}})}{K(\boldsymbol{\eta}^{(t)}; \sigma^{2(t)}, \tilde{\mathbf{Y}})} \frac{q(\boldsymbol{\eta}^{(t)}; \log(\boldsymbol{\eta}^*), c_0^2 \mathbf{I})}{q(\boldsymbol{\eta}^*; \log(\boldsymbol{\eta}^{(t)}), c_0^2 \mathbf{I})} \right),$$

where $q(\boldsymbol{\eta}, \log(\boldsymbol{\eta}^{(t)}), c_0^2 \mathbf{I})$ denotes the density of $\text{LN}_{R_N}(\log(\boldsymbol{\eta}^{(t)}), c_0^2 \mathbf{I})$ as defined in (4.4). If accepted, we set $\boldsymbol{\eta}^{(t+1)} = \boldsymbol{\eta}^*$, otherwise, $\boldsymbol{\eta}^{(t+1)} = \boldsymbol{\eta}^{(t)}$.

In the following, we summarize the sampling algorithm used in our Bayesian methodology.

MCMC Procedure. *Given a fixed order of Bernstein polynomials N ,*

1. *Begin the MCMC algorithm by initializing the values of parameters $\boldsymbol{\eta}^{(0)}$ and $\sigma^{2(0)}$. For instance, one may use the least squares estimator of $\boldsymbol{\eta}$ subject to constraint $\boldsymbol{\eta} \geq \mathbf{0}$ obtained by solving the quadratic programming. That is,*

$$\begin{aligned} \hat{\boldsymbol{\eta}}^{(0)} &= \arg \min_{\boldsymbol{\eta} \geq \mathbf{0}} (\tilde{\mathbf{y}} - \tilde{\mathbf{W}}\boldsymbol{\eta})^\top (\tilde{\mathbf{y}} - \tilde{\mathbf{W}}\boldsymbol{\eta}) \\ \hat{\sigma}^{2(0)} &= \frac{1}{n - R_N} (\tilde{\mathbf{y}} - \tilde{\mathbf{W}}\hat{\boldsymbol{\eta}}^{(0)})^\top (\tilde{\mathbf{y}} - \tilde{\mathbf{W}}\hat{\boldsymbol{\eta}}^{(0)}) \end{aligned}$$

2. Start iterations to update the parameters for $t = 1, \dots, n_{iter}$ iterations, where n_{iter} denotes the total number of MCMC iterations. At the t -th iteration, the samples of $\sigma^{2(t)}$ and $\boldsymbol{\eta}^{(t)}$ are obtained as follows

$$\begin{aligned}\sigma^{2(t)} | \boldsymbol{\eta}^{(t-1)}, \tilde{\mathbf{y}} &\sim \text{IG}(a + \frac{n}{2}, b + \frac{1}{2} \|\tilde{\mathbf{y}} - \widetilde{\mathbf{W}}\boldsymbol{\eta}^{(t-1)}\|_2^2), \\ \boldsymbol{\eta}^{(t)} &= I(u \leq \pi(\boldsymbol{\eta}^* | \boldsymbol{\eta}^{(t-1)}))\boldsymbol{\eta}^* + I(u \geq \pi(\boldsymbol{\eta}^* | \boldsymbol{\eta}^{(t-1)}))\boldsymbol{\eta}^{(t-1)}, \\ &\text{when } u \sim U(0, 1) \text{ and } \boldsymbol{\eta}^* \sim \text{LN}_{R_N}(\log(\boldsymbol{\eta}^{(t-1)}), c_0^2 \mathbf{I}).\end{aligned}$$

We implement the above MCMC procedure using the WinBUGS software where the Metropolis Step uses the so-called current point Metropolis based updating univariately. WinBUGS uses first 400 iterations to select c_0 such that the acceptance rate is between 20% and 40%.

Selecting the Order of Polynomials

The description of MCMC procedure in previous subsections is based on the fixed order of Bernstein basis polynomials denoted by N . We next discuss how to select a value of N using information criteria. Initially, a large value of N (e.g., $N = \lceil n^{2/5} \rceil$ where n is the sample size of observations) is used to run the MCMC procedure as described above. When the MCMC iterations complete, we obtain the estimated effective number of parameters in the posterior distribution, pD , which is defined as half the posterior variance of the deviance¹ (Gelman et al. 2003). We then set $N = \lfloor pD \rfloor$, the integer part of pD , and run the MCMC procedure with this new value of N . This iterative process of updating N stops when the difference of N is less than 2. We outline the process of selecting $\hat{N}$ as follows.

Algorithm to Choose $\hat{N}$. Suppose $N^{(l)}, pD^{(l)}$ are the current states at the l -th iteration. We initialize $N^{(0)}$ by a large value, e.g., $N^{(0)} = \lceil n^{2/5} \rceil$. Given $N^{(0)}$, the MCMC procedure is performed and consequently $pD^{(0)}$ is obtained. Set $N^{(1)} = \lfloor pD^{(0)} \rfloor$. For $l = 1, 2, \dots$

¹Another version of pD is defined as the posterior mean of the deviance minus the deviance of the posterior means by Spiegelhalter et al. (2002). Both definitions of pD are widely used in the Bayesian method. We choose the one introduced by Gelman et al. (2003) in order to guarantee the positivity of pD .

Step 1. Given $N^{(l)}$, run the MCMC procedure as described above.

Step 2. When the MCMC procedure is complete, compute $pD^{(l)} = \frac{1}{2}\text{Var}(\text{deviance}|\text{data})$, and set $N^{(l+1)} = [pD^{(l)}]$.

Step 3. If $|N^{(l)} - N^{(l+1)}| < 2$, stop iteration and set $\hat{N} = N^{(l)}$. Otherwise, go to Step 1 with $N^{(l+1)}$.

In our regression model presented in Section 4.3.1, we assume the support of predictor variables to be the domain $[0, 1]^d$. Notice that in general the predictors may not be observed to lie in $[0, 1]^d$. To satisfy the domain restriction, we use the following “linear” transformation for our empirical applications,

$$\tilde{x}_{ij} = \frac{x_{ij} - x_{(1)j} + \delta_j}{x_{(n)j} - x_{(1)j} + 2\delta_j}, \quad \text{for } j = 1, 2, \dots, d,$$

where $x_{(1)j} = \min_{1 \leq i \leq n} x_{ij}$ and $x_{(n)j} = \max_{1 \leq i \leq n} x_{ij}$ denote the minimum and maximum order statistics of the j^{th} predictor respectively. Finally, by denoting s_j to be the sample standard deviation of $\{x_{ij} : i = 1, \dots, n\}$, we choose $\delta_j = s_j$ for $j = 1, \dots, d$ to allow the estimated function to predict outside of range of the observed x_{ij} ’s. The transformations are linear in the x -values and hence keep essential feature of the x -shape, however, if other (possibly nonlinear) transformations are considered, the shape restrictions may not necessarily be satisfied by a simple set of linear restrictions expressed by $\mathbf{A}_N \boldsymbol{\beta}_N \geq 0$ and hence we caution against nonlinear transformations.

4.3.2 Simulation Studies

To evaluate the performance of the proposed Bayesian Bernstein polynomial estimator, we implemented our approach through several simulation studies and compared with a few competing nonparametric estimators proposed by Chang et al. (2007), Dette et al. (2006) and Birke and Dette (2007) respectively. The estimator by Chang et al. (2007) is also based on Bernstein

polynomials and fitted in a Bayesian framework. Their estimator can be fitted in monotonic regression functions and convex regression functions. Dette et al. (2006) proposed a kernel-based monotonic estimator using density regression method, and Birke and Dette (2007) extended the method to estimate convex regression functions. Notice that these competing estimators are proposed specifically for estimating regression functions of single predictors. Although our method can be easily applied to estimate a regression function of any fixed dimensional predictor subject to various popular shape restrictions, in order to make the comparison more convincing and comparable to these estimators we only present in this section by studying four univariate functions (i.e., two monotonic functions and two convex functions), which have been studied in Chang et al. (2007).

Data Generation and Computational Details

We generated $n = 100$ observations for each of the scenarios $y_i = m_s(x_i) + \epsilon_i$ (for $s = 1, 2, 3, 4$ and $i = 1, \dots, 100$), where $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$ with $\sigma \in \{0.1, 1\}$ and $x_i \stackrel{iid}{\sim} \text{Unif}(0, 1)$. Each of the regression functions $m_s(\cdot)$ is defined on $[0, 1]$. Out of these four functions, the first two are monotonic increasing functions, taking the following forms:

$$m_1(x) = \sin\left(\frac{\pi}{2}x\right),$$

$$m_2(x) = \begin{cases} 2x & 0 \leq x \leq 0.25, \\ 0.5 & 0.25 < x \leq 0.75, \\ 2x - 1 & 0.75 < x \leq 1. \end{cases}$$

The other two are convex functions, taking the following forms:

$$m_3(x) = (16/9)(x - 1/4)^2,$$

$$m_4(x) = \begin{cases} -4x + 1 & 0 \leq x \leq 0.25, \\ 0 & 0.25 < x \leq 0.75, \\ 4x - 3 & 0.75 < x \leq 1. \end{cases}$$

Notice that, $m_2(x)$ and $m_4(x)$ have a great portion of flat area. Since in our Bayesian methodology, we use log-normal distribution as the prior for $\boldsymbol{\eta} = (\eta_1, \dots, \eta_{R_N})$, samples of $\eta_j > 0$ for $j = 1, \dots, R_N$ will be generated, and thus the boundary values of $\eta_j = 0$, which help construct the flat curve, are not carefully captured (see Curtis and Ghosh (2011) for an extension). Thus, it is expected that our method might not perform well when estimating functions $m_2(x)$ and $m_4(x)$. The data generation and subsequent estimation are repeated 200 times. The above simulation settings are exactly reproduced as in the simulation scenarios in Chang et al. (2007).

We used the same three criteria as suggested by Chang et al. (2007) to evaluate the performance of the estimators: L_1 -norm, L_2^2 -norm, and sup-norm of $\hat{m} - m$ defined as follows:

$$\begin{aligned} L_1\text{-norm} &= \frac{1}{n_{grid}} \sum_{k=1}^{n_{grid}} |\hat{m}(x_k) - m(x_k)|, \\ L_2^2\text{-norm} &= \frac{1}{n_{grid}} \sum_{k=1}^{n_{grid}} (\hat{m}(x_k) - m(x_k))^2, \\ \text{sup-norm} &= \max_{1 \leq k \leq n_{grid}} |\hat{m}(x_k) - m(x_k)|, \end{aligned}$$

where $m(\cdot)$ denotes the true regression function, and the fixed equidistant design $x_k = \frac{k}{50}$ on $[0, 1]$ are the $n_{grid} = 51$ prediction points where $k = 0, \dots, 50$. The estimator $\hat{m}(\cdot)$ is obtained through posterior mean of the regression function, namely,

$$\hat{m}(x) = E[\mathbf{W}(x)^\top \boldsymbol{\beta} | \mathbf{y}] = \mathbf{W}(x)^\top E[\boldsymbol{\beta} | \mathbf{y}] = \frac{1}{n_{iter}} \sum_{t=1}^{n_{iter}} \hat{m}^{(t)}(x),$$

where n_{iter} is total number of MCMC iterations during one Monte Carlo simulation, and

$\beta^{(t)}$'s are the sampled values of β from the t -th MCMC iteration. The simulation results L_1 -norm, sup-norm, and L_2^2 -norm (shown as MSE) of the estimators proposed by Chang et al. (2007), Dette et al. (2006) and Birke and Dette (2007) are directly reproduced from Chang et al. (2007)'s Tables 1-4.

To fit our Bayesian model expressed as (4.6), we let $\mu_0 = \mathbf{0}$, $\Sigma_0 = 0.001 \cdot \mathbf{I}$, $a = 0.01$ and $b = 0.01$. The model was fit using a Gibbs sampler and Metropolis-Hastings algorithm constructed with the WinBUGS software (Spiegelhalter et al. 1999). We set up three MCMC chains, each with 10000 draws to obtain the posterior mean after 1000 draws of burn-in. One chain was initialized by the least squares estimator of η subject to constraint and σ^2 as suggested in the MCMC procedure of Section 4.3.1, and the other two chains were initialized by arbitrary values (e.g., $\eta^{(0)} \in \{0.5 \cdot \mathbf{1}_{R_N \times 1}, \mathbf{1}_{R_N \times 1}\}$, $\sigma^{2(0)} \in \{1, 4\}$). In order to select the proper order of Bernstein polynomials N , we started with the initial $N^{(0)} = 10$ (one may use with $\lfloor n^{2/5} \rfloor \approx 6$ or another arbitrary positive integer as starting value of N) and followed the procedure outlined in the computational details of Section 4.3.1 to choose N for each simulated data set.

Results

The estimation accuracy is evaluated by the mean of three criteria (i.e., L_1 -norm, sup-norm, and L_2^2 -norm) values over 200 Monte Carlo simulations. We report the comparative results in Tables 4.1 - 4.4, each corresponding to one of the four regression functions. In the tables, we mark the minimum results in bold for the sake of easy comparison. These numerical results clearly suggest that our Bayesian method performs much better than the kernel-based method proposed by Dette et al. (2006) and Birke and Dette (2007) in almost all simulation scenarios. Compared with the Bayesian method proposed by Chang et al. (2007), our method performs comparably well (even better in some simulation scenarios).

Table 4.1 gives the results where the true regression function $m_1(x)$ is strictly increasing. As expected, we beat the other two competing estimators, as the values of all three evaluation metrics from our method are smallest under both noise levels. Notice that the differences are

statistically significant. When $\sigma = 0.1$, we consider the method by Chang et al. (2007) perform comparably well if the result is measured by L_2^2 -norm only.

Table 4.2 gives the results where the true regression function $m_2(x)$ is increasing with a flat portion in the middle (see Figure 4.2). Since our method does not use specialized priors for flat portions in the regression (as in Curtis and Ghosh (2011)), it is expected that our method may not perform as well as when estimating strictly increasing functions such as $m_1(x)$. When $\sigma = 0.1$, the method by Chang et al. (2007) outperformed others in terms of all three criteria, although our results were close to theirs. But our method still produced much more estimation accuracy than the one by Dette et al. (2006). Interestingly, when $\sigma = 1$, our Bayesian method again outperforms the other two competitors, although the differences are not statistically significant.

Table 4.3 gives the results where the true regression function $m_3(x)$ is convex. Since the true regression function is a polynomial function, it is expected that the two estimators (i.e., our proposed estimator and the one by Chang et al. (2007)) based on Bernstein polynomials perform well. The results shown in Table 4.3 is consistent with our expectation. When the noise is small, two estimators based on Bernstein polynomials perform comparably well, and both perform better than the third estimator proposed by Birke and Dette (2007). When the noise is large, the performances of all three methods are performed equally well as each method outperformed the other two in terms of one criterion.

Table 4.4 gives the results where the true regression function $m_4(x)$ is convex with flat portion. Although we did not expect the great performance of our proposed method, our Bayesian estimator was able to provide the best estimation accuracy in terms of all three criteria when $\sigma = 0.1$. When the noise is large, our method gave the smallest sup-norm, while the estimator proposed by Chang et al. (2007) gave the smallest L_1 -norm and L_2^2 -norm. Both estimators outperformed the estimator proposed by Birke and Dette (2007).

In addition to numerical results, we displayed the estimations of four regression functions in Figure 4.2. It was observed that $\sigma = 0.1$ or 1 provided similar results, so we only presented

Table 4.1: Simulation study for $m_1(x) = \sin(\frac{\pi}{2}x)$. The results are averaged over 200 Monte Carlo replications. Monte Carlo standard error is displayed in the parentheses.

σ	$ \hat{m} - m $	Our Bayesian Estimator	Method by Chang et al.(2007)	Method by Dette et al. (2006)
0.1	L_1 -norm	0.0157 (0.0003)	0.0161	0.0238
	Sup-norm	0.0500 (0.0012)	0.0532	0.0874
	L_2^2 -norm	0.0004 (0.00002)	0.0004	0.0010
1	L_1 -norm	0.1111 (0.0031)	0.1148	0.1194
	Sup-norm	0.2634 (0.0088)	0.2795	0.2651
	L_2^2 -norm	0.0206 (0.0032)	0.0215	0.0233

Table 4.2: Simulation study for $m_2(x) = 2x \cdot I(x \in [0, 0.25]) + 0.5 \cdot I(x \in [0.25, 0.75]) + (2x - 1) \cdot I(x \in [0.75, 1])$. The results are averaged over 200 Monte Carlo replications. Monte Carlo standard error is displayed in the parentheses.

σ	$ \hat{m} - m $	Our Bayesian Estimator	Method by Chang et al.(2007)	Method by Dette et al. (2006)
0.1	L_1 -norm	0.0375 (0.0002)	0.0353	0.0413
	Sup-norm	0.1046 (0.0009)	0.1002	0.1415
	L_2^2 -norm	0.0020 (0.00003)	0.0019	0.0030
1	L_1 -norm	0.1214 (0.0027)	0.1255	0.1267
	Sup-norm	0.2905 (0.0056)	0.3035	0.3509
	L_2^2 -norm	0.0232 (0.0011)	0.0244	0.0256

the results under the small noise level (i.e., $\sigma = 0.1$). The plots on the top panel give our estimation results regarding increasing regression functions, and the plots on the bottom panel give our results regarding convex regression functions. Attached within each plot, we also report the average value of selected polynomial orders N over 200 data sets in each simulation scenario. In our simulation study, we found that $\hat{N} = 7, 12, 6, 10$ are good enough (in terms of DIC metric) to fit the four regression functions, respectively. For each simulated data, we were able to obtain the posterior median of the regression function at x_k as the point estimator (i.e., $\hat{m}(x_k) = \text{Median}[m_N^{(t)}(x_k)|\mathbf{y}]$, for $t = 1, \dots, n_{iter}$), posterior 2.5% quantile and posterior 97.5% quantile as 95% credible interval at each of prediction points $x_k = (0, \frac{1}{50}, \dots, \frac{49}{50}, 1)$. The estimators (shown as dashed lines) in Figure 4.2 have been averaged over 200 simulations, and

Table 4.3: Simulation study for $m_3(x) = (16/9)(x - 1/4)^2$. The results are averaged over 200 Monte Carlo replications. Monte Carlo standard error is displayed in the parentheses.

σ	$ \hat{m} - m $	Our Bayesian Estimator	Method by Chang et al.(2007)	Method by Birke and Dette (2007)
0.1	L_1 -norm	0.0171 (0.0004)	0.0157	0.0775
	Sup-norm	0.0422 (0.0013)	0.0525	0.2296
	L_2^2 -norm	0.0004 (0.00002)	0.0004	0.0101
1	L_1 -norm	0.1393 (0.0037)	0.1362	0.1366
	Sup-norm	0.3549 (0.0132)	0.4247	0.3643
	L_2^2 -norm	0.0324 (0.0020)	0.0319	0.0292

Table 4.4: Simulation study for $m_4(x) = (-4x+1) \cdot I(x \in [0, 0.25]) + 0 \cdot I(x \in [0.25, 0.75]) + (4x-3) \cdot I(x \in [0.75, 1])$. The results are averaged over 200 Monte Carlo replications. Monte Carlo standard error is displayed in the parentheses.

σ	$ \hat{m} - m $	Our Bayesian Estimator	Method by Chang et al.(2007)	Method by Birke and Dette (2007)
0.1	L_1 -norm	0.0276 (0.0013)	0.0405	0.1311
	Sup-norm	0.0956 (0.0054)	0.1381	0.4663
	L_2^2 -norm	0.0020 (0.0009)	0.0026	0.0282
1	L_1 -norm	0.1591 (0.0045)	0.1394	0.2265
	Sup-norm	0.3635 (0.0178)	0.4603	0.7010
	L_2^2 -norm	0.0396 (0.0024)	0.0338	0.0793

similarly the 95% pointwise credible areas (shown with shade) have been averaged as well. Except for Function $m_2(\cdot)$, the pointwise posterior intervals contain the curve.

[insert Figures 4.2 here.]

