

Shape Restricted Nonparametric Regression Based on Multivariate Bernstein Polynomials

Jiangdian Wang and Sujit K. Ghosh
Department of Statistics
North Carolina State University

Last revised on: September 18, 2011

NC State University Department of Statistics Technical Report# 2640

Abstract

There has been increasing interest in estimating a multivariate regression function subject to shape restrictions, such as nonnegativity, isotonicity, convexity and concavity. 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 article considers a suitable class of multivariate Bernstein polynomials and proposes a sieved estimator obtained from a nested sequence of shape-restricted multivariate Bernstein polynomials. The proposed nonparametric estimator is 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 estimator compared to an existing estimator in terms of various goodness of fit metrics.

Key Words: Bernstein Polynomial, Multivariate Predictors, Nonparametric Methods, Shape Restricted Regression.

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 1 to illustrate regression functions with certain shape restrictions. The first one, depicted in Figure 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 1(b) shows another example of a regression function which is restricted to follow a nondecreasing (and concave) trend in both dimensions.

[insert Figure 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; Ait-Sahalia and Duarte 2003; Dette, Neumeyer and Pilz 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 (2007) 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 article, 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 a nested sequence of multivariate Bernstein polynomials of all degrees. Univariate Bernstein polynomial estimators have played important roles in nonparametric 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 restric-

tion (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 paper is organized as follows. Section 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 and the technical details of the proofs are deferred to the Appendix. Section 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 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 6, we conclude with a brief discussion of our findings and present ideas for further extensions.

2 Modelling Shape Restrictions Using 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 $D_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ 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.,

$$m(\cdot) = \arg \min_{f \in \mathcal{F}} \mathbf{E}\{(f(\mathbf{X}) - Y)^2\}. \quad (2.1)$$

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 D_n .

In this paper, 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 3.1 in Section 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 2.3 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:

$$\mathcal{F}_N = \{B_N(\mathbf{x}) \equiv \sum_{\mathbf{k} \in \mathbb{M}_d(\mathbf{N})} \beta_{\mathbf{k}} \cdot b_{\mathbf{k}}(\mathbf{x}, \mathbf{N}) : \mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{0}, \text{ and } \sum_{\mathbf{k}} |\beta_{\mathbf{k}}| \leq L_N\}, \quad (2.2)$$

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 3.1) 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 (2005), 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 (2.2) can therefore be written as $B_N(\mathbf{x}) = \mathbf{b}_N^T(\mathbf{x}) \boldsymbol{\beta}_N$. We will give examples of one-dimensional and two-dimensional cases to illustrate how to construct \mathcal{F}_N and \mathbf{A}_N in Sections 2.1 and 2.2. 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 $D_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$, the estimator m_N is selected from the Bernstein polynomial sieve \mathcal{F}_N defined in (2.2) that minimizes the empirical L_2 risk. That is:

$$m_N(\cdot) = \arg \min_{\boldsymbol{\beta}_N : \mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{0}, \sum |\beta_{\mathbf{k}}| \leq L_N} \frac{1}{n} \sum_{i=1}^n (\mathbf{b}_N^T(\mathbf{X}_i) \boldsymbol{\beta}_N - Y_i)^2. \quad (2.3)$$

Note that the existence of m_N follows by the compactness of \mathcal{F}_N and requires only a finite-dimensional optimization (see Section 2.3 for computational details).

2.1 Constraint Bernstein Polynomial Sieve in One-Dimensional Case

Let $m(x)$ be the regression function subject to some shape restrictions, i.e., $m(\cdot) \in \mathcal{F}$. Then the N^{th} sieve as defined in (2.2) reduces to:

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

We provide a few illustrative examples of regression functions $m(x)$, and construct corresponding restriction matrices \mathbf{A}_N and their ranks R_N that enforce the following shapes.

(i) *Nonnegativity*

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:

$$\mathbf{A}_N \boldsymbol{\beta}_N \equiv \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_N \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

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, $B_N(x) \equiv \sum_{k=0}^N \beta_k \cdot b_k(x, N)$ are also nonnegative. This implies $\mathcal{F}_N \subset \mathcal{F}$.

(ii) *Monotone Increasing*

Without loss of any generality, the monotonicity in this article will refer to only the increasing monotonicity. Decreasing monotonicity can be simply obtained by reversing the inequalities. Let $m(x)$ be a real-valued function in the space $\mathcal{F} = \{f \in C[0, 1] : f(x_1) \leq f(x_2), \forall 0 \leq x_1 \leq x_2 \leq 1\}$. Then the shape restriction for increasing monotonicity is obtained as follows:

$$\mathbf{A}_N \boldsymbol{\beta}_N \equiv \begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots \\ & & \ddots & & \\ 0 & \dots & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_N \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \quad (2.4)$$

In this case, \mathbf{A}_N is an $N \times (N+1)$ matrix with rank $R_N = N$. Obviously, the constraint shown in (2.4) is equivalent to $\beta_0 \leq \beta_1 \leq \dots \leq \beta_N$. In addition, we note that the derivatives $B'_N(x) = N \sum_{k=0}^{N-1} (\beta_{k+1} - \beta_k) b_k(x, N-1) \geq 0$. As a consequence, for any $B_N(\cdot) \in \mathcal{F}_N$, we have $B_N(\cdot) \in \mathcal{F}$ and hence $\mathcal{F}_N \subset \mathcal{F}$.

(iii) *Convexity*

Let $m(x)$ be a real-valued function in the function space $\mathcal{F} = \{f \in C[0, 1] : 2f(\frac{x_1+x_2}{2}) \leq$

$f(x_1) + f(x_2), \forall x_1, x_2 \in [0, 1]$. We define the restriction on the coefficients as follows:

$$\mathbf{A}_N \boldsymbol{\beta}_N \equiv \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ & & \ddots & & & \\ 0 & \dots & 0 & 1 & -2 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_N \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

In this case \mathbf{A}_N is an $(N - 1) \times (N + 1)$ matrix with rank $R_N = N - 1$. Note that the second derivatives of $B_N(\mathbf{x})$ can be written as $B_N''(\mathbf{x}) = N(N - 1) \sum_{k=0}^{N-2} (\beta_{k+2} - 2\beta_{k+1} + \beta_k) b_k(x, N - 2)$, the restriction ensures $B_N''(\cdot) \geq 0$ for all N , and therefore $\mathcal{F}_N \subset \mathcal{F}$. Hence many other similar shape restrictions studied by Chang et al. (2007) can all be represented as $\mathbf{A}_N \boldsymbol{\beta}_N \geq 0$ for a suitable choice of the restriction matrix \mathbf{A}_N with full row rank. See also the set of shape restrictions discussed in (Curtis and Ghosh 2011) for further generalizations to more than one predictor.

2.2 Constraint Bernstein Polynomial Sieve in Two-Dimensional Case

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 (2.2) 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:

$$\mathbf{A}_N^{(1)} = \begin{pmatrix} -1 & 0 & \dots & 0 & 1 \\ & -1 & 0 & \dots & 0 & 1 \\ & & & \ddots & & \\ & & & & -1 & 0 & \dots & 0 & 1 \end{pmatrix}_{N(N+1) \times (N+1)^2}, \quad (2.5)$$

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

$$\mathbf{A}_N^{(2)} = \begin{pmatrix} \mathbf{B} & & & \\ & \mathbf{B} & & \\ & & \ddots & \\ & & & \mathbf{B} \end{pmatrix}_{N(N+1) \times (N+1)^2} \quad \text{with } \mathbf{B} = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots \\ & & & \ddots & \\ 0 & \dots & 0 & -1 & 1 \end{pmatrix}_{N \times (N+1)} \quad (2.6)$$

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 (2.5). 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 (2.6). 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 $\mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{0}$ 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 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.

