Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Robust Estimation of Additive Boundaries With Quantile Regression and Shape Constraints

Yan Fang , Lan Xue , Carlos Martins-Filho & Lijian Yang

To cite this article: Yan Fang , Lan Xue , Carlos Martins-Filho & Lijian Yang (2020): Robust Estimation of Additive Boundaries With Quantile Regression and Shape Constraints, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2020.1847123

To link to this article: https://doi.org/10.1080/07350015.2020.1847123

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2020, VOL. 00, NO. 0, 1–14
https://doi.org/10.1080/07350015.2020.1847123

Robust Estimation of Additive Boundaries With Quantile Regression and Shape Constraints

Yan Fang[a], Lan Xue[b], Carlos Martins-Filho[c], and Lijian Yang[d]

[a]School of Finance, Shanghai University of International Business and Economics, Shanghai, China; [b]Department of Statistics, Oregon State University, Corvallis, OR; [c]Department of Economics, University of Colorado, Boulder, CO; [d]Center for Statistical Science, Tsinghua University, Beijing, China

ABSTRACT

We consider the estimation of the boundary of a set when it is known to be sufficiently smooth, to satisfy certain shape constraints and to have an additive structure. Our proposed method is based on spline estimation of a conditional quantile regression and is resistant to outliers and/or extreme values in the data. This work is a desirable extension of existing works in the literature and can also be viewed as an alternative to existing estimators that have been used in empirical analysis. The results of a Monte Carlo study show that the new method outperforms the existing methods when outliers or heterogeneity are present. Our theoretical analysis indicates that our proposed boundary estimator is uniformly consistent under a set of standard assumptions. We illustrate practical use of our method by estimating two production functions using real-world datasets.

ARTICLE HISTORY
Received July 2020
Accepted November 2020

KEYWORDS
Polynomial spline; Robust estimation; Uniform consistency

1. Introduction

The estimation of the boundary of a set Ψ given a subset of $n \in \mathbb{N}$ observations of its elements has been the object of a large literature in Statistics and Econometrics. The construction and evaluation of the estimators that have emerged depend broadly on three types of restrictions: (a) on the topological structure of the space that contains Ψ; (b) on the properties that characterize the boundary of Ψ; and (c) on the sampling assumptions governing the generation of the subset of observations to be used in estimation. Important earlier contributions to this literature include Korostelev, Simar, and Tsybakov (1995) where $\Psi = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq g(x)\} \subseteq [0, 1] \times [0, 1]$, g is a monotone non-decreasing function and $S_n = \{(X_i, Y_i)\}_{i=1}^n$ is a set of independently drawn observations from a uniform density f whose support is the closure of Ψ. Härdle, Park, and Tsybakov (1995) relaxed the assumption that f is uniform and assumes that g belongs to a Hölder class. Hall, Park, and Stern (1998) assumed g is continuously differentiable and S_n is generated by a Poisson process. Park, Simar, and Weiner (2000) and Kneip, Simar, and Wilson (2008) assumed Ψ is a closed and strictly convex subset of $\mathbb{R}_+^q \times \mathbb{R}_+^d$, S_n is a set of independently drawn observations from a continuous density f and the boundary of Ψ is sufficiently smooth.

At a general level, estimation of set boundaries is made difficult by the fact that the observations in S_n lie largely, by nature, in the interior of Ψ. This fact produces two undesirable properties of many estimators that have emerged in the literature, viz., negative biases and significant sensitivity to extreme sample values (outliers). The latter characteristic resulting from the fact that most proposed estimators "envelope" the set of observations in S_n. Solutions to these problems have mostly come, respectively, in the form of bias correction mechanisms (see Gijbels et al. 1999; Park, Simar, and Weiner 2000; Jeong and Simar 2006) and through sharper or alternative specification of the boundary as in Cazals, Florens, and Simar (2002), Aragon, Daouia, and Thomas-Agnan (2005), Martins-Filho and Yao (2007), or Daouia, Noh, and Park (2016). In addition to inherent bias and sensitivity to outliers, most boundary estimators proposed in the extant literature suffer from the "curse of dimensionality," that is, rates of convergence to the boundary that decrease with the dimensionality of Ψ. This is true, for example, for the popular FDH (free disposal hull) and DEA (data envelopment analysis) estimators first proposed by Deprins, Simar, and Tulkens (1984) and Charnes, Cooper, and Rhodes (1978) and studied in Park, Simar, and Weiner (2000) and Kneip, Simar, and Wilson (2008). This is particularly problematic in empirical settings, where most relevant applications involve Ψ being a subset of spaces with dimension at least greater than two.

In this article, we propose a smooth boundary estimator that helps mitigate the aforementioned difficulties of the extant literature and, in addition, allows for the imposition of commonly assumed shape constraints on the boundary, such as monotonicity and concavity. The main idea is to combine the boundary model proposed by Martins-Filho and Yao (2007) with the additive specification and spline estimation procedure of Wang, Xue, and Yang (2020). In particular, with little loss of generality, we let $\Psi = \{(\mathbf{x}, y) : \mathbf{x} \in [0, 1]^d \subset \mathbb{R}_+^d, 0 \leq y \leq g(\mathbf{x}) = g_0 + \sum_{l=1}^d g_l(x_l)\}$ and consider the estimation of g based on a set of independently drawn observations $S_n = \{(Y_i, \mathbf{X}_i)\}_{i=1}^n \subset \Psi$ with

$$Y_i = g(\mathbf{X}_i)R_i, \tag{1}$$

CONTACT Lan Xue xuel@stat.oregonstate.edu Department of Statistics, Oregon State University, Corvallis, OR 97331.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/UBES.

where R_i is an unobserved random variable taking values in [0,1] and g belongs to a suitably defined class of additive functions that may include shape constraints. The additive structure for g not only eliminates "the curse of dimensionality" when d is large, but also enables separate estimation of the shape and location of the boundary. Following Martins-Filho and Yao (2007), our robust boundary estimator is made possible by first estimating the boundary shape using a nonparametric quantile regression, then estimating the location using a robust procedure based on the pseudo residuals. The estimation of both shape and location are resistant to outliers.

Incorporating boundary shape constraints, such as monotonicity and concavity, into boundary estimation has been the object of a number of recent articles. Parmeter and Racine (2013) introduced constrained kernel regression to obtain a smooth boundary estimate with shape constraints. Daouia, Noh, and Park (2016) used polynomial splines to obtain smooth shape constrained boundary estimators and Wang et al. (2014) proposed a kernel based shape-restricted $\tau \in (0,1)$-quantile regression estimator. The methods developed in these articles cannot be easily applied to the case where $d \geq 2$. Here, we adopt the one-step backfitted polynomial spline estimator with shape constraints proposed in Wang and Xue (2015) to estimate an additive boundary function. It uses polynomial splines to approximate nonparametric additive functions and ensures shape constraints of the boundary estimates by imposing constraints on the spline coefficients. The proposed estimation method takes advantage of linear programming and is very easy to solve numerically.

We provide consistency of our proposed estimator and establish its rate of uniform convergence in probability. Since, in essence, we consider a constrained quantile regression estimator, the theoretical tools needed are different from the least squares approach adopted in Wang, Xue, and Yang (2020). Derivation of the asymptotic distributional theory is challenging due to increasing number of parameters involved in spline approximation, nonlinear form of the boundary function, and shape constrained estimation.

The remainder of this article is organized as follows. Section 2 describes the additive boundary model in detail. In Section 3, we give a detailed description of our three-step polynomial quantile spline method to estimate the additive boundary and provide the main theorems establishing the asymptotic properties of our estimation procedure. Section 4 introduces a multiplicative boundary model that allows for interactions among covariates. Then, we briefly introduce procedures for knot selection and outlier deletion in Section 5. In Section 6, we apply our method in both simulations and real data analysis, and comparisons between our proposed method and the one in Wang, Xue, and Yang (2020) are provided. Section 7 concludes this article. The lemmas and proofs are provided in the supplementary materials.

2. Additive Boundary Model

We consider the boundary model described in Equation (1) with

$$g(\mathbf{X}_i) = g_0 + g_1(X_{i1}) + \cdots + g_d(X_{id}), \tag{2}$$

where g_0 is an unknown constant and each g_l is an unknown nonparametric function defined on [0,1] for $l = 1, \ldots, d$. For identification and estimation, we assume that each g_l is theoretically centered with $E\left(g_l(X_l)\right) = 0$. The additive structure for the boundary g described in Equation (2) has a number of desirable properties. Besides imposing weaker restrictions than a parametric model for the boundary, as will be apparent when we discuss the theoretical properties of our estimator, the additive structure eliminates the "curse of dimensionality" that would emerge with $d \geq 2$. One potential disadvantage of the additive structure is the fact that g varies with x_l only through the component function g_l. In particular, if g is differentiable, its partial derivative with respect to x_l is functionally independent of x_m for $m \neq l$. This can be problematic in certain applications in Economics where g may be interpreted as a production function and it is desirable to have "marginal products" $\frac{\partial g}{\partial x_l}(\mathbf{x})$ depend on all of $\mathbf{x}$ rather than only x_l. In Section 4, we introduce a multiplicative model that accommodates this situation.

Wang, Xue, and Yang (2020) proposed a two-step procedure to estimate the additive boundary functions in (2). They used polynomial splines to approximate nonparametric functions and employed the least squares regression method to estimate the spline coefficients. However, least squares estimation is highly sensitive to outliers and skewed distributions commonly associated with real data in fields such as Economics and Finance. Alternatively, quantile regression has proved advantageous compared to regular mean regression to accommodate these issues. Therefore, in this article, we mainly focus on the additive quantile regression approach for robust and stable boundary estimation. In particular, we assume that for $\tau \in (0,1)$ the quantile function $Q_\tau(R_i|\mathbf{X}_i) = Q_\tau(R_i) = \mu_{R_\tau} \in (0,1)$, for $i = 1, \ldots, n$, where $Q_\tau(R) = \inf\{r \in [0,1] : \tau \leq F_R(r)\}$, F_R is the marginal distribution function for variable R, and τ is the order of the quantile. For example, the median function corresponds to $\tau = 0.5$. Then based on model (1), the conditional quantile function of order τ of Y_i given $\mathbf{X}_i$ can be written as

$$Q_\tau(Y_i|\mathbf{X}_i) = g(\mathbf{X}_i)\,\mu_{R_\tau} = m_\tau(\mathbf{X}_i), \tag{3}$$

with

$$m_\tau(\mathbf{X}_i) = m_{0,\tau} + m_{1,\tau}(X_{i1}) + \cdots + m_{d,\tau}(X_{id}), \tag{4}$$

where $m_{0,\tau} = g_0\mu_{R_\tau}$, $m_{l,\tau}(X_{il}) = g_l(X_{il})\mu_{R_\tau}$ is a function defined on [0,1] and gives the relative shape of the boundary component g_l for $l = 1, \ldots, d$, and the constant μ_{R_τ} determines the location of the boundary function. An advantage of model (1) is that it allows for the separate estimation of the shape and the location of the boundary functions. An estimation strategy that explores this separation was first proposed in Martins-Filho and Yao (2007) and adopted by Wang, Xue, and Yang (2020).