4.3.3 Real Data Application

We illustrate the methodology proposed in Section 4.3.1 by analyzing the rabbit data studied in Birke and Dette (2007). The data records the relation between the dry weight of eye lens (in milligrams) and age (in days) of $n = 71$ rabbits in Australia. More details about the data can be found in the website (<http://www.statsci.org/data/oz/rabbit.html>). We are interested in the estimation of the underlying increasing trend of eye lens weight as a function of age. To

apply our methodology, we first scaled the predictor variable (age) to the interval $[0, 1]$ by the transformation formula defined in Section 4.3.1.

To fit our Bayesian model, we let $\boldsymbol{\mu}_0 = \mathbf{0}$, $\Sigma_0 = 0.001 \cdot \mathbf{I}$, $a = 0.01$ and $b = 0.01$. The results are based on MCMC procedure conducted in WinBUGS. We set up one MCMC chains with 30000 draws after 3000 draws of burn-in. The chain was initialized by the least squares estimator of $\boldsymbol{\eta}$ subject to constraint and σ^2 as described in the MCMC procedure of Section 4.3.1. In this study, we use the increasing constraint in the model. To select the proper order of Bernstein polynomials N , we started with the initial $N^{(0)} = 18$ and chose $\hat{N} = 13$ by following the procedure outlined in Section 4.3.1. In Figure 4.3, we plot the posterior median as the estimator, and the posterior quantiles as the 95% credible interval. Recall that we fit the same data using the frequentist method is Chapter 2. As shown in the figure, the estimated curve obtained in the Bayesian framework is similar to the estimators obtained in the frequentist framework (see Page 42).

[insert Figure 4.3 here.]

4.4 Conclusion

The bootstrap method is applied to make inferences about parameters of interest based on estimators that may not be available in explicit forms. Different from the traditional parametric approaches to make inferences, the bootstrap method creates a large number of replicated re-samples using the original data as population, and generates a bootstrapped distribution to approximate the sampling distribution of the estimator. One can compute a confidence interval or perform a hypothesis test using this approximate bootstrapped sampling distribution of the estimator. The bootstrap method is computationally intensive but very easy to implement. A major drawback of bootstrap method is that the inference is based on large sample theory, and thus the general finite sample properties can not be explored.

Another alternative method would be to use a Bayesian framework to obtain posterior in-

ference for the estimator. One of the big advantages to use the MCMC based Bayesian method is that the posterior interval bands can be obtained simultaneously with the estimator without making efforts to derive asymptotic-based inference. Therefore, in the second part of this chapter, we presented a Bayesian method based on Bernstein polynomials to obtain finite sample performance of the shape restricted regression problem. We applied the Metropolis-Hastings algorithm within Gibbs sampler and derived a sampling scheme to generate approximate posterior samples of the model parameters. We use the numerical examples to demonstrate the effectiveness of our methodology. However, in this method, we still choose the order of polynomials using iterative steps when applying the MCMC procedure. Thus, the method remains computationally intensive as we need to run the MCMC method for each selected values of N (via the pD of DIC).

One of the key features of a Bayesian approach is its flexibility to incorporate trans-dimensional modelling by imposing prior distributions for parameters and updating this distribution with data. Since the dimension of the parameter space is determined by the orders of Bernstein polynomials used in the proposed model, one can take advantage of this feature of Bayesian approach, and model the uncertainty of the number of parameters through reversible jump Markov Chain Monte Carlo (RJ-MCMC) pioneered by Green (1995). In other words, a prior distribution of the orders of Bernstein polynomials can be imposed to obtain the posterior probability of several possible dimensions of the parameter space. This work will be discussed as our future work in Chapter 7.

Figure 4.1: Histogram of the Bootstrap Distribution of 10000 Resampled IMSPE's.

Figure 4.2: Simulation results over 200 simulation data sets under $\sigma = 0.1$. Scatter plot: randomly generated data from one Monte Carlo simulation out of 200 simulations. Solid line: true regression curve. Dashed line: (average) Bayesian regression posterior median curve. Shaded region: (average) 95 percent pointwise credible (confidence) region for the Bayesian fit. Top Panel: increasing functions $m_1(x) = \sin(\frac{\pi}{2}x)$; $m_2(x) = 2x \cdot I(x \in [0, 0.25]) + 0.5 \cdot I(x \in [0.25, 0.75]) + (2x - 1) \cdot I(x \in [0.75, 1])$. Bottom Panel: convex functions $m_3(x) = (16/9)(x - 1/4)^2$; $m_4(x) = (-4x + 1) \cdot I(x \in [0, 0.25]) + 0 \cdot I(x \in [0.25, 0.75]) + (4x - 3) \cdot I(x \in [0.75, 1])$.

Figure 4.3: The Rabbit data alongside the Bayesian increasing Bernstein polynomial regression posterior mean curve (solid line) and 95 percent credible intervals (dotted line).

Chapter 5

Mixed Effects Model Subject to Shape Restrictions

5.1 Introduction

Nonparametric regression models have been studied extensively in the literature. The main benefit of nonparametric over parametric model is that the number and nature of the parameters are flexible and not fixed in advance. There have been many proposed approaches to nonparametric modelling, including kernel methods, smoothing splines, penalized splines, polynomial regression, regression splines, and so on (Friedman and Tibshirani 1984; Mukerjee 1988; Mammen et al. 2001; Aït-Sahalia and Duarte 2003; Dette et al. 2006). When the nonparametric framework is applied, the shape of the functional relationship between predictor variables and response variables is determined by the data. However, it is widely believed that in many practical settings the predictors and the response are known to preserve certain shape restrictions, such as monotonicity, concavity etc. The question of how we carefully use this predetermined knowledge in the nonparametric regression has received increasing attentions.

In this chapter we focus on isotonic nonparametric regression models using Bernstein polynomials with a mixed model framework. The need for a mixed effects model is motivated by a

real data on the growth of Sitka spruces under different environmental conditions. Even when each tree is grown under a given experimental condition (e.g. ozone-enriched), the growth rate of trees would exhibit heterogeneous behaviors; which in turn motivates random effects model. The data displayed in Figure 5.1 shows the growth measurements of 79 Sitka spruce trees grown in natural (control) or ozone-enriched environments over a period of 674 days. The ozone-enriched group (shown on the right panel) consists of 54 trees, while the remaining 25 trees compose the control group (shown on the left panel). This data gives us a typical example of longitudinal studies where the goal is often to compare the growth curve of two or more groups. The data has been previously analyzed by Crainiceanu et al. (2005) using an additive mixed effects model. As growth curves can only be non-decreasing, it is natural to fit a shape restricted mixed effects model, such that the regression functions in each group are restricted to be non-decreasing. It is also clear from these spaghetti plots that trees within a group exhibit similar but heterogeneous growth rate. We will give more details on the data analysis in Section 5.4.

[insert Figure 5.1 here.]

Longitudinal study is an important research strategy implemented in agricultural and biomedical research. It usually involves repeated observations or measurements on the same subjects over long periods of time. Many approaches are available to analyze continuous longitudinal data. Linear mixed effects model is one of the most commonly used methods to analyze the longitudinal data. Linear mixed effects models are an important class of statistical models which contain both fixed effects and random effects (and hence the name mixed effects). In general, since there is no closed-form analytical solution to obtain parameter estimation within a linear mixed model, iterative algorithms are generally used by finding progressively better approximations to the exact solution. For instance, the popular Expectation-Maximization (EM) method is often applied to determine these estimators by treating random effects as the “missing data” (Dempster et al. 1977). Furthermore, when linear constraints are added to fixed effects, and/or the error variances are assumed heterogeneous, the computation of max-

imum likelihood becomes more complicated and prohibitively computationally expensive. In this scenario, other iterative algorithms based on estimating equations are useful and easy to implement. Estimating equations approach is robust against the specification of model error distributions. Here we propose an iterative algorithm which does not require the derivation of the log-likelihood function. Instead, this proposed iterative method obtains the estimates by implementing weighted least squares and hence is robust against many parametric error distributions.

We introduce the regression model and provide specific computational details of our algorithm in Section 5.2, and present results from simulation study in the following section. The application of restricted Bernstein polynomial method through linear mixed models to the Sitka spruces data is presented in Section 5.4. The chapter ends with the brief discussion in Section 5.5.

5.2 Models and Algorithms

Consider a longitudinal study involving I subjects of interest, indexed by i . Suppose we obtain J_i number of measurements obtained at time points $t_{i1}, \dots, t_{iJ_i}$ for each of $i = 1, \dots, I$ subjects. Let y_{ij} be the observed response measured on the subject i at time t_{ij} . We may consider the following model for the data $\{(y_{ij}, t_{ij}); i = 1, \dots, I; j = 1, \dots, J_i\}$:

$$y_{ij} = m_i(t_{ij}) + \sigma_i \varepsilon_{ij}, \quad i = 1, \dots, I; j = 1, \dots, J_i, \quad (5.1)$$

where $m_i(\cdot)$ represents the unknown mean trajectory trend of the i th subject, and the ε_{ij} are independent measurement errors with $E(\varepsilon_{ij}) = 0$ and $E(\varepsilon_{ij}^2) = 1$. We also assume that measurement errors are independent with subject-specific random functions $m_i(\cdot)$'s. Note that each of the random regression functions $m_i(\cdot)$ satisfies certain shape restrictions, and belongs to a class of continuous functions subject to a given class of shape restrictions $\mathcal{F}$ as described in Chapter 2.

Following the method of sieves developed in Chapter 2, we construct the constrained Bernstein polynomial sieve $\{\mathcal{F}_N^{(i)}\}$ for each subject i as follows. Given an order N of the Bernstein polynomials, we approximate $m_i(t)$ for $t \in [0, 1]$ as

$$\begin{aligned} B_N^{(i)}(t) &= \sum_{k=0}^N \beta_{ik} \cdot b_k(t, N) \\ &= \sum_{k=0}^N \beta_{ik} \binom{N}{k} t^k (1-t)^{N-k}. \end{aligned} \quad (5.2)$$

In order to capture the heterogeneity among the subjects, we treat the coefficients $\beta_i = (\beta_{i0}, \dots, \beta_{iN})^T$ as subject-specific random effects, which in turn produce random functions $B_N^{(i)}(\cdot)$'s to approximate the $m_i(\cdot)$ functions. Often, it is convenient to split $\beta_i = \tilde{\beta} + \mathbf{b}_i$, where $\tilde{\beta} = (\beta_0, \dots, \beta_N)$ denotes the fixed effects, and $\mathbf{b}_i = (b_{i0}, \dots, b_{iN})$ denotes the random effects with the restriction $E[\mathbf{b}_i] = \mathbf{0}$ and $\text{Var}[\mathbf{b}_i] = \Sigma$. We also assume that $\mathbf{b}_i$'s are uncorrelated across i . Thus, we obtain

$$B_N^{(i)}(t) = \sum_{k=0}^N (\tilde{\beta}_k + b_{ik}) \cdot b_k(t, N), \text{ for } t \in [0, 1], \quad (5.3)$$

where we use the coefficients for the fixed effect $\tilde{\beta}$ to generate population level estimator of the regression function $m(\cdot) = E[m_i(\cdot)]$, which is assumed to belong to the restricted class $\mathcal{F}$ of continuous functions. Consequently, we construct the constrained Bernstein polynomial random sieves $\{\mathcal{F}_N^{(i)}\}$ as follows,

$$\begin{aligned} \mathcal{F}_N^{(i)} &= \left\{ B_N^{(i)}(t) = \sum_{k=0}^N (\tilde{\beta}_k + b_{ik}) \cdot b_k(t, N) : \mathbf{A}_N \tilde{\beta} \geq \mathbf{c} \text{ and } \mathbf{b}_i \stackrel{iid}{\sim} D_N(\mathbf{0}, \Sigma) \right\}, \\ &= \left\{ B_N^{(i)}(t) = \sum_{k=0}^N \tilde{\beta}_k \cdot b_k(t, N) + \sum_{k=0}^N b_{ik} \cdot b_k(t, N) : \right. \\ &\quad \left. \mathbf{A}_N \tilde{\beta} \geq \mathbf{c} \text{ and } \mathbf{b}_i \stackrel{iid}{\sim} D_N(\mathbf{0}, \Sigma) \right\} \end{aligned} \quad (5.4)$$

for $N = 1, 2, \dots$, where the inequality $\mathbf{A}_N \tilde{\beta} \geq \mathbf{c}$ is satisfied componentwise and $D_N(\mathbf{0}, \Sigma)$

denotes a $(N + 1)$ -dimension distribution with mean $\mathbf{0}$ and variance (matrix) Σ . The order of the polynomial, N , grows with the total number of observations. The matrix $\mathbf{A}_N$ represents a full row rank restriction matrix chosen carefully to ensure that the mean of each (function) member in the sieve $\mathcal{F}_N^{(i)}$ preserves the desired shape restrictions.

Thus, to approximate the model (5.1), we use (5.2) - (5.4) and fit the following linear mixed effects models:

$$y_{ij} = \sum_{k=0}^N \tilde{\beta}_k \cdot b_k(t_{ij}, N) + \sum_{k=0}^N b_{ik} \cdot b_k(t_{ij}, N) + \sigma_i \varepsilon_{ij} \quad (5.5)$$

where $i = 1, \dots, I; j = 1, \dots, J_i$, $\mathbf{A}_N \tilde{\boldsymbol{\beta}} \geq \mathbf{c}$ and $\mathbf{b}_i \stackrel{iid}{\sim} D_N(\mathbf{0}, \Sigma)$. Vectorizing (5.5) at the subject level and writing $\mathbf{y}_i = (y_{i1}, \dots, y_{iJ_i})$, $\mathbf{W}_i = ((b_k(t_{ij}, N)))_{J_i \times (N+1)}$, we obtain

$$\mathbf{y}_i = \mathbf{W}_i \tilde{\boldsymbol{\beta}} + \mathbf{W}_i \mathbf{b}_i + \sigma_i \boldsymbol{\varepsilon}_i, \quad (5.6)$$

where $\sigma_i > 0$ denote subject level residual variances. Model (5.6) is a standard expression of linear mixed model with heteroscedastic measurement error variance. With the shape restriction given by $\mathbf{A}_N \tilde{\boldsymbol{\beta}} \geq \mathbf{c}$, we consider this mixed model subject to linear constraints. To obtain the appropriate order of polynomials within the sieve, one may the leave-one-out cross-validation techniques proposed in Wu and Zhang (2002) or V -fold cross-validation technique (see our example in Section 5.4) at the subject levels to choose the best value of N .

In the following, we first introduce linear mixed models with heteroscedastic measurement error variances, and propose our estimating equation based iterative solution to obtain the parameter estimation of a linear mixed model. Our proposed iterative method extends the ideas of Foulley and Quaas (1994) which were based on Gaussian errors. We further extend the iterative method to estimate parameters subject to linear constraints.

5.2.1 Linear Mixed Models

We consider a general linear mixed model allowing for heteroscedastic residual errors,

$$\begin{aligned}
\mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \quad \text{for } i = 1, \dots, I, \\
\mathbb{E}[\mathbf{b}_i] &= \mathbf{0} \text{ and } \text{Var}[\mathbf{b}_i] = \Sigma \\
\mathbb{E}[\boldsymbol{\varepsilon}_i] &= \mathbf{0} \text{ and } \text{Var}[\boldsymbol{\varepsilon}_i] = \sigma_i^2 \mathbf{I}_{J_i} \\
\text{Cov}(\mathbf{b}, \boldsymbol{\varepsilon}) &= \mathbf{0}.
\end{aligned} \tag{5.7}$$

where $\mathbf{y}_i$ is the $J_i \times 1$ response vector for observations measured at J_i time points for the i th subject, $\mathbf{X}_i$ is the $J_i \times p$ design matrix representing predictor values measured at the baseline for the i th subject, $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed-effects, $\mathbf{Z}_i$ is the $J_i \times q$ design matrix capturing the within subject variation for observations in the i th subject, $\mathbf{b}_i$ is the $q \times 1$ vector of random-effects for the i th group, $\boldsymbol{\varepsilon}_i$ is the $J_i \times 1$ vector of residual errors for observations in the i th subject, Σ is the $q \times q$ covariance matrix for the random effects, and $\sigma_i^2 \mathbf{I}_{J_i}$ is the $J_i \times J_i$ covariance matrix for the residual errors in the i th subject. Although we have assumed the residual errors are uncorrelated within the measurements of the i th subject, we may relax this assumption and replace $\sigma_i^2 \mathbf{I}_{J_i}$ by $\sigma_i^2 \Lambda_{J_i}$ where Λ_{J_i} could capture possible autocorrelations among the measurement errors of the i th subject. Notice that even when $\text{Var}[\boldsymbol{\varepsilon}_i] = \sigma_i^2 \mathbf{I}_{J_i}$, we have $\text{Var}[\mathbf{y}_i | \mathbf{x}_i] = \mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \sigma_i^2 \mathbf{I}_{J_i}$ and hence any autocorrelations within the $y_{i1}, \dots, y_{iJ_i}$ is captured by Σ and σ_i^2 's.

Let $\mathbf{Y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_m^T)^T$, $\mathbf{b} = (\mathbf{b}_1^T, \dots, \mathbf{b}_m^T)^T$, and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^T, \dots, \boldsymbol{\varepsilon}_m^T)^T$ each be a $n \times 1$ vector where $n = \sum_{i=1}^I J_i$. Let $\mathbf{X} = (\mathbf{X}_1^T, \dots, \mathbf{X}_I^T)^T$ be a $n \times p$ matrix, and $\mathbf{Z}$ be a $n \times qm$ block diagonal matrix with $(\mathbf{Z}_1, \dots, \mathbf{Z}_m)$ as main diagonal blocks matrixes. It then follows that the equation in (5.7) can be expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \tag{5.8}$$

where $E(\varepsilon) = \mathbf{0}$, $\text{Var}(\varepsilon) = \text{diag}(\sigma_1^2 \mathbf{I}_{J_1}, \sigma_2^2 \mathbf{I}_{J_2}, \dots, \sigma_I^2 \mathbf{I}_{J_I}) \equiv D(\boldsymbol{\sigma})$, and $E(\mathbf{b}) = \mathbf{0}$, $\text{Var}(\mathbf{b}) = \text{diag}(\Sigma, \dots, \Sigma) \equiv \tilde{D}(\Sigma)$. Hence, it follows that

$$\begin{aligned} E(\mathbf{Y}) &= \mathbf{X}\boldsymbol{\beta}, \\ \text{Var}(\mathbf{Y}) &= \Omega(\boldsymbol{\sigma}, \Sigma)^{-1} \\ &= (D(\boldsymbol{\sigma}) + \mathbf{Z}\tilde{D}(\Sigma)\mathbf{Z}^T). \end{aligned}$$

Based on above moment equations, we develop estimation of $\boldsymbol{\beta}$, $\boldsymbol{\sigma}$ and Σ using a sequence of iterative steps, where the over goal is to find $(\boldsymbol{\beta}, \boldsymbol{\sigma}, \Sigma)$ that minimizes $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \Omega(\boldsymbol{\sigma}, \Sigma)(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. Clearly we can not derive a closed-form analytic solution for the optimizer as the above weighted quadratic form is non-linear in $\boldsymbol{\sigma}$ and Σ .

5.2.2 Iterative Method for Linear Mixed Model with Heteroscedastic Errors

For the whole population focus, the interest is in the fixed effects (i.e. $\boldsymbol{\beta}$). Since the model includes heteroscedastic errors as well as the random effect, there is no close-form solution to estimate $\boldsymbol{\beta}$. We propose an iterative algorithm to estimate the parameters. The initial values of the fixed effect $\hat{\boldsymbol{\beta}}^{(0)}$ and the random effect $\hat{\mathbf{b}}^{(0)}$ are obtained from the ordinary least squares method. The initial values of variances are also computed as unbiased estimators through the least squares methods. During the iteration loops, the $\hat{\boldsymbol{\beta}}^{(l)}$ is treated as the iterative re-weighted least squares estimator based on the estimated variances at the $(l - 1)$ -th step.