2.3 Computation of the Sieved Estimator

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

$$\begin{aligned}\text{minimize} \quad & \frac{1}{n} \sum_{i=1}^n (\mathbf{b}_N^\top(\mathbf{X}_i) \boldsymbol{\beta}_N - Y_i)^2 \text{ w.r.t. } \boldsymbol{\beta}_N \\ \text{subject to:} \quad & \mathbf{A}_N \boldsymbol{\beta}_N \geq 0, \sum_{\mathbf{k}} |\beta_{\mathbf{k}}| \leq L_N\end{aligned}\tag{2.7}$$

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 (2.3). \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 (2.7) essentially

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

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); Gourieroux, 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 Tibshirani 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 (2.7) with varying N , the $CV(N)$ for our BP estimator takes the following form:

$$CV(N) = \frac{1}{V} \sum_{v=1}^V \sum_{i \in I_{-v}} (Y_i - \hat{B}_N(\mathbf{X}_i))^2 \quad (2.8)$$

where $\hat{B}_N(\mathbf{x}) = \mathbf{b}_N^\top(\mathbf{x})\hat{\boldsymbol{\beta}}_N$ with $\hat{\boldsymbol{\beta}}_N$'s obtained based on the training data set by solving the quadratic programming defined in (2.7).

We computed the cross validation function $CV(N)$ defined in (2.8) 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 $\sum_{i=1}^n \mathbf{b}_N(\mathbf{X}_i)\mathbf{b}_N^\top(\mathbf{X}_i)$ remains empirically invertible. The optimal value \hat{N} is chosen to minimize (2.8), 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 $m(\mathbf{x}) = m_1(x_1) + \dots + m_d(x_d)$. 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 $\mathbf{A}_N^{(j)}\boldsymbol{\beta}_N^{(j)} \geq 0$ as shown in Section 2.1. 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 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 (2.1). Recall that in Sections 2.1 and 2.2, 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 (2.3).

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 3.1. *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 3.2. $\bigcup_{N=1}^{\infty} \mathcal{F}_N$ *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 3.1. *Let $\mathcal{F}_N = \mathcal{F}_N(D_n)$ be a class of functions $f_N : \mathcal{R}^d \rightarrow \mathcal{R}$ depending on the data $D_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ as defined in (2.2). If m_N satisfies equation (2.3), 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 3.2 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 3.2. *Let \mathcal{F}_N be the N^{th} sieve defined in equation (2.2) and $m_N(\cdot)$ be the estimator defined in equation (2.3). 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*

$$\lim_{N \rightarrow \infty} \inf_{B_N \in \mathcal{F}_N} \int (B_N(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) = 0 \text{ a.s.}, \quad (3.1)$$

$$\begin{aligned} & \lim_{N \rightarrow \infty} \sup_{B_N \in T_L \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 - \mathbf{E}(B_N(\mathbf{X}) - Y_L)^2 \right| \\ & = 0 \text{ a.s. for all } L > 0, \quad (3.2) \end{aligned}$$

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

$$\lim_{N \rightarrow \infty} \mathbf{E} \left\{ \inf_{B_N \in \mathcal{F}_N} \int (B_N(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) \right\} = 0, \quad (3.3)$$

$$\lim_{N \rightarrow \infty} \mathbf{E} \left\{ \sup_{B_N \in T_L \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 - \mathbf{E}(B_N(\mathbf{X}) - Y_L)^2 \right| \right\} = 0 \text{ for all } L > 0, \quad (3.4)$$

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 3.1 and Lemma 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 (2.2) and the regression function estimator (2.3) are described in the following theorem.

Theorem 3.1. *Let \mathcal{F}_N be the sieve space defined in equation (2.2) satisfying Properties 3.1 and 3.2, and $m_N(\cdot)$ be the restricted regression function estimator given by equation (2.3). 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

$$R_N \rightarrow \infty, L_N \rightarrow \infty, \text{ and } \frac{R_N L_N^4 \log L_N}{n} \rightarrow 0, \quad (3.5)$$

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 (3.5), and additionally for some $\delta > 0$, L_N satisfies

$$\frac{L_N^4}{N^{1-\delta}} \rightarrow 0, \quad (3.6)$$

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 3.2. In other words, using conditions (3.5) and (3.6), we will show conditions (3.1) - (3.4) in Lemma 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 (3.5) and (3.6) 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 Sections 2.1 and 2.2 all satisfy these choices. Thus, when n approaches infinity, both R_N and L_N will tend to infinity. To satisfy condition (3.5) in Theorem 3.1, we need to show

$$\frac{R_N L_N^4 \log L_N}{n} = C \cdot \frac{n^{kr} n^{4dlk} \log n}{n} = C \cdot \frac{\log n}{n^{1-kr-4dlk}} \rightarrow 0, \quad (3.7)$$

where $C > 0$ is a universal constant. The limit in (3.7) 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 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 (2007).

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

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:

- (a) $m_1(x_1, x_2) = \min(x_1, x_2)$,
- (b) $m_2(x_1, x_2) = \begin{cases} (x_1 - \frac{1}{2})(x_2 - \frac{1}{2}) + x_1x_2 & \text{if } x_1 \geq \frac{1}{2}, x_2 \geq \frac{1}{2} \text{ and} \\ x_1x_2 & \text{otherwise,} \end{cases}$
- (c) $m_3(x_1, x_2) = x_1^{1/3} x_2^{2/3}$.

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 2.3). 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 including 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.

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:

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

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.

4.2.1 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 Figure 2). 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 Figure 2 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 Figure 2 here.]

Figures 3-5 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, 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 4 and 5, 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 Table 1).

[insert Figure 3-5 here.]

The entries in columns 3 and 4 in Table 1 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 Table 1, 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 Table 1 here.]

4.2.2 Local Measure of Performances

In addition to global performance, we also investigated the local estimation 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) is not differentiable at the point $(.5, .5)$ and the test function (b) do not have a continuous derivative at this interior point. Table 2 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 Table 2 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 Table 2 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 6. 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 6 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.

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, Figures 8 and 9). 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 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 2.3. We also used 7-fold cross validation as described in Section 2.3 to choose N , and the optimal order $\hat{N} = 4$ produced the minimum cross validation score (see Figure 7). 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., $RMSPE = \sqrt{\frac{1}{n} \sum_i^n (\hat{y}_i - y_i)^2}$) 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.

We present 3-D scatter plots of the real data as well as the estimates from two approaches in Figure 8. In addition, the 2-D scatter plots relating the estimated response to two predictors respectively, are presented in Figure 9. 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 10, we plot the estimated values against observed values to compare predictive performances. Figure 10 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 11. 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 8 - 11 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

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

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 12, 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 12).

[insert Figure 12 here.]

6 Discussion

In this article, 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 2.3, 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 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 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.

References

- Aït-Sahalia, Y. and Duarte, J. (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. (2007), “Nonparametric Estimation of Regression Functions under Restrictions on Partial Derivatives,” *Tech. Report*, Duke University.
- 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.
- Brunk, H.D. (1955), “Maximum Likelihood Estimates of Monotone Parameters,” *The Annals of Mathematical Statistics*, 26(4), 607-616.
- 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.