It should be noted that in some settings it may be desirable to allow $Q_\tau(R_i|\mathbf{X}_i)$ to functionally depend on $\mathbf{X}_i$ or other observable random variables. In fact, in the frontier estimation literature in Economics, it is often assumed that a set of "environmental variables" different from $\mathbf{X}_i$ impacts the distribution of R and some of its functionals, such as conditional expectations, variances, and quantiles (see, e.g., Caudill, Ford, and Gropper 1995; Alvarez et al. 2006; Simar and Wilson 2007; Parmeter, Wang, and Kumbhakar 2017; Simar, van Keilegom, and Zelenyuk 2017). In these settings, the separate estimation of boundary shape and location as proposed herein is not possible, since our two-step

estimation procedure depends critically on $Q_\tau(R_i|\mathbf{X}_i) = \mu_{R_\tau}$. For example, when estimating production functions one may assume that certain observable random variables capturing geographic conditions, institutional environment or other factors that are not under the control of the firm, impact the joint distribution of $(Y_i, \mathbf{X}_i)$ through the conditional distribution of R_i on these variables, exposing some of the limitations of our approach.

In the following section, we propose a three-step procedure to robustly estimate the additive functions in (2). In the first step, we use quantile regression to estimate $\{m_{l,\tau}\}_{l=1}^d$. In the second step, we estimate the location parameter μ_{R_τ}. Finally, in the last step, we incorporate the estimation results from the first two steps to estimate the additive boundary functions using the fact that $g_l = m_{l,\tau}/\mu_{R_\tau}$. Among these three steps, the first step is the most important. Therefore, in the following we focus on the estimation of the additive quantile functions $\{m_{l,\tau}\}_{l=1}^d$.

3. Methodology and Theory

3.1. Methodology

We start by describing a procedure for estimating $\{m_{l,\tau}\}_{l=1}^d$ in model (4) by polynomial splines. Let $t_n = \{0 = t_0 \le t_1 \le \cdots \le t_{N_n} \le t_{N_n+1} = 1\}$ be a partition of $[0,1]$ with N_n interior knots. Polynomial splines of degree p are polynomial functions with degree p (or less) on each partitioned interval and which are globally $(p-1)$-times differentiable on [0,1]. Denote the space of p-times continuously differentiable real-valued functions on [0,1] by $\mathcal{C}^p[0,1]$ and the space of polynomial splines with degree p by $\mathcal{G}^p = \mathcal{G}^p([0,1], t_n)$. In addition, let the B-spline basis of $\mathcal{G}^p$ be given by $\widetilde{\mathbf{B}}(x) = \left(\widetilde{B}_1(x), \ldots, \widetilde{B}_{J_n+1}(x)\right)^T$, where $J_n = N_n + p$. Here, without loss of generality, we concentrate on the first J_n basis due to the fact that $\sum_{j=1}^{J_n+1} \widetilde{B}_j(x) = 1$. For each $l = 1, \ldots, d$ let $B_{lj}(x_l) = \widetilde{B}_{lj}(x_l) - n^{-1}\sum_{i=1}^n \widetilde{B}_{lj}(x_{il})$, and $\mathbf{B}_l(x_l) = \left(B_{l1}(x_l), \ldots, B_{lJ_n}(x_l)\right)^T$. Then $\mathbf{B}_l(x_l)$ defines a centered B-spline basis for variable X_l, and $\mathbf{B}(\mathbf{x}) = \left(1, \mathbf{B}_1^T(x_1), \ldots, \mathbf{B}_d^T(x_d)\right)^T$ is a set of B-spline basis for estimating the additive quantile functions in (4). The B-spline bases are centered using their sample averages to consistently estimate the additive functions, which are centered with respect to their theoretical means in (4). The centering only affects the constant ascribed to each additive function and does not change the functional shape of each additive component.

If functions $\{m_{l,\tau}\}_{l=1}^d$ are smooth, we can approximate each of them using spline functions that can be represented as linear combinations of the centered B-splines basis. That is, one can write $m_{l,\tau}(x_l) \approx \mathbf{B}_l^T(x_l)\beta_l$, where $\beta_l = (\beta_{l1}, \ldots, \beta_{lJ_n})^T$. Letting $\mathbf{B}(\mathbf{X}_i) = \left(1, \mathbf{B}_1^T(X_{i1}), \ldots, \mathbf{B}_d^T(X_{id})\right)^T$ for $i = 1, \ldots, n$, the traditional polynomial spline estimators (Stone 1985; He and Shi 1998) of the unknown coefficients $\beta = (\beta_0, \beta_1^T, \ldots, \beta_d^T)^T$ are obtained as

$$\widetilde{\beta} = \underset{\beta \in \mathbb{R}^{dJ_n+1}}{\arg\min} \sum_{i=1}^n \rho_\tau\left(Y_i - \mathbf{B}^T(\mathbf{X}_i)\beta\right), \tag{5}$$

where ρ_τ is the "check function" defined as $\rho_\tau(u) = u(\tau - I(u<0))$ with I being the indicator function of the set $u < 0$. Then the traditional polynomial spline estimators of the unknown quantile function $m_{l,\tau}$ at x_l can be written as

$$\widetilde{m}_{l,\tau}(x_l) = \mathbf{B}_l^T(x_l)\widetilde{\beta}_l, \text{ for } l = 1, \ldots, d. \tag{6}$$

The traditional polynomial spline estimator in (6) does not incorporate any shape constraints. Wang and Xue (2015) developed a one-step backfitted constrained polynomial spline method to estimate monotone additive functions. Sufficient conditions for a polynomial spline $m_{l,\tau}(x_l) = \mathbf{B}_l^T(x_l)\beta_l$ to be monotonically increasing is that the coefficients β_l satisfy $\beta_{l,1} \ge 0$ and $\beta_{l,j} \ge \beta_{l,j-1}$, for $j = 2, \ldots, J_n$. In addition, Lemmas 2 and 3 of Wang, Xue, and Yang (2020) showed that sufficient conditions for a polynomial spline $m_{l,\tau}(x_l) = \mathbf{B}_l^T(x_l)\beta_l$ with degree $p \ge 2$ to be concave are that the coefficients β_l satisfy $\beta_{l,2} - \beta_{l,1} \le 2\beta_{l,1}$; $\beta_{l,j} - \beta_{l,j-1} \le j\beta_{l,j-1}/(j-1) - \beta_{l,j-2}$ for $j = 3, \ldots, p-1$; $\beta_{l,p} - \beta_{l,p-1} \le p\beta_{l,p-1}/(p-1) - \beta_{l,p-2}$; $\beta_{l,j} - \beta_{l,j-1} \le \beta_{l,j-1} - \beta_{l,j-2}$ for $j = p+1, \ldots, N_n+1$; $\beta_{l,N_n+2} - \beta_{l,N_n+1} \le (p-1)\left(\beta_{l,N_n+1} - \beta_{l,N_n}\right)/p$; and for $j = N_n+3, \ldots, N_n+p$, $\beta_{l,j} - \beta_{l,j-1} \le \left(N_n - j + p + 1\right)\left(\beta_{l,j-1} - \beta_{l,j-2}\right)/\left(N_n - j + p + 2\right)$. When $p = 1$, sufficient conditions for a linear spline to be concave are that if $N_n = 1$, $\beta_2 - \beta_1 \le \beta_1$; otherwise, $\beta_2 - \beta_1 \le \beta_1$ and $\beta_j - \beta_{j-1} \le \beta_{j-1} - \beta_{j-2}$ for $j = 3, \ldots, N_n + 1$. For simplicity, let $\mathcal{C}_M$ and $\mathcal{C}_C$ denote the set of spline coefficients that satisfy the monotone increasing conditions and concave conditions, respectively.

Using the traditional polynomial spline estimator $\{\widetilde{m}_{l,\tau}\}_{l=1}^d$ as an initial estimator and letting $\widetilde{m}_{0,\tau} := \widetilde{\beta}_0$, the first component of $\widetilde{\beta}$, we define $Y_{i,-l} = Y_i - \widetilde{m}_{0,\tau} - \sum_{l' \ne l} \widetilde{m}_{l',\tau}(X_{il'})$ as the pseudo-responses associated with the lth direction and $Y_{-l} = \left(Y_{1,-l}, \ldots, Y_{n,-l}\right)^T$ for $l = 1, \ldots, d$. Then, Y_{-l} can be regarded as an approximation of $\left(m_{l,\tau}(X_{1l}), \ldots, m_{l,\tau}(X_{nl})\right)^T$.

To obtain spline estimators that follow shape constraints, we consider a one-step backfitted procedure. For each $l = 1, \ldots, d$, the spline coefficients $\widehat{\beta}_l$ are obtained as

$$\widehat{\beta}_l = \underset{\beta_l \in C_\ell}{\arg\min} \sum_{i=1}^n \rho_\tau\left(Y_{i,-l} - \mathbf{B}_l^T(X_{il})\beta_l\right). \tag{7}$$

We note that the minimization is over the constrained set $\mathcal{C}_\ell$. For example, it gives monotone functions if $\mathcal{C}_\ell = \mathcal{C}_M$, and monotone and concave functions if $\mathcal{C}_\ell = \mathcal{C}_M \cap \mathcal{C}_C$. The resulting shape constrained polynomial spline estimator of $m_{l,\tau}$ at x_l is given by

$$\widehat{m}_{l,\tau}(x_l) = \mathbf{B}_l^T(x_l)\widehat{\beta}_l, \tag{8}$$

for $l = 1, \ldots, d$, and $m_\tau(\mathbf{x})$ is estimated as $\widehat{m}_\tau(\mathbf{x}) = \widehat{m}_{0,\tau} + \sum_{l=1}^d \widehat{m}_{l,\tau}(x_l)$.

Similar to Martins-Filho and Yao (2007) and Wang, Xue, and Yang (2020), model (1) immediately gives $\frac{Y_i}{m_\tau(\mathbf{X}_i)} = \frac{g(\mathbf{X}_i)R_i}{g(\mathbf{X}_i)\mu_{R_\tau}} = \frac{R_i}{\mu_{R_\tau}}$. Therefore, the location of the boundary μ_{R_τ} can be estimated by

$$\widehat{\mu}_{R_\tau} = \left(\max_{1 \le i \le n}\left(Y_i/\widehat{m}_\tau(\mathbf{X}_i)\right)\right)^{-1}, \tag{9}$$

since R_i is bounded above by 1. Although $\widehat{m}_\tau(\mathbf{X}_i)$ is robust to outliers, $\widehat{\mu}_{R_\tau}$ is not, since Y_i is used in the calculation of the

maximum. In Section 5, we introduce a robust extension of $\widehat{\mu}_{R_\tau}$ which is more useful when the data contains outliers. Finally, since $g_l = m_{l,\tau}/\mu_{R_\tau}$, $g_l(x_l)$ can be directly estimated as

$$\widehat{g}_l(x_l) = \widehat{m}_{l,\tau}(x_l)/\widehat{\mu}_{R_\tau}. \tag{10}$$

The additive boundary g at $\mathbf{x}$ is estimated by $\widehat{g}(\mathbf{x}) = \widehat{g}_0(\mathbf{x}) + \sum_{l=1}^{d} \widehat{g}_l(x_l)$.

Compared with the constrained regression spline method given in Wang, Xue, and Yang (2020), our proposed method not only inherits the desirable properties of computational efficiency and spline approximation, but also allows us to avoid assumptions about the distribution of the error process. Furthermore, the proposed constrained quantile regression method provides a highly robust estimator which is more resistant to extreme observations or outliers than classical regression.