After $\hat{\boldsymbol{\beta}}^{(l)}$ being updated, $\hat{\mathbf{b}}^{(l)}$ is obtained as the expected value of $\mathbf{b}$ given $\mathbf{Y} = \mathbf{y}$. Based on our model assumption, we have $\mathbf{b} \sim N(\mathbf{0}, \tilde{D}(\Sigma))$ and $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \Omega(\boldsymbol{\sigma}, \Sigma)^{-1})$. The covariance of $\mathbf{b}$ and $\mathbf{Y}$ is equal to $\text{cov}(\mathbf{b}; \mathbf{y}) = \text{cov}(\mathbf{b}; \mathbf{Z}\mathbf{b} + \varepsilon) = \mathbf{Z} \cdot \text{var}(\mathbf{b}) = \mathbf{Z}\tilde{D}(\Sigma)$. Consequently, the joint distribution of $(\mathbf{b}, \mathbf{Y})^T$ can be written as

$$\begin{pmatrix} \mathbf{b} \\ \mathbf{Y} \end{pmatrix} \sim N \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{X}\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \tilde{D}(\Sigma) & \tilde{D}(\Sigma)\mathbf{Z}^T \\ \mathbf{Z}\tilde{D}(\Sigma) & \Omega(\boldsymbol{\sigma}, \Sigma)^{-1} \end{bmatrix} \right). \quad (5.9)$$

From (5.9), it is well known that the conditional distribution (often called the posterior distribution) of $\mathbf{b}$ given $\mathbf{Y} = \mathbf{y}$ is

$$\mathbf{b}|\mathbf{Y} = \mathbf{y} \sim N\left(\tilde{D}(\Sigma)\mathbf{Z}^T\Omega(\boldsymbol{\sigma}, \Sigma)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tilde{D}(\Sigma) - \tilde{D}(\Sigma)\mathbf{Z}^T\Omega(\boldsymbol{\sigma}, \Sigma)\mathbf{Z}\tilde{D}(\Sigma)\right). \quad (5.10)$$

At the l -th iteration, $\hat{\mathbf{b}}^{(l)} = E[\mathbf{b}|\mathbf{Y} = \mathbf{y}]$ is updated with $\hat{\boldsymbol{\beta}}^{(l)}$ (from the current iteration status), $\hat{\boldsymbol{\sigma}}^{(l-1)}$ and $\hat{\Sigma}^{(l-1)}$ (both from the previous iteration status) being plugged into (5.10). We then update the covariance matrixes $\hat{\boldsymbol{\sigma}}^{(l)}$ and $\hat{\Sigma}^{(l)}$ using $\hat{\mathbf{b}}^{(l)}$ and $\hat{\boldsymbol{\beta}}^{(l)}$. The iterative procedure of the algorithm is summarized as follows.

Algorithm 1. Suppose $\hat{\boldsymbol{\beta}}^{(l)}, \hat{\mathbf{b}}^{(l)}, \hat{\Sigma}^{(l)}$ and $\hat{\boldsymbol{\sigma}}^{(l)}$ are the current values at the l -th iteration and we set the initial values to be $\hat{\boldsymbol{\beta}}^{(0)}, \hat{\mathbf{b}}^{(0)}, \hat{\Sigma}^{(0)}$ and $\hat{\boldsymbol{\sigma}}^{(0)}$.

1. Initialize the parameters by values obtained from least squares estimation.

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(0)} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ \hat{\mathbf{b}}^{(0)} &= (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(0)}) \\ \hat{\Sigma}^{(0)} &= \frac{1}{m-1} \sum_1^m (\hat{\mathbf{b}}_i^{(0)})(\hat{\mathbf{b}}_i^{(0)})^T, \text{ where } \hat{\mathbf{b}}^{(0)} = (\hat{\mathbf{b}}_1^{(0)}, \dots, \hat{\mathbf{b}}_m^{(0)}) \\ \hat{\sigma}_i^{2(0)} &= \frac{1}{J_i-1} (\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{(0)} - \mathbf{Z}_i\hat{\mathbf{b}}_i^{(0)})^T (\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{(0)} - \mathbf{Z}_i\hat{\mathbf{b}}_i^{(0)}) \\ \hat{\boldsymbol{\sigma}}^{(0)} &= (\hat{\sigma}_1^{(0)}, \dots, \hat{\sigma}_m^{(0)})^T. \end{aligned}$$

2. Start iterations of updating the parameters. During the l -th iteration status, update the

parameters as follows.

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{(l)} &= (\mathbf{X}^T \Omega(\hat{\boldsymbol{\sigma}}^{(l-1)}, \hat{\Sigma}^{(l-1)}) \mathbf{X})^{-1} \mathbf{X}^T \Omega(\hat{\boldsymbol{\sigma}}^{(l-1)}, \hat{\Sigma}^{(l-1)}) \mathbf{y} \\
\hat{\mathbf{b}}^{(l)} &= \tilde{D}(\hat{\Sigma}^{(l-1)}) \mathbf{Z}^T \Omega(\hat{\boldsymbol{\sigma}}^{(l-1)}, \hat{\Sigma}^{(l-1)}) (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}^{(l)}) \\
\hat{\Sigma}^{(l)} &= \frac{1}{m-1} \sum_1^m (\hat{\mathbf{b}}_i^{(l)} - \bar{\mathbf{b}}^{(l)}) (\hat{\mathbf{b}}_i^{(l)} - \bar{\mathbf{b}}^{(l)})^T, \text{ where } \hat{\mathbf{b}}^{(l)} = (\hat{\mathbf{b}}_1^{(l)}, \dots, \hat{\mathbf{b}}_m^{(l)}) \\
\hat{\sigma}_i^{2(l)} &= \frac{1}{J_i - 1} (\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}^{(l)} - \mathbf{Z}_i \hat{\mathbf{b}}_i^{(l)})^T (\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}^{(l)} - \mathbf{Z}_i \hat{\mathbf{b}}_i^{(l)}) \\
\hat{\boldsymbol{\sigma}}^{(l)} &= (\hat{\sigma}_1^{(l)}, \dots, \hat{\sigma}_m^{(l)})^T.
\end{aligned}$$

Here $\bar{\mathbf{b}}^{(l)} = \frac{1}{m} \sum_{i=1}^m \hat{\mathbf{b}}_i^{(l)}$.

3. Repeat Step 2 until convergence.

Notice that the initial estimators are generated from least squares method. It is easy to show $\hat{\boldsymbol{\beta}}^{(0)}$ and $\hat{\mathbf{b}}^{(0)}|\mathbf{b}$ are unbiased estimators. $E[\hat{\boldsymbol{\beta}}^{(0)}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E[\mathbf{y}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \boldsymbol{\beta}$; $E[\hat{\mathbf{b}}^{(0)}|\mathbf{b}] = E[(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}^{(0)})|\mathbf{b}] = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T E[(\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}^{(0)})|\mathbf{b}] = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{X} \boldsymbol{\beta} + \mathbf{Z} \mathbf{b} - \mathbf{X} \boldsymbol{\beta}) = \mathbf{b}$. Similarly, $\hat{\Sigma}^{(0)}$ and $\hat{\sigma}_i^{2(0)}$ are unbiased estimates of Σ and σ^2 . During the iteration, the estimators are updated through the re-weighted least squares method, which guarantees $\hat{\boldsymbol{\beta}}^{(l)}$ are the best linear unbiased estimators at the given iteration step and hence the converged value of $\hat{\boldsymbol{\beta}}$ essentially solved the optimization problem of minimizing $(\mathbf{Y} - \mathbf{X} \boldsymbol{\beta})^T \Omega(\boldsymbol{\sigma}, \Sigma) (\mathbf{Y} - \mathbf{X} \boldsymbol{\beta})$ if $\boldsymbol{\sigma}, \Sigma$ are known.

It follows that monitoring the convergence of $\hat{\boldsymbol{\sigma}}^{(l)}$ and $\hat{\Sigma}^{(l)}$ is crucial as their convergence essentially provides convergence of $\hat{\boldsymbol{\beta}}^{(l)}$'s. Hence, to ensure the convergence of the algorithm, we set up the following stopping criteria. The iteration procedure stops at the l th step, if

$$\max \left\{ \left\| \hat{\Sigma}^{(l)} (\hat{\Sigma}^{(l-1)})^{-1} - \mathbf{I}_q \right\|_2, \quad \left\| \frac{\hat{\boldsymbol{\sigma}}^{(l)}}{\hat{\boldsymbol{\sigma}}^{(l-1)}} - \mathbf{1}_m \right\|_2 \right\} \leq \delta, \quad (5.11)$$

where δ is an arbitrary small number, and $\|\cdot\|_2$ denotes Euclidean norm. Recall that for a $n \times n$ matrix $A = ((a_{ij}))$, the norm $\|A\|_2$ is defined by $\|A\|_2 = \sqrt{\sum_i \sum_j a_{ij}^2}$ (also known as the Frobenius norm). Similarly, for a $n \times 1$ vector $\mathbf{x} = (x_1, \dots, x_n)^T$, the norm $\|\mathbf{x}\|_2$ is defined by

$\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}$. As before we use the conventional notation $\frac{\mathbf{x}}{\mathbf{y}}$ to denote the componentwise ratio, i.e., $\frac{\mathbf{x}}{\mathbf{y}} = (\frac{x_1}{y_1}, \dots, \frac{x_n}{y_n})^\top$.

5.2.3 Extension to Shape Restricted Mixed Regression Model

Next we revisit our estimation problem of estimating regression function within the framework described in (5.6). We consider again the model such as

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \\ \mathbf{A}\boldsymbol{\beta} &\geq \mathbf{c}, \end{aligned} \tag{5.12}$$

where $\mathbf{A}$ is the constraint matrix, and $\mathbf{c}$ is an arbitrary vector, and $\mathbf{Z}, \mathbf{b}, \boldsymbol{\varepsilon}$ are defined in Section 5.2.1. Recall that, assuming $\mathbf{Z}, \mathbf{b}$ are fixed in the current iteration, the regression model in (5.12) is equivalent to

$$\begin{aligned} \mathbf{E}(\mathbf{Y}) &= \mathbf{X}\boldsymbol{\beta} \\ \mathbf{Var}(\mathbf{Y}) &= \Omega(\boldsymbol{\sigma}, \Sigma)^{-1} = (D(\boldsymbol{\sigma}) + \mathbf{Z}\tilde{D}(\Sigma)\mathbf{Z}^\top) \\ \text{where } \mathbf{A}\boldsymbol{\beta} &\geq \mathbf{c}, \end{aligned} \tag{5.13}$$

where $\Omega(\boldsymbol{\sigma}, \Sigma)$ is the inverse of $\mathbf{Y}$'s variance defined in Section 5.2.1. In this case, the least squares is not an appropriate technique to estimate $\boldsymbol{\beta}$, but the estimation can be obtained as the solution to a quadratic programming problem. At each iteration, $\hat{\boldsymbol{\beta}}^{(l)}$ is obtained through quadratic programming, and the estimations for other parameters keep unchanged as described in Section 5.2.2.

The iterative method to estimate parameters in a linear mixed model with linear constraints is outlined as follows.

Algorithm 2. Suppose $\hat{\boldsymbol{\beta}}^{(l)}, \hat{\mathbf{b}}^{(l)} = (\hat{\mathbf{b}}_1^{(l)}, \dots, \hat{\mathbf{b}}_m^{(l)}), \hat{\Sigma}^{(l)}, \hat{\boldsymbol{\sigma}}^{(l)}$ and $\hat{\sigma}_i^{2(l)}$ for $i = 1, \dots, m$ are the current states at the l -th iteration and we set the initial values to be $\hat{\boldsymbol{\beta}}^{(0)}, \hat{\mathbf{b}}^{(0)} = (\hat{\mathbf{b}}_1^{(0)}, \dots, \hat{\mathbf{b}}_m^{(0)}), \hat{\Sigma}^{(0)},$

$\hat{\boldsymbol{\sigma}}^{(0)}$ and $\hat{\sigma}_i^{2(0)}$ for $i = 1, \dots, m$.

1. Initialize the parameters by values obtained from least squares estimation.

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(0)} &= \arg \min (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \text{ subj. to } \mathbf{A}\boldsymbol{\beta} \geq \mathbf{c} \\ \hat{\mathbf{b}}^{(0)} &= (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(0)}) \\ \hat{\Sigma}^{(0)} &= \frac{1}{m-1} \sum_1^m (\hat{\mathbf{b}}_i^{(0)})(\hat{\mathbf{b}}_i^{(0)})^T \\ \hat{\sigma}_i^{2(0)} &= \frac{1}{J_i-1} (\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}^{(0)} - \mathbf{Z}_i \hat{\mathbf{b}}_i^{(0)})^T (\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}^{(0)} - \mathbf{Z}_i \hat{\mathbf{b}}_i^{(0)}) \\ \hat{\boldsymbol{\sigma}}^{(0)} &= (\hat{\sigma}_1^{(0)}, \dots, \hat{\sigma}_m^{(0)})^T.\end{aligned}$$

2. Start iterations of updating the parameters. During the l -th iteration status, update the parameters as follows.

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(l)} &= \arg \min (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \Omega(\hat{\boldsymbol{\sigma}}^{(l-1)}, \hat{\Sigma}^{(l-1)}) (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &\text{subj. to } \mathbf{A}\boldsymbol{\beta} \geq \mathbf{c} \\ \hat{\mathbf{b}}^{(l)} &= \tilde{D}(\hat{\Sigma}^{(l-1)}) \mathbf{Z}^T \Omega(\hat{\boldsymbol{\sigma}}^{(l-1)}, \hat{\Sigma}^{(l-1)}) (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(l)}) \\ \hat{\Sigma}^{(l)} &= \frac{1}{m-1} \sum_1^m (\hat{\mathbf{b}}_i^{(l)} - \bar{\mathbf{b}}^{(l)})(\hat{\mathbf{b}}_i^{(l)} - \bar{\mathbf{b}}^{(l)})^T \\ \hat{\sigma}_i^{2(l)} &= \frac{1}{J_i-1} (\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}^{(l)} - \mathbf{Z}_i \hat{\mathbf{b}}_i^{(l)})^T (\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}^{(l)} - \mathbf{Z}_i \hat{\mathbf{b}}_i^{(l)}) \\ \hat{\boldsymbol{\sigma}}^{(l)} &= (\hat{\sigma}_1^{(l)}, \dots, \hat{\sigma}_m^{(l)})^T.\end{aligned}$$

Here $\bar{\mathbf{b}}^{(l)} = \frac{1}{m} \sum_{i=1}^m \hat{\mathbf{b}}_i^{(l)}$.

3. Repeat Step 2 until convergence.

The convergence is assessed using the same metric as in (5.11). Notice that $\text{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \Omega(\hat{\boldsymbol{\sigma}}, \hat{\Sigma}) \mathbf{X})^{-1} \equiv \hat{V}_{\boldsymbol{\beta}}$, and inference about $\tilde{m}(x) = \sum_{k=0}^N \tilde{\beta}_k b_k(x, N)$ can be obtained as $\text{Var}(\tilde{m}(x)) = \mathbf{W}(x)^T \hat{V}_{\boldsymbol{\beta}} \mathbf{W}(x)$, where $\mathbf{W}(x) = (b_0(x, N), \dots, b_N(x, N))$. As a consequence, we can compute pointwise 95% confidence intervals.

5.3 A Simulation Study

In this section, we present a simple simulation study to validate the performance of our iterative method on linear mixed model with linear constraints. Note that we are interested in the estimation of $\hat{\beta}$. The estimation of other parameters are by-products of our algorithm.

The simulation study involves the following linear regression model with random effects. Let y_{ij} denote the j -th observation on subject i , for $j = 1, \dots, J_i = 5$ and $i = 1, \dots, I = 80$. We assume observations on different subjects (or clusters) are independent. The repeated measurements conducted over a given subject i share a subject specific random effect, denoted by $\mathbf{b}_i$, where $\mathbf{b}_i \stackrel{iid}{\sim} \mathbf{N}_3(\mathbf{0}, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 1 \end{pmatrix}.$$

Conditional on the subject specific effects, the measurements satisfy

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i,$$

where $\mathbf{X}_i$ is a 5×4 matrix, within which x_{ij} is generated from a standard normal distribution (note that the first column of the design matrix $\mathbf{X}$ is set to be $\mathbf{1}$, corresponding to the intercept column). $\mathbf{Z}_i$ is a 5×3 matrix, within which z_{ij} is generated from $\mathbf{N}(0, 0.2^2)$. For the i th subject, the measurement errors $\boldsymbol{\varepsilon}_i$ are generated from a multivariate normal distribution $\boldsymbol{\varepsilon}_i \sim \mathbf{N}_5(\mathbf{0}, \sigma^2 \mathbf{I}_5)$, where $\sigma^2 = \frac{i}{80}$. One hundred data sets were generated according to this scheme with the true fixed effect $\boldsymbol{\beta} = (1, 2, 3, 4)^T$, which satisfies the given constraint $\mathbf{A}\boldsymbol{\beta} \geq \mathbf{c}$ with

$$\mathbf{A} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix},$$

and $\mathbf{c} = (0, 1, 0)^T$.

For each data set we applied our proposed iterative algorithm described in Section 5.2.3. The stopping rule defined in (5.11) was applied and $\delta = 10^{-2}$ was chosen in the study. In this simulation, the algorithm converged within 20 iterations most of the time. The summary of $\hat{\beta}$ estimations are given in Table 5.1. Similar results can also be found in Figure 5.2. As can be seen from both the table and the figure, it is clearly evident that our proposed iterative method achieves excellent results in this simulation study. Out of these four estimates, $\hat{\beta}_2$ gives largest estimation bias. However, even the largest relative ratio bias over the true β value that comes from $\hat{\beta}_2$ is less than 1.65%.

Table 5.1: Summary of $\hat{\beta}$ estimations. The results are based on 100 MC repetitions.

	$\beta_1 = 1$	$\beta_2 = 2$	$\beta_3 = 3$	$\beta_4 = 4$
Min $\hat{\beta}$	0.8005	1.716	2.886	3.858
Q ₁ $\hat{\beta}$	0.9443	1.926	2.966	3.947
Median $\hat{\beta}$	0.9903	1.970	3.013	4.003
Q ₃ $\hat{\beta}$	1.0520	2.013	3.060	4.052
Max $\hat{\beta}$	1.1930	2.130	3.238	4.155
Bias $\hat{\beta}$	-0.0035	-0.0329	0.0164	0.0008
MCSE($\hat{\beta}$)	0.0831	0.0696	0.0665	0.0713

[insert Figure 5.2 here.]

5.4 Real Data Application

As mentioned in Section 5.1, for the real data example considered in this section we use the growth rate of Sitka spruces data. The data consists of 79 trees, within which 54 are grown in an ozone-enriched atmosphere and the rest 25 are grown in a natural atmosphere. Growth measurements (tree sizes) of each of these 79 trees were measured on a logarithmic scale at the same time points up to 13 times over a period of 2 years. The data, which was first reported

and analyzed by Diggle et al. (2002), is accessible in the R package *SemiPar*.

Crainiceanu et al. (2005) studied this data using an additive mixed model in a Bayesian framework. Their additive model is given as follows,

$$\begin{aligned} y_{ij} &= \gamma_{0i} + \gamma_1 \omega_i + f(x_{ij}) + \varepsilon_{ij}, \\ \gamma_{0i} &\sim \mathbf{N}(0, \sigma_\gamma^2), \end{aligned}$$

where y_{ij} , is the log size of subject i , $1 \leq i \leq 79$, at the j -th measurement time, $1 \leq j \leq 13$, taken on day x_{ij} . In their model, γ_{0i} are random intercepts for each tree, ω_i is the ozone exposure indicator (1 = ozone, 0 = no ozone), γ_1 is a constant ozone effect, and ε_{ij} are random errors. Notice that under this model $E[y_{ij}|\omega_i, x_{ij}] = \gamma_1 \omega_i + f(x_{ij})$ and the (fixed) growth rate's difference is essentially captured by $\gamma_1 = E[y_{ij}|\omega_i = 1, x_{ij}] - E[y_{ij}|\omega_i = 0, x_{ij}]$. Hence, despite the flexibility of modelling the nonlinear growth rate of trees under two environmental conditions, the model cannot capture time varying growth rate differences. Low-rank thin-plate splines were applied to fit the growth rate model $f(\cdot)$ (see Figure 5.5), and obtained the estimate of growth rate difference $\hat{\gamma}_1 = -0.31$ with 95% credible interval $[-0.61, -0.007]$.