- 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>.
- 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.
- Dykstra, R.J. (1983), "An Algorithm for Restricted Least Squares Regression," *Journal of the American Statistical Association*, 78, 837-842.
- Efron, B. and Tibshirani, R. (1997), "Improvements on Cross-Validation: The .632 + Bootstrap Method," *Journal of the American Statistical Association*, 92(438), 548-560.
- 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.
- Gallant, A.R. and Golub, G.H. (1984), "Imposing Curvature Restrictions on Flexible Functional Forms," *Journal of Econometrics*, 38, 295-321.
- Geman, S. and Hwang C. (1982), "Nonparametric Maximum Likelihood Estimation by the Method of Sieves," *The Annals of Statistics*, 19(2), 401-414.
- 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.
- 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.
- Grenander, U. (1981), *Abstract Inference (Probability & Mathematical Statistics)*, John Wiley and Sons, New York.
- 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.
- He X. and Shi P. (1998), "Monotone B-Spline Smoothing," *Journal of the American Statistical Association*, 93, 643-650.
- Hildreth, C. (1954), "Point Estimate of Ordinates of Concave Functions," *Journal of the American Statistical Association*, 49, 598-619.

- Judge, G.G. and Takayama, T. (1966), "Inequality Restrictions in Regression Analysis," *Journal of the American Statistical Association*, 61, 166-181.
- 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., 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.
- 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.
- Silvapulle, M.J. and Sen, P.K. (2006), *Constrained Statistical Inference*, John Wiley and Sons, New York.
- Stadtmuller, 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 Minimality 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, X. and Li, F. (2008), “Isotonic Smoothing Spline Regression,” *Journal of Computational and Graphical Statistics*, 17(1), 21-37.

Appendix

A Proofs

A.1 Proof of Property 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

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

(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

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

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 3.2

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. □

A.3 Proof of Lemma 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

$$\begin{aligned} & \{\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n\}^{\frac{1}{2}} - \{\mathbf{E}(m(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \\ &= \left(\{\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n\}^{\frac{1}{2}} - \inf_{B_N \in \mathcal{F}_N} \{\mathbf{E}(B_N(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \right) \\ &+ \left(\inf_{B_N \in \mathcal{F}_N} \{\mathbf{E}(B_N(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} - \{\mathbf{E}(m(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \right). \end{aligned} \quad (\text{A.5})$$

The second term of equation (A.5) goes to zero by the triangle inequality and condition (3.1):

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

Next, we show that the first term of equation (A.5) 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,

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

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,

$$\begin{aligned}
& \{\mathbf{E}(m_N(\mathbf{X}) - Y)^2 | D_n\}^{\frac{1}{2}} - \inf_{B_N \in \mathcal{F}_N} \{\mathbf{E}(B_N(\mathbf{X}) - Y)^2\}^{\frac{1}{2}} \\
\leq & 2 \cdot \{\mathbf{E}(Y - Y_L)^2\}^{\frac{1}{2}} + 2 \cdot \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - Y_{i,L})^2 \right\}^{\frac{1}{2}} \\
& + 2 \cdot \sup_{B_N \in T_L \mathcal{F}_N} \left| \left\{ \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 \right\}^{\frac{1}{2}} - \{\mathbf{E}(B_N(\mathbf{X}) - Y_L)^2\}^{\frac{1}{2}} \right|. \quad (\text{A.6})
\end{aligned}$$

By condition (3.2) and the strong law of large numbers, we get formulae (A.6) $\leq 4\{\mathbf{E}(Y - Y_L)^2\}^{\frac{1}{2}} \rightarrow 0$ *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

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

Note that

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

The second term on the right hand side of above inequality converges to 0 by the triangle inequality and condition (3.3). 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

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

Therefore,

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

where we use conditions (3.3) and (3.4) 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,

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

This completes the proof of part (b). \square

A.4 Proof of Theorem 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 Gyorfi et al. (2002). According to Lemma 3.2, it is sufficient to show that conditions (3.5) and (3.6) imply conditions (3.1) - (3.4) in Lemma 3.2. Notice that the conditions (3.1) and (3.3) follow from Properties 3.1

and 3.2. Let $\varepsilon > 0$. Since $\mathbf{E}(Y^2) < \infty$, we have regression function $m \in L_2(\mu)$. By Property 3.2 (see Section 3),

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

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 $\boldsymbol{\beta}^* = (\beta_0^*, \dots, \beta_{N^*}^*)^\top$ where $\mathbf{A}_{N^*} \boldsymbol{\beta}^* \geq \mathbf{0}$ 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 $N \geq N^*$,

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

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 (3.1) and (3.3) 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 $Z = (\mathbf{X}, Y)$, $Z_1 = (\mathbf{X}_1, Y_1)$, ..., $Z_n = (\mathbf{X}_n, Y_n)$, 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

$$\begin{aligned} & \mathbf{P} \left\{ \sup_{B_N \in T_L \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 - \mathbf{E}[(B_N(\mathbf{X}) - Y_L)^2] \right| > \varepsilon \right\} \\ &= \mathbf{P} \left\{ \sup_{h \in \mathcal{H}_N} \left| \frac{1}{n} \sum_{i=1}^n h(Z_i) - \mathbf{E}[h(Z)] \right| \right\} \\ &\leq 8\mathbf{E}\mathcal{N}_1\left(\frac{\varepsilon}{8}, \mathcal{H}_N, \mathbf{Z}^{(n)}\right) e^{-\frac{n\varepsilon^2}{128(4L_N^2)^2}} \\ &\leq 8\mathbf{E}\mathcal{M}_1\left(\frac{\varepsilon}{8}, \mathcal{H}_N, \mathbf{Z}^{(n)}\right) e^{-\frac{n\varepsilon^2}{128(4L_N^2)^2}} \end{aligned} \tag{A.7}$$

where $\mathcal{N}_1(\frac{\epsilon}{8}, \mathcal{H}_N, \mathbf{Z}^{(n)})$ is the $\frac{\epsilon}{8}$ -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 $\mathcal{M}_1(\frac{\epsilon}{8}, \mathcal{H}_N, \mathbf{Z}^{(n)})$ is the $\frac{\epsilon}{8}$ -packing number of \mathcal{H}_N with respect to L_1 norm (Gyorfi 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,

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n |h_1(Z_1) - h_2(Z_2)| = \frac{1}{n} \sum_{i=1}^n |h_1(\mathbf{x}_1, y_1) - h_2(\mathbf{x}_2, y_2)| \\
&= \frac{1}{n} \sum_{i=1}^n |(B_N^{(1)}(\mathbf{x}_i) - T_L y_i)^2 - (B_N^{(2)}(\mathbf{x}_i) - T_L y_i)^2| \\
&= \frac{1}{n} \sum_{i=1}^n |B_N^{(1)}(\mathbf{x}_i) - B_N^{(2)}(\mathbf{x}_i)| \cdot |B_N^{(1)}(\mathbf{x}_i) - 2T_L y_i + B_N^{(2)}(\mathbf{x}_i)| \\
&\leq 4L_N \frac{1}{n} \sum_{i=1}^n |B_N^{(1)}(\mathbf{x}_i) - B_N^{(2)}(\mathbf{x}_i)|. \tag{A.8}
\end{aligned}$$

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 Gyorfi et al. (2002), we have :

$$\begin{aligned}
& \mathcal{M}_1\left(\frac{\epsilon}{8}, \mathcal{H}_N, \mathbf{Z}^{(n)}\right) \\
&\leq \mathcal{M}_1\left(\frac{\epsilon}{32L_N}, T_L \mathcal{F}_N, \mathbf{X}^{(n)}\right) \\
&\leq 3 \left(\frac{2e(2L_N)}{\frac{\epsilon}{32L_N}} \log \left(\frac{3e(2L_N)}{\frac{\epsilon}{32L_N}} \right) \right)^{V_{T_L \mathcal{F}_N^+}} \\
&= 3 \left(\frac{128eL_N^2}{\epsilon} \log \left(\frac{192eL_N^2}{\epsilon} \right) \right)^{V_{T_L \mathcal{F}_N^+}} \tag{A.9}
\end{aligned}$$