3.2. Asymptotic Properties

In what follows, $a_n \sim b_n$ means that there are constants $0 < a \le b < \infty$ such that $a \le a_n/b_n \le b$ for all n and $|\cdot|$ denotes the Euclidean norm of a vector or the absolute value of a real number according to the context. To characterize some of the asymptotic properties of our estimator, we make the following assumptions.

A1 The components of the sequence of random vectors $\{\mathbf{X}_i\}_{i=1}^n$ are independent and identically distributed taking values in $[0,1]^d$. The common joint density function of $\mathbf{X}_i$, denoted by f_X, is absolutely continuous with respect to Lebesgue measure and satisfies $0 < c_1 \le f_X(\mathbf{x}) \le c_2 < \infty$ for $\mathbf{x} \in [0,1]^d$ and for some constants c_1 and c_2.

A2 The components of the sequence of random variables $\{R_i\}_{i=1}^n$ are independent and identically distributed with $Q_\tau(R|\mathbf{X} = \mathbf{x}) = Q_\tau(R) = \mu_{R_\tau} \in (0,1)$ for almost every $\mathbf{x} \in [0,1]^d$. Their common marginal distribution function F_R is absolutely continuous with density f_R such that $F_R(0) = 0$ and $F_R(1) = 1$. In addition, we assume f_R is a strictly positive function with a uniformly bounded first-order derivative.

A3 The sequence of interior knots in $t_n = \{0 = t_0 \le t_1 \le \cdots \le t_{N_n} \le t_{N_n+1} = 1\}$ is equally spaced on [0,1], with $t_j = j/(N_n+1)$ for $j = 0, 1, \ldots, N_n+1$.

A4 For every $l = 1, \ldots, d$, the function g_l is $(p+1)$-times continuously differentiable on $[0,1]$ for some integer $p \ge 1$.

A5 The number of interior knots N_n satisfies $N_n^2\sqrt{\log n/n} \to 0$, $N_n^{p+2}\sqrt{\log n/n} \to \infty$, $n \to \infty$.

A6 For every $l = 1, \ldots, d$, the function g_l is monotone increasing and there exists a constant $c_3 > 0$, such that $g_l^{(1)}(x_l) \ge c_3$ for all $x_l \in [0,1]$, where $g_l^{(1)}(x_l) = \frac{d}{dx_l} g_l(x_l)$.

A6* For every $l = 1, \ldots, d$, the function g_l is concave and there exists a constants $c_4 < 0$, such that $g_l^{(2)}(x_l) \le c_4$ for all $x_l \in [0,1]$, where $g_l^{(2)}(x_l) = \frac{d^2}{dx_l^2} g_l(x_l)$.

Assumption (A1) restricts the density of $\mathbf{X}_i$ to have bounded support, which coincides with Condition 1 in Wang, Xue, and Yang (2020). Assumption (A2) ensures a common distribution for the efficiency variable. It also requires that f_R be bounded away from 0, which is needed for the rate of convergence as in Horowitz and Lee (2005). Assumption (A3) is about equal spacing of the interior knots, which is the same as the one in Xue and Yang (2006) and Wang, Xue, and Yang (2020), and assumption (A4) imposes a restriction on the rate of growth on the number of knots. For example, the choice of $N_n \sim (\log n/n)^{-1/(2p+3)}$ satisfies the condition. Assumption (A5) requires that the additive functions be smooth. Assumptions (A6) and (A6*) restrict the additive functions to be strictly monotone increasing or concave, respectively. Similar assumptions are also used in Wang, Xue, and Yang (2020). These assumptions are common in the nonparametric literature and are reasonable in a wide variety of applications.

For the traditional spline estimator $\widetilde{m}_\tau$ defined in (6), we have the following result, which follows from Theorem 1 of Horowitz and Lee (2005).

Lemma 1. Suppose Assumptions (A1)–(A5) hold. Then, for $l = 1, \ldots, d$,

$$\sup_{x_l \in [0,1]} \left|\widetilde{m}_{l,\tau}(x_l) - m_{l,\tau}(x_l)\right| = O_p\left(N_n/\sqrt{n}\right), \tag{11}$$

and

$$|\widetilde{m}_{0,\tau} - m_{0,\tau}| = O_p\left(N_n/\sqrt{n}\right) \tag{12}$$

as $n \to \infty$.

To obtain the asymptotic results for $\widehat{m}_{l,\tau}$, we first consider the one-step backfitted *unconstrained* estimator $\check{m}_{l,\tau}$. To be more specific, for each $l = 1, \ldots, d$, $\check{m}_{l,\tau}$ is defined as

$$\check{m}_{l,\tau}(x_l) = \mathbf{B}_l^T(x_l)\check{\beta}_l, \tag{13}$$

with

$$\check{\beta}_l = \underset{\beta_l \in \mathbb{R}^{J_n}}{\arg\min} \sum_{i=1}^{n} \rho_\tau\left(Y_{i,-l} - \mathbf{B}_l^T(x_{il})\beta_l\right), \tag{14}$$

where $Y_{i,-l} = Y_i - \tilde{m}_{0,\tau} - \sum_{l' \ne l} \tilde{m}_{l',\tau}(X_{il'})$. Note that $\{\check{m}_{l,\tau}\}_{l=1}^d$ are one-step backfitted estimators defined similarly as $\{\widehat{m}_{l,\tau}\}_{=1}^d$ in (8), but without any shape constraints. In the following, their corresponding estimators of μ_{R_τ} are denoted as $\check{\mu}_{R_\tau}$ and $\widehat{\mu}_{R_\tau}$, respectively, while the corresponding estimators of g_l are denoted as $\check{g}_l$ and $\widehat{g}_l$.

Theorem 1. Suppose Assumptions (A1)–(A5) hold. Then, for $l = 1, \ldots, d$, as $n \to \infty$,

$$\sup_{x_l \in [0,1]} \left|\check{m}_{l,\tau}(x_l) - m_{l,\tau}(x_l)\right| = O_p\left(N_n\sqrt{\log n/n} + N_n^{-(p+1)}\right), \tag{15}$$

and for $p \ge 1$,

$$\sup_{x_l \in [0,1]} \left|\check{m}_{l,\tau}^{(1)}(x_l) - m_{l,\tau}^{(1)}(x_l)\right| = O_p\left(N_n^2\sqrt{\log n/n} + N_n^{-p}\right),$$

and for $p \ge 2$,

$$\sup_{x_l \in [0,1]} \left|\check{m}_{l,\tau}^{(2)}(x_l) - m_{l,\tau}^{(2)}(x_l)\right| = O_p\left(N_n^3\sqrt{\log n/n} + N_n^{-p+1}\right).$$

The L_2 convergence rate of polynomial spline estimators for quantile additive models has been established in He and Shi (1994, 1996). Similar uniform convergence results were obtained in Theorem 1 of Horowitz and Lee (2005) for a traditional polynomial spline estimator. Here, uniform convergence results are established for the one-step backfitted estimator and their first- and second-order derivatives instead. Because of backfitting, the techniques used in the proof are quite different from those in the existing literature. In addition, our results can be viewed as an extension of Theorem 1 in He and Shi (1998), which established the uniform rate of convergence for spline estimators in univariate nonparametric quantile regression. However, He and Shi (1998) assumes the regression errors are identically distributed, while the multiplicative model (1) implies that our error terms are heteroscedastic.

Theorem 2. Under regularity conditions for the monotone constrained estimator (i.e., (A1)–(A6)), or under regularity conditions for the concave constrained estimator (i.e., (A1)–(A5), (A6*)), as $n \to \infty$,

$$\sup_{x_l \in [0,1]} \left|\widehat{m}_{l,\tau}(x_l) - m_{l,\tau}(x_l)\right| + \left|\widehat{\mu}_{R_\tau} - \mu_{R_\tau}\right| + \sup_{\mathbf{x}\in[0,1]^d} \left|\hat{g}(\mathbf{x}) - g(\mathbf{x})\right| = O_p\left(N_n\sqrt{\log n/n} + N_n^{-(p+1)}\right)$$

for $p \le 3$ and $l = 1, \ldots, d$.

Under condition (A6) or (A6*), Theorem 1 implies that the unconstrained estimators actually satisfy the shape constraints of monotonicity or concavity asymptotically when the sample size is sufficiently large. However, the unconstrained spline estimators do not necessarily satisfy the set of linear constraints on the spline coefficients proposed in Section 3, since these constraints in general are only sufficient, but not necessary. But when $p \le 3$, Wang and Xue (2015) showed these linear constraints are necessary and sufficient conditions. Therefore, the unconstrained and constrained estimators are asymptotically equivalent and enjoy the same asymptotic properties for $p \le 3$.

4. Multiplicative Boundary Model

The additive boundary model in (4) is easy to interpret and allows us to circumvent the "curse-of-dimensionality" when estimating multivariate nonparametric functions. However, the additive structure in model (4) does not allow the derivative of a component function to depend on the argument of other component functions. In this section, we consider a multiplicative boundary model that allows interactions among input variables. The model is inspired by the well-known Cobb–Douglas function with g in (1) taking the form

$$g(\mathbf{X}) = \alpha_0 X_1^{\alpha_1} \cdots X_d^{\alpha_d}, \tag{16}$$

where α_0 and $\{\alpha_l\}_{l=1}^d$ are unknown parameters and quantify the responsiveness of the boundary to changes in its arguments. The function in (16) is multiplicative on its domain. Equivalently, it is additive in log-scale with $\log(g(\mathbf{X})) = \log(\alpha_0) + \sum_{l=1}^d \alpha_l \log(X_l)$. As a nonparametric extension, we consider $\log(g(\mathbf{X})) = m_0 + m_1(X_1) + \cdots + m_d(X_d)$, where m_0 is an unknown constant and $m_1, \ldots, m_d$ are unknown functions. As such, the nonparametric multiplicative model contains the Cobb–Douglas model in (16) as a special case. In addition, due to the monotonicity of the log-transformation, the shape constraints on each of the additive components m_j entail the corresponding shape constraints on the derivative with respect to each component direction in the domain. For example, the monotonicity of m_1 implies that the derivative with respect to X_1 is also monotone. Therefore, we extend the shape constrained polynomial spline method proposed in Section 3 to estimate the nonparametric multiplicative model.

To estimate the unknown components in the multiplicative boundary model, we observe that $\log(Y) = m_0 + m_1(X_1) + \cdots + m_d(X_d) + \log(R)$. Therefore,

$$\begin{aligned} m_{\tau,\log(Y)}(\mathbf{x}) &= Q_\tau\left(\log(Y)|\mathbf{X} = \mathbf{x}\right) \\ &= m_0^* + m_1(x_1) + \cdots + m_d(x_d), \end{aligned} \tag{17}$$

where $m_0^* = m_0 + \mu_{\tau,\log(R)}$ with $\mu_{\tau,\log(R)}$ being the τth-quantile of $\log(R)$. Therefore, the quantile function is different from the boundary function only by the location constant. In addition, $g(\mathbf{x}) = \exp(Q_\tau\left(\log(Y)|\mathbf{X} = \mathbf{x}\right))/\exp(\mu_{\tau,\log(R)})$.