[insert Figure 5.5 here.]

We studied the same data, and fitted the unknown regression function with Bernstein polynomial subject to the constraint that growth rate can not be decreasing. The model can be expressed as

$$\begin{aligned} y_{ij} &= \sum_{k=0}^N \beta_{ik} b_k(x_{ij}, N) + \sigma_i \varepsilon_{ij}, \quad j = 1, \dots, J_i = 13 \\ \boldsymbol{\beta}_i &= (\beta_{i0}, \dots, \beta_{iN})^T, \quad i = 1, \dots, I = 79, \end{aligned} \tag{5.14}$$

where N is the order of polynomials. Writing (5.14) at the subject level, we have

$$\mathbf{y}_i = \mathbf{W}_i \boldsymbol{\beta}_i + \sigma_i \boldsymbol{\varepsilon}_i,$$

where $\mathbf{W}_i$ is a $13 \times (N + 1)$ matrix with each element equal to Bernstein basis polynomials $b_k(x_{ij}, N)$. Let $\boldsymbol{\beta}_i = \boldsymbol{\alpha}_0 + \omega_i \boldsymbol{\alpha}_1 + \mathbf{b}_i$ where $\mathbf{b}_i \stackrel{iid}{\sim} \mathbf{N}(0, \Sigma)$ represent the random effects, and $\boldsymbol{\alpha}_0 = (\alpha_{00}, \dots, \alpha_{0N})^T$, $\boldsymbol{\alpha}_1 = (\alpha_{10}, \dots, \alpha_{1N})^T$ represent the fixed effects. We can write the above model as

$$\begin{aligned} \mathbf{y}_i &= \mathbf{W}_i(\boldsymbol{\alpha}_0 + \omega_i \boldsymbol{\alpha}_1 + \mathbf{b}_i) + \sigma_i \boldsymbol{\varepsilon}_i \\ &= \mathbf{W}_i \boldsymbol{\alpha}_0 + \omega_i \mathbf{W}_i \boldsymbol{\alpha}_1 + \mathbf{W}_i \mathbf{b}_i + \sigma_i \boldsymbol{\varepsilon}_i. \end{aligned}$$

Notice that the monotonic non-decreasing constraint can be implemented as follows,

$$\begin{aligned} \alpha_{00} &\leq \alpha_{01} \leq \dots \leq \alpha_{0N}, \\ \text{and } \alpha_{00} + \alpha_{10} &\leq \alpha_{01} + \alpha_{11} \leq \dots \leq \alpha_{0N} + \alpha_{1N}. \end{aligned} \quad (5.15)$$

The above constraint can be conveniently expressed using the following matrix equation $\mathbf{A}\boldsymbol{\alpha} \geq 0$ where

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_0 & \mathbf{0} \\ \mathbf{A}_0 & \mathbf{A}_0 \end{pmatrix}, \quad \text{with } \mathbf{A}_0 = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ & & \ddots & & \\ 0 & 0 & \dots & -1 & 1 \end{pmatrix}_{N \times (N+1)}$$

and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0^T, \boldsymbol{\alpha}_1^T)^T$. Letting $\mathbf{X}_i = (\mathbf{W}_i, \omega_i \mathbf{W}_i)$ and $\mathbf{Z}_i = \mathbf{W}_i$, we express the model as

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\alpha} + \mathbf{Z}_i \mathbf{b}_i + \sigma_i \boldsymbol{\varepsilon}_i \quad (5.16)$$

Subject to $\mathbf{A}\boldsymbol{\alpha} \geq \mathbf{0}$.

Equation (5.16) gives a typical linear mixed effects model, and we applied the iterative algorithm developed in Section 5.2.3. We chose the optimal value of the tuning parameter by 5-fold cross validation method conducted at the subject level. Since the data includes observations of

subjects from the ozone-enriched and natural groups, we first partitioned the subjects from each group into 5 non-overlapping subsets, each including an almost equal number of subjects. Hence, the ozone-enriched group was partitioned into 5 subsets: each of the first four subsets includes 11 subjects, and the fifth subset consists of the rest 10 subjects. Similarly, the natural group was partitioned into 5 subsets, each with 5 subjects. Then, we generated five subgroups of sample, each being composed of one subset randomly selected (without replacement) from the ozone-enriched group and one subset randomly selected (without replacement) from the natural group. For a given value of N , each subgroups would be assigned exactly once as the validation set, and the rest four subgroups together were regarded as training set (see Figure 5.3). The cross validation function, targeting mean curve estimation at the population level, is defined as

$$CV(N) = \frac{1}{V} \sum_{v=1}^V \left(\sum_{i \in I_{-v}} \sum_{j=1}^{J_i} (y_{ij} - \mathbf{X}_{i_j} \hat{\boldsymbol{\alpha}}^{(v)})^2 / \sum_{i \in I_{-v}} J_i \right) \quad (5.17)$$

where $V = 5$, $\mathbf{X}_{i_j}$ is the j th row in the matrix $\mathbf{X}_i$ defined in (5.16), and $\hat{\boldsymbol{\alpha}}^{(v)}$ represents the estimator of $\boldsymbol{\alpha}$ obtained based on the training set (denoted by I_v) and evaluated in the corresponding validation set (denoted by I_{-v}). The optimal order $\hat{N} = 7$ is produced as the minimizer of $CV(N)$ (see Figure 5.4).

[insert Figures 5.4 and 5.3 here.]

Figure 5.6 displays the estimation results for the Sitka spruces data. On the left panel, we plotted the estimated mean growth rate curve at the population level for both ozone-enriched and natural groups. It shows the ozone-enriched environment slows down the growth of the Sitka spruces slightly, which is consistent with the original data findings. On the right panel, we plotted the time-related ozone effect shown as the growth rate difference at various time points, along with 95% pointwise confidence intervals. The dashed line in the plot represents the fixed ozone effect computed in Crainiceanu et al. (2005).

[insert Figure 5.6 here.]

In fact, the model proposed by Crainiceanu et al. (2005) (i.e., C-model) is a special case of our model. Notice that in C-model,

$$\begin{aligned} E[y_{ij}|w_i = 1, x_{ij}] &= \gamma_1 + f(x_{ij}), \\ E[y_{ij}|w_i = 0, x_{ij}] &= f(x_{ij}), \end{aligned}$$

hence the ozone effect, represented by γ_1 , is assumed to be fixed over time. On the other hand, in our model,

$$\begin{aligned} E[y_{ij}|w_i = 1, x_{ij}] &= \sum_{k=0}^N \alpha_{0k} b_k(x_{ij}, N), \\ E[y_{ij}|w_i = 0, x_{ij}] &= \sum_{k=0}^N (\alpha_{0k} + \alpha_{1k}) b_k(x_{ij}, N). \end{aligned}$$

Hence, the ozone effect expressed as the growth rate difference

$$E[y_{ij}|w_i = 1, x_{ij}] - E[y_{ij}|w_i = 0, x_{ij}] = \sum_{k=0}^N \alpha_{1k} \cdot b_k(x_{ij}, N)$$

is not a constant unless $\alpha_{1k} = \gamma_1$ for all k . Figure 5.6(b) clearly indicates that the ozone effect is changing with time, and hence that our model is far better than C-model. Since the estimates $(\hat{\Sigma}, \hat{\sigma})$ (see Appendix B) are the bi-products of our method, we could compute the variances (and pointwise confidence intervals) of the estimated value $\tilde{m}(x)$ at any x point by following the equations derived at the end of Section 5.2.3. We therefore constructed the 95% confidence intervals at 13 observation days using $\text{Var}(\tilde{m}(x))$ and plotted these intervals in Figure 5.6(b). In addition, we tested the validity of C-model which is equivalent to test the null hypothesis $H_0 : \alpha_{0k} = \dots = \alpha_{Nk}$. As all 95% confidence intervals cover the constant ozone's slowing-down effect (i.e., dashed line) calculated by Crainiceanu et al. (2005), it demonstrates that their model assumption about fixed ozone effect is acceptable. However, notice that the zero reference line

is also included by a few confidence intervals where x (days) < 300 . This indicates that the ozone-enriched environment might not have negative growth effects on baby trees, but definitely slows down the growth rate of trees that are more than one-year-old.

5.5 Discussion

We proposed an iterative algorithm to estimate the linear mixed models. The proposed method can be extended to use in a nonparametric Bernstein polynomial-based estimator combined with linear mixed models. A simulation study demonstrates the effectiveness of the proposed algorithm. In the real data analysis, in which we predict new curves of interest, we show the extension of this algorithm to estimate a nonparametric Bernstein polynomial-based estimator successfully.

Figure 5.1: Spaghetti plot: each line represents the growth of an individual Sitka spruce tree over time. Left: control group. Right: ozone-enriched group.

Figure 5.2: Box plots of β estimation. Reference lines in red represent true β 's.

Figure 5.3: 5-fold CV at the subject level

Figure 5.4: CV score vs. N for Sitka spruce data.

Figure 5.5: Thin-plate spline fit by Crainiceanu et al. (2005) for the function for Sitka spruce data (solid) with 95% credible sets.

Figure 5.6: Restricted Bernstein polynomial fit for Sitka spruces date: (a) mean growth rate fit by restricted Bernstein polynomial model; (b) ozone effect shown as growth rate difference at different time points. The fixed ozone effect obtained from Crainiceanu et al. (2005) is shown as dashed line.

Chapter 6

Shape Restricted Nonparametric Regression with Heteroscedastic Errors

6.1 Introduction

In Chapters 2 and 3, the shape restricted regression function estimators that we proposed were obtained under the assumption of homoscedastic variance (i.e., variance function does not change with predictors), yet the real data application in Section 3.5 demonstrates that this assumption may not always hold in reality. For the data set that we analyzed, one may expect a non-constant variance in the observed responses and the conditional variance may vary smoothly as a function of predictor values. This motivates us to obtain estimates of the shape-restricted regression function by allowing the conditional variance to vary with the predictor vector.

The problem of variance estimation in nonparametric regression has received considerable attentions since 1980s. The study of understanding the local variability of the data is important in many practical fields, such as monitoring the risk or volatility in financial time series (Gal-

lant and Tauchen 1997; Härdle and Shi 1997; Fan and Yao 1998), measuring the reliability in prediction (Yao and Tong 1994), making confidence interval for turbulence modelling (Ruppert et al. 1997), constructing correlation structures of spatial data (Opsomer et al. 1999) and so on. Among various nonparametric methods, the kernel-based algorithms have remained one of the most popular approaches to estimate variance function in both univariate and multivariate nonparametric regressions, and their asymptotic properties have been thoroughly studied (Brown and Levine 2007; Wang et al. 2008; Cai et al. 2009).

In this chapter, we construct the nonparametric estimation using Bernstein polynomials as base functions. According to the result by Carnicer and Pena (1993), Bernstein polynomial preserves the shape property best among all approximation polynomials. Moreover, all of its derivatives possess the same convergence properties (Lorentz 1986). Using Bernstein polynomial, we are able to convert a fairly general shape restricted functional regression problem into a (finite-dimensional) least squares problem with (only) linear constraints (for a multitude of different shapes) which can be solved easily using well established quadratic programming method. Following popular practice, our proposed nonparametric regression model adopts a two-stage iterative method which aims at estimating a regression function as well as the variance function simultaneously. We focus on the univariate scenario only in this chapter, the extension to multivariate cases is cumbersome but straightforward by following the approach proposed in Chapter 3. The underline basic idea for the estimation of heteroscedasticity variance (and volatility) function is adopted from the work by Fan and Yao (1998), and is extended to accommodate shape restrictions on the regression function.

The chapter is organized as follows. Section 6.2 presents the regression model and proposes the iterative method to simultaneously estimate the regression function and the variance function. Section 6.3 demonstrates empirical performance of the proposed estimation method using simulation studies. The performance is compared with the estimation based on the homoscedastic model based method proposed in Chapter 2. Section 6.4 provides an illustration of real data application using the Infant Mortality Rate data example studied in Section 3.5.

The chapter ends with a brief conclusion in Section 6.5.

6.2 Models and Algorithms

Let $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ be the set of observations which are assumed to be independently and identically distributed (i.i.d) as (X, Y) . Let $\mathcal{F}$ be a class of smooth functions subject to certain shape restrictions. Let $m(x) = E(Y|X = x) \in \mathcal{F}$ and $V(x) = \text{Var}(Y|X = x) > 0$ represent the (conditional mean) regression function and the conditional variance function respectively. We write the nonparametric regression model of Y_i on X_i as

$$Y_i = m(X_i) + \sqrt{V(X_i)}\epsilon_i, \quad i = 1, \dots, n, \quad (6.1)$$

where $E(\epsilon_i|X_i) = 0$, $\text{Var}(\epsilon_i|X_i) = 1$, the regression function $m(\cdot)$ preserves certain shape restrictions, and the local volatility function (Engle 1982; Fan and Yao 1998) $\sqrt{V(\cdot)}$ can be expressed as a positive valued continuous function of the predictor variable X . Throughout the chapter, we assume that predictor variables are suitably transformed to lie in the unit interval $[0, 1]$ (see Section 2.2.2 in Chapter 2 for further details). From equation (6.1) it easily follows that

$$\begin{aligned} (Y_i - m(X_i))^2 &= V(X_i)\epsilon_i^2, \\ \text{and hence } E[(Y_i - m(X_i))^2] &= V(X_i), \\ \text{and } E[e_i^2] &= V(X_i), \end{aligned} \quad (6.2)$$

where $e = Y - m(X)$ and $E(e^2|X = x) = V(x)$ are the residual error and variance, respectively.

6.2.1 Estimator

For a continuous regression function $m(\cdot)$, its approximating Bernstein polynomial of order N is given by

$$B_N(x; m) = \sum_{k=0}^N m\left(\frac{k}{N}\right) \cdot b_k(x, N),$$

where $b_k(x, N) = \binom{N}{k} x^k (1-x)^{N-k}$ are Bernstein basis polynomials. By the Weierstrass theorem, $B_N(\cdot; m) \rightarrow m(\cdot)$ uniformly over $[0, 1]$ as N goes to infinity (Lorentz 1986).

Using method of sieves, we construct the constrained Bernstein polynomial sieve $\{\mathcal{F}_N^{(\beta)}\}$ for the estimation of the regression function as follows:

$$\mathcal{F}_N^{(\beta)} = \{B_N(x) = \sum_{k=0}^N \beta_k \cdot b_k(x, N) = \mathbf{b}_N^\top(x) \boldsymbol{\beta}_N : \mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{0}\}, \quad N = 1, 2, \dots \quad (6.3)$$

where $\mathbf{b}_N(x) = (b_0(x, N), \dots, b_N(x, N))^\top$ and $\boldsymbol{\beta}_N = (\beta_{N,0}, \dots, \beta_{N,N})^\top$. The order of polynomial N grows as the sample size n increases, e.g., $N = o(n^k)$ with a suitably chosen $k > 0$. The shape constraints are satisfied by $\mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{0}$, within which the matrix $\mathbf{A}_N$ is chosen in such a way that each (function) member in the sieve $\mathcal{F}_N^{(\beta)}$ preserves the desired shape restrictions.

Note that, if the regression function $m(\cdot)$ is known, we can regard the estimating of variance function as a nonparametric regression problem. In practice, $m(\cdot)$ is unknown and can usually be substituted by a nonparametric regression estimator, e.g., the one from the above proposed sieve $\mathcal{F}_N^{(\beta)}$. This motivates us to propose a two-step estimation procedure. Here we construct another constrained Bernstein polynomial sieve $\{\mathcal{F}_M^{(\gamma)}\}$ for the estimation of the variance function as follows

$$\mathcal{F}_M^{(\gamma)} = \{V_M(x) = \sum_{k=0}^M \gamma_k \cdot b_k(x, M) = \mathbf{b}_M^\top(x) \boldsymbol{\gamma}_M : \mathbf{I}_M \boldsymbol{\gamma}_M \geq \mathbf{0}\}, \quad M = 1, 2, \dots, \quad (6.4)$$

where $\mathbf{I}_M$ denotes the M -dimensional identity matrix.

The estimators of the regression function and variance function are obtained through the

intersection of two function spaces $\mathcal{F}_N^{(\beta)} \cap \mathcal{F}_M^{(\gamma)}$. Let $B_{N_1}(\cdot) \in \mathcal{F}_N^{(\beta)}$ and $V_{N_2}(\cdot) \in \mathcal{F}_M^{(\gamma)}$ be the Bernstein polynomials expansion to approximate model (6.1) which can be written as

$$\begin{aligned} B_{N_1}(x_i) &= \sum_k \beta_k \cdot b_k(x_i, N_1), \text{ such that } \mathbf{A}_{N_1} \boldsymbol{\beta}_{N_1} \geq 0, \\ V_{N_2}(x_i) &= \sum_k \gamma_k \cdot b_k(x_i, N_2), \text{ such that } \boldsymbol{\gamma}_{N_2} \geq 0, \end{aligned} \quad (6.5)$$

where $\boldsymbol{\beta}_{N_1} = (\beta_0, \dots, \beta_{N_1})^\top$ and $\boldsymbol{\gamma}_{N_2} = (\gamma_0, \dots, \gamma_{N_2})^\top$. The goal would be to minimize mean weight squared errors to obtain the estimates of $\boldsymbol{\beta}_{N_1}$ and $\boldsymbol{\gamma}_{N_2}$ and hence the corresponding estimators of $\hat{B}_{N_1}(\cdot)$ and $\hat{V}_{N_2}(\cdot)$.

However, in practice as there is no closed-form analytic solution to obtain parameter estimation expressed as the weighted least squares optimization problem, we next introduce an iterative algorithm to seek successive approximations to the exact solution $(\hat{m}, \hat{V})$.

6.2.2 Computation of the Estimator

To simplify the notation, we assume $N_1 = N_2 = N$, then (6.5) can be written in a matrix form as follows

$$\begin{aligned} B_N(\mathbf{x}) &= \mathbf{W}(\mathbf{x})^\top \boldsymbol{\beta}_N, \\ V_N(\mathbf{x}) &= \mathbf{W}(\mathbf{x})^\top \boldsymbol{\gamma}_N, \end{aligned} \quad (6.6)$$

where $\mathbf{W}(x) = (W_0(x), \dots, W_N(x))^\top$ with $W_k(x) = b_k(x, N)$, $\boldsymbol{\beta}_N = (\beta_0, \dots, \beta_N)^\top$ and $\boldsymbol{\gamma}_N = (\gamma_0, \dots, \gamma_N)^\top$. Once the estimators $(\hat{\boldsymbol{\beta}}_N, \hat{\boldsymbol{\gamma}}_N)$ are achieved, the estimated regression and variance functions can be obtained through (6.6). To obtain the estimators $(\hat{m} = \hat{B}_N(\cdot), \hat{V} = \hat{V}_N(\cdot))$, it

is equivalent to solve the following optimization problem:

$$\begin{aligned}
& \text{minimize} && \sum_{i=1}^n \frac{(y_i - B_N(x_i))^2}{V_N(x_i)} && \text{with respect to } (\boldsymbol{\beta}_N, \boldsymbol{\gamma}_N) \\
& \text{subject to:} && \mathbf{A}_N \boldsymbol{\beta}_N \geq 0 \\
& && \mathbf{I}_N \boldsymbol{\gamma}_N \geq 0.
\end{aligned} \tag{6.7}$$

Since the sum of weighted squares is a nonlinear function in $(\boldsymbol{\beta}_N, \boldsymbol{\gamma}_N)$, we can no longer use the straightforward quadratic programming (QP) algorithm to obtain the estimate of $(\hat{\boldsymbol{\beta}}_N, \hat{\boldsymbol{\gamma}}_N)$. We propose a two-step iterative algorithm to approximate the exact solution. Each step requires solving a weighted least squared (WLS) subject to linear constraint and hence can be considered as a QP problem. In the first step, we assume the variance function has been approximated by $\hat{V}_N(x)$. In other words, $\boldsymbol{\gamma}_N$ is treated as a known parameter and (6.7) can consequently be simplified to a WLS optimization problem with respect to only $\boldsymbol{\beta}_N$. We implement a QP algorithm to compute the regression estimator $\hat{B}_N(x)$ subject to a given shape constraint (i.e., Step 1) as

$$\hat{B}_N(x) = \arg \min_{\boldsymbol{\beta}_N: \mathbf{A}_N \boldsymbol{\beta}_N \geq 0} \sum_{i=1}^n \frac{(y_i - B_N(x_i))^2}{\hat{V}_N(x_i)} = \arg \min_{\boldsymbol{\beta}_N: \mathbf{A}_N \boldsymbol{\beta}_N \geq 0} \sum_{i=1}^n \frac{(y_i - \mathbf{W}(x_i)^\top \boldsymbol{\beta}_N)^2}{\hat{V}_N(x_i)},$$

where $\mathbf{W}(x) = (W_0(x), \dots, W_N(x))^\top$ with $W_k(x) = b_k(x, N)$ and $\boldsymbol{\beta}_N = (\beta_0, \dots, \beta_N)^\top$. Notice that $\hat{V}_N(x_i)$ are “known” and hence the above optimization can be solved using a QP algorithm.