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 $T_L \mathcal{F}_N^+$ (Gyorfi 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 Gyorfi 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

$$\begin{aligned}
\{B_N(\mathbf{x}) + \alpha t : B_N \in \mathcal{F}_N, \alpha \in \mathbb{R}^-\} &= \left\{ \sum_{\mathbf{k} \in \mathbb{M}^d(\mathbf{N})} \beta_{\mathbf{k}} \cdot b_{\mathbf{k}}(\mathbf{x}, \mathbf{N}) + \alpha : \mathbf{A}_N \boldsymbol{\beta}_N \geq \mathbf{0}, \text{ and } \alpha \in \mathbb{R}^- \right\} \\
&= \left\{ \sum_{\mathbf{k} \in \mathbb{M}^d(\mathbf{N})} \beta_{\mathbf{k}} \cdot b_{\mathbf{k}}(\mathbf{x}, \mathbf{N}) + \alpha : \tilde{\mathbf{A}}_N \tilde{\boldsymbol{\beta}}_N \geq \mathbf{0} \right\}, \tag{A.10}
\end{aligned}$$

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.10) is a linear vector space of dimension $R_N + 1$. We obtain from Theorem 9.5 in Györfi et al. (2002) that

$$V_{\mathcal{F}_N^+} \leq R_N + 1. \quad (\text{A.11})$$

From inequalities (A.8) to (A.11), we further bound equation (A.7) by

$$\begin{aligned} & \mathbf{P} \left\{ \sup_{B_N \in T_L \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 - \mathbf{E}[(B_N(\mathbf{X}) - Y_L)^2] \right| > \varepsilon \right\} \\ & \leq 24 \left(\frac{128eL_N^2}{\varepsilon} \log \left(\frac{192eL_N^2}{\varepsilon} \right) \right)^{R_N+1} e^{-\frac{n\varepsilon^2}{2048L_N^4}} \\ & \leq 24 \left(\frac{192eL_N^2}{\varepsilon} \right)^{2(R_N+1)} e^{-\frac{n\varepsilon^2}{2048L_N^4}}, \end{aligned} \quad (\text{A.12})$$

and

$$\begin{aligned} & \sum_{n=1}^{\infty} \mathbf{P} \left\{ \sup_{B_N \in T_L \mathcal{F}_N} \left| \frac{1}{n} \sum_{i=1}^n (B_N(\mathbf{X}_i) - Y_{i,L})^2 - \mathbf{E}[(B_N(\mathbf{X}) - Y_L)^2] \right| > \varepsilon \right\} \\ & \leq \sum_{n=1}^{\infty} 24 \exp \left(2(R_N + 1) \log \frac{192eL_N^2}{\varepsilon} - \frac{n\varepsilon^2}{2048L_N^4} \right), \\ & = \sum_{n=1}^{\infty} 24 \exp \left(-n^\delta \frac{n^{1-\delta}}{L_N^4} \left(\frac{\varepsilon^2}{2048} - \frac{2(R_N + 1)L_N^4 \log \frac{192eL_N^2}{\varepsilon}}{n} \right) \right) \\ & \leq \infty. \end{aligned} \quad (\text{A.13})$$

Last step of inequality follows from conditions (3.5) and (3.6). By the Borel-Contelli lemma, we have

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

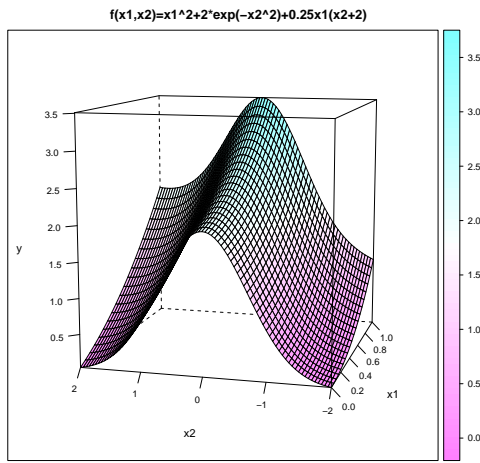
Since ε is arbitrary, the above is equivalent to condition (3.2) in Lemma 3.2.

For any nonnegative random variable T and an arbitrary constant $\varepsilon > 0$, $\mathbf{E}[T] =$

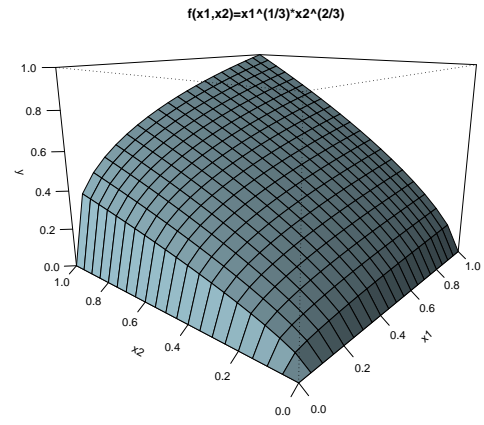
$\int_0^\infty \mathbf{P}\{T > t\}dt \leq \epsilon + \int_\epsilon^\infty \mathbf{P}\{T > t\}dt$. Using this fact and inequality (A.12), we obtain:

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

where the second term goes to 0 as $n \rightarrow \infty$ (and thus, $N \rightarrow \infty$) when condition (3.5) holds. Therefore, the above expression converges to 0 as ϵ goes to 0, and this proves condition (3.4) in Lemma 3.2 and hence completes the proof. \square



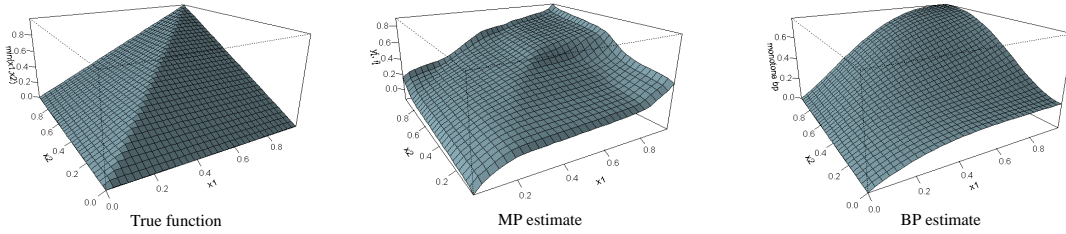
(a) Increasing in x_1 , concave in x_2



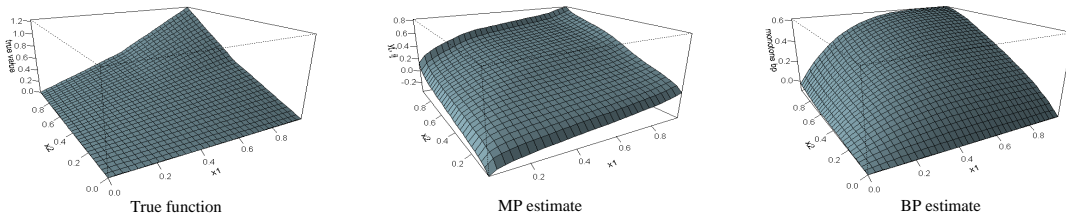
(b) Increasing and concave in x_1 and x_2

Figure 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).