Similar to the estimation method proposed in Section 3, the two-step constrained spline method can be used to estimate the multiplicative model. In the first step, the backfitted constrained polynomial spline method in Section 3 can be used to estimate $m_{\tau,\log(Y)}(\mathbf{x})$, but replacing Y_i with $\log(Y_i)$. In particular, Equation (17) indicates that the additive functions $\{m_l\}_{l=1}^d$ can be directly estimated by $\left\{\hat{m}_l(x_l)\right\}_{l=1}^d$, which are defined similarly as $\hat{m}_{l,\tau}(x_l)$ in (8) but with $\hat{\beta}$ obtained using $\log Y_i$ as response variables in (7) instead. In the second step, let $\eta = \exp(\mu_{\tau,\log(R)})$, and since $R_i/\exp(\mu_{\tau,\log(R)}) = Y_i/\exp\left(Q_\tau\left(\log(Y)|\mathbf{X}\right)\right)$, taking the maximum over $i = 1, \ldots, n$ on both sides and setting $\max_{1\le i\le n} R_i = 1$ we obtain

$$\hat{\eta} = \left\{\max_{i=1,\ldots,n} \frac{Y_i}{\exp\left(\hat{m}_{\tau,\log(Y)}(\mathbf{X}_i)\right)}\right\}^{-1}.$$

Combining the previous two estimation results, the boundary function can be estimated by $\hat{g}(\mathbf{x}) = \exp\left(\hat{m}_{\tau,\log(Y)}(\mathbf{X})\right)\hat{\eta}^{-1}$. Similar to the additive boundary model, we have the following uniform convergence result under the multiplicative model.

Theorem 3. Under regularity conditions for the monotone constrained estimator (i.e., (A1)–(A6)), and under regularity conditions for the concave constrained estimator (i.e., (A1)–(A5), (A6*)), as $n \to \infty$,

$$\sup_{x_l\in[0,1]} |\widehat{m}_l(x_l) - m_l(x_l)| + |\widehat{\eta} - \eta| + \sup_{\mathbf{x}\in[0,1]^d} \left|\hat{g}(\mathbf{x}) - g(\mathbf{x})\right| = O_p\left(N_n\sqrt{\log n/n} + N_n^{-(p+1)}\right),$$

for $p \le 3$ and $l = 1, \ldots, d$.

5. Implementation

5.1. Knot Number Selection

Polynomial splines are popular nonparametric regression tools because of their good approximation properties (de Boor

2001). However, it is well understood that the selection of knot sequence plays an important role in the numerical performance of spline estimators. To reduce the computational complexity, the same knot sequences are used for both traditional and one-step backfitted polynomial spline estimation procedures. In addition, for each input variable, the knot sequences are equally spaced with the same number of interior knots N_n. The optimal N_n, denoted $\widehat{N}_n^{\text{opt}}$, is selected using the Bayesian information criterion (BIC) (He and Shi 1998). Specifically, denote the estimator for the ith response Y_i by $\widehat{Y}_i(N_n)$, $i = 1, \ldots, n$. Then, $\widehat{N}_n^{\text{opt}}$ is given by $\widehat{N}_n^{\text{opt}} = \arg\min_{N_n} \text{BIC}(N_n)$, where $\text{BIC}(N_n) = \log\left\{\sum_{i=1}^n \rho_\tau\left[Y_i - \widehat{Y}_i(N_n)\right]\right\} + \left\{d(N_n + p) + 1\right\}\log n/n$.

5.2. Outliers Deletion

In (9), the estimator for the location of the boundary is $\widehat{\mu}_{R_\tau} = \left\{\max_{i\in\{1,\ldots,n\}}\left[Y_i/\widehat{m}_\tau(\mathbf{X}_i)\right]\right\}^{-1}$, which is highly sensitive to the outliers in the data. This can produce a deterioration of the accuracy of the boundary estimation. To get a more robust estimator of μ_{R_τ}, it is necessary to screen out or remove outliers in $\widehat{\mu}_{R_\tau}$. One approach is to use inter-quartile range (IQR), which is defined as $\text{IQR} = Q_3 - Q_1$ with Q_1 and Q_3 being the first and third quartile of $Y_i/\widehat{m}_\tau(\mathbf{X}_i)$ and $i = 1, \ldots, n$. Define the adjusted interval $\text{AIQR} = [Q_1 - 1.5\text{IQR}, Q_3 + 1.5\text{IQR}]$, and data points that fall out of AIQR are considered outliers. Accordingly, the adjusted robust estimator of μ_{R_τ} is defined as $\widehat{\mu}_{R_\tau} = \left[\max_{i:Y_i/\widehat{m}_\tau(\mathbf{X}_i)\in\text{AIQR}} Y_i/\widehat{m}_\tau(\mathbf{X}_i)\right]^{-1}$.

6. Simulations and Empirical Results

In this section, we investigate the numerical performance of our estimators by applying the proposed methods to both simulated and real data. Simulations are carried out for both univariate and multivariate cases. In the simulation study, the observations are simulated mainly according to the data generating scheme used in Wang, Xue, and Yang (2020) so that we can compare our proposed methods with the ones in Wang, Xue, and Yang (2020). Specifically, we consider six different estimation methods. Three are from Wang, Xue, and Yang (2020) including the unconstrained regression spline (URS), the monotone constrained regression spline (MCRS), and the monotone and concave constrained regression spline (MCCRS). The remaining three estimators are our proposed methods using the quantile approach, namely, the unconstrained quantile spline (UQS), the monotone constrained quantile spline (MCQS), and both monotone and concave constrained quantile spline (MCCQS) as in (10). For all six methods, we have used linear splines ($p = 1$) with the same knot sequence, where the number of knots is selected using BIC described above.

The performance of each method is assessed by the averaged integrated squared error (AISE) and the median integrated squared error (MISE). Let $\hat{g}$ be an estimator of g. The integrated squared error (ISE) of $\hat{g}$ is defined as $\text{ISE}(\hat{g}) = \frac{1}{n_{\text{grid}}}\sum_{k=1}^{n_{\text{grid}}}\left(\hat{g}(x_k) - g(x_k)\right)^2$, where $\{x_k\}_{k=1}^{n_{\text{grid}}}$ are the grid points. Suppose there are N replications, and ISE_r is computed for each replication $r = 1, \ldots, N$. Then, the AISE is defined as $\text{AISE} = \frac{1}{N}\sum_{r=1}^N \text{ISE}_r$, while the MISE is defined as the median value of all ISE_r $(r = 1, \ldots, N)$.

In Section 6.2, we apply the proposed estimators to estimate a production function using Norwegian farm data. In the supplementary materials to this article, we estimate a production function using U.S. high technology firm data.

6.1. Simulation Study

6.1.1. Univariate Case

Data are independently generated from a boundary model with $Y = g(X)R$, where the variable X is uniformly distributed on the interval $[1, 2]$ and $g(x) = 3(x - 1.5)^3 + 0.25x + 1.1125$ is monotone increasing. To explore the robustness of quantile regression, the variable R is generated from the following two distributions:

D1: Exponential distribution: $R = \exp(-Z)$ with $Z \sim \exp(1/3)$;

D2: Mixed normal distribution: $R = \frac{\exp(Z)}{(1+\exp(Z))}$ with $Z \sim 0.3N\left(-2, 0.5^2\right) + 0.7N\left(2, 0.5^2\right)$.

For each distribution, we consider sample sizes n = 100, 250 or 500, and the number of replications N = 1000. To estimate g, the median regression with $\tau = 0.5$ is used in the first step of the proposed method. The estimation methods are URS, UQS, MCRS, MCQS, MCCRS, and MCCQS. For simplicity, the linear spline ($p = 1$) is used in our estimation. The optimal number of interior knots is selected by using BIC as mentioned in Section 5.1, and knots are equally placed in the range of input variables.

The simulation results are summarized in Tables 1 and 2, which report AISE and MISE values from different methods. They clearly show that, in general, both AISEs and MISEs from the proposed quantile methods decrease as the sample size n increases, which supports the asymptotic convergence results given in Section 3. In addition, the shape constraints generally enhance the performance of boundary estimators with both MCRS and MCCRS (both MCQS and MCCQS) giving smaller AISEs or MISEs than URS (UQS) for both exponential and mixed normal distributions.

Tables 1 and 2 suggest that mean regression methods are slightly and consistently better than the proposed quantile regression method for the exponential distribution. However, it is obvious that the quantile regression is more robust for the mixed normal distribution where the distribution of R is polarized with heavier tails on both ends of the support [0,1].

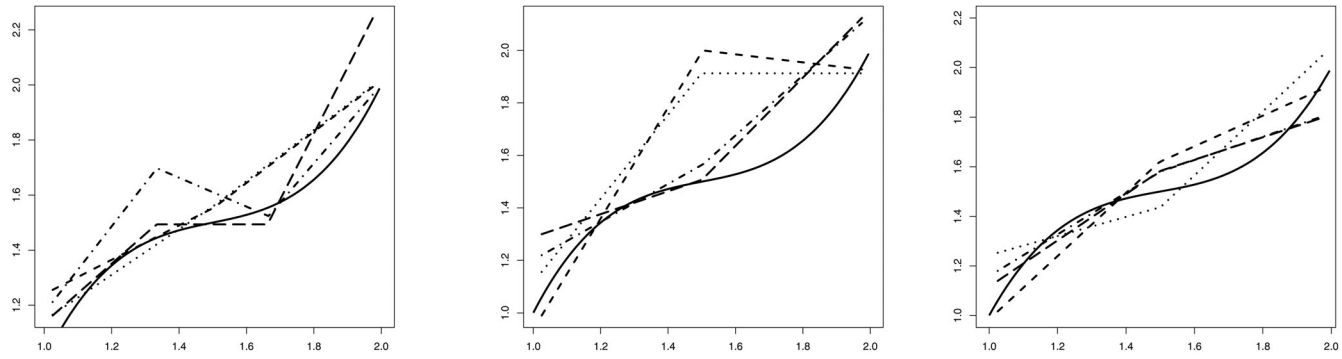
To provide a visual presentation of our simulation results, we plot the typical estimated boundary functions using four different methods (URS, UQS, MCRS, and MCQS) when the sample size is n = 250. For each estimation method, the typical estimate is the one with its ISE being the median among 1000 ISEs from replications. Figures 1(a) and 1(c) plot the estimated boundary functions from both exponential and mixed normal distribution using URS (dashed), MCRS (dotted), UQS (dot-dashed), and MCQS (long-dashed) for each distribution, along with the true boundary (solid). It shows that all methods estimate the boundary function reasonably well. In addition, estimated boundaries

Table 1. AISEs of boundary function estimators under both exponential without/with outliers and mixed normal distributions.

Distributions	n	URS	UQS	MCRS	MCQS	MCCRS	MCCQS
Exponential	100	0.01772	0.03163	0.01585	0.02486	0.01602	0.02473
	250	0.01411	0.02173	0.01330	0.01899	0.01510	0.02135
	500	0.01136	0.01529	0.01087	0.01442	0.01511	0.01941
Mixed normal	100	0.05276	0.08121	0.02863	0.02230	0.02546	0.02213
	250	0.02129	0.00620	0.01642	0.00602	0.01538	0.00642
	500	0.01213	0.00406	0.00998	0.00402	0.01139	0.00634
Exponential with outliers	100	0.17105	0.04401	0.10567	0.04106	0.10778	0.03862
	250	0.12384	0.05979	0.09752	0.05901	0.09687	0.05950
	500	0.11802	0.07804	0.10953	0.07713	0.10683	0.07808

Table 2. MISEs of boundary function estimators under both exponential without/with outliers and mixed normal distributions.