In the second step, we update $\hat{\boldsymbol{\gamma}}_N$ using the estimated residuals from Step 1. We consider the residual-based estimation of the conditional variance function (equivalent to the estimation of $\hat{\boldsymbol{\gamma}}_N$ by (6.6)). Following the approach proposed by Fan and Yao (1998), we first create estimated squared residual denoted by $\hat{e}_i^2 = \{Y_i - \hat{B}_N(X_i)\}^2$ where $\hat{B}_N(X_i)$ is obtained by Step 1. From (6.2), we have the relation $E[e^2|X = x] = V(x)$, where $e = Y - m(X)$. This leads us

to the residual-based estimator $\hat{V} = \hat{V}_N(x)$ by computing

$$\hat{V}_N(x) = \arg \min_{\gamma_N: \mathbf{I}_N \gamma_N \geq 0} \sum_{i=1}^n (\hat{e}_i^2 - V_N(x_i))^2 = \arg \min_{\gamma_N: \mathbf{I}_N \gamma_N \geq 0} \sum_{i=1}^n (\hat{e}_i^2 - W_i \gamma_N)^2$$

which can again be solved using the efficient QP algorithm.

We summarize the proposed two-step iterative algorithm to estimate $\hat{m}$ and $\hat{V}$ as follows.

Algorithm. *Given the order of polynomials N , suppose $\hat{\beta}_N^{(l)}, \hat{\gamma}_N^{(l)}$, and $\mathbf{D}^{(l)} = \text{diag}(W \gamma_N^{(l)})$ are the current states at the l -th iteration.*

1. *Initialize the γ parameters by setting following values.*

$$\begin{aligned} \hat{\gamma}_N^{(0)} &= \mathbf{1} \\ \mathbf{D}^{(0)} &= \text{diag}(W \gamma_N^{(0)}) \end{aligned}$$

2. *Next for $l = 1, 2, \dots$ compute*

$$\begin{aligned} \text{(i) } \mathbf{D}^{(l)} &= \text{diag}(W \gamma_N^{(l-1)}) \\ \hat{\beta}_N^{(l)} &= \arg \min_{\beta_N} (\mathbf{y} - \mathbf{W} \beta_N)^T \{\mathbf{D}^{(l)}\}^{-1} (\mathbf{y} - \mathbf{W} \beta_N) \\ &\quad \text{subj. to } \mathbf{A}_N \beta_N \geq \mathbf{0} \\ \text{(ii) } \mathbf{e}^{(l)} &= \mathbf{y} - \mathbf{W} \hat{\beta}_N^{(l)} \\ \hat{\gamma}_N^{(l)} &= \arg \min_{\gamma_N} (\{\mathbf{e}^{(l)}\}^2 - \mathbf{W} \gamma_N)^T (\{\mathbf{e}^{(l)}\}^2 - \mathbf{W} \gamma_N) \\ &\quad \text{subj. to } \mathbf{I} \gamma_N \geq \mathbf{0} \end{aligned}$$

Repeat Steps (i) and (ii) above until convergence. Notice that for both Steps (i) and (ii) we can use QP algorithm to obtain $\hat{\beta}_N^{(l)}$ and $\hat{\gamma}_N^{(l)}$.

To determine the convergence of the algorithm, we set up the following stopping criteria:

$$\begin{aligned}
& \text{Compute for} \quad l = 1, 2, \dots \\
& \varepsilon^{(l)} = \left\| \frac{\mathbf{W}\boldsymbol{\beta}_N^{(l)} - \mathbf{W}\boldsymbol{\beta}_N^{(l-1)}}{\mathbf{W}\boldsymbol{\beta}_N^{(l-1)}} \right\|_{\infty} \\
& \kappa^{(l)} = \left\| \frac{\mathbf{W}\boldsymbol{\gamma}_N^{(l)} - \mathbf{W}\boldsymbol{\gamma}_N^{(l-1)}}{\mathbf{W}\boldsymbol{\gamma}_N^{(l-1)}} \right\|_{\infty} \\
& \text{and stop if} \quad \max\{\varepsilon^{(l)}, \kappa^{(l)}\} \leq \delta.
\end{aligned}$$

The threshold $\delta > 0$ is an arbitrary small number, $\|\cdot\|_{\infty}$ denotes the maximum norm defined as $\|\mathbf{a}\|_{\infty} = \max_{0 \leq k \leq N} |a_k|$. In above, we use the conventional notation $\frac{\mathbf{a}}{\mathbf{b}}$ to denote the componentwise ratio, i.e., $\frac{\mathbf{a}}{\mathbf{b}} = (\frac{a_0}{b_0}, \dots, \frac{a_N}{b_N})^{\top}$.

Finally we use a V -fold cross-validation method to choose a optimal order of Bernstein polynomials N . Our cross-validation method is based on minimizing weighted squares defined in (6.7). Details of the V -fold cross-validation can be found in Section 2.2.2 in Chapter 2.

6.3 Simulations

In this section, we explore the performance of the proposed estimation method using simulated data sets obtained from a heteroscedastic model. We also explore the consequences of violating the assumption of heteroscedasticity by fitting a homoscedastic model to a data set obtained from a heteroscedastic model.

To generate data we used the regression (or mean) function: $m(x) = \log(x)$, and the variance function: $V(x) = \exp(x-2) + x^3$ for $x \in [0, 1]$. The mean function is continuous on the support $[0, 1]$ with an increasing shape constraint. We simulated 401 Monte Carlo replication of samples of size $n = 50$ from $Y_i = m(X_i) + \sqrt{V(X_i)}\epsilon_i$, where $X_i \stackrel{iid}{\sim} U(0, 1)$, and $\epsilon_i \stackrel{iid}{\sim} N(0, 1)$. For each generated data set, the proposed estimation methods based on heteroscedastic variance assumptions and homoscedastic variance assumptions were applied to obtain the regression function estimators. The order of Bernstein polynomials N has been selected using a V -fold

cross validation method with $V = 5$ for each simulated sample. The MC average values of $\hat{N}$ were equal to 10.15 and 12.32 for the heteroscedastic model and the homoscedastic model respectively.

To compare the prediction accuracy, we consider the measurement metrics taken over the fixed equidistant prediction points $x_k = \frac{k}{n_{grid}} \in [0, 1]$ where $n_{grid} = 99$ and $k = 1, \dots, n_{grid}$. The performances of the estimators of the regression function are evaluated by two measurement metrics, namely the root integrated squared error (RISE) and the integrated absolute error (IAE) defined as

$$RISE = \sqrt{n_{grid}^{-1} \sum_{k=1}^{n_{grid}} (\hat{m}(x_k) - m(x_k))^2},$$

$$IAE = n_{grid}^{-1} \sum_{k=1}^{n_{grid}} |\hat{m}(x_k) - m(x_k)|,$$

where $\hat{m}(\cdot)$ represents the estimator of the true regression function $m(\cdot)$. For each estimation method, we report the average values of the above two measurement metrics based on 401 Monte Carlo simulations in Table 6.1. We observe from the table that both estimators of the regression (mean) function give similar results, while the heteroscedastic estimator performs slightly better in terms of both criteria, although the improvement is not statistically significant.

Table 6.1: Performance of regression estimator under heteroscedastic and homoscedastic error assumptions based on 401 simulations. Monte Carlo standard errors are displayed in the parentheses.

Model	MRISE	MIAE
hetero	0.2167 (0.0034)	0.1460 (0.0017)
homo	0.2263 (0.0036)	0.1493 (0.0019)

Figures 6.1 and 6.2 give the estimated regression functions obtained from the heteroscedastic method and the homoscedastic method respectively. The 95% confidence band of the estimated regression function is also depicted in the figures. The typical data sample shown as scatter

plots in the figures was selected in such a way that the corresponding ISE equals to its median among all 401 simulations. Both estimators for regression functions and variance functions shown in the figures are averaged over 401 simulations. It is clear that these two estimation methods to approximate the shape-restricted regression function perform equally well in this simulation study. In terms of the estimation for variance function, the heteroscedastic method is able to capture the increasing trend of underline variance function very well, as we observe the increasing width of 95% confidence bands along with predictor's values in Figure 6.1. To get further insights of variance estimation, in Figure 6.3 we plot the estimated conditional variances with comparison to the true conditional variance function along with 95% pointwise band based on Monte Carlo runs. We observe that in general, the estimation of variance function $\hat{V}$ fits the true variance function very well, while when $x \in [0, 0.1) \cup (0.9, 1]$ it shows some boundary effect.

[insert Figures 6.1 - 6.3 here.]

6.4 Infant Mortality Rate Data Example

In Section 3.5 of Chapter 3, we studied the Infant Mortality Rate data and found that the variance in this example may not be constant as the value of a predictor variable increases. More details about IMR data can be referred to Section 3.5. In this section, we are interested in the impact of the predictor variable the level of Adult Literacy Rate (ALR) on the response variable Infant Mortality Rate (IMR). Our goal is to re-analyze this data and estimate the regression function and the variance function under the heteroscedasticity assumption. The 176 observations of the IMR data is depicted as the scatter plot in Figure 6.4(a). The scatter plot of the data suggests the decreasing trend, and the decreasing shape-restriction is applied in the model (6.7).

The proposed heteroscedastic model has been implemented to fit this data. The order of Bernstein polynomials N was selected by 7-fold cross-validation and we found the optimal

value $\hat{N} = 6$. We provided the estimated regression function $\hat{m}_N$ on the left panel Figure 6.4. The right panel depicted the absolute residuals as well as the estimated conditional volatility function $\sqrt{\hat{V}}$ along with the values of predictor variable ALR. The estimated regression function captures the decreasing trend in the data, while the estimated volatility function shows a reverse U-shaped structure.

[insert Figures 6.4 here.]

6.5 Conclusion

In this chapter, we proposed an iterative method to estimate the regression function and variance function under the heteroscedastic variance assumption. The estimation of variance function, that applies another nonparametric regression to the squared residuals, is a data-driven and nonparametric method. We construct the nonparametric estimation using Bernstein polynomials as base functions. According to the result by Carnicer and Pena (1993), Bernstein polynomial preserves the shape property best among all approximation polynomials. Moreover, all of its derivatives possess the same convergence properties (Lorentz 1986). Using Bernstein polynomial, we are able to convert a fairly general shape restricted functional regression problem into a (finite-dimensional) least squares problem with (only) linear constraint (for a multitude of different shapes) which can be solved easily using well established quadratic programming method. Following popular practice, our proposed nonparametric regression model adopts a two-stage iterative method which aims at estimating a regression function as well as the variance function simultaneously. Numerical studies from both simulations and real data application show the method performs well. We focus on the univariate scenario only in this chapter, the extension to multivariate cases is cumbersome but straightforward by following the approach proposed in Chapter 3. The underline basic idea for the estimation of heteroscedasticity variance (and volatility) function is adopted from the work by Fan and Yao (1998), and is extended to accommodate shape restrictions on the regression function.

Figure 6.1: The heteroscedastic estimation: the estimated regression function (black dashed), the true function (red solid), 95% confidence interval (gray dashed).

Figure 6.2: The homoscedastic estimation: the estimated regression function (black dashed), the true function (red solid), 95% confidence interval (gray dashed).

Figure 6.3: The estimation of conditional variance function (black dashed) vs. the true conditional variance function (red solid), along with 95% pointwise band based on MC runs (gray dashed).

Figure 6.4: Infant mortality data (a) Raw data and the estimated regression function; (b) Absolute residuals and their estimated volatility function.

Chapter 7

Discussions and Future Work

In this thesis, we proposed nonparametric sieved estimators based on a nested sequence of shape-restricted Bernstein polynomials. The proposed nonparametric estimator has been shown to be easily adapted to accommodate many popular shape restrictions and computationally attractive. Numerical results derived from simulated data sets and real data analysis were used to illustrate the superior performance of the proposed estimator compared to a few other available estimator in terms of various goodness of fit metrics. Since the standard methods to derive asymptotic distributions may not be applicable directly to the estimator, we derived the the inference through the bootstrap method and the Bayesian method.

Furthermore, we extended our work and proposed its application in the linear mixed effects model framework. The implementation was obtained using iterative methods. We have shown the effectiveness of our methodology through both simulation studies and the real data study. In addition, we proposed the estimator of the shape-restricted regression function under the assumption of heteroscedastic variance. We adopted a two-stage iterative method which simultaneously estimates the conditional mean function and the conditional variance function. The proposed method has been applied in the analysis of several simulation data and real data examples.

We next propose some directions about future research, particularly regarding the Bayesian

methodology, which previously leaves many interesting open areas for our further inquiries.

7.1 Bayesian Model with Varying Dimensions

In Chapter 4, we discussed a Bayesian approach to estimate nonparametric shape-restricted regression problem. In the proposed approach, we did not consider the prior distribution on the order of Bernstein polynomial expansions. The literature suggests that Reversible-jump Markov chain Monte Carlo (RJMCMC) allows simulation of the posterior distribution on spaces of varying dimensions. Instead of choosing the optimal order of polynomials N (using DIC or other informative criteria), we could use the reversible jump methods to incorporate the uncertainty in choosing N .

The general formulation of RJMCMC is presented by Green (1995). Chapter 11 in Robert and Casella (2004) provides more implementation details. RJMCMC is a generalized version of the Metropolis-Hastings algorithm. Let $\{\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_N, \dots\}$ be a countable set of all possible function spaces, within which the index integer N denotes the order of Bernstein polynomials in the proposed estimator for a shape-restricted function. Let $\mathcal{M} = \bigcup_{N=2}^{\infty} \mathcal{M}_N$ denotes the whole state space. Suppose (t, β_t) denotes the current state in the Markov chain, where β_t denotes the parameter vector in the $(N+1)$ -th dimensional model $\mathcal{M}_N$. RJMCMC algorithm draws samples from the whole function space $\mathcal{M}$. If the model moves in a Markov chain from the state (N, β_N) to another state $(N', \beta_{N'})$ where $N \neq N'$, we call that the model space varies trans-dimensionally. In order to use RJMCMC, we need to place a prior distribution $Pr(N)$ to the order N , and apply the similar sampling scheme of Metropolis-Hastings algorithm. The nontrivial issue nontrivial in this direction of extension is that MCMC convergence might be too slow and the diagnosis of MCMC convergence is not straightforward.

7.2 Testing for Shape Restrictions

Based on the potential improvement of our Bayesian method proposed in Section 7.1, another possible extension is that we can develop a probability model to quantify the appropriate shape constraints in the application. Note that, in the current study we always assume that we collect proactive information about the shape constraint either from the expertise opinion or from the data itself. In this extension, we are not going to pre-assume any shape restrictions to the regression model. Instead, the proposed method will give the posterior probabilities of any possible shape restrictions.

Using the formulation described in Chapters 2 and 3 the infinite dimensional model $\mathbf{Y} = m(\mathbf{X}) + \epsilon$ can conveniently be reduced to a finite dimensional model and can be expressed as

$$\mathbf{Y} = \mathbf{W}_N \boldsymbol{\beta}_N + \epsilon,$$

where $\mathbf{W}_N$ is the Bernstein polynomial matrix where each element represents Bernstein polynomial basis at $\mathbf{x}$ points, and $\boldsymbol{\beta}_N$ is the vector of coefficients. For the moment, suppose we disregard any prior information that we may have $m(\cdot)$ or equivalently about restricting on $\boldsymbol{\beta}_N$. Within a Bayesian approach, we consider priors on the parameters in the model,

$$\begin{aligned} \mathbf{Y} | \boldsymbol{\beta}_N, \sigma^2, N &\sim N_n(\mathbf{W} \boldsymbol{\beta}_N, \sigma^2 \mathbf{I}_n), \\ \sigma^2 &\sim \text{IG}(a, b), \\ N &\sim \text{Poi}(\lambda) \end{aligned}$$

Following RJMCMC, we will obtain samples of $\hat{N}^{(l)}$ and $\hat{\boldsymbol{\beta}}_N^{(l)}$ for the l -th iteration. Record $(\hat{N}^{(l)}, \hat{\boldsymbol{\beta}}_N^{(l)})$ for all iterations $l = 1, 2, \dots, L$. Suppose a certain shape constraint is represented by $\mathbf{R}_N \boldsymbol{\beta}_N \geq \mathbf{c}$, where $\mathbf{R}_N$ is the relevant shape constraint matrix as described in Chapters 2 and 3. Then the posterior probability of this shape constraint, which is given by $Pr(\mathbf{R}_N \boldsymbol{\beta}_N > \mathbf{c} | \mathbf{y}, \mathbf{x})$,

can be approximated by

$$\frac{1}{L} \sum_{l=1}^L I(\mathbf{R}_{N^{(l)}} \boldsymbol{\beta}_{N^{(l)}} \geq \mathbf{c}),$$

where $\mathbf{R}_{N^{(l)}}$ represents the shape matrix corresponding to the order $N^{(l)}$ at the l -th sample of the Markov chain, and $I(\cdot)$ is the indicator function. Notice that once $\{(\boldsymbol{\beta}_{N^{(l)}}, N^{(l)}), l = 1, \dots, L\}$ have been obtained after discarding few initial samples, we can compute the posterior probabilities of any given shape by approximately choosing the restriction matrixes $\mathbf{R}$'s. Although the above method is simple to implement, the assessment and diagnosis of MCMC convergence need to be critically examined before we use this method in practice.

REFERENCES

- Aït-Sahalia and Duarte (2003), "Nonparametric option pricing under shape restrictions," *Journal of Econometrics*, 116, 9-47.
- Bacchetti, P. (1989), "Additive isotonic models," *Journal of the American Statistical Association*, 84, 289-294.
- Barlow, R., Bartholomew, D., Bremner, J. and Brunk, H. (1972), *Statistical Inference under Order Restrictions*, John Wiley and Sons Inc.
- Beresteanu, A. (2004), "Nonparametric Estimation of Regression Functions under Restrictions on Partial Derivatives," *Tech. Report*, Duke University.
- Bernstein, S. (1912), "émonstration du théorém de Weierstrass, fondée sur le calcul des probabilités," *Communications of Kharkov Mathematical Society*, 13, 1-2.
- Birke, M. and Dette, H. (2007), "Estimating a convex function in nonparametric regression," *Scandinavian Journal of Statistics*, 34, 384-404.
- Bollaerts, K., Eilers, P. and van Mechelen I. (2006), "Simple and Multiple P-splines Regression with Shape Constraints," *The British Psychological Society*, 59, 451-469.
- Brown, L.D. and Levine, M. (2007), "Variance estimation in nonparametric regression via the difference sequence method," *The Annals of Statistics*, 35, 2219C2232.
- Brunk, H.D. (1955), "Maximum Likelihood Estimates of Monotone Parameters," *The Annals of Mathematical Statistics*, 26(4), 607-616.
- Cai, T., Levine, M. and Wang, L. (2009), "Variance function estimation in multivariate nonparametric regression with fixed design," *Journal of Multivariate Analysis*, 100, 126-136.
- Carnicer, J.M. and Pena, J.M. (1993), "Shape Preserving Representations and Optimality of the Bernstein Basis," *Advances in Computational Mathematics*, 1, 173-196.
- Chak, P.M., Madras, N. and Smith, B. (2001), "Flexible Functional Forms: Bernstein Polynomials," *Tech. Report*, York University.
- Chak, P.M., Madras, N. and Smith, B. (2005), "Semi-nonparametric Estimation with Bernstein Polynomials," *Economics Letters*, 89, 153-156.
- Chang, I.S., Chien, L.C., Hsiung, C.A., Wen, C.C. and Wu, Y.J. (2007), "Shape Restricted Regression with Random Bernstein Polynomials," *Lecture Notes-Monograph Series*, 54, 187-202.
- Chib, S. and Greenberg, E. (1995), "Understanding the MetropolisCHastings Algorithm", *American Statistician*, 49(4), 327C335.