(a) $m_1(x_1, x_2) = \min(x_1, x_2)$



(b) $m_2(x_1, x_2) = \begin{cases} (x_1 - \frac{1}{2})(x_2 - \frac{1}{2}) + x_1x_2 & \text{if } x_1 \geq \frac{1}{2}, x_2 \geq \frac{1}{2} \\ x_1x_2 & \text{O.W.} \end{cases}$



(c) $m_3(x_1, x_2) = x_1^{\frac{1}{3}} x_2^{\frac{2}{3}}$

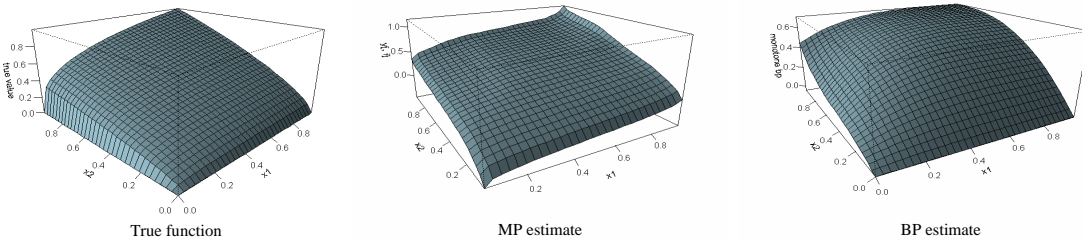


Figure 2: 3-D surface plots of $m_i(x_1, x_2)$ when $\sigma = 0.1$ and $n = 400$. *Note:* The estimated surfaces (by MP and BP methods) are based on a data set randomly selected from 500 simulated data sets. From left to right: surface for true function, estimated surface by the monoProc (MP) method, and estimated surface by the restricted Bernstein Polynomial (BP) method.

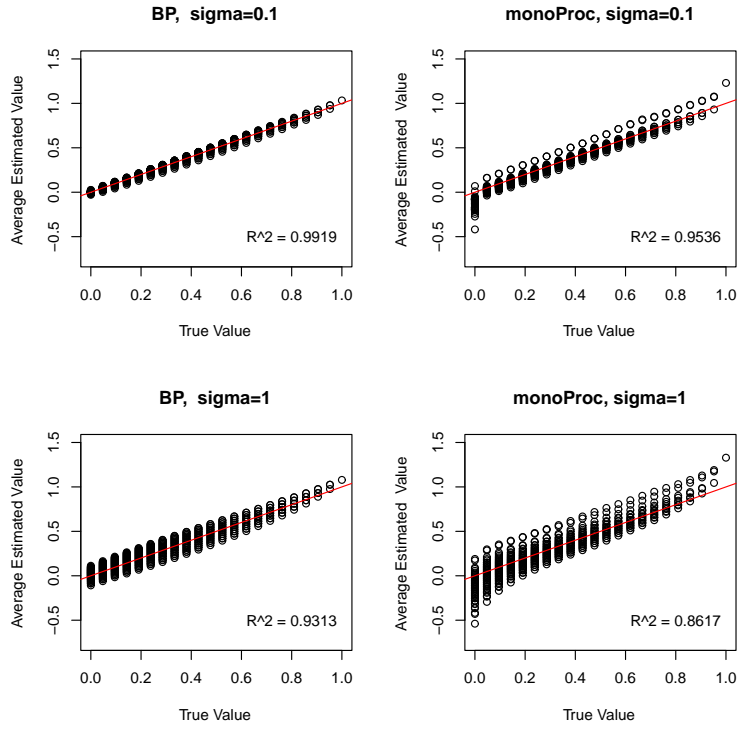


Figure 3: Plot of average predicted values vs. true values for $m_1(x_1, x_2) = \min(x_1, x_2)$. Solid line is 45° reference line.

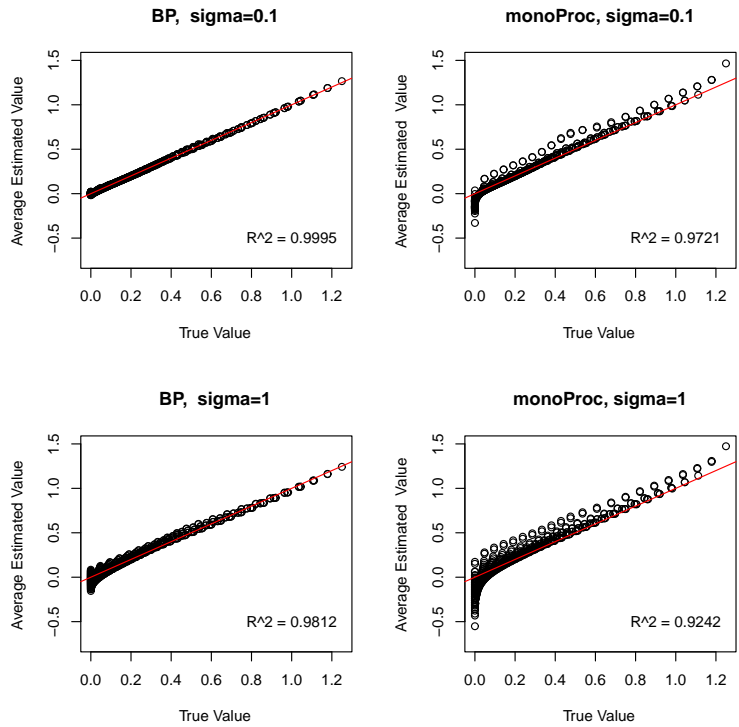


Figure 4: Plot of average predicted values vs. true values for $m_2(x_1, x_2) = x_1x_2I(x_1 \leq 0.5, x_2 \leq 0.5) + \{(x_1 - 0.5)(x_2 - 0.5) + x_1x_2\}I(x_1 > 0.5, x_2 > 0.5)$. Solid line is 45° reference line.

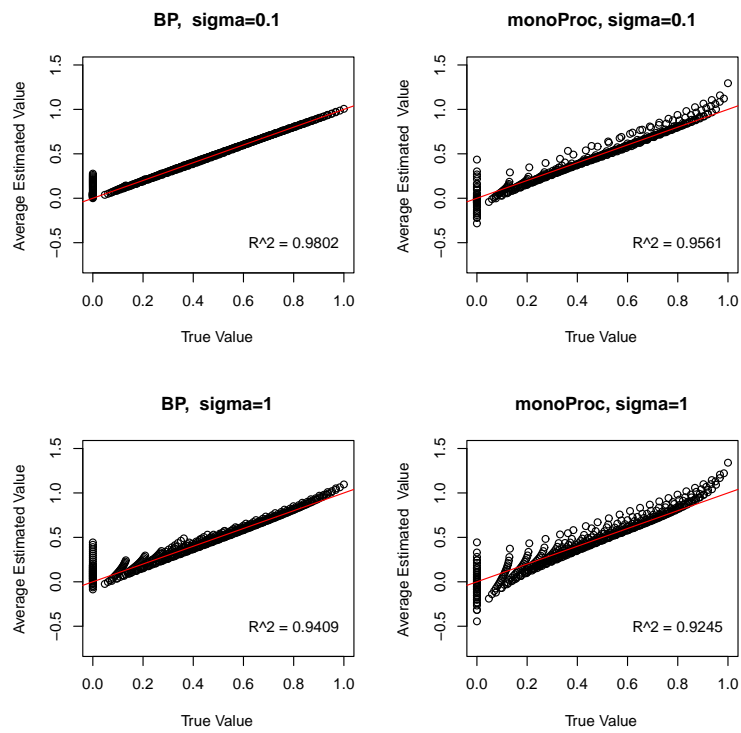


Figure 5: Plot of average predicted values vs. true values for $m_3(x_1, x_2) = x_1^{\frac{1}{3}} x_2^2$. Solid line is 45° reference line.