Distributions	n	URS	UQS	MCRS	MCQS	MCCRS	MCCQS
Exponential	100	0.01292	0.01840	0.01194	0.01640	0.01147	0.01603
	250	0.01168	0.01563	0.01091	0.01442	0.01236	0.01589
	500	0.01030	0.01269	0.00996	0.01201	0.01294	0.01563
Mixed normal	100	0.01958	0.00537	0.01641	0.00523	0.01290	0.00491
	250	0.01215	0.00437	0.01040	0.00430	0.00888	0.00455
	500	0.00719	0.00325	0.00594	0.00320	0.00727	0.00481
Exponential with outliers	100	0.10965	0.02051	0.06529	0.01978	0.06978	0.01752
	250	0.08838	0.02286	0.06868	0.02148	0.07309	0.02124
	500	0.07982	0.02707	0.07386	0.02653	0.07745	0.02808



(a) Exponential without outliers (b) Exponential with outliers (c) Mixed Normal

Figure 1. Simulation results for $n = 250$. In each plot, the solid black line represents the true curve. The dashed (– – –) and dotted (·····) lines denote the boundaries estimated using URS and MCRS, while the dot-dashed (–·–·–) and long-dashed (— —) lines represent the corresponding boundary estimated using UQS and MCQS.

under shape constraints are closer to the true boundary than those without shape constraints, implying incorporating shape constraint is desirable in boundary estimation.

To investigate the effect of outliers on boundary estimation, we simulate data from the same boundary model as before with the variable $R = \exp(-Z)$ and $Z \sim \exp(1/3)$. But additional outliers are randomly and uniformly generated on the interval $[\max(Y), 5\max(Y)]$ with the number of outliers being 5% of the sample size. Furthermore, to get an efficient and robust estimation of the boundary function, we implement the outliers deletion introduced in Section 5.2. From the AISE and MISE values reported in Tables 1 and 2, the quantile regression is more resistant to outliers than the mean regression approach. For example, when sample size $n = 100$, the AISE of URS (or UQS) from exponential with outliers is almost 10 times (or 1.5 times) larger than the one without outliers, and the AISE (or MISE) of UQS is only 26% (or 19%) of that from URS. But one notices that both AISEs and MISEs do not decrease as the sample size increases. This is due to the fact that the same percentage of outliers is added for all sample sizes.

Figure 1(b) plots the typical estimated boundary functions for exponential distribution with outliers using URS, UQS, MCRS, and MCQS. It reveals that quantile spline is more resistant to outliers than regression spline. In addition, shape constraints also improve the performance of boundary estimation. In summary, our simulation study confirms that the proposed quantile methods have better numerical performances than their mean regression counterparts when the distribution of R has heavier tails on [0,1] and/or there are outliers in the data. Moreover, considering shape constraints not only provides a more interpretable estimator for boundary functions, but also improves the estimation accuracy.

6.1.2. Multivariate Additive Case

Experiments were also carried out with $d = 4$. More specifically, we consider the additive model $Y = [g_0 + g_1(X_1) + g_2(X_2) +$

$g_3(X_3) + g_4(X_4)\big]R$, where $g_0 = 8$, $g_1(x) = 2x - 1$, $g_2(x) = 2x + [\sin(2\pi x)]/2\pi - 1$, $g_3(x) = 3x^{1/3} - 9/4$, and $g_4(x) = \big(\log(x) + 1\big)/2$. Each of the variables X_l $(l = 1, 2, 3, 4)$ are independently and uniformly distributed on $[0, 1]$. R is generated according to D2 in Section 6.1.1. A total of $N = 1000$ replications with sample sizes $n = 100, 250$, and 500 are considered. Similarly, the order of the quantile regression τ is set to be 0.5, that is, the median regression.

The additive boundary functions are estimated using URS, MCRS MCCRS, MQS, MCQS, and MCCQS. Again, linear splines ($p = 1$) will be used in our multivariate analysis, and the optimal number of interior knots is also selected by using BIC. For simplicity, all input variables have the same number of interior knots, and knots are equally spaced in the range of each input variable X_l $(l = 1, 2, 3, 4)$. Both AISEs and MISEs are computed for each estimated function $\hat{g}_l$ $(l = 1, 2, 3, 4)$. In addition, we evaluate the overall performances of the boundary function estimation by using both regression spline and quantile spline. Results are shown in Tables 3 and 4.

Table 3. AISEs of multivariate additive boundary functions estimators with mixed normal distribution.

n	Method	$\hat{g}_1(x)$	$\hat{g}_2(x)$	$\hat{g}_3(x)$	$\hat{g}_4(x)$	$\hat{g}(\mathbf{x})$
100	URS	0.48320	0.47491	0.47773	0.46661	2.53886
	UQS	0.23289	0.19860	0.19719	0.16461	1.15042
	MCRS	0.32190	0.27417	0.28519	0.26144	2.00370
	MCQS	0.16533	0.12386	0.12644	0.10802	0.94372
	MCCRS	0.30092	0.23461	0.22834	0.22038	1.83743
	MCCQS	0.15853	0.11205	0.11749	0.10332	0.93337
250	URS	0.19861	0.20988	0.20795	0.20680	1.64059
	UQS	0.03948	0.03287	0.03488	0.04762	0.37463
	MCRS	0.15345	0.14706	0.14961	0.13614	1.34049
	MCQS	0.03110	0.03094	0.03162	0.03780	0.34073
	MCCRS	0.14615	0.11826	0.12296	0.11565	1.25386
	MCCQS	0.02976	0.02640	0.03051	0.03745	0.34321
500	URS	0.10418	0.10270	0.10478	0.11193	1.17211
	UQS	0.01355	0.01559	0.02096	0.03054	0.21670
	MCRS	0.09041	0.08307	0.08355	0.08525	0.96522
	MCQS	0.01348	0.01553	0.02061	0.02903	0.21367
	MCCRS	0.08468	0.06730	0.07197	0.07516	0.92844
	MCCQS	0.01267	0.01343	0.02050	0.02897	0.21754

Table 4. MISEs of multivariate additive boundary functions estimators with mixed normal distribution.

n	Method	$\hat{g}_1(x)$	$\hat{g}_2(x)$	$\hat{g}_3(x)$	$\hat{g}_4(x)$	$\hat{g}(\mathbf{x})$
100	URS	0.31996	0.31414	0.30594	0.31045	2.04145
	UQS	0.04731	0.04508	0.05010	0.05667	0.46583
	MCRS	0.15656	0.13678	0.15545	0.14713	1.33625
	MCQS	0.04658	0.04472	0.03820	0.04159	0.38344
	MCCRS	0.15195	0.12347	0.11489	0.12426	1.20808
	MCCQS	0.04002	0.03303	0.03370	0.03712	0.37979
250	URS	0.13057	0.14690	0.14126	0.13768	1.34887
	UQS	0.01699	0.01945	0.02391	0.03127	0.21084
	MCRS	0.08768	0.09002	0.09308	0.08238	1.04597
	MCQS	0.01681	0.01922	0.02278	0.02541	0.19754
	MCCRS	0.08529	0.07073	0.06333	0.06212	0.95540
	MCCQS	0.01493	0.01476	0.02222	0.02489	0.19629
500	URS	0.06701	0.07335	0.07675	0.08139	1.07930
	UQS	0.00817	0.01182	0.01630	0.02597	0.15842
	MCRS	0.06126	0.05508	0.05184	0.05808	0.78028
	MCQS	0.00820	0.01185	0.01619	0.02417	0.15157
	MCCRS	0.05999	0.04712	0.03989	0.04727	0.71384
	MCCQS	0.00738	0.00978	0.01623	0.02416	0.15568

Similar to the results from the univariate approach, as the sample size increases, both AISEs and MISEs decrease for all six estimation methods. Clearly, the methods with shape constraints improve the accuracy of boundary estimation not only for individual additive component but also for the boundary function. In general, the quantile regression consistently and significantly improves estimation accuracy. For example, when $n = 250$ and for input variable X_1, the AISE (or MISE) of MCQS is only 20% of that from MCRS.

For each input variable, we plot the estimated directional boundary using URS, UQS, MCCRS, and MCCQS. We randomly generate $N = 1000$ datasets with sample size $n = 250$, and then compute the ISE values for the additive functions. The dataset which corresponds to the median value of ISE will be used as the dataset to generate the corresponding estimated boundary for each input variable. Figure 2 illustrates curve estimates from URS, UQS, MCCRS, and MCCQS. It shows that all methods give reasonable curve estimates, while the ones from the quantile regression are generally better than their mean regression counterpart. For example, UQS (MCCQS) gives better curve estimates than URS (MCCRS). In addition, the estimated curves using shape constraints are closer to the true boundary than the ones without shape constraints.

To investigate the performance of the proposed boundary estimators under different quantiles, we consider the estimation of the multivariate additive model given above using the quantile regressions with $\tau = 0.25$ and 0.75, and compare their performances with the median regression ($\tau = 0.5$). Tables 5 and 6 compare the AISEs and MISEs of MCCQS at different quantiles. Similar simulation results (not reported here) are observed for other methods such as MQS and MCQS. It is clear that as the sample sizes increase, both AISEs and MISEs decrease for all quantiles. In addition, $\tau = 0.25$ gives the worst estimation results at all sample sizes. It is due to the fact that the location parameter μ_τ has smaller value with smaller quantile τ, and the inverse of $\hat{\mu}_\tau$, and the estimated boundary can be unstable if the quantity $\hat{\mu}_\tau$ is small.

6.1.3. Multivariate Multiplicative Case

We generate data from the following multiplicative model,

$$\log(Y) = \big[g_0 + g_1(X_1) + g_2(X_2) + g_3(X_3) + g_4(X_4)\big] + \log(R), \tag{18}$$

where $g_0 = 3$, $g_1(x) = x/2 - 1/4$, $g_2(x) = \big(2x + [\sin(2\pi x)]/(2\pi) - 1\big)/4$, $g_3(x) = x^{1/3} - 3/4$, and $g_4(x) = \big(\log(x) + 2\big)/10$. The set-up for both the input variables $(X_1, \ldots, X_4)$ and R are the same as the ones considered in Section 6.1.2. Similarly, $N = 1000$ replications with sample sizes $n = 100$, 250, and 500 and the median regression are considered.