- Corbeil, R. R. and Searle, S. R. (1976), "Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model", *Technometrics*, 18, 31-38.
- Crainiceanu, C.M., Ruppert, D. and Wand, M.P. (2005), "Bayesian Analysis for Penalized Spline Regression Using WinBUGS", *Journal of Statistical Software*, 14, 1-24.
- Curtis, S.M. and Ghosh, S.K. (2011), "A Variable Selection Approach to Monotonic Regression with Bernstein Polynomials," *Journal of Applied Statistics*, <http://dx.doi.org/10.1080/02664761003692423>.
- Davidian, M. and Carroll, R.J. (1987), "Variance Function Estimation," *Journal of the American Statistical Association*, 82, 1079-1091.
- Dempster, A. P., Laird, N. M. and Rubin, D.B. (1977), "Maximum likelihood from incomplete data via the EM algorithm", *Journal of Royal Statistical Society B*, 39, 1-38.
- Dette, H. Neumeyer, N., and Pilz, K.F. (2006), "A simple nonparametric estimator of a monotone regression function," *Bernoulli*, 12, 469-490.
- Dette, H. and Scheder, R. (2007), "Strictly Monotone and Smooth Nonparametric Regression for Two or More Variables," *Canadian Journal of Statistics*, 34(4), 535-561.
- Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002), *Analysis of Longitudinal Data*, 2nd Edition, Oxford University Press.
- Dykstra, R.J. (1983), "An Algorithm for Restricted Least Squares Regression," *Journal of the American Statistical Association*, 78, 837-842.
- Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife", *The Annals of Statistics*, 7(1), 1C26.
- Efron, B, and Tibshirani, R.J. (1993). *An introduction to the bootstrap*, Chapman and Hall, London.
- Efron, B. and Tibshirani, R.J. (1997), "Improvements on Cross-Validation: The .632 + Bootstrap Method," *Journal of the American Statistical Association*, 92(438), 548-560.
- Efron, B. (2011), "The bootstrap and Markov chain Monte Carlo," Stanford University, <http://www-stat.stanford.edu/~ckirby/brad/papers/2011BootstrapMCMC.pdf>.
- Engle, R.F. (1982), "Autoregressive conditional heteroscedasticity with estimates of variance of U.K. inflation", *Econometrica*, 50, 987-1008.
- Fan, J. and Yao, Q. (1998), "Efficient estimation of conditional variance functions in stochastic regression," *Biometrika*, 85(3), 645-660.
- Foulley, J.L. and Quaas, R.L. (1994), "Statistical analysis of heterogeneous variances in Gaussian linear mixed models", *Proceedings of the Fifth World Congress on Genetics Applied to Livestock Production*, University of Guelph, Guelph, ON, Canada, 18, 341-348.

- Fraser, D.A.S. and Massam, H. (1989), "A Mixed Primal-Dual Bases Algorithm for Regression under Inequality Constraints: Application to Concave Regression," *Scandinavian Journal of Statistics*, 16, 65-74.
- Friedman, J. and Tibshirani, R. (1984), "The Monotone Smoothing of Scatterplots," *Technometrics*, 26(3), 243-250.
- Gal, S.G. (2008), *Shape-Preserving Approximation by Real and Complex Polynomials*, Birkhäuser Boston.
- Gallant, A.R. and Golub, G.H. (1984), "Imposing Curvature Restrictions on Flexible Functional Forms," *Journal of Econometrics*, 38, 295-321.
- Gallant, A.R. and Tauchen, G. (1997), "Estimation of Continuous Time Models for Stock Returns and Interest Rates," *Macroeconomic Dynamics*, 1, 135-168 .
- Gelfand, A.E. and Smith, A.F.M. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 398-409.
- Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003), *Bayesian Data Analysis*, 2nd edition, CRC Press.
- Geman, A. (2006), "Prior distributions for variance parameters in hierarchical models," *Bayesian Analysis*, 1(3), 515-533.
- Geman, S. and Hwang C. (1982), "Nonparametric Maximum Likelihood Estimation by the Method of Sieves," *The Annals of Statistics*, 19(2), 401-414.
- Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine*, 6, 721-741.
- Goldfarb D. and Idnani A. (1982), "Dual and Primal-Dual Methods for Solving Strictly Convex Quadratic Programs," in *Numerical Analysis* (Lecture Notes in Mathematics, No. 909), eds. J.P. Hennart, Springer-Verlag, Berlin, pp. 226-239.
- Goldfarb D. and Idnani A. (1983), "A Numerically Stable Dual Method for Solving Strictly Convex Quadratic Programs," *Mathematical Programming*, 27, 1-33.
- Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*, 3rd edition. Johns Hopkins University Press.
- Gourieroux, C. Holly, A., and Monfrot, A. (1982), "Likelihood Ratio Test, Wald Test, and Kuhn-Tucker Test in Linear Models with Inequality Constraints on the Regression Parameters," *Econometrica*, 50(1), 63-80.
- Green, P.J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82(4), 711-732.
- Grenander, U. (1981), *Abstract Inference (Probability & Mathematical Statistics)*, John Wiley and Sons, New York.

- Groeneboom, P. Jongbloed, G., and Wellner, J.A. (2001), "Estimation of convex functions: characterizations and asymptotic theory," *The Annual of Statistics*, 26, 1653-1698.
- Gyorfi, L., Kohler, M., Krzyzak, A. and Walk, H. (2002), *A Distribution-Free Theory of Nonparametric Regression*, Springer-Verlag, New York.
- Hall, P. and Huang, L. (2001), "Nonparametric kernel regression subject to monotonicity constraints," *Annals of Statistics*, 29, 624-647.
- Han, S.P. (1988), "A successive projection method," *Mathematical Programming*, 40(1), 1-14.
- Hansen, B.E. (2010), *Econometrics*. Revised: January 10, 2010. Retrieved from <http://www.ssc.wisc.edu/bhansen/econometrics/>.
- Hanson, P.L. and Pledger, G. (1976), "Consistency in concave regression," *The Annals of Statistics*, 4, 1038-1050.
- Härdle, W. and Tsybakov, A. (1997), "Local polynomial estimators of the volatility function in nonparametric autoregression," *Journal of Econometrics*, 81, 223-242.
- He, X. and Shi, P. (1998), "Monotone B-Spline Smoothing," *Journal of the American Statistical Association*, 93, 643-650.
- Herzog, F. and Hill, J. D. (1946), "The Bernstein polynomials for discontinuous functions", *American Journal of Mathematics*, 68, 109-124.
- Hildreth, C. (1954), "Point Estimate of Ordinates of Concave Functions," *Journal of the American Statistical Association*, 49, 598-619.
- Holmes, C.C. and Heard, N.A. (2003), "Generalised monotonic regression using random change points", *Statistics in Medicine*, 22, 623-638.
- Joy, K.I. (2000), "Bernstein Polynomials" *On-Line Geometric Modeling Notes*. Retrieved from <http://idav.ucdavis.edu/education/CAGDNotes/Bernstein-Polynomials.pdf>.
- Judge, G.G. and Takayama, T. (1966), "Inequality Restrictions in Regression Analysis," *Journal of the American Statistical Association*, 61, 166-181.
- Lavine, M. and Mockus, A. (1995), "A nonparametric Bayes method for isotonic regression," *Journal of Statistical Planning and Inference*, 46, 235-248.
- Leitenstorfer, F. and Tutz, G. (2007), "Generalized Monotonic Regression based on B-splines with an Application to Air Pollution data," *Biostatistics*, 8(3), 654-673.
- Liew, C.K. (1976), "Inequality Constrained Least-Squares Estimation," *Journal of the American Statistical Association*, 71, 746-751.
- Lorentz, Z.Z. (1986), *Bernstein Polynomials*, 2nd edition, Chelsea Publishing Company, New York.

- Mammen, E. (1991), "Nonparametric regression under qualitative smoothness assumptions," *The Annual of Statistics*, 19), 741-795.
- Mammen, E., Marron, J.S., Turlach, B.A. and Wand, M.P. (2001), "A General Projection Framework for Constrained Smoothing," *Statistical Science*, 16(3), 232-248.
- Meyer M.C. (1999), "An Extension of the Mixed Primal-Dual Bases Algorithm to the Case of More Constraints than Dimensions," *Journal of Statistical Planning and Inference*, 81, 13-31.
- Meyer M.C. (2008), "Inference using Shape-Restricted Regression Splines," *Annals of Applied Statistics*, 2(3), 1013-1033.
- Molitor, J. and Sun, D. (2002), "Bayesian Analysis under Ordered Functions of Parameters," *Environmental and Ecological Statistics*, 9(2), 179-193.
- Mukerjee, H. (1988), "Monotone Nonparametric Regression," *The Annals of Statistics*, 16(2), 741-750.
- Neelon, B. and Dunson, D. B. (2004), "Bayesian isotonic regression and trend analysis," *Biometrics*, 60, 398-406.
- Opsomer, J., Ruppert, D., Wand, M., Holst, U. and Hössjer, O. (1999), "Kriging with nonparametric variance function estimation," *Biometrics*, 55, 704C710.
- Pal, J., Woodroffe, M. and Meyer, M. (2007), "Estimating a Polya Frequency Function," *IMS Lecture Notes Monograph Series*, 54, 239-249.
- Picard, R. and Cook, D. (1984), "Cross-Validation of Regression Models," *Journal of the American Statistical Association*, 79(387), 575-583.
- Ramsay, J.O. (1988), "Monotone Regression Splines in Action," *Statistical Science*, 3(4), 425-461.
- Robert, C.P. and Casella, G. (2004), *Monte Carlo Statistical Methods*, 2nd edition, Springer.
- Ruppert, D., Wand, M., Holst, U. and Hössjer, O. (1997), "Local polynomial variance-function estimation," *Technometrics*, 39, 262C273.
- Sauer, T. (1999), "Multivariate Bernstein polynomials, convexity and related shape properties". In Pena, J. M., editor, *Shape preserving representations in Computer Aided Design*, Nova Science Publishers.
- Silvapulle, M.J. and Sen, P.K. (2006), *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions*, John Wiley and Sons, New York.
- Spiegelhalter, D. J., Thomas, A. and Best, N. G. (1999), *WinBUGS Version 1.2 User Manual*, MRC Biostatistics Unit.
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde A. (2002), "Bayesian Measures of Complexity and Fit", *Journal of the Royal Statistical Society, Series B*, 64, 583-639.

- Stadtmüller, U. (1986), "Asymptotic Properties of Nonparametric Curve Estimates," *Periodica Mathematica Hungarica*, 17(2), 83-108.
- Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society B*, 36(1), 111-147.
- Stone, M. (1977), "Asymptotics for and against Cross-validation," *Biometrika*, 64, 29-35.
- Tenbusch, A. (1994), "Two-dimensional Bernstein Polynomial Density estimators," *Metrika*, 41, 233-253.
- Tenbusch, A. (1997), "Nonparametric Curve Estimation with Bernstein Estimates," *Metrika*, 45, 1-30.
- Terrell, D. (1996), "Incorporating Monotonicity and Concavity Conditions in Flexible Function Forms," *Journal of Applied Econometrics*, 11, 179-194.
- Turlach, B. A. and Weingessel, A. (2010), *quadprog* (Version 1.5-3), R package.
- van Eeden, C. (2006), *Restricted Parameter Space Estimation Problems: Admissibility and Minimaxity Properties*, Springer, New York.
- Villalobos M. and Wahba G. (1987), "Inequality-constrained Multivariate Smoothing Splines with Application to the Estimation of Posterior Probabilities," *Journal of the American Statistical Association*, 82(397), 239-248.
- Wang, L., Brown, L., Cai, T. and Levine, M. (2008), "Effect of mean on variance function estimation in nonparametric regression," *The Annals of Statistics*, 36, 646-664.
- Wang, X. and Li, F. (2008), "Isotonic Smoothing Spline Regression," *Journal of Computational and Graphical Statistics*, 17(1), 21-37.
- Wu, C.F. (1982), "Some algorithms for concave and isotonic regression," *Studies in the Management Sciences*, 19, 105-116.
- Wu, H. and Zhang, J.-T. (2002), "Local Polynomial Mixed-Effects Models for Longitudinal Data," *Journal of the American Statistical Association*, 97, 883-897.
- Yao, Q. and Tong, H. (1994), "Quantifying the influence of initial values on nonlinear prediction," *Journal of the Royal Statistical Society Series B*, 56, 701-725.

APPENDICES

Appendix A

Detailed Proofs of Asymptotic Properties

A.1 Proof of Property 3.3.1

Proof. We only give the proof for the one dimensional case of $m(\cdot)$. To establish the stated property it is sufficient to show that:

(1) $\mathcal{F}_N \subset L_2[0, 1]$, $\forall N \in \mathbb{N}$, which follows from the fact that Bernstein basis polynomials $b_k(x, N) = \binom{N}{k} x^k (1-x)^{N-k}$ belong to $C[0, 1]$, and hence also their linear combinations.

(2) $\mathcal{F}_N \subset \mathcal{F}_{N+1}$, $\forall N \in \mathbb{N}$.

To establish the above fact, we use the iterative property of the Bernstein Polynomial (Lorentz 1986):

$$b_k(x, N-1) = \frac{N-k}{N} b_k(x, N) + \frac{k+1}{N} b_{k+1}(x, N) \text{ for } k = 0, 1, 2, \dots, N-1. \quad (\text{A.1})$$

By using the property in equation (A.1), any function $B_N(x) \in \mathcal{F}_N$ can be written as:

$$\begin{aligned}
B_N(x) &= \sum_{k=0}^N \beta_k \cdot b_k(x, N) \\
&= \sum_{k=0}^N \beta_k \cdot \left\{ \frac{N+1-k}{N+1} \cdot b_k(x, N+1) + \frac{k+1}{N+1} \cdot b_{k+1}(x, N+1) \right\} \\
&= \sum_{k=0}^N \beta_k \frac{N+1-k}{N+1} \cdot b_k(x, N+1) + \sum_{k=0}^N \beta_k \frac{k+1}{N+1} \cdot b_{k+1}(x, N+1) \\
&= \sum_{k=0}^N \beta_k \frac{N+1-k}{N+1} \cdot b_k(x, N+1) + \sum_{k=1}^N \beta_{k-1} \frac{k}{N+1} \cdot b_k(x, N+1) \\
&= \beta_0 \cdot b_0(x, N+1) + \sum_{k=1}^N \left(\beta_k \frac{N+1-k}{N+1} + \beta_{k-1} \frac{k}{N+1} \right) b_k(x, N+1) \\
&\quad + \beta_N \cdot b_{N+1}(x, N+1).
\end{aligned}$$

Define $\tilde{\beta}_k$ as new coefficients of $\{b_k(x, N+1)\}_{k=0, \dots, N+1}$. We have:

$$B_N(x) = \sum_{k=0}^{N+1} \tilde{\beta}_k \cdot b_k(x, N+1) = \tilde{B}_{N+1}(x),$$

where

$$\begin{aligned}
\tilde{\beta}_0 &= \beta_0, \\
\tilde{\beta}_k &= \beta_k \frac{N+1-k}{N+1} + \beta_{k-1} \frac{k}{N+1}, \text{ when } k = 1, \dots, N, \\
\tilde{\beta}_{N+1} &= \beta_N.
\end{aligned} \tag{A.2}$$

(a) **Nonnegativity:** In this case, $B_N(x) \in \mathcal{F}_N = \{B_N(x) \equiv \sum_{k=0}^N \beta_k \cdot b_k(x, N) : \beta_k \geq 0, \forall k\}$. By using the relationships in (A.2), it is obvious that if $\beta_k \geq 0, \forall k$, then $\tilde{\beta}_k \geq 0, \forall k$, and hence $\mathcal{F}_N \subset \mathcal{F}_{N+1} = \{\tilde{B}_{N+1}(x) \equiv \sum_{k=0}^{N+1} \tilde{\beta}_k \cdot b_k(x, N+1) : \tilde{\beta}_k \geq 0, \forall k\}$.

(b) **Monotonicity:** Since $B_N(x) \in \mathcal{F}_N = \{\sum_{k=0}^N \beta_k \cdot b_k(x, N) : \beta_{k-1} \leq \beta_k, 1 \leq k \leq N\}$, it follows that $\beta_0 \leq \beta_1 \leq \dots \leq \beta_N$. Using this, we show $\tilde{\beta}_0 \leq \tilde{\beta}_1 \leq \dots \leq \tilde{\beta}_{N+1}$ as follows:

- First, $\tilde{\beta}_0 = \beta_0 \leq \tilde{\beta}_1 = \beta_1 - \frac{1}{N+1}\beta_1 + \frac{1}{N+1}\beta_0$
- Second, $\tilde{\beta}_N = \beta_N - \frac{N}{N+1}(\beta_N - \beta_{N-1}) \leq \beta_N = \tilde{\beta}_{N+1}$

- Finally, for $k = 2, \dots, N-1$, we want to show $\tilde{\beta}_{k-1} \leq \tilde{\beta}_k$. Notice that

$$\begin{aligned}
& \tilde{\beta}_{k-1} \leq \tilde{\beta}_k \\
\Leftrightarrow & \beta_{k-1} - \frac{k-1}{N+1}\beta_{k-1} + \frac{k-1}{N+1}\beta_{k-2} \leq \beta_k - \frac{k}{N+1}\beta_k + \frac{k}{N+1}\beta_{k-1} \\
\Leftrightarrow & (N-k+2)\beta_{k-1} + (k-1)\beta_{k-2} \leq (N+1-k)\beta_k + k \cdot \beta_{k-1} \\
\Leftrightarrow & (N-k+1)\beta_{k-1} + (k-1)\beta_{k-2} \leq (N+1-k)\beta_k + (k-1)\beta_{k-1} \quad (\text{A.3})
\end{aligned}$$

Inequality (A.3) is obviously satisfied, because $\beta_{k-2} \leq \beta_{k-1} \leq \beta_k$. Hence, $\tilde{\beta}_{k-1} \leq \tilde{\beta}_k$.

We have shown above that $B_N(x) = \sum_{k=0}^{N+1} \tilde{\beta}_k \cdot b_k(x, N+1) = \tilde{B}_{N+1}(x)$ and $\tilde{\beta}_0 \leq \tilde{\beta}_1 \leq \dots \leq \tilde{\beta}_{N+1}$, therefore we conclude $B_N(x) \in \mathcal{F}_{N+1}$. Since $B_N(\cdot)$ is any arbitrary function in $\mathcal{F}_N$, we conclude $\mathcal{F}_N \subset \mathcal{F}_{N+1} = \{B_{N+1}(x) : \beta_{k-1} \leq \beta_k, 1 \leq k \leq N+1\}$.