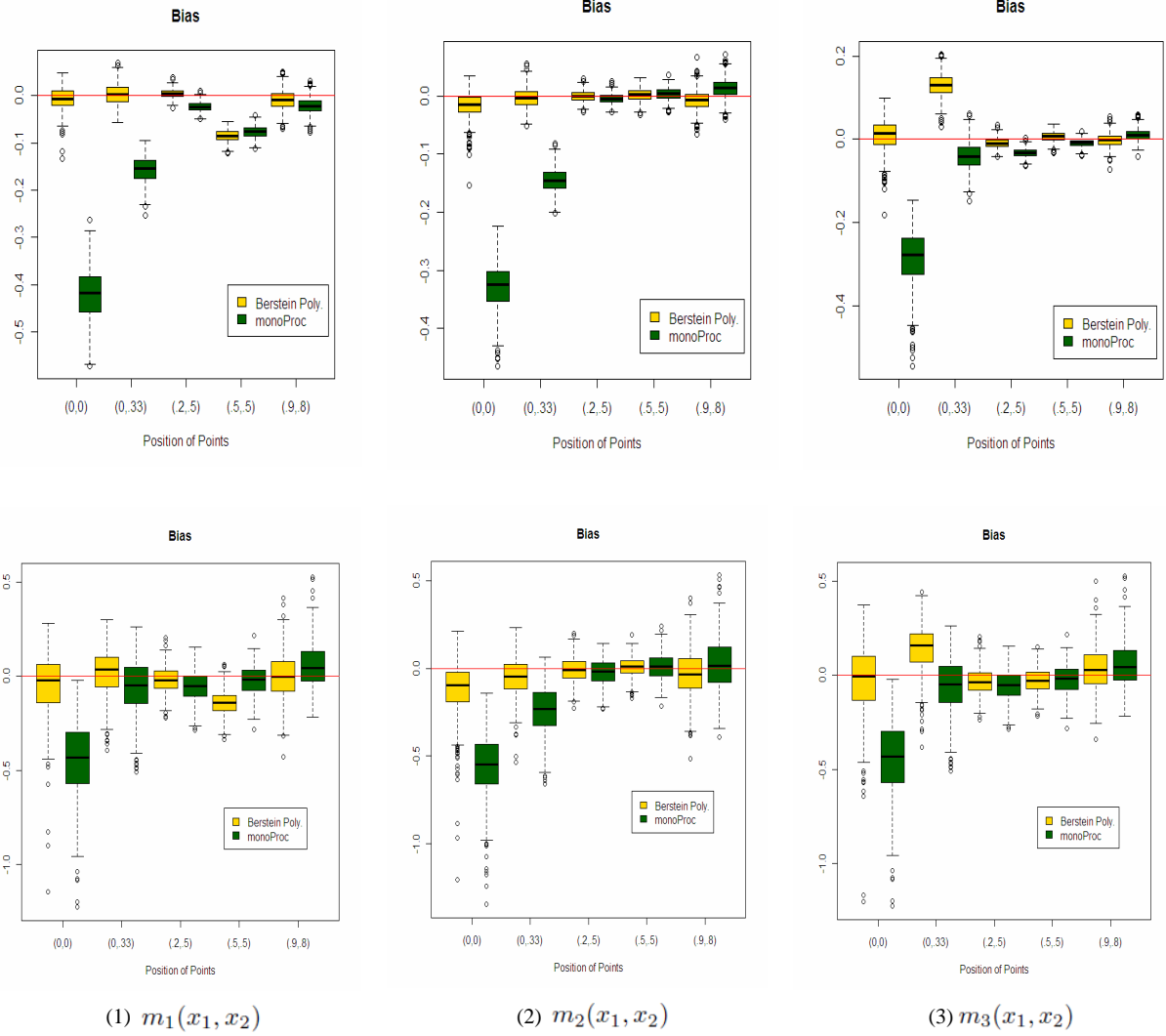


Figure 6: Box plots of bias at five points using the BP and the MP method with sample size $n=400$ and 500 replications. *Note:* Simulation scenarios are described as follows. From top to bottom: $\sigma = 0.1$ and $\sigma = 1.0$; From left to right: $m_1(x_1, x_2) = \min(x_1, x_2)$, $m_2(x_1, x_2) = x_1 x_2 I(x_1 \leq 0.5, x_2 \leq 0.5) + \{(x_1 - 0.5)(x_2 - 0.5) + x_1 x_2\} I(x_1 > 0.5, x_2 > 0.5)$, and $m_3(x_1, x_2) = x_1^{\frac{1}{2}} x_2^{\frac{1}{2}}$.

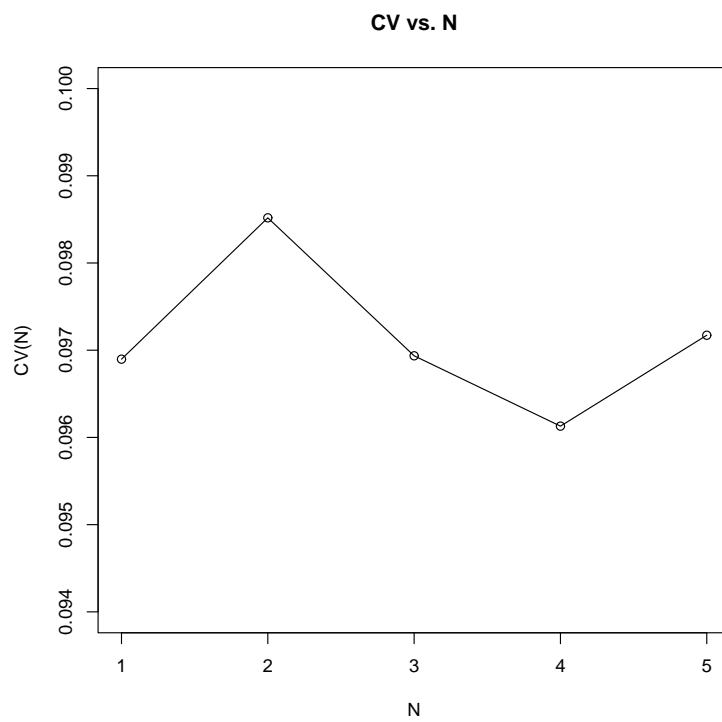
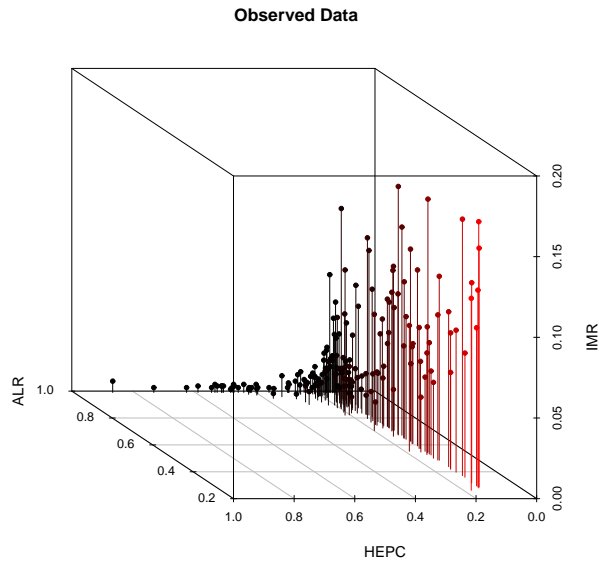
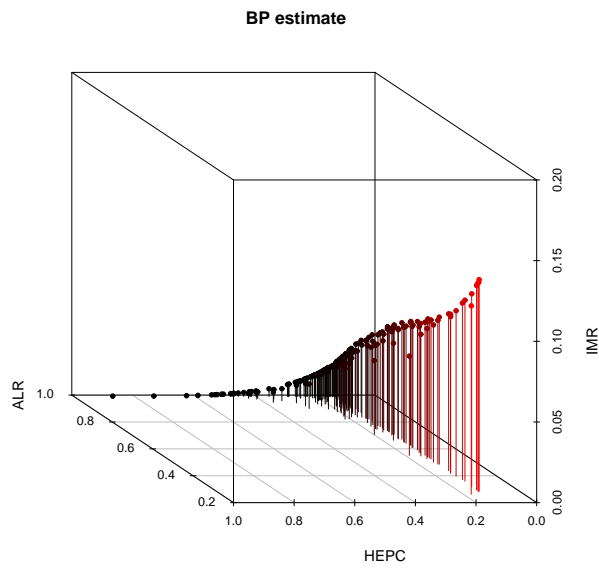