URS, MCRS MCCRS, MQS, MCQS, and MCCQS are used to estimate the multiplicative boundary functions. The linear splines ($p = 1$) will be used in our analysis, and the optimal number of interior knots is also selected by using BIC. The same number of interior knots is used for each input variable, and knots are equally spaced in the range of each input variable. Both AISEs and MISEs are computed for each estimated function $\hat{g}_l$ $(l = 1, 2, 3, 4)$. In addition, we evaluate the overall performances of boundary function estimation by using

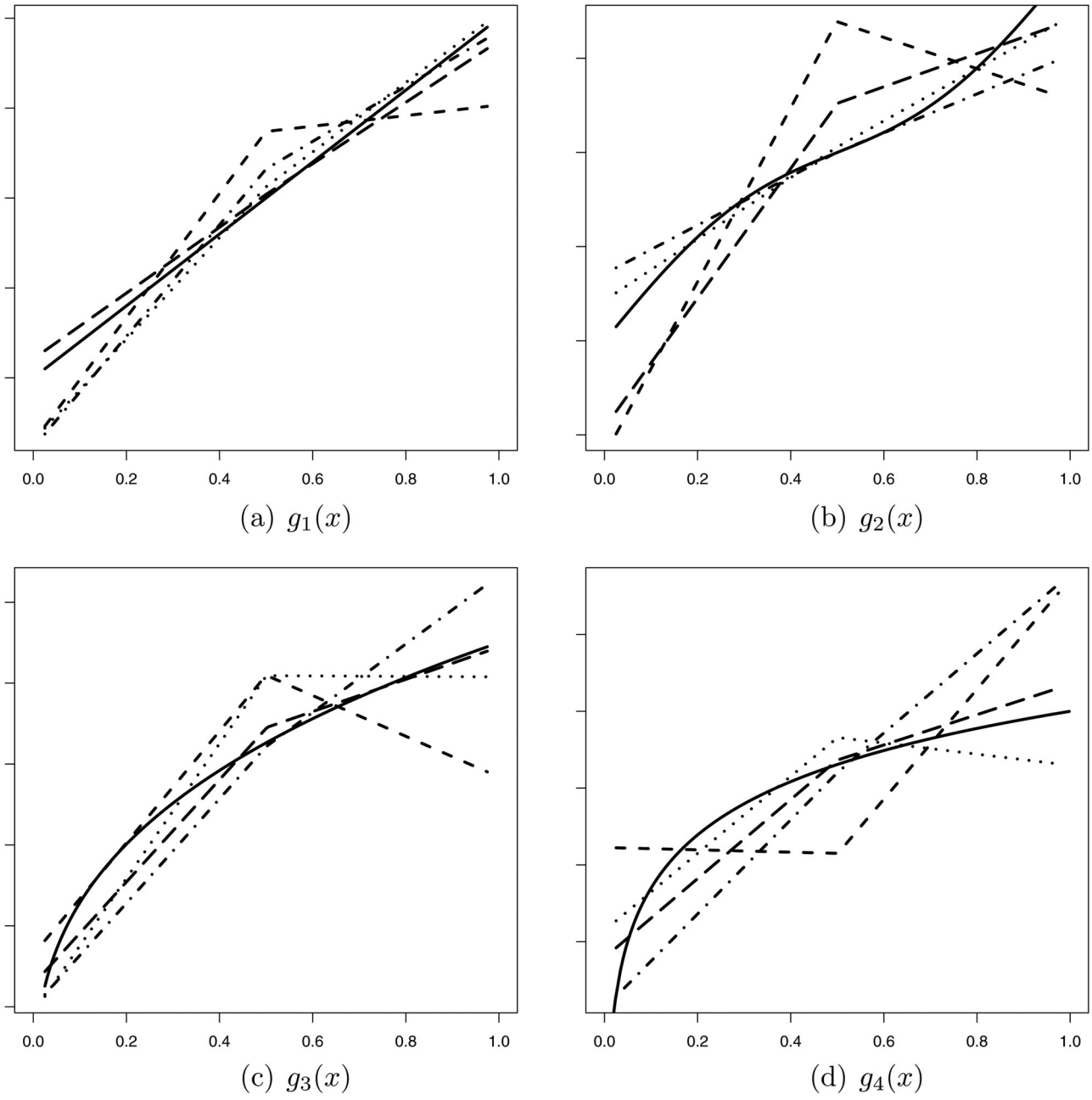


Figure 2. Plots of the estimated boundaries in Equation (10) for $n = 250$. In each plot, the solid line represents the true curve, while the dashed (– – –), dot-dashed (–·–·–), dotted (·····), and long-dashed (— —) lines represent typically fitted curves using URS, MCCRS, UQS, and MCCQS, respectively.

Table 5. AISEs of multivariate additive boundary functions estimators with mixed normal distribution using MCCQS under different τ.

n	τ	$\hat{g}_1(x)$	$\hat{g}_2(x)$	$\hat{g}_3(x)$	$\hat{g}_4(x)$	$\hat{g}(\mathbf{x})$
100	0.25	3.25470	2.49609	2.49433	2.27860	19.79113
	0.50	0.15853	0.11205	0.11749	0.10332	0.93337
	0.75	0.01672	0.01808	0.02072	0.02442	0.18752
250	0.25	1.76886	1.58171	1.65198	1.53507	11.32516
	0.50	0.02976	0.02640	0.03051	0.03745	0.34321
	0.75	0.00616	0.00977	0.01256	0.01623	0.11250
500	0.25	1.21707	1.00144	0.98613	1.08339	7.49704
	0.50	0.01267	0.01343	0.02050	0.02897	0.21754
	0.75	0.00334	0.00721	0.00958	0.01411	0.08992

Table 6. MISEs of multivariate additive boundary functions estimators with mixed normal distribution using MCCQS under different τ.

n	τ	$\hat{g}_1(x)$	$\hat{g}_2(x)$	$\hat{g}_3(x)$	$\hat{g}_4(x)$	$\hat{g}(\mathbf{x})$
100	0.25	1.01359	0.21218	0.27843	0.16527	11.08011
	0.50	0.04002	0.03303	0.03370	0.03712	0.37979
	0.75	0.01091	0.01280	0.01713	0.01921	0.15926
250	0.25	0.37429	0.21218	0.27843	0.16527	5.73473
	0.50	0.01493	0.01476	0.02222	0.02489	0.19629
	0.75	0.00440	0.00785	0.01062	0.01449	0.10263
500	0.25	0.30395	0.21217	0.27842	0.16527	3.86783
	0.50	0.00738	0.00978	0.01623	0.02416	0.15568
	0.75	0.00242	0.00628	0.00849	0.01312	0.08401

Table 7. AISEs of multivariate multiplicative boundary functions estimators with mixed normal distribution.

n	Method	$\hat{g}_1(x)$	$\hat{g}_2(x)$	$\hat{g}_3(x)$	$\hat{g}_4(x)$	$\hat{g}(\mathbf{x})$
100	URS	0.01916	0.01981	0.02023	0.03101	571.029
	UQS	0.00714	0.00855	0.00820	0.01823	164.497
	MCRS	0.01285	0.01187	0.01437	0.02195	430.972
	MCQS	0.00544	0.00536	0.00550	0.01553	112.571
	MCCRS	0.01129	0.01035	0.01203	0.02022	304.044
	MCCQS	0.00520	0.00488	0.00515	0.01523	98.635
250	URS	0.00734	0.00828	0.00855	0.01928	193.622
	UQS	0.00380	0.00405	0.00455	0.01549	80.343
	MCRS	0.00590	0.00606	0.00700	0.01653	155.710
	MCQS	0.00308	0.00302	0.00347	0.01415	60.585
	MCCRS	0.00493	0.00496	0.00607	0.01580	114.134
	MCCQS	0.00285	0.00268	0.00341	0.01405	54.444
500	URS	0.00388	0.00406	0.00465	0.01570	90.970
	UQS	0.00220	0.00247	0.00305	0.01426	52.978
	MCRS	0.00350	0.00347	0.00414	0.01447	79.053
	MCQS	0.00194	0.00205	0.00268	0.01369	44.999
	MCCRS	0.00278	0.00280	0.00379	0.01424	60.806
	MCCQS	0.00171	0.00182	0.00266	0.01366	41.687

Table 8. MISEs of multivariate multiplicative boundary functions estimators with mixed normal distribution.

n	Method	$\hat{g}_1(x)$	$\hat{g}_2(x)$	$\hat{g}_3(x)$	$\hat{g}_4(x)$	$\hat{g}(\mathbf{x})$
100	URS	0.01287	0.01387	0.01365	0.02533	319.590
	UQS	0.00233	0.00292	0.00348	0.01442	68.713
	MCRS	0.00784	0.00706	0.00964	0.01793	187.318
	MCQS	0.00233	0.00289	0.00300	0.01360	37.566
	MCCRS	0.00680	0.00608	0.00634	0.01695	128.057
	MCCQS	0.00196	0.00209	0.00285	0.01338	31.715
250	URS	0.00502	0.00554	0.00614	0.01680	126.035
	UQS	0.00165	0.00196	0.00243	0.01358	51.435
	MCRS	0.00420	0.00395	0.00466	0.01488	81.774
	MCQS	0.00165	0.00196	0.00242	0.01296	34.003
	MCCRS	0.00337	0.00325	0.00369	0.01410	59.709
	MCCQS	0.00129	0.00141	0.00237	0.01287	29.662
500	URS	0.00272	0.00291	0.00338	0.01476	63.728
	UQS	0.00106	0.00142	0.00211	0.01343	40.039
	MCRS	0.00272	0.00287	0.00303	0.01369	50.968
	MCQS	0.00106	0.00142	0.00211	0.01297	33.490
	MCCRS	0.00173	0.00200	0.00269	0.01328	38.955
	MCCQS	0.00073	0.00110	0.00209	0.01294	30.184

both regression spline and quantile spline. Results are shown in Tables 7 and 8.

Similar to the additive case, as the sample size increases, both AISEs and MISEs decrease for all six estimation methods. The methods with shape constraints improve the accuracy of boundary estimation not only for individual additive components but also for the boundary function. Quantile regression consistently and significantly improves the estimation accuracy. For example, when $n = 250$ and for input variable X_1, the AISE (or MISE) of MCQS is only 52% (or 39%) of that from MCRS.

For each input variable, we also plot $\{\hat{g}_i\}_{i=1}^d$ in Equation (18) using four different methods with URS, UQS, MCCRS, and MCCQS. Figure 3 illustrates curve estimates for functions $\{g_i\}_{i=1}^d$ from URS, UQS, MCCRS, and MCCQS in the multivariate multiplicative case. They follow the same patterns except for the difference in scale. It shows all methods give reasonable curve estimates, while the ones from the quantile regression are generally better than their mean regression counterpart. For example, UQS (MCCQS) gives better curve estimates than URS (MCCRS). In addition, the estimated curves using shape constraints are closer to the true boundary than the ones without shape constraints.

6.2. Applications of Quantile Regression

In this section, we present an application of our proposed estimation methods to analyze Norwegian farm data. The supplementary materials contain an additional application of the proposed method using High Technology Firm data. The median ($\tau = 0.5$) regression and linear ($p = 1$) splines are used in the following analysis.

The Norwegian Farm data is from Kumbhakar, Lien, and Hardaker (2014). The dataset contains observations from 151 grain farms in Norway for year 2007. The same dataset was also analyzed in Wang and Xue (2015) and Wang, Xue, and Yang (2020). The objective is to assess the relative efficiency of these farms, that is, $Y_i/g(\mathbf{X}_i)$. The variable of interest is the farm revenue measured in Norwegian krone. We use the log of farm revenue denoted as Y in our boundary/frontier model. The input variables include the total number of hours worked (labor) on the farm (X_1), the productive variable in hectares (X_2), the variable farm inputs (X_3), and the fixed farm input and capital costs (X_4). We consider the additive frontier model $Y = [g_0 + g_1(X_1) + g_2(X_2) + g_3(X_3) + g_4(X_4)]R$ to quantify the farm revenue based on the given input variables.