(c) **Convexity:** Since $B_N(x) \in \mathcal{F}_N = \{\sum_{k=0}^N \beta_k \cdot b_k(x, N) : \beta_{k+1} - \beta_k \leq \beta_{k+2} - \beta_{k+1}, 0 \leq k \leq N-2\}$, it follows $\beta_{k+2} - 2\beta_{k+1} + \beta_k \geq 0, k = 0, \dots, N-2$. Using this, we will show $\tilde{\beta}_{k+2} - 2\tilde{\beta}_{k+1} + \tilde{\beta}_k \geq 0$, for $k = 0, \dots, N-1$ as follows:

- First, $\tilde{\beta}_2 - 2\tilde{\beta}_1 + \tilde{\beta}_0 = \frac{N-1}{N+1} \cdot \beta_2 - \frac{N-1}{N+1} \cdot 2\beta_1 + \frac{N-1}{N+1} \cdot \beta_0 = \frac{N-1}{N+1} \cdot (\beta_2 - 2\beta_1 + \beta_0) \geq 0$.
- Second, $\tilde{\beta}_{N+1} - 2\tilde{\beta}_N + \tilde{\beta}_{N-1} = \beta_N - 2\beta_N + 2 \cdot \frac{N}{N+1} \cdot (\beta_N - \beta_{N-1}) + \frac{2}{N+1} \cdot \beta_{N-1} + \frac{N-1}{N+1} \tilde{\beta}_{N-2}$
 $= \beta_N \cdot \frac{N-1}{N+1} - \beta_{N-1} \cdot \frac{2N-2}{N+1} + \beta_{N-2} \cdot \frac{N-1}{N+1} = \frac{N-1}{N+1} (\beta_N - 2\beta_{N-1} + \beta_{N-2}) \geq 0$

- Finally, for $k = 1, \dots, N - 2$, we want to show $\tilde{\beta}_{k+2} - 2\tilde{\beta}_{k+1} + \tilde{\beta}_k \geq 0$. Notice that

$$\begin{aligned}
& \tilde{\beta}_{k+2} - 2\tilde{\beta}_{k+1} + \tilde{\beta}_k \\
&= \beta_{k+2} \cdot \frac{N+1-(k+2)}{N+1} - \beta_{k+1} \cdot \frac{k+2}{N+1} - \beta_{k+1} \cdot \frac{2N}{N+1} - \beta_k \cdot \frac{2(k+1)}{N+1} \\
&\quad + \beta_k \cdot \frac{N+1-k}{N+1} + \beta_{k-1} \frac{k}{N+1} \\
&= \beta_{k+2} \frac{N-(k+1)}{N+1} - \beta_{k+1} \frac{2N-(k+2)}{N+1} + \beta_k \frac{N+1-2(k+1)}{N+1} + \beta_{k-1} \frac{k}{N+1} \\
&= \beta_{k+2} \frac{N-(k+1)}{N+1} - \beta_{k+1} \frac{2N-2(k+1)}{N+1} + \beta_k \frac{N-(k+1)}{N+1} \\
&\quad + \beta_{k+1} \frac{k}{N+1} - \beta_k \frac{2k}{N+1} + \beta_{k-1} \frac{k}{N+1} \\
&= \frac{N-(k+1)}{N+1} (\beta_{k+2} - 2\beta_{k+1} + \beta_k) + \frac{k}{N+1} (\beta_{k+1} - 2\beta_k + \beta_{k-1}) \tag{A.4}
\end{aligned}$$

As both terms in (A.4) are greater than 0, one gets $\tilde{\beta}_{k+2} - 2\tilde{\beta}_{k+1} + \tilde{\beta}_k \geq 0$, and hence $\mathcal{F}_N \subset \mathcal{F}_{N+1} = \{B_{N+1}(x) : \beta_{k+2} - 2\beta_{k+1} + \beta_k \geq 0, \forall k\}$.

□

A.2 Proof of Property 3.3.2

Proof. Again, we only give the proof for the one dimensional case. Assume that $f(x) \in C[0, 1] \cap \mathcal{F}$. We take $\hat{\beta}_k = f(\frac{k}{N})$, for $k = 0, 1, \dots, N$. Note that,

1. If $\mathcal{F} = \{f \in C[0, 1] : f(x) \geq 0, \forall x \in [0, 1]\}$, we have $\hat{\beta}_k = f(\frac{k}{N}) \geq 0, \forall k$.
2. If $\mathcal{F} = \{f \in C[0, 1] : f(x_1) \leq f(x_2), \forall 0 \leq x_1 \leq x_2 \leq 1\}$, then $f(x)$ is nondecreasing for $x \in [0, 1]$, i.e., $f(\frac{0}{N}) \leq f(\frac{1}{N}) \leq \dots \leq f(\frac{N}{N})$. Since $\forall k, \hat{\beta}_k = f(\frac{k}{N})$, we have $\hat{\beta}_0 \leq \hat{\beta}_1 \leq \dots \leq \hat{\beta}_N$.
3. If $\mathcal{F} = \{f \in L_2[0, 1] : 2f(\frac{x_1+x_2}{2}) \leq f(x_1) + f(x_2), \forall x_1, x_2 \in [0, 1]\}$, then $f(x)$ is convex for $x \in [0, 1]$, i.e., $2f(\frac{k+1}{N}) = 2f(\frac{k/N + (k+2)/N}{2}) \leq f(\frac{k}{N}) + f(\frac{k+2}{N})$ for all k . Taking $\hat{\beta}_k = f(\frac{k}{N})$, it is easy to see that $\hat{\beta}_{k+1} - \hat{\beta}_k \leq \hat{\beta}_{k+2} - \hat{\beta}_{k+1}$ for $k = 0, 1, \dots, N - 2$.

Now define $\hat{B}_N(x) = \sum_{k=0}^N \hat{\beta}_k \cdot b_k(x, N) = \sum_{k=0}^N f(\frac{k}{N}) \cdot b_k(x, N) \in \mathcal{F}_N \subset \bigcup_{j=1}^{\infty} \mathcal{F}_j$. By Stone-Weierstrass approximation theorem, $\hat{B}_N(x)$ converges uniformly to $f(x)$ (Lorentz 1986), and this completes the proof. $\square$

A.3 Proof of Lemma 3.3.1

Proof. The proof builds on the proof for Lemma 10.1 in Györfi et al. (2002).

$\forall f : \mathcal{R}^d \rightarrow \mathcal{R}$, it has been shown that

$$\begin{aligned} \mathbf{E}\{(f(\mathbf{X}) - Y)^2\} &= \mathbf{E}\{(f(\mathbf{X}) - m(\mathbf{X}))^2\} + \mathbf{E}\{(m(\mathbf{X}) - Y)^2\} \\ &= \int_{\mathcal{R}^d} (f(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) + \mathbf{E}\{(m(\mathbf{X}) - Y)^2\}. \end{aligned} \quad (\text{A.5})$$

where μ denotes the distribution of X . Notice that:

$$\begin{aligned} \mathbf{E}\{(m_N(\mathbf{X}) - Y)^2 | D_n\} &= \mathbf{E}\{(m_N(\mathbf{X}) - m(\mathbf{X}))^2\} + \mathbf{E}\{(m(\mathbf{X}) - Y)^2\} \\ &= \int_{\mathcal{R}^d} (m_N(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) + \mathbf{E}\{(m(\mathbf{X}) - Y)^2\}. \end{aligned} \quad (\text{A.6})$$

Therefore, by equation (A.6), we have:

$$\begin{aligned} \int_{\mathcal{R}^d} (m_N(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) &= \mathbf{E}\{(m_N(\mathbf{X}) - Y)^2 | D_n\} - \mathbf{E}\{(m(\mathbf{X}) - Y)^2\} \\ &= \left(\mathbf{E}\{(m_N(\mathbf{X}) - Y)^2 | D_n\} - \inf_{B_N \in \mathcal{F}_n} \mathbf{E}\{(B_N(\mathbf{X}) - Y)^2\} \right) \\ &\quad + \left(\inf_{B_N \in \mathcal{F}_n} \mathbf{E}\{(B_N(\mathbf{X}) - Y)^2\} - \mathbf{E}\{(m(\mathbf{X}) - Y)^2\} \right). \end{aligned} \quad (\text{A.7})$$

Following equation (A.5), we can rewrite the second term on the right hand side of equation (A.7) as follows:

$$\inf_{B_N \in \mathcal{F}_N} \mathbf{E}\{(B_N(\mathbf{X}) - Y)^2\} - \mathbf{E}\{(m(\mathbf{X}) - Y)^2\} = \inf_{B_N \in \mathcal{F}_N} \int (B_N(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}).$$

We derive an upper bound for the first term of equation (A.7) as follows. Note that $\frac{1}{n} \sum_{i=1}^n (m_N(X_i) - Y_i)^2 = \min_{B_N \in \mathcal{F}_N} \frac{1}{n} \sum_{i=1}^n (B_N(X_i) - Y_i)^2$ by the definition of m_N .

$$\begin{aligned}
& \mathbf{E}\{(m_N(\mathbf{X}) - Y)^2 | D_n\} - \inf_{B_N \in \mathcal{F}_N} \mathbf{E}\{(B_N(\mathbf{X}) - Y)^2\} \\
&= \sup_{B_N \in \mathcal{F}_N} \left(\mathbf{E}\{(m_N(\mathbf{X}) - Y)^2 | D_n\} - \frac{1}{n} \sum_{i=1}^n (m_N(\mathbf{X}_i) - Y_i)^2 \right. \\
&\quad \left. + \frac{1}{n} \sum_{i=1}^n (m_N(X_i) - Y_i)^2 - \frac{1}{n} \sum_{i=1}^n (B_N(X_i) - Y_i)^2 \right. \\
&\quad \left. + \frac{1}{n} \sum_{i=1}^n (B_N(X_i) - Y_i)^2 - \mathbf{E}\{(B_N(\mathbf{X}) - Y)^2\} \right) \\
&\leq \sup_{B_N \in \mathcal{F}_N} \left(\mathbf{E}\{(m_N(\mathbf{X}) - Y)^2 | D_n\} - \frac{1}{n} \sum_{i=1}^n (m_N(\mathbf{X}_i) - Y_i)^2 \right. \\
&\quad \left. + \frac{1}{n} \sum_{i=1}^n (B_N(X_i) - Y_i)^2 - \mathbf{E}\{(B_N(\mathbf{X}) - Y)^2\} \right) \\
&\leq 2 \sup_{B_N \in \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(X_i) - Y_i)^2 - \mathbf{E}\{(B_N(\mathbf{X}) - Y)^2\} \right|.
\end{aligned}$$

□

A.4 Proof of Lemma 3.3.2

Proof. The main steps to establish the result are essentially based on the proof of Theorem 10.2 in Györfi et al. (2002). Note that, as $N = N(n)$ is function of n , $N \rightarrow \infty$ as $n \rightarrow \infty$. Since we have

$$\int_{\mathcal{R}^d} (m_N(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) = \mathbf{E}\{(m_N(\mathbf{X}) - Y)^2 | D_n\} - \mathbf{E}\{(m(\mathbf{X}) - Y)^2\},$$

it is sufficient to show

$$\{\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n\}^{\frac{1}{2}} - \{\mathbf{E}(m(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \rightarrow 0 \text{ a.s.}$$

We have

$$\begin{aligned}
& \{\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n\}^{\frac{1}{2}} - \{\mathbf{E}(m(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \\
&= \left(\{\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n\}^{\frac{1}{2}} - \inf_{B_N \in \mathcal{F}_N} \{\mathbf{E}(B_N(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \right) \\
&+ \left(\inf_{B_N \in \mathcal{F}_N} \{\mathbf{E}(B_N(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} - \{\mathbf{E}(m(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \right). \tag{A.8}
\end{aligned}$$

The second term of equation (A.8) goes to zero by the triangle inequality and condition (3.8):

$$\begin{aligned}
& \inf_{B_N \in \mathcal{F}_N} \{\mathbf{E}(B_N(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} - \{\mathbf{E}(m(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \\
&\leq \inf_{B_N \in \mathcal{F}_N} \left| \{\mathbf{E}(B_N(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} - \{\mathbf{E}(m(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \right| \\
&\leq \inf_{B_N \in \mathcal{F}_N} \{\mathbf{E}[(B_N(\mathbf{X}) - Y) - (m(\mathbf{X}) - Y)]^2\}^{\frac{1}{2}} \\
&= \inf_{B_N \in \mathcal{F}_N} \left\{ \int (B_N(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) \right\}^{\frac{1}{2}} \rightarrow 0 \text{ a.s. when } N \rightarrow \infty.
\end{aligned}$$

Next, we show that the first term of equation (A.8) is bound by 0. Let $L > 0$ be arbitrary.

Because $\lim_{N \rightarrow \infty} L_N = \infty$, we assume $L_N > L$ without loss of generality. Then,

$$\begin{aligned}
& \{\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n\}^{\frac{1}{2}} - \inf_{B_N \in \mathcal{F}_N} \{\mathbf{E}(B_N(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \\
&= \sup_{B_N \in \mathcal{F}_N} \left\{ \{\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n\}^{\frac{1}{2}} - \{\mathbf{E}(B_N(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \right\} \\
&\leq \sup_{B_N \in \mathcal{F}_N} \left\{ \{\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n\}^{\frac{1}{2}} - \{\mathbf{E}(m_N(\mathbf{X}) - Y_L)^2 | D_n\}^{\frac{1}{2}} \right. \\
&\quad + \{\mathbf{E}(m_N(\mathbf{X}) - Y_L)^2 | D_n\}^{\frac{1}{2}} - \left\{ \frac{1}{n} \sum_{i=1}^n (m_N(\mathbf{X}_i) - Y_{i,L})^2 \right\}^{\frac{1}{2}} \\
&\quad + \left\{ \frac{1}{n} \sum_{i=1}^n (m_N(\mathbf{X}_i) - Y_{i,L})^2 \right\}^{\frac{1}{2}} - \left\{ \frac{1}{n} \sum_{i=1}^n (\tilde{m}_N(\mathbf{X}_i) - Y_{i,L})^2 \right\}^{\frac{1}{2}} \\
&\quad \left\{ \frac{1}{n} \sum_{i=1}^n (\tilde{m}_N(\mathbf{X}_i) - Y_{i,L})^2 \right\}^{\frac{1}{2}} - \left\{ \frac{1}{n} \sum_{i=1}^n (\tilde{m}_N(\mathbf{X}_i) - Y_i)^2 \right\}^{\frac{1}{2}} \\
&\quad \left\{ \frac{1}{n} \sum_{i=1}^n (\tilde{m}_N(\mathbf{X}_i) - Y_i)^2 \right\}^{\frac{1}{2}} - \left\{ \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_i)^2 \right\}^{\frac{1}{2}} \\
&\quad + \left\{ \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_i)^2 \right\}^{\frac{1}{2}} - \left\{ \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 \right\}^{\frac{1}{2}} \\
&\quad + \left\{ \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 \right\}^{\frac{1}{2}} - \{\mathbf{E}(B_N(\mathbf{X}) - Y_L)^2\}^{\frac{1}{2}} \\
&\quad \left. + \{\mathbf{E}(B_N(\mathbf{X}) - Y_L)^2\}^{\frac{1}{2}} - \{\mathbf{E}(B_N(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \right\}.
\end{aligned}$$

where T_L is the truncation operation $T_L y = y \cdot I(|y| \leq L) + L \cdot \text{sign}(y) \cdot I(|y| > L)$, $\tilde{m}_N = T_L m_N$, and $T_L \mathcal{F}_N = \{T_L f : f \in \mathcal{F}_N\}$ is a class of truncated functions. The second and seventh term on the right hand side above are bounded by

$$\sup_{B_N \in T_L \mathcal{F}_N} \left| \left\{ \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 \right\}^{\frac{1}{2}} - \{\mathbf{E}(B_N(\mathbf{X}) - Y_L)^2\}^{\frac{1}{2}} \right|.$$

The third and fifth term on the right hand side above are bounded by 0, as $\tilde{m}_N$ is truncated version of m_N . Therefore,

$$\begin{aligned}
& \{\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n\}^{\frac{1}{2}} - \inf_{B_N \in \mathcal{F}_N} \{\mathbf{E}(B_N(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \\
& \leq 2 \cdot \{\mathbf{E}(Y - Y_L)^2\}^{\frac{1}{2}} + 2 \cdot \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - Y_{i,L})^2 \right\}^{\frac{1}{2}} \\
& \quad + 2 \cdot \sup_{B_N \in T_L \mathcal{F}_N} \left| \left\{ \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 \right\}^{\frac{1}{2}} - \{\mathbf{E}(B_N(\mathbf{X}) - Y_L)^2\}^{\frac{1}{2}} \right|. \quad (\text{A.9})
\end{aligned}$$

By condition (3.9) and the strong law of large numbers, we get formulae (A.9) $\leq 4\{\mathbf{E}(Y - Y_L)^2\}^{\frac{1}{2}} \rightarrow 0$ *a.s.* when $L \rightarrow \infty$. This completes the proof of part (a).

Next we prove part (b). We start with the following decomposition.

$$\begin{aligned}
& \int_{\mathcal{R}^d} (m_N(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) \\
& = \left(\{\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n\}^{\frac{1}{2}} - \{\mathbf{E}(m(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \right)^2 \\
& \quad + 2\{\mathbf{E}(m(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \left(\{\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n\}^{\frac{1}{2}} - \{\mathbf{E}(m(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \right),
\end{aligned}$$

and thus it is sufficient to show

$$\mathbf{E} \left(\left(\{\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n\}^{\frac{1}{2}} - \{\mathbf{E}(m(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \right)^2 \right) \rightarrow 0 \text{ when } n \rightarrow \infty.$$

Note that

$$\begin{aligned}
& \mathbf{E} \left(\left(\{\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n\}^{\frac{1}{2}} - \{\mathbf{E}(m(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \right)^2 \right) \\
& \leq 2\mathbf{E} \left\{ \left(\{\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n\}^{\frac{1}{2}} - \inf_{B_N \in \mathcal{F}_N} \{\mathbf{E}(B_N(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \right)^2 \right\} \\
& \quad + 2\mathbf{E} \left\{ \left(\inf_{B_N \in \mathcal{F}_N} \{\mathbf{E}(B_N(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} - \{\mathbf{E}(m(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \right)^2 \right\}.
\end{aligned}$$

The second term on the right hand side of above inequality converges to 0 by the triangle

inequality and condition (3.10). That is,

$$\begin{aligned}
& 2\mathbf{E} \left\{ \inf_{B_N \in \mathcal{F}_N} (\mathbf{E}(B_N(\mathbf{X}) - Y)^2)^{\frac{1}{2}} - (\mathbf{E}(m(\mathbf{X}) - Y)^2)^{\frac{1}{2}} \right\}^2 \\
& \leq 2\mathbf{E} \left\{ \inf_{B_N \in \mathcal{F}_N} (\mathbf{E}(B_N(\mathbf{X}) - Y)^2)^{\frac{1}{2}} \right\}^2 \\
& = 2\mathbf{E} \left\{ \inf_{B_N \in \mathcal{F}_N} (\mathbf{E}(B_N(\mathbf{X}) - Y)^2) \right\} \rightarrow 0 \text{ when } n \rightarrow \infty.
\end{aligned}$$

We next show

$$\mathbf{E} \left\{ (\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n)^{\frac{1}{2}} - \inf_{B_N \in \mathcal{F}_N} (\mathbf{E}(B_N(\mathbf{X}) - Y)^2)^{\frac{1}{2}} \right\}^2 \rightarrow 0 \text{ when } N \rightarrow \infty.$$