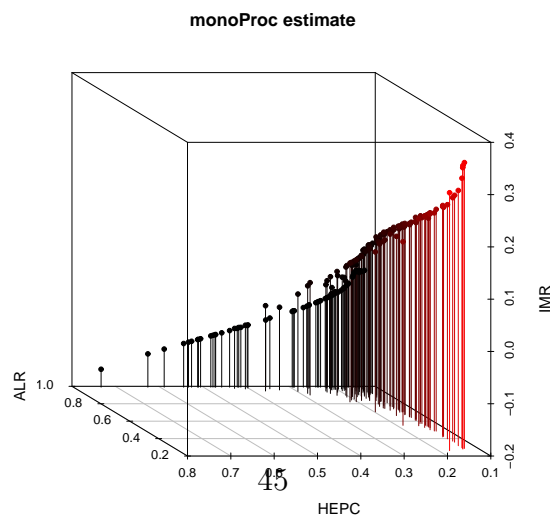
Figure 7: CV score vs. N for the infant mortality data.



(a) Observed data



(b) BP estimate



(c) monoProc estimate

Figure 8: 3-D scatter plots of the infant mortality data. *Predictors*: ALR, HEPC; *Response*: IMR .

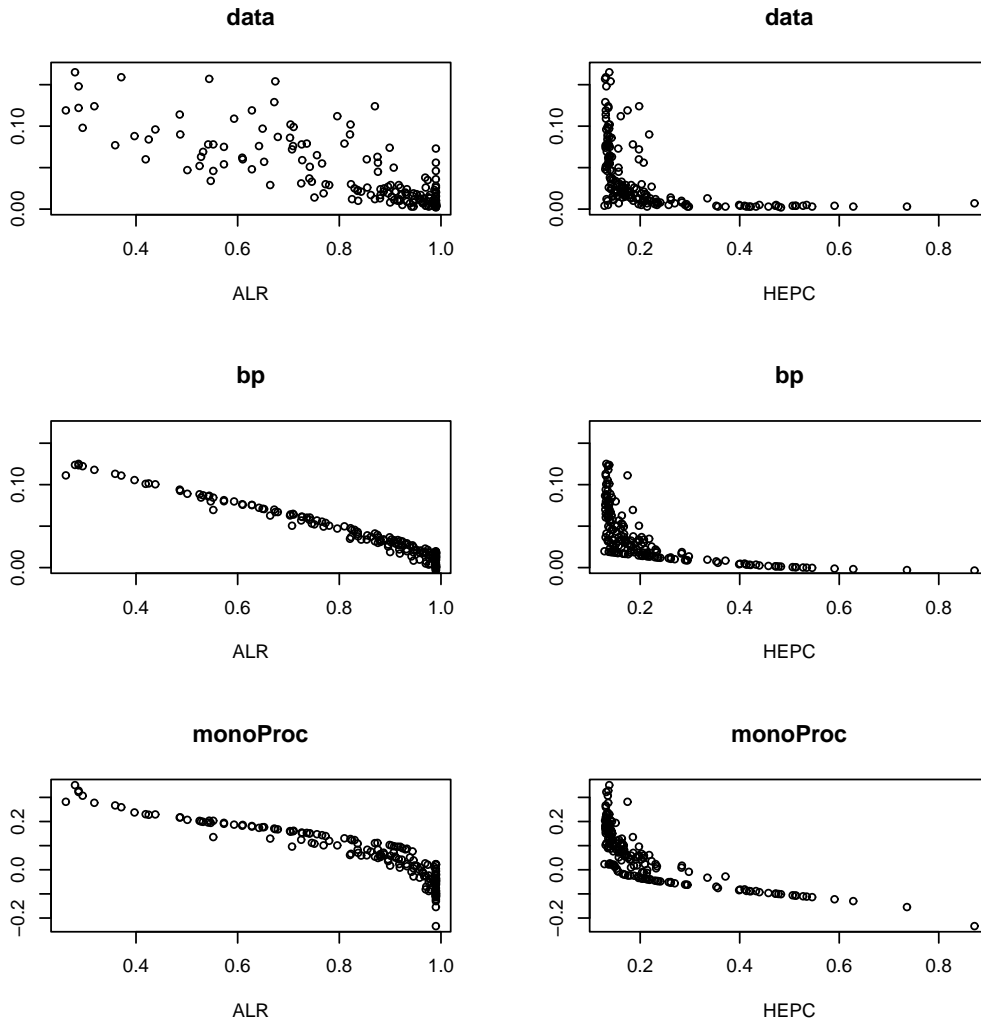


Figure 9: Predicted values of $\hat{m}(x_{1i}, x_{2i})$ vs. predictors x_{1i} and x_{2i} . $i=1, \dots, 176$. Note: $X_1=ALT$, $X_2=HEPC$, $Y=IMR$. Top Panel: observed y_i vs. x_{1i} and x_{2i} ; Middle Panel: predicted values obtained by Bernstein polynomial method vs. predictors; Bottom: predicted values obtained by monoProc vs. predictors.

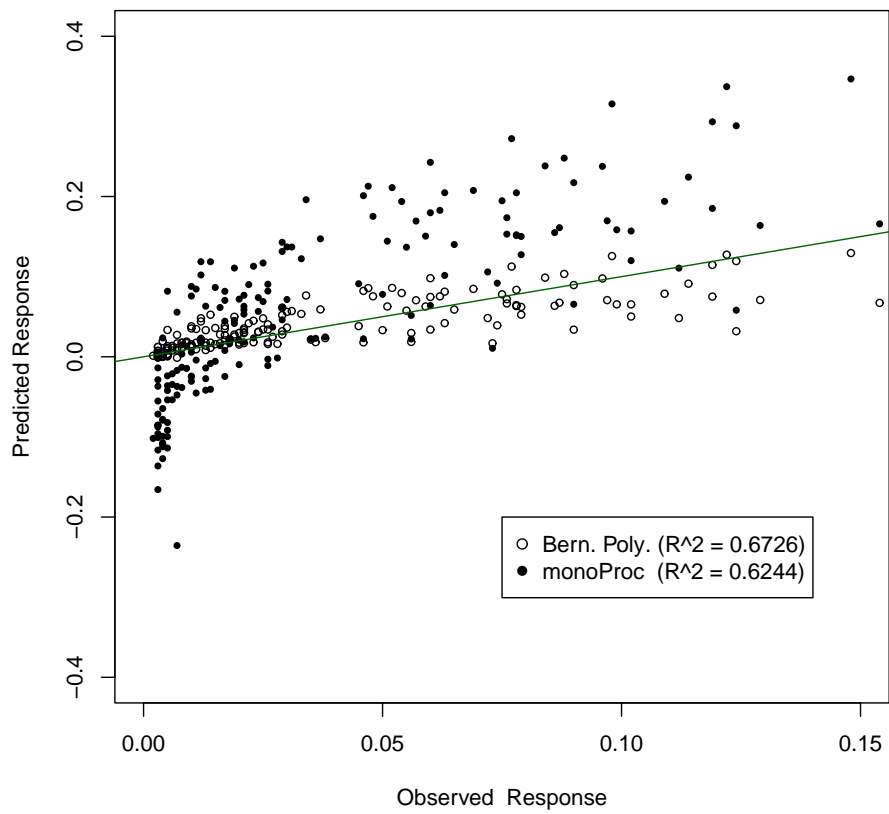


Figure 10: Comparison of observed and predicted values of the response. Solid line represents the 45° reference line, i.e., $y=x$.