Figure 4 shows the relationship between farm revenue (in log scale) and inputs. It shows that all input variables have monotonic effects on the farm revenue. In addition, as the input variable increases, the rate of change slightly decreases. Thus, it seems reasonable to estimate the farm production frontier with both monotone and concave constraints for Norwegian Farm data. Therefore, we consider monotone and concave constrained estimation using both regression spline (MCCRS) and quantile spline (MCCQS). In addition, unconstrained methods (URS and UQS) are also considered for comparison purposes. For simplicity, the number of interior knots N_n is set to be the integer part of $n^{1/(2p+3)}$, and the knots are equally spaced in the range of observed values for each input variable.

The estimation results are plotted in Figure 5, where the circles are pseudo observations, and the dashed and long-dashed lines denote the estimated quantile functions using UQS and MCCQS, respectively. In Figure 5, we also plotted 95% bootstrapped point-wise confidence intervals (dotted lines) using UQS, where the lower and upper bounds are calculated as 2.5% and 97.5% sample quantiles of the UQS estimates from 100 bootstrapped samples. Because there are no obvious outliers in the Norwegian Farm data, the regression spline and quantile spline methods give very similar results.

We estimate the frontier functions and evaluate the production efficiency of each farm. The estimated frontier functions using different methods are displayed in Figure 6(a), where the solid circles are the observed farm revenues, and the solid, dotted, dot-dashed and long-dashed lines denote the estimated maximum revenue using URS, MCCRS, UQS, and MCCQS, respectively. It appears that these four estimation approaches

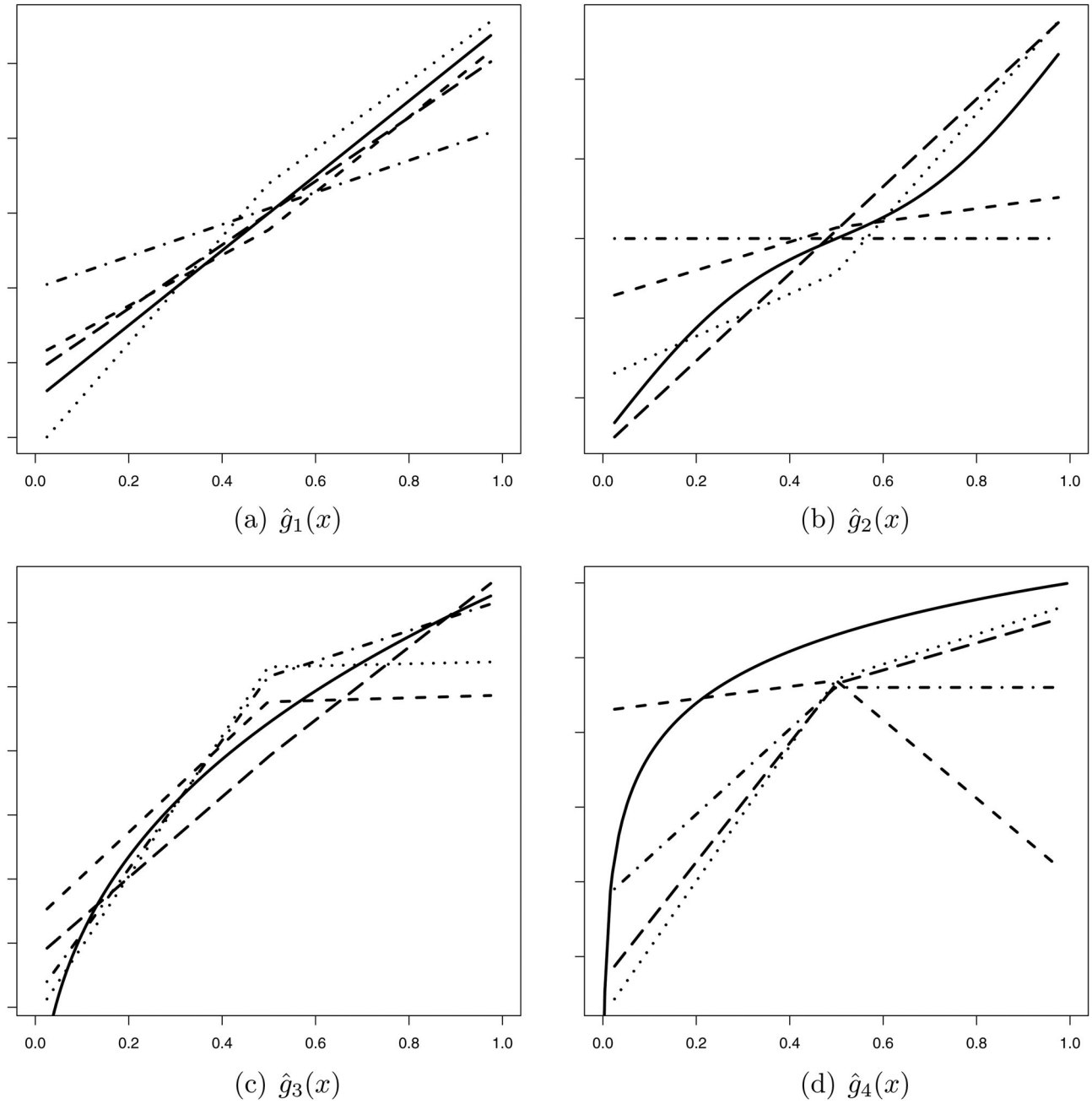


Figure 3. Plots of curve estimates for functions $\{g_i\}_{i=1}^{d}$ in Equation (18) for $n = 250$. In each plot, the solid line represents the true curve, while the dashed (– – –), dot-dashed (–·–·–), dotted (·····), and long-dashed (— —) lines represent typically fitted curves using URS, MCCRS, UQS, and MCCQS, respectively.

give very similar results with four frontier functions overlapping with each other. Figure 6(b) plots the kernel density of the estimated farm production efficiency, and the line specifications are the same as in Figure 6(a). It shows that the majority of farms have estimated efficiency higher than 0.95, indicating that most farms are fairly efficient in their production.

Furthermore, to provide some insights on what potentially causes the differences in farm production efficiency, we run a linear regression using the off-farm income share, the coupled subsidy income share, the environmental subsidy income share, the farmer's experience and the farm's education level as explanatory variable to explain farm production efficiency individually. The definition of each explanatory variables is the same as Wang, Xue, and Yang (2020). The efficiency obtained from the MCCQS method is used as the dependent variable.

The estimated coefficients are given in Table 9. Similar to the results in Wang, Xue, and Yang (2020), both coupled subsidy income share and environmental subsidy income share have a significant negative effect on the farm production efficiency, while the off-farm income share and the farm's experience have no significant effects on the farm production efficiency. In addition, we use one-way ANOVA analysis to examine the effect of education on farm productivity. We divide farmers' educations into three groups, that is, such as, primary, secondary, and high, however, the result shows there is no relationship between them.

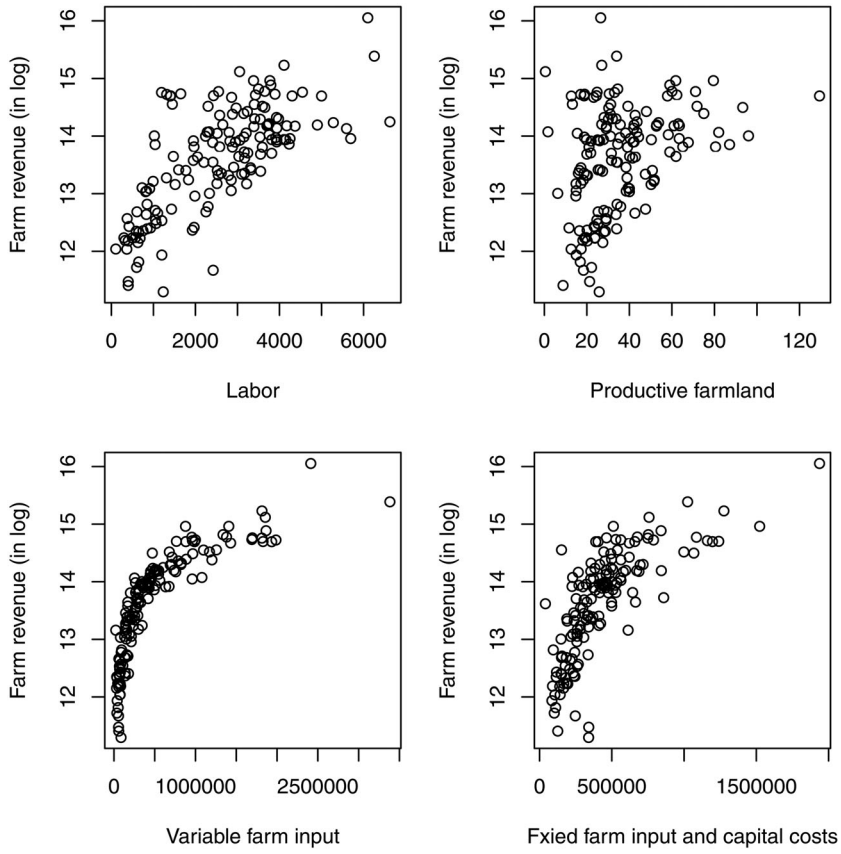


Figure 4. Scatterplots showing the correlation between farm revenue and inputs in Norwegian farm Data.

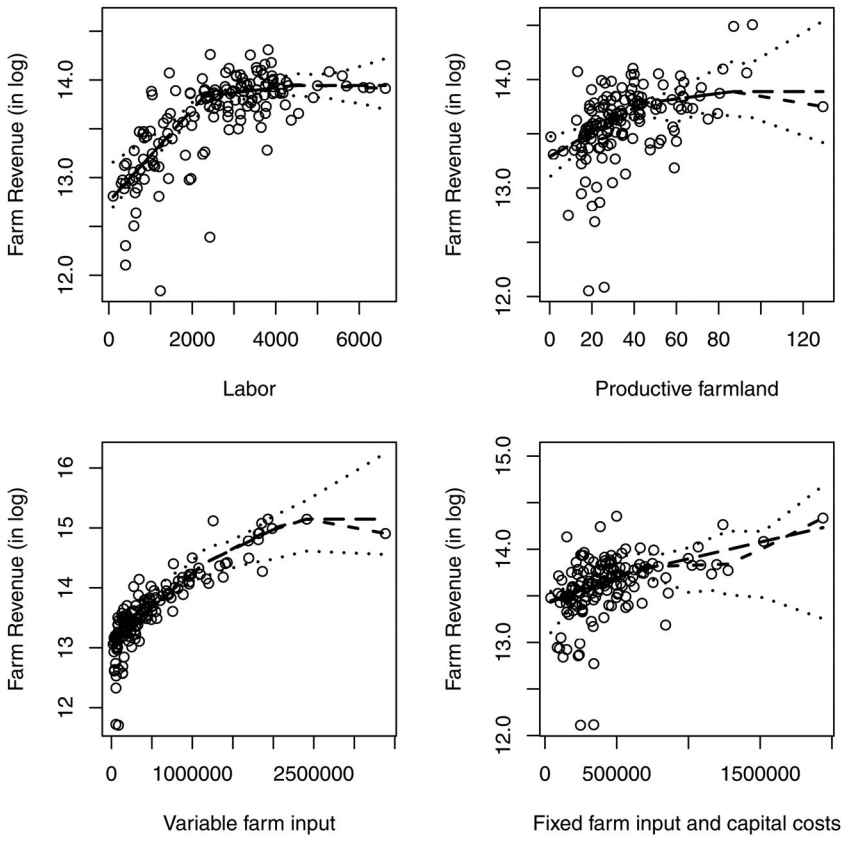


Figure 5. The nonparametric estimate of the expected median on the farm revenue. The circles are pseudo observations for each input variable, the dashed (– – –) and long-dashed (— —) lines denote the estimated nonparametric regression based on UQS and MCCQS, respectively. The dotted (·····) lines describe the 95% point-wise confidence interval from 100 bootstrap samples using the UQS method.