From the proof of part (a), we have

$$\begin{aligned}
& (\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n)^{\frac{1}{2}} - \inf_{B_N \in \mathcal{F}_N} (\mathbf{E}(B_N(\mathbf{X}) - Y)^2)^{\frac{1}{2}} \\
& \leq 2(\mathbf{E}(Y - Y_L)^2)^{\frac{1}{2}} + 2 \left(\frac{1}{n} \sum_{i=1}^n (Y_i - Y_{i,L})^2 \right)^{\frac{1}{2}} \\
& \quad + 2 \sup_{B_N \in T_L \mathcal{F}_N} \left| \left(\frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 \right)^{\frac{1}{2}} - (\mathbf{E}(B_N(\mathbf{X}) - Y_L)^2)^{\frac{1}{2}} \right|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
0 &\leq \mathbf{E} \left\{ (\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n)^{\frac{1}{2}} - \inf_{B_N \in \mathcal{F}_N} (\mathbf{E}(B_N(\mathbf{X}) - Y)^2)^{\frac{1}{2}} \right\}^2 \\
&\leq \mathbf{E} \left\{ 2(\mathbf{E}(Y - Y_L)^2)^{\frac{1}{2}} + 2 \left(\frac{1}{n} \sum_{i=1}^n (Y_i - Y_{i,L})^2 \right)^{\frac{1}{2}} \right. \\
&\quad \left. + 2 \sup_{B_N \in T_L \mathcal{F}_N} \left| \left(\frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 \right)^{\frac{1}{2}} - (\mathbf{E}(B_N(\mathbf{X}) - Y_L)^2)^{\frac{1}{2}} \right| \right\}^2 \\
&\leq \mathbf{E} \left\{ 3 \cdot 2^2 \cdot \mathbf{E}(Y - Y_L)^2 + 3 \cdot 2^2 \cdot \frac{1}{n} \sum_{i=1}^n (Y_i - Y_{i,L})^2 \right. \\
&\quad \left. + 3 \cdot 2^2 \cdot \sup_{B_N \in T_L \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 - \mathbf{E}(B_N(\mathbf{X}) - Y_L)^2 \right| \right\} \\
&= 12\mathbf{E}(Y - Y_L)^2 + 12\mathbf{E} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - Y_{i,L})^2 \right\} \\
&\quad + 12\mathbf{E} \left\{ \sup_{f_n \in T_L \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(X_i) - Y_{i,L})^2 - \mathbf{E}(B_N(\mathbf{X}) - Y_L)^2 \right| \right\} \\
&\rightarrow 24\mathbf{E}(Y - Y_L)^2 \text{ a.s. when } N \rightarrow \infty
\end{aligned}$$

where we use conditions (3.10) and (3.11) as well as the strong law of large numbers. Since L is arbitrary, $24\mathbf{E}(Y - Y_L)^2 \rightarrow 0$ when L grows to infinity. Therefore,

$$\mathbf{E} \left\{ (\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n)^{\frac{1}{2}} - \inf_{B_N \in \mathcal{F}_n} (\mathbf{E}(B_N(\mathbf{X}) - Y)^2)^{\frac{1}{2}} \right\}^2 \rightarrow 0.$$

This completes the proof of part (b). □

A.5 Proof of Theorem 3.3.1

Proof. We first present a brief overview of the concepts of ϵ -covering number and ϵ -packing number (Gyorfi et al. 2002) which will be used in the subsequent parts of our proof. Let $\epsilon > 0$, $\mathcal{G}$ be a class of functions $\mathcal{R}^d \rightarrow R$, $z_1^n = (z_1, \dots, z_n)$ be n fixed points in $\mathcal{R}^d$ and ν_n be the

corresponding empirical measure. Let $\|f\|_{L_p(\nu_n)} = \left\{ \frac{1}{n} \sum_i |f(z_i)^p| \right\}^{1/p}$. The ϵ -covering number of $\mathcal{G}$ w.r.t. $\|\cdot\|_{L_p(\nu_n)}$, which is denoted by $\mathcal{N}_p(\epsilon, \mathcal{G}, z_1^n)$, is the minimal $N \in \mathcal{N}$ such that $\|g - g_j\|_{L_p(\nu_n)} < \epsilon$. Similarly, the ϵ -packing number, which is denoted by $\mathcal{M}_p(\epsilon, \mathcal{G}, z_1^n)$, is the maximal $N \in \mathcal{N}$ such that $\|g - g_j\|_{L_p(\nu_n)} \geq \epsilon$.

The proof builds on the proof of Theorem 10.3 in Györfi et al. (2002). According to Lemma 3.3.2, it is sufficient to show that conditions (3.12) and (3.13) imply conditions (3.8) - (3.11) in Lemma 3.3.2. Notice that the conditions (3.8) and (3.10) follow from Properties 3.3.1 and 3.3.2. Let $\varepsilon > 0$. Since $\mathbf{E}(Y^2) < \infty$, we have regression function $m \in L_2(\mu)$. By Property 3.3.2,

$$\bigcup_{N=1}^{\infty} \mathcal{F}_N = \bigcup_{N=1}^{\infty} \{B_N \equiv \sum_{\mathbf{k} \in \mathbb{M}_d(\mathbf{N})} \beta_{\mathbf{k}} \cdot b_{\mathbf{k}}(\mathbf{x}, \mathbf{N}) : \mathbf{A}_N \beta_N \geq \mathbf{0}, \text{ and } \sum |\beta_{\mathbf{k}}| \leq L_N\}$$

is dense in $L_2(\mu)$, where μ denotes the distribution of $\mathbf{X} \in R^d$. Hence there exist $N^* \in \mathcal{N} = \{1, 2, \dots\}$ satisfying $\beta^* = (\beta_0^*, \dots, \beta_{N^*}^*)^\top$ where $\mathbf{A}_{N^*} \beta^* \geq \mathbf{0}$ such that

$$\int \left[\sum_{\mathbf{k} \in \mathbb{M}_d(\mathbf{N}^*)} \beta_{\mathbf{k}}^* \cdot b_{\mathbf{k}}(\mathbf{x}, \mathbf{N}^*) - m(\mathbf{x}) \right]^2 \mu(d\mathbf{x}) < \varepsilon.$$

Since $\forall \mathbf{k} \in \mathbb{M}_d(\mathbf{N}^*)$, we have $0 \leq b_{\mathbf{k}}(\mathbf{x}, \mathbf{N}^*) \leq 1$. It follows that

$$\sup_{\mathbf{x} \in R^d} \left| \sum_{\mathbf{k} \in \mathbb{M}_d(\mathbf{N}^*)} \beta_{\mathbf{k}}^* \cdot b_{\mathbf{k}}(\mathbf{x}, \mathbf{N}^*) \right| \leq L_N < \infty.$$

Using the fact $R_N \rightarrow \infty$ (as $N \rightarrow \infty$) and $L_N \rightarrow \infty$ (as $N \rightarrow \infty$), we have, for all $N \geq N^*$,

$$\sum_{\mathbf{k} \in \mathbb{M}_d(\mathbf{N}^*)} \beta_{\mathbf{k}}^* \cdot b_{\mathbf{k}}(\mathbf{x}, \mathbf{N}^*) \in \tilde{\mathcal{F}}_N = \{B_N : \|B_N\|_{\infty} \leq L_N, \mathbf{A}_N \beta_N \geq \mathbf{0}\} \subset \mathcal{F}_N,$$

where $\|\cdot\|_\infty$ is the sup-norm. Hence, for $N \geq N^*$,

$$\inf_{B_N \in \mathcal{F}_N} \int (B_N(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) \leq \int \left[\sum_{\mathbf{k} \in \mathbb{M}_d(\mathbf{N}^*)} \beta_{\mathbf{k}}^* \cdot b_{\mathbf{k}}(\mathbf{x}, \mathbf{N}^*) - m(\mathbf{x}) \right]^2 \mu(d\mathbf{x}) < \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this implies conditions (3.8) and (3.10) in Lemma 2.

Let $L > 0$ be arbitrary. Because of $L_N \rightarrow \infty$ we assume $L \leq L_N$ for sufficiently large N .

Set $Z = (\mathbf{X}, Y)$, $Z_1 = (\mathbf{X}_1, Y_1), \dots, Z_n = (\mathbf{X}_n, Y_n)$, and

$$\mathcal{H}_N = \{h : R^d \times R \rightarrow R : \exists B_N \in T_{L_N} \mathcal{F}_N \text{ such that } h(\mathbf{x}, y) = (B_N(\mathbf{x}) - T_L y)^2\},$$

where T_L is the truncation operation $T_L y = y \cdot I(|y| \leq L) + L \cdot \text{sign}(y) \cdot I(|y| > L)$, and $T_L \mathcal{F}_N = \{T_L f : f \in \mathcal{F}_N\}$ is the corresponding class of truncated functions. Then $h \in \mathcal{H}_N$ satisfies

$$0 \leq h(\mathbf{x}, y) \leq 2L_N^2 + 2L^2 \leq 4L_N^2.$$

By Theorem 9.1 and Lemma 9.2 in Györfi et al. (2002), given any $\varepsilon > 0$, we have

$$\begin{aligned} & \mathbf{P} \left\{ \sup_{B_N \in T_L \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 - \mathbf{E}[(B_N(\mathbf{X}) - Y_L)^2] \right| > \varepsilon \right\} \\ &= \mathbf{P} \left\{ \sup_{h \in \mathcal{H}_N} \left| \frac{1}{n} \sum_{i=1}^n h(Z_i) - \mathbf{E}[h(Z)] \right| \right\} \\ &\leq 8\mathbf{E}\mathcal{N}_1\left(\frac{\varepsilon}{8}, \mathcal{H}_N, \mathbf{Z}^{(n)}\right) e^{-\frac{n\varepsilon^2}{128(4L_N^2)^2}} \\ &\leq 8\mathbf{E}\mathcal{M}_1\left(\frac{\varepsilon}{8}, \mathcal{H}_N, \mathbf{Z}^{(n)}\right) e^{-\frac{n\varepsilon^2}{128(4L_N^2)^2}} \end{aligned} \tag{A.10}$$

where $\mathcal{N}_1(\frac{\varepsilon}{8}, \mathcal{H}_N, \mathbf{Z}^{(n)})$ is the $\frac{\varepsilon}{8}$ -covering number of $\mathcal{H}_N$ on the points $\mathbf{Z}^{(n)} = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ with respect to L_1 norm, and $\mathcal{M}_1(\frac{\varepsilon}{8}, \mathcal{H}_N, \mathbf{Z}^{(n)})$ is the $\frac{\varepsilon}{8}$ -packing number of $\mathcal{H}_N$ with respect to L_1 norm (Györfi et al. 2002).

Take any two functions h_1, h_2 , where $h_i(\mathbf{x}, y) = (B_N^{(i)}(\mathbf{x}) - T_L y)^2$ for some $B_N^{(i)} \in \mathcal{F}_N$ for

$i = 1, 2..$ Then it can be shown that,

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n |h_1(Z_1) - h_2(Z_2)| = \frac{1}{n} \sum_{i=1}^n |h_1(\mathbf{x}_1, y_1) - h_2(\mathbf{x}_2, y_2)| \\
&= \frac{1}{n} \sum_{i=1}^n |(B_N^{(1)}(\mathbf{x}_i) - T_L y_i)^2 - (B_N^{(2)}(\mathbf{x}_i) - T_L y_i)^2| \\
&= \frac{1}{n} \sum_{i=1}^n |B_N^{(1)}(\mathbf{x}_i) - B_N^{(2)}(\mathbf{x}_i)| \cdot |B_N^{(1)}(\mathbf{x}_i) - 2T_L y_i + B_N^{(2)}(\mathbf{x}_i)| \\
&\leq 4L_N \frac{1}{n} \sum_{i=1}^n |B_N^{(1)}(\mathbf{x}_i) - B_N^{(2)}(\mathbf{x}_i)|.
\end{aligned} \tag{A.11}$$

Therefore, $\{B_N^{(i)}\}$ could be an $\frac{\epsilon}{8 \times 4L_N}$ -packing of $T_L \mathcal{F}_N$ on $\mathbf{X}^{(n)} = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$. Along with Theorem 9.4 in Györfi et al. (2002), we have :

$$\begin{aligned}
& \mathcal{M}_1\left(\frac{\epsilon}{8}, \mathcal{H}_N, \mathbf{Z}^{(n)}\right) \\
&\leq \mathcal{M}_1\left(\frac{\epsilon}{32L_N}, T_L \mathcal{F}_N, \mathbf{X}^{(n)}\right) \\
&\leq 3 \left(\frac{2e(2L_N)}{\frac{\epsilon}{32L_N}} \log \left(\frac{3e(2L_N)}{\frac{\epsilon}{32L_N}} \right) \right)^{V_{T_L \mathcal{F}_N^+}} \\
&= 3 \left(\frac{128eL_N^2}{\epsilon} \log \left(\frac{192eL_N^2}{\epsilon} \right) \right)^{V_{T_L \mathcal{F}_N^+}}
\end{aligned} \tag{A.12}$$

where $\mathcal{F}_N^+ = \{(\mathbf{x}, t) : B_N(\mathbf{x}) - t \geq 0 \text{ and } B_N \in \mathcal{F}_N\}$; $V_{T_L \mathcal{F}_N^+}$ is defined as the largest positive integer such that there exists a set of n points in $\mathbb{R}^d$ which can be shattered by $T_L \mathcal{F}_N^+$ (Györfi et al. 2002). Note that $T_L \mathcal{F}_N^+$ shattering a set of points implies $\mathcal{F}_N^+$ can also shatter the same set of points. Therefore, $V_{T_L \mathcal{F}_N^+} \leq V_{\mathcal{F}_N^+}$. Next, we need to bound $V_{\mathcal{F}_N^+}$ through Theorem 9.5 in Györfi et al. (2002). Since $\mathcal{F}_N^+ = \{(\mathbf{x}, t) : B_N(\mathbf{x}) - t \geq 0 \text{ and } B_N \in \mathcal{F}_N\} \subseteq \{(\mathbf{x}, t) : B_N(\mathbf{x}) + \alpha t \geq 0 \text{ and } B_N \in \mathcal{F}_N, \alpha \in \mathbb{R}^-\}$, and

$$\begin{aligned}
\{B_N(\mathbf{x}) + \alpha t : B_N \in \mathcal{F}_N, \alpha \in \mathbb{R}^-\} &= \left\{ \sum_{\mathbf{k} \in \mathbb{M}^d(\mathbf{N})} \beta_{\mathbf{k}} \cdot b_{\mathbf{k}}(\mathbf{x}, \mathbf{N}) + \alpha : \mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{0}, \text{ and } \alpha \in \mathbb{R}^- \right\} \\
&= \left\{ \sum_{\mathbf{k} \in \mathbb{M}^d(\mathbf{N})} \beta_k \cdot b_{\mathbf{k}}(\mathbf{x}, \mathbf{N}) + \alpha : \tilde{\mathbf{A}}_N \tilde{\boldsymbol{\beta}}_N \geq \mathbf{0} \right\},
\end{aligned} \tag{A.13}$$

where

$$\tilde{\mathbf{A}}_N \tilde{\boldsymbol{\beta}}_N = \begin{pmatrix} -1 & 0 \\ 0 & \mathbf{A}_N \end{pmatrix} \begin{pmatrix} \alpha \\ \boldsymbol{\beta}_N \end{pmatrix}.$$

Recall that R_N is the rank of the restriction matrix $\mathbf{A}_N$, then the subset of sieve space given by (A.13) is a linear vector space of dimension $R_N + 1$. We obtain from Theorem 9.5 in Györfi et al. (2002) that

$$V_{\mathcal{F}_N^+} \leq R_N + 1. \quad (\text{A.14})$$

From inequalities (A.11) to (A.14), we further bound equation (A.10) by

$$\begin{aligned} & \mathbf{P} \left\{ \sup_{B_N \in T_L \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 - \mathbf{E}[(B_N(\mathbf{X}) - Y_L)^2] \right| > \varepsilon \right\} \\ & \leq 24 \left(\frac{128eL_N^2}{\epsilon} \log \left(\frac{192eL_N^2}{\epsilon} \right) \right)^{R_N+1} e^{-\frac{n\epsilon^2}{2048L_N^4}} \\ & \leq 24 \left(\frac{192eL_N^2}{\epsilon} \right)^{2(R_N+1)} e^{-\frac{n\epsilon^2}{2048L_N^4}}, \end{aligned} \quad (\text{A.15})$$

and

$$\begin{aligned} & \sum_{n=1}^{\infty} \mathbf{P} \left\{ \sup_{B_N \in T_L \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 - \mathbf{E}[(B_N(\mathbf{X}) - Y_L)^2] \right| > \varepsilon \right\} \\ & \leq \sum_{n=1}^{\infty} 24 \exp \left(2(R_N + 1) \log \frac{192eL_N^2}{\epsilon} - \frac{n\epsilon^2}{2048L_N^4} \right), \\ & = \sum_{n=1}^{\infty} 24 \exp \left(-n^\delta \frac{n^{1-\delta}}{L_N^4} \left(\frac{\epsilon^2}{2048} - \frac{2(R_N + 1)L_N^4 \log \frac{192eL_N^2}{\epsilon}}{n} \right) \right) \\ & \leq \infty. \end{aligned} \quad (\text{A.16})$$

Last step of inequality follows from conditions (3.12) and (3.13). By the Borel-Contelli lemma,

we have

$$\mathbf{P} \left\{ \sup_{B_N \in T_L \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 - \mathbf{E}[(B_N(\mathbf{X}) - Y_L)^2] \right| > \varepsilon \right\} \rightarrow 0$$

Since ε is arbitrary, the above is equivalent to condition (3.9) in Lemma 3.3.2.

For any nonnegative random variable T and an arbitrary constant $\epsilon > 0$, $\mathbf{E}[T] = \int_0^\infty \mathbf{P}\{T > t\} dt \leq \epsilon + \int_\epsilon^\infty \mathbf{P}\{T > t\} dt$. Using this fact and inequality (A.15), we obtain:

$$\begin{aligned} & \mathbf{E} \left\{ \sup_{B_N \in T_L \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 - \mathbf{E}[(B_N(\mathbf{X}) - Y_L)^2] \right| \right\} \\ & \leq \epsilon + \int_\epsilon^\infty \mathbf{P} \left\{ \sup_{B_N \in T_L \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 - \mathbf{E}[(B_N(\mathbf{X}) - Y_L)^2] \right| > t \right\} dt \\ & \leq \epsilon + \int_\epsilon^\infty 24 \cdot \left(\frac{192eL_N^2}{t} \right)^{2(R_N+1)} \cdot \exp \left(-\frac{nt^2}{2048L_N^4} \right) dt \\ & \leq \epsilon + 24 \cdot \left(\frac{192eL_N^2}{\epsilon} \right)^{2(R_N+1)} \cdot \left[-\frac{2048L_N^4}{n\epsilon} \cdot \exp \left(-\frac{n\epsilon \cdot t}{2048L_N^4} \right) \right]_{t=\epsilon}^\infty \\ & = \epsilon + 24 \cdot \left(\frac{192eL_N^2}{\epsilon} \right)^{2(R_N+1)} \cdot \frac{2048L_N^4}{n\epsilon} \cdot \exp \left(-\frac{n\epsilon^2}{2048L_N^4} \right) \\ & = \epsilon + 24 \cdot \frac{2048L_N^4}{n\epsilon} \cdot \exp \left(2(R_N+1) \cdot \log \left(\frac{192eL_N^2}{\epsilon} \right) - \frac{n\epsilon^2}{2048L_N^4} \right) \rightarrow \epsilon, \end{aligned}$$

where the second term goes to 0 as $n \rightarrow \infty$ (and thus, $N \rightarrow \infty$) when condition (3.12) holds.

Therefore, the above expression converges to 0 as ϵ goes to 0, and this proves condition (3.11) in Lemma 3.3.2 and hence completes the proof. $\square$

Appendix B

Display of Variance Estimation in Section 5.4

B.1 Estimated Variance of Subject Level Measurement Errors

Figure B.1: The estimated variance $\hat{\sigma}_i^2$ for subject i .

B.2 Estimated Variance of Random Effects

$$\hat{\Sigma} = \begin{pmatrix} 35.65254 & -52.87452 & 82.94142 & -99.36036 & 108.1730 & -96.53045 & 79.23547 & -57.23705 \\ -52.87452 & 78.42475 & -122.80921 & 147.05371 & -159.6683 & 142.23617 & -116.40426 & 84.04162 \\ 82.94142 & -122.80921 & 197.19139 & -237.65903 & 267.9423 & -244.41952 & 208.11039 & -151.29770 \\ -99.36036 & 147.05371 & -237.65903 & 286.90492 & -326.4882 & 299.51255 & -257.34216 & 187.37858 \\ 108.17295 & -159.66826 & 267.94227 & -326.48820 & 390.8275 & -369.19597 & 331.81142 & -243.40170 \\ -96.53045 & 142.23617 & -244.41952 & 299.51255 & -369.1960 & 354.36046 & -325.92325 & 239.96001 \\ 79.23547 & -116.40426 & 208.11039 & -257.34216 & 331.8114 & -325.92325 & 309.51205 & -228.99965 \\ -57.23705 & 84.04162 & -151.29770 & 187.37858 & -243.4017 & 239.96001 & -228.99965 & 169.55590 \end{pmatrix}$$