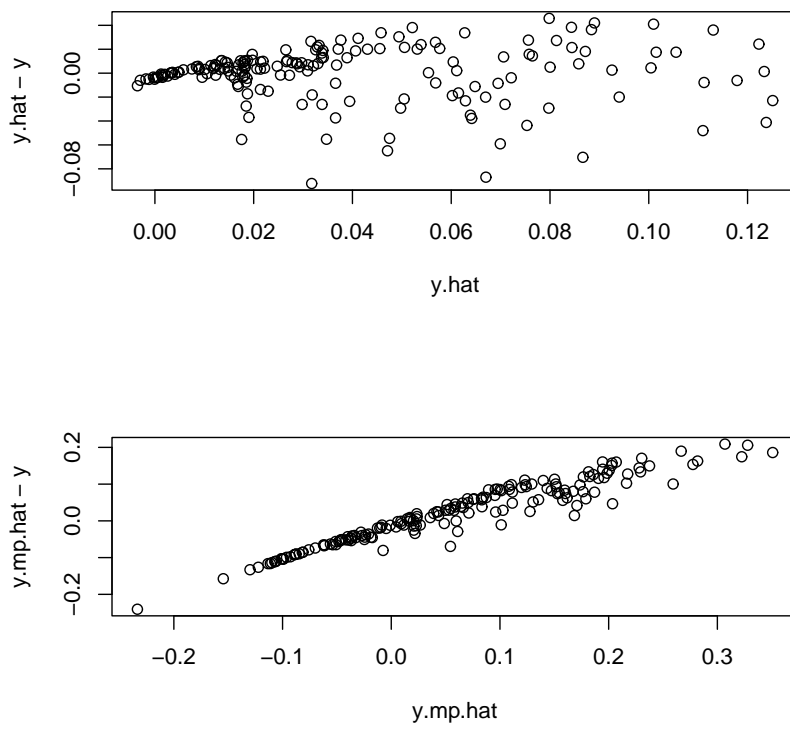
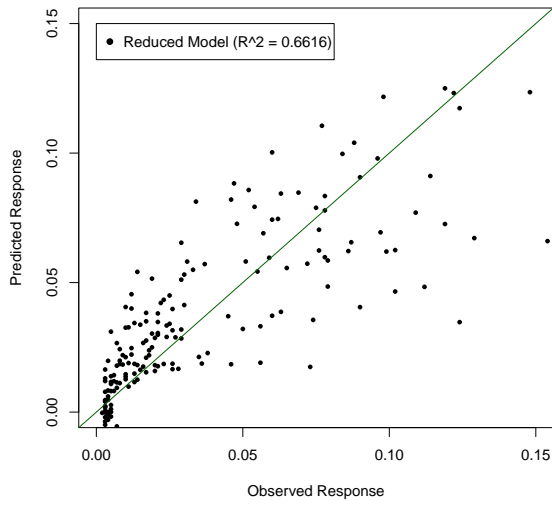
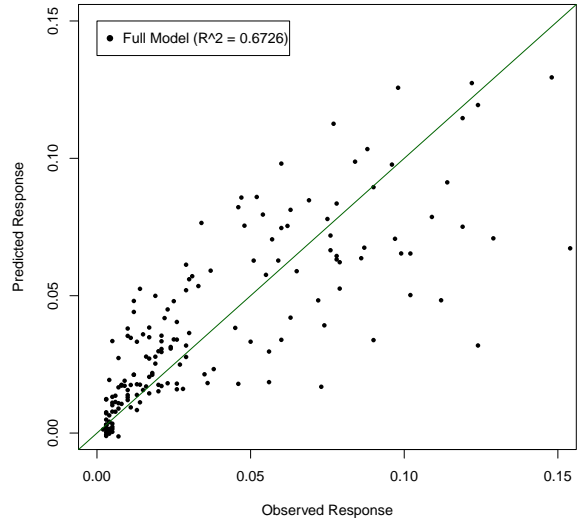


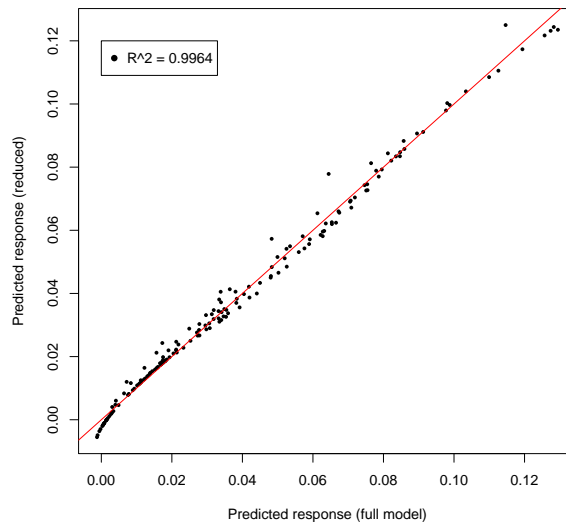
Figure 11: Residual vs. predicted values. Residual = predicted value - observed response. Top panel: Bernstein polynomial based estimate. Bottom panel: monoProc estimate.



(a) Reduced Model



(b) Full Model



(c) Full Model vs. Reduced Model

Figure 12: 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., $y=x$.

Table 1: RMISE and MIAE ($\times 100$) based on 500 Monte Carlo replications as global measurements of performance. *Note:* Standard error is displayed in the parentheses. $m_1(x_1, x_2) = \min(x_1, x_2)$; $m_2(x_1, x_2) = x_1 x_2 I(x_1 \leq 0.5, x_2 \leq 0.5) + \{(x_1 - 0.5)(x_2 - 0.5) + x_1 x_2\} I(x_1 > 0.5, x_2 > 0.5)$; $m_3(x_1, x_2) = x_1^{\frac{1}{3}} x_2^{\frac{2}{3}}$.

(a) $m_1(x_1, x_2)$				
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

(b) $m_2(x_1, x_2)$				
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

(c) $m_3(x_1, x_2)$				
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 2: Average prediction biases. *Note:* $m_1(x_1, x_2) = \min(x_1, x_2)$; $m_2(x_1, x_2) = x_1 x_2 I(x_1 \leq 0.5, x_2 \leq 0.5) + \{(x_1 - 0.5)(x_2 - 0.5) + x_1 x_2\} I(x_1 > 0.5, x_2 > 0.5)$; $m_3(x_1, x_2) = x_1^{1/3} x_2^{2/3}$. Columns 3 and 4: Mean bias (S.E. of bias); Columns 5 and 6: [2.5th percentile, 97.5th percentile] of the biases.

(a) $m_1(x_1, x_2)$					
σ	Point	BP	MP	BP	MP
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]

(b) $m_2(x_1, x_2)$					
σ	Point	BP	MP	BP	MP
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]

(c) $m_3(x_1, x_2)$					
σ	Point	BP	MP	BP	MP
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: 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