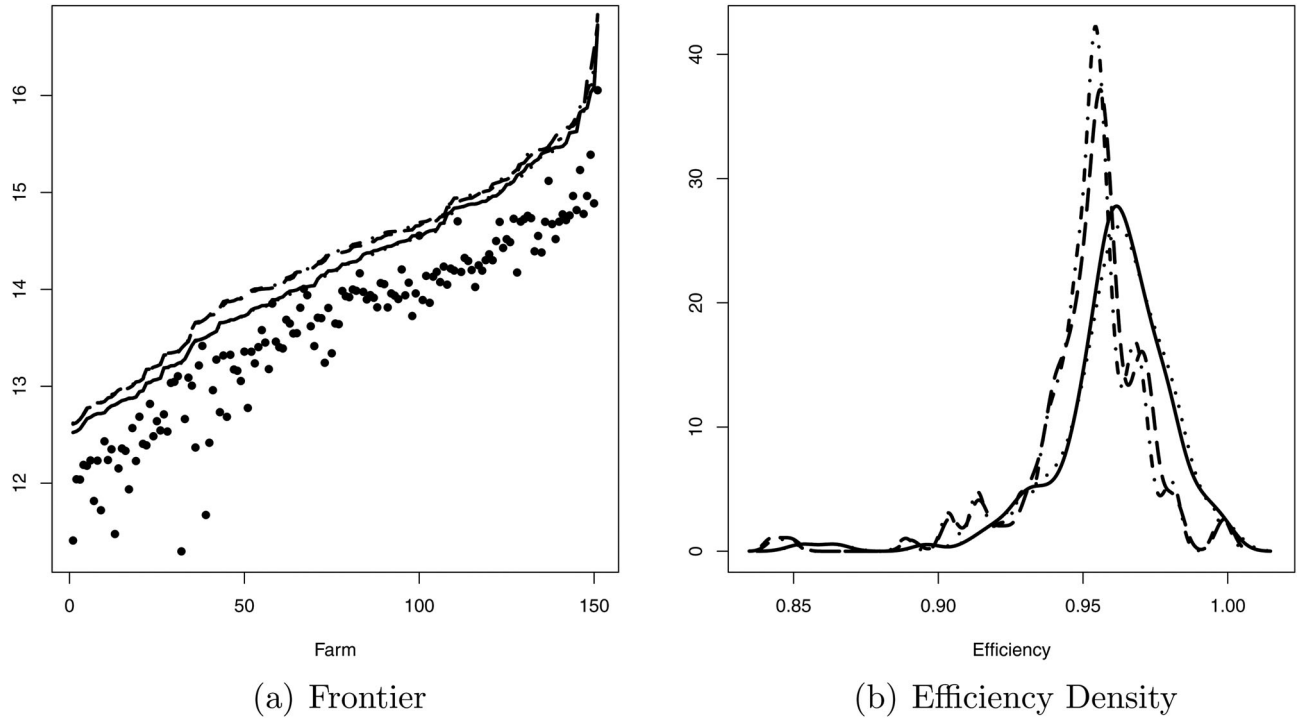


Figure 6. Panel (a) plots the estimated maximum farm log revenues, while Panel (b) gives the kernel density distribution of the relative efficiency estimates, where the solid (——), dotted (·····), dot-dashed (–·–·), and long-dashed (— —) lines represent the density estimate using URS, MCCRS, UQS, and MCCQS, respectively. The solid circles in Panel (a) are the observed true farm log revenues.

Table 9. The estimated regression coefficients with standard errors in parentheses.

Variables	Coefficients
Off-farm income share	−0.0007(0.0070)
Coupled subsidy income share	−0.0890(0.0151)
Environmental subsidy income share	−0.2250(0.0437)
Farmer's experience	0.0000(0.0002)

7. Conclusion

In this article, we employ a two-stage estimation strategy for additive boundary functions. A one-step backfitted quantile regression is used to estimate the shape of the frontier and a robust method using pseudo-residuals is proposed for the location of the boundary. Our method inherits the robustness property of quantile regression and is resistant to skewed distributions and/or outliers in the data. In addition, we also impose shape constraints on the estimated boundary through a set of simple linear constraints on spline coefficients. Our results show that the proposed method not only takes advantage of linear programming and is computationally efficient, but also enjoys desirable theoretical properties. We show that our shape constrained estimator is asymptotically equivalent to, and thus enjoys the same asymptotic properties as, the unconstrained one. For future research, it is worth developing confidence bands for the proposed estimators as in Wang and Yang (2009).

Supplementary Materials

The supplementary materials contain an additional real data application of the proposed method, as well as the relevant lemmas and detailed proofs of the theorems.

Acknowledgments

The authors thank the reviewers, the associate editor, and the co-editor for their helpful suggestions and comments.

Funding

This research is supported by National Natural Science Foundation for Young Scholars of China award 11501355 (Fang), Simons Foundation award 272556 (Xue), National Science Foundation award DMS-1812258 (Xue), National Natural Science Foundation of China award 11771240 (Yang), and Research Fund for the Doctoral Program of Higher Education of China award 20133201110002 (Yang).

References

Alvarez, A., Amsler, C., Orea, L., and Schmidt, P. (2006), "Interpreting and Testing the Scaling Property in Models Where Inefficiency Depends on Firm Characteristics," *Journal of Productivity Analysis*, 25, 201–212. [2]

Aragon, Y., Daouia, A., and Thomas-Agnan, C. (2005), "Nonparametric Frontier Estimation: A Conditional Quantile-Based Approach," *Econometric Theory*, 21, 358–389. [1]

Caudill, S., Ford, J., and Gropper, D. (1995), "Frontier Estimation and Firm-Specific Inefficiency Measures in the Presence of Heteroscedasticity," *Journal of Business and Economic Statistics*, 13, 105–111. [2]

Cazals, C., Florens, J.-P., and Simar, L. (2002), "Nonparametric Frontier Estimation: A Robust Approach," *Journal of Econometrics*, 106, 1–25. [1]

Charnes, A., Cooper, W. W., and Rhodes, E. (1978), "Measuring the Efficiency of Decision Making Units," *European Journal of Operational Research*, 2, 429–444. [1]

Daouia, A., Noh, H., and Park, B. U. (2016), "Data Envelope Fitting With Constrained Polynomial Splines," *Journal of the Royal Statistical Society*, Series B, 78, 3–30. [1,2]

de Boor, C. (2001), *A Practical Guide to Splines*, New York: Springer. [6]

Deprins, D., Simar, L., and Tulkens, H. (1984), "Measuring Labor Efficiency in Post Offices," in *The Performance of Public Enterprises: Concepts and Measurement*, eds. E. M. Airoldi, D. Blei, E. Erosheva, and S. E. Fienberg, Amsterdam: North Holland, pp. 243–267. [1]

Gijbels, I., Mammen, E., Park, B., and Simar, L. (1999), "On Estimation of Monotone and Concave Frontier Functions," *Journal of the American Statistical Association*, 94, 220–228. [1]

Hall, P., Park, B. U., and Stern, S. E. (1998), "On Polynomial Estimators of Frontiers and Boundaries," *Journal of Multivariate Analysis*, 66, 71–98. [1]

Härdle, W., Park, B. U., and Tsybakov, A. B. (1995), "Estimation of Non-Sharp Support Boundaries," *Journal of Multivariate Analysis*, 55, 205–218. [1]

He, X., and Shi, P. (1994), "Convergence Rate of B-Spline Estimators of Nonparametric Conditional Quantile Functions," *Journal of Nonparametric Statistics*, 3, 299–308. [5]

——— (1996), "Bivariate Tensor-Product B-Splines in a Partly Linear Model," *Journal of Multivariate Analysis*, 58, 162–181. [5]

——— (1998), "Monotone B-Spline Smoothing," *Journal of the American Statistical Association*, 93, 643–650. [3,5,6]

Horowitz, J., and Lee, S. (2005), "Nonparametric Estimation of an Additive Quantile Regression Model," *Journal of the American Statistical Association*, 100, 1238–1249. [4,5]

Jeong, S.-O., and Simar, L. (2006), "Linearly Interpolated FDH Efficiency Score for Nonconvex Frontiers," *Journal of Multivariate Analysis*, 97, 2141–2161. [1]

Kneip, A., Simar, L., and Wilson, P. (2008), "Asymptoics and Consistent Bootstraps for DEA Estimators in Non-Parametric Frontiers," *Econometric Theory*, 24, 1663–1697. [1]

Korostelev, A. P., Simar, L., and Tsybakov, A. B. (1995), "Efficient Estimation of Monotone Boundaries," *The Annals of Statistics*, 23, 476–489. [1]

Kumbhakar, S., Lien, G., and Hardaker, J. (2014), "Technical Efficiency in Competing Panel Data Models: A Study of Norwegian Grain Farming," *Journal of Productivity Analysis*, 41, 321–337. [10]

Martins-Filho, C., and Yao, F. (2007), "Nonparametric Frontier Estimation via Local Linear Regression," *Journal of Econometrics*, 141, 283–319. [1,2,3]

Park, B. U., Simar, L., and Weiner, C. (2000), "The FDH Estimator for Productivity Efficient Scores: Asymptotic Properties," *Econometric Theory*, 16, 855–877. [1]

Parmeter, C., and Racine, J. (2013), "Smooth Constrained Frontier Analysis," in *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, eds. X. Chen and N. Swanson, New York: Springer, pp. 463–488. [2]

Parmeter, C., Wang, H.-J., and Kumbhakar, S. (2017), "Nonparametric Estimation of the Determinants of Inefficiency," *Journal of Productivity Analysis*, 47, 205–221. [2]

Simar, L., van Keilegom, I., and Zelenyuk, V. (2017), "Nonparametric Least Squares Methods for Stochastic Frontier Models," *Journal of Productivity Analysis*, 47, 189–204. [2]

Simar, L., and Wilson, P. (2007), "Estimation and Inference in Two-Stage, Semiparametric Models of Production Processes," *Journal of Econometrics*, 136, 31–64. [2]

Stone, C. (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 13, 689–705. [3]

Wang, J., and Yang, L. (2009), "Polynomial Spline Confidence Bands for Regression Curves," *Statistica Sinica*, 19, 325–342. [13]

Wang, L., and Xue, L. (2015), "Constrained Polynomial Spline Estimation of Monotone Additive Models," *Journal of Statistical Planning and Inference*, 167, 27–40. [2,3,5,10]

Wang, L., Xue, L., and Yang, L. (2020), "Estimation of Additive Frontier With Shape Constraints," *Journal of Nonparametric Statistics*, 32, 262–293. [1,2,3,4,6,10,11]

Wang, Y., Wang, S., Dang, C., and Ge, W. (2014), "Nonparametric Quantile Frontier Estimation Under Shape Restriction," *European Journal of Operational Research*, 232, 671–678. [2]

Xue, L., and Yang, L. (2006), "Additive Coefficient Modeling via Polynomial Spline," *Statistica Sinica*, 16, 1423–1446. [